Received: 25 February 2019 | Revised: 19 August 2019 | Accepted: 10 September 2019

DOI: 10.1002/gdj3.81

DATA PAPER

Geoscience Data Journal | RMetS | WILEY

The EUSTACE global land station daily air temperature dataset

Yuri Brugnara[1,2] | **Elizabeth Good**[3] | **Antonello A. Squintu**[4] | **Gerard van der Schrier**[4] | **Stefan Brönnimann**[1,2]

[1]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

[2]Institute of Geography, University of Bern, Bern, Switzerland

[3]Met Office Hadley Centre, Exeter, UK

[4]Royal Netherlands Meteorological Institute, De Bilt, the Netherlands

Correspondence
Yuri Brugnara, Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland.
Email: yuri.brugnara@giub.unibe.ch

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 640171

Abstract

We describe a global dataset of quality-controlled in situ daily air temperature observations covering the period 1850–2015, developed in the framework of the EUSTACE (EU Surface Temperature for All Corners of Earth) project (www.eustaceproject.org). The dataset includes a total of 35,364 daily series of maximum and minimum temperature obtained from seven different collections. About 97% of the series are publicly available in a common format, while the remaining 3% can be obtained from the original data providers. Unlike other similar products, duplicates have been removed without blending of series, which simplifies data traceability and improves the temporal homogeneity of the individual series at the cost of a smaller average length. Residual artificial signals (breakpoints) in the series caused by station relocations, changes in instrumentation, etc., have been detected by means of the combination of four breakpoint detection tests, four variables and three temporal aggregations. The combined results give not only the most probable position of the breakpoints, but also a measure of their likelihood. The reliability of the detection was estimated for each year of each target series, based on the number of reference series and on their correlation with the target series. Moreover, its general performance was evaluated through a benchmark of synthetic series. This product will be combined with datasets of marine and ice in situ air temperature observations and with measurements from satellite to produce the first complete global statistical reconstruction of daily near-surface air temperature.

KEYWORDS

air temperature, breakpoints, extremes, land stations

Dataset https://doi.org/10.5285/7925ded722d743fa8259a93acc7073f2

Creator: Institute of Geography, University of Bern, Switzerland

Title: EUSTACE: Global land station daily air temperature measurements with non-climatic discontinuities identified, for 1850–2015

Publisher: Centre for Environmental Data Analysis (CEDA), Science and Technology Facilities Council, Natural Environmental Research Council

Publication year: 2019

Resource type: Dataset

Version: 1.0

1 | INTRODUCTION

Station-based meteorological observations are a fundamental data source for studying recent climate change and the only one providing daily resolution information on secular timescale. Despite the still non-optimal management of these observations (see Thorne *et al*., 2017), numerous efforts in recent years created large collections of daily and sub-daily data that are freely and easily accessible to the scientific community (e.g., Klein Tank *et al*., 2002; Dunn *et al*., 2012; Menne *et al*., 2012; Rennie *et al*., 2014).

Station data, however, have several caveats. One of these is the inconsistency of the measurements over time, due mainly to relocations, changes in the instruments (in particular from liquid-in-glass to electronic) and changes in the standard procedures to observe meteorological variables that have occurred throughout history. Unfortunately, these changes (which are station dependent) are rarely documented in global datasets. The inhomogeneities that they introduce are, in most cases, step-like signals (so-called 'breakpoints') that can be detected statistically when their magnitude is sufficiently large.

Dozens of methods have been developed to detect breakpoints, each giving different results (Venema *et al*., 2012). In global datasets, the large amount of data limits the usability of most of these methods. The Pairwise Homogeneity Assessment (Menne and Williams Jr, 2009) is one of the few examples of breakpoint detection methods that have been applied to global datasets (e.g., Lawrimore *et al*., 2011; Dunn *et al*., 2014; Thorne *et al*., 2016).

Another limitation of station data is their uneven spatial distribution. Instrumental observations have started in Europe in the 17th century and have then spread in the rest of the world following colonization and commercial routes (Brönnimann and Wintzer, 2018). Even nowadays data coverage depends strongly on socio-economic factors, meaning that some areas of the world are still poorly monitored. In the last few decades, however, satellites have provided a revolutionary alternative to in situ observation, allowing a nearly complete coverage of the planet's surface.

The EUSTACE (EU Surface Temperature for All Corners of Earth) project was funded to combine, in a statistical way, station and satellite observations in order to reconstruct daily fields of near-surface air temperature at every point on Earth since 1850. The dataset described in this paper is an intermediate product of EUSTACE that will be combined with consistent satellite-based air temperature estimates in future work. It represents a state-of-the-art global collection of daily temperature observations, controlled for quality and homogeneity issues. In particular, we applied for the first time multiple breakpoint detection methods to a global dataset. This reduces the impact of shortcomings in a certain method and, at the same time, provides useful information on the breakpoints from the agreement between different methods. The strategies adopted within the EUSTACE project to adjust the detected inhomogeneities are described in separate papers (e.g., Squintu *et al*., 2019), and the adjusted data will be published as separate datasets.

The aim of this paper is to describe the advantages of this dataset, as well as its limitations, so that it can be effectively used by the research community. We start by describing the data sources and how the raw data were processed (Section 2). In Section 4, we evaluate the performance of the breakpoint detection and show some examples of the information that can be derived from it. We finally summarize the main characteristics of the dataset in Section 5.

2 | DATA PRODUCTION METHODS

2.1 | Data sources

The data used to assemble the EUSTACE land station dataset come from seven collections that include the main public databases and a selection of smaller datasets that were produced by different research projects or provided by National Weather Services (NWSs). The variables represented are daily maximum and minimum temperature, and the period covered is 1850–2015. The amount of pre-1850 temperature data in digital form is currently insufficient for the aims of the EUSTACE project; therefore, those data were not included in the dataset.

The philosophy behind the EUSTACE station dataset is to reach an acceptable compromise between data quantity and quality. Although data coverage was improved locally (e.g. parts of South America) with respect to other global datasets, for most regions this is not necessarily the case. Individual data sources were critically evaluated and excluded if there were clear indications of general quality and/or consistency issues. In particular, one of the requirements was for observations to be actual daily extremes (i.e. measured with a max/min thermometer or equivalent). To decrease series inhomogeneity, we tried to avoid blended series. Despite this, we took advantage of the massive collection work already done for the GHCN-Daily dataset, whose data have been through a blending procedure (see Menne *et al*., 2012). For other sources, we always chose the 'non-blended' version when possible. We also collected information on the time of observation and on the original data source in case of blended series.

The vast majority of the series (97%) have an open data policy, and the rest cannot be redistributed (i.e. they can be provided exclusively by the original sources, typically NWSs). Table 1 lists the collections and the number of series that we included in the final dataset, as well as their policy. We did not include obvious duplicates (e.g. ECA&D series contained in GHCN-Daily), plus a number of problematic subsets. Global Summary of the Day (GSOD) and HadISD

TABLE 1 List of sources used to assemble the EUSTACE global land station dataset

Name	Reference	Public series	Non-redistr. series	Excluded subsets
GHCN-Daily v3.22	Menne *et al*. (2012)	29,023	0	GSOD
European Climate Assessment & Dataset (ECA&D) non-blended	Klein Tank *et al*. (2002)	3,627	1,007	
International Surface Temperature Initiative (ISTI) v1.00 – stage 2	Rennie *et al*. (2014)	1,327	0	Brazil, GSOD, HadISD
DECADE	Hunziker *et al*. (2017)	338	0	
Servicio Meteorologico Nacional Argentina	-	0	23	
ERA-CLIM	Stickler *et al*. (2014)	15	0	
Southern Alps homogenized	Brugnara *et al*. (2016)	0	4	
Total		34,330	1,034	

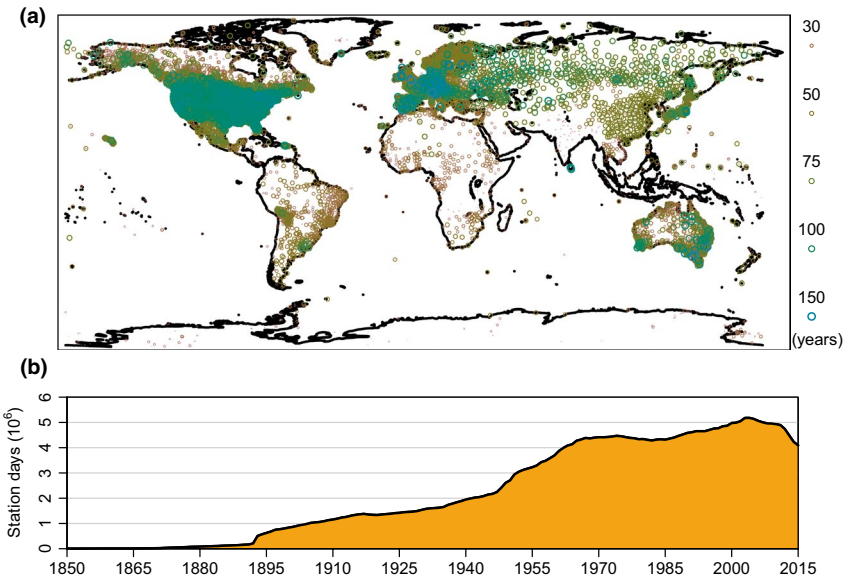

FIGURE 1 (a) Map of the stations included in the EUSTACE dataset, where the size and colour of the points depend on the length of the series. Longer series are plotted on top of shorter series. (b) Temporal evolution of the total amount of data

daily data, in particular, were excluded because they are often calculated from a few synoptic observations and would introduce a systematic underestimation of the diurnal temperature range. Moreover, in several cases we found frequent, unrealistically large differences between GSOD daily values and the corresponding official daily records provided by NWSs.

The geographical and temporal distribution of the series – after duplicates have been removed (see Section 2.2) – are summarized in Figures 1 and 2. Figure 3 gives an overview of how the data are processed (the single steps are described in detail in the rest of this section).

2.2 | Duplicates

The preliminary dataset contained a large number of duplicates. Two data points are duplicates when they originate from the same instrument at the same station. However, detecting when this is the case is not always straightforward. Station name and coordinates of two duplicates can differ for many reasons (different languages, different precision, manipulation errors, etc.), while data manipulation can introduce differences in the temperature values. International identifiers such as the WMO number exist only for some of the stations, and not all sources report them.

We applied a simple duplicate detection algorithm that analyses the daily temperature values one by one. Each year of each record is compared with data of the same year from all stations within a 200-km radius, by looking at daily differences. When the absolute value of the differences does not reach 0.8 K for 60 days in a row (excluding missing values), the data in that year are considered duplicated. A year that has <60 available days is considered duplicated if all observations in that year are exactly identical to those of another series (at least 15 observations required). If the years $i - 1$ and $i + 1$ (or either one at the end and beginning of a series, respectively) are considered to be duplicated,

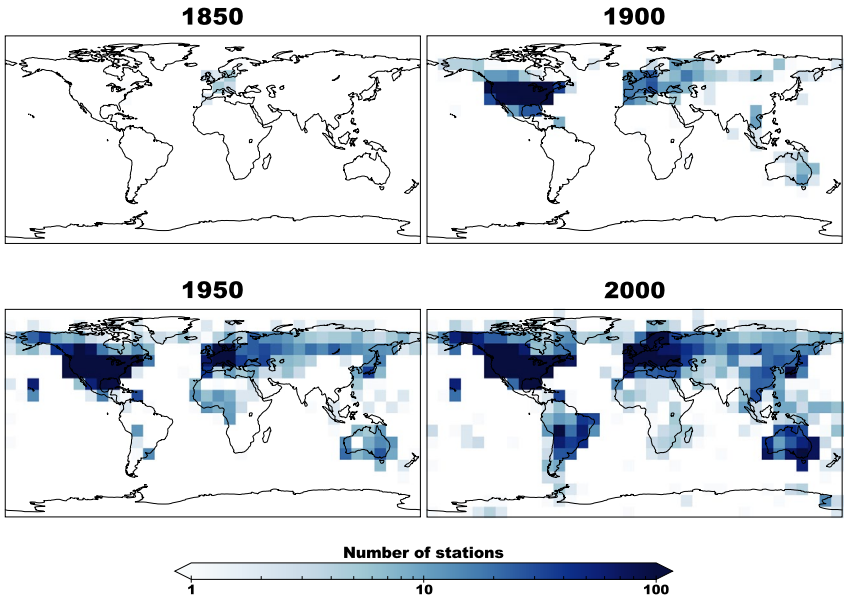

FIGURE 2 Map of station availability in four selected years, shown on a grid of 10 × 10 degree boxes

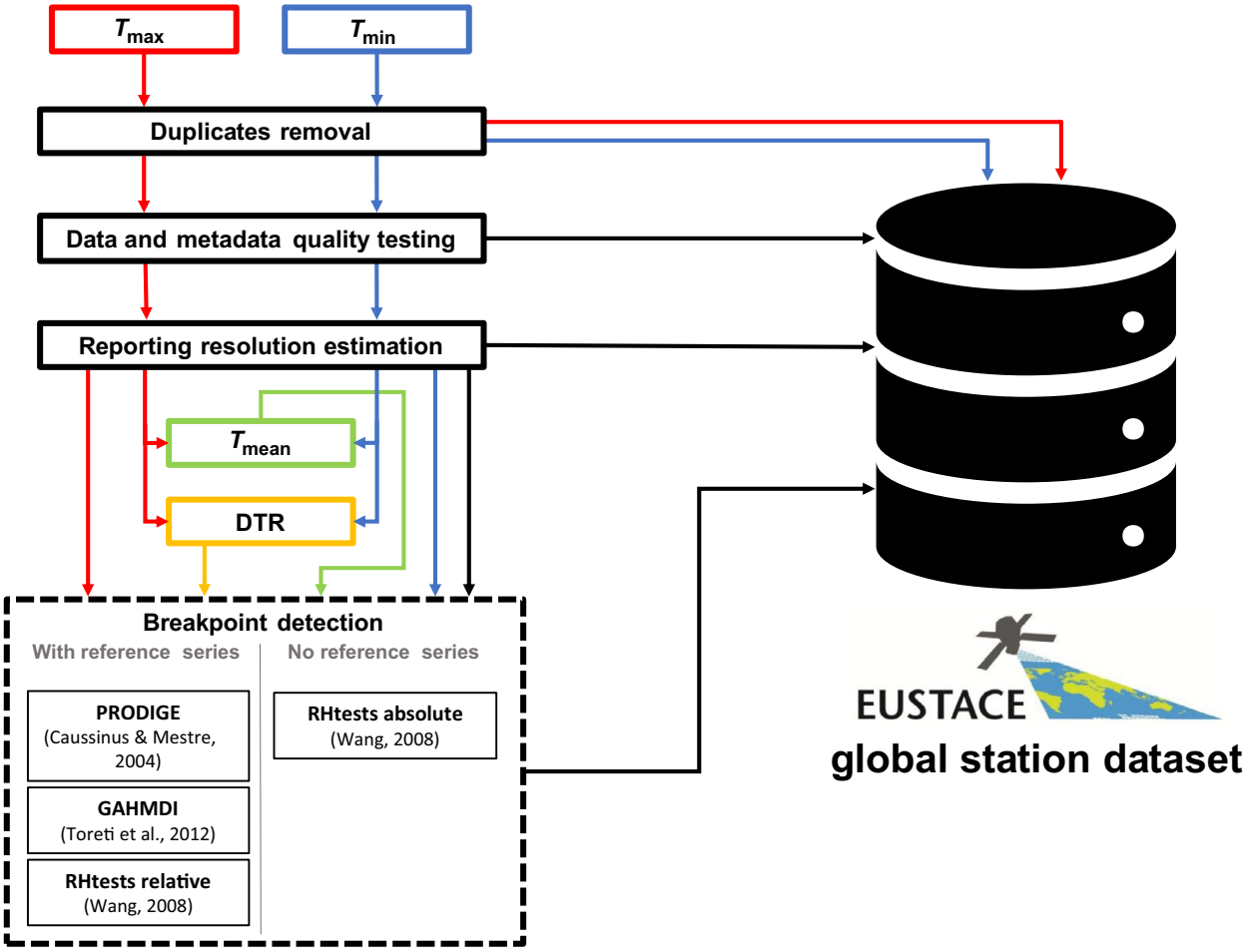

FIGURE 3 Flow chart that summarizes the process followed to build the EUSTACE global station dataset

then the year *i* is considered duplicated too, independently from the outcome of the duplicate detection. We neglect the detection if less than three duplicated years are found. The algorithm is run independently for each of the two variables.

The threshold of 0.8 K encompasses most roundings and conversions, while 60 days are required instead of the whole year to allow for isolated large differences caused by data manipulation or digitization errors. Possible temporal shifts due to different reporting conventions are taken into account.

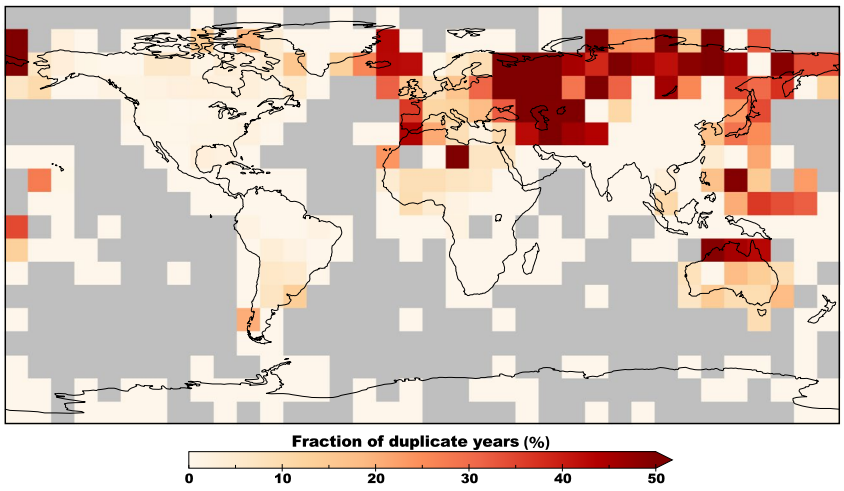

FIGURE 4 Fraction of duplicate years for maximum temperature, shown on a grid of 10 × 10 degree boxes. The grey boxes denote the areas where there are no stations

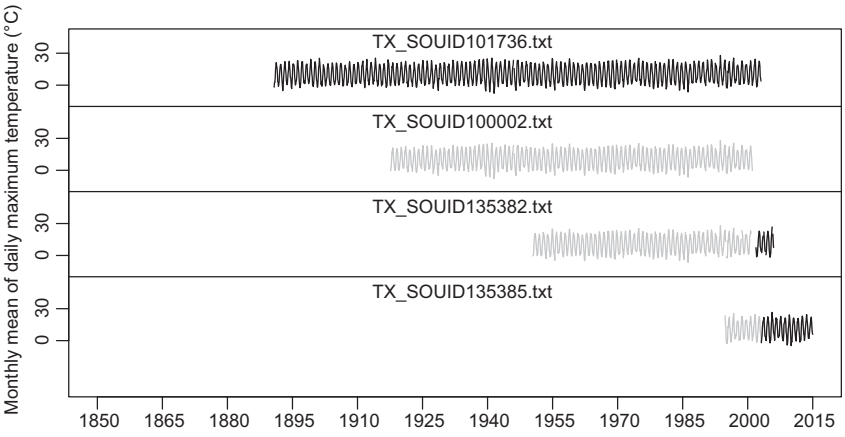

FIGURE 5 Example of handling of multiple duplicates in the EUSTACE dataset (station of Vaexjoe, Sweden). The parts in grey received the duplicate flag, and the entirely grey series was removed from the dataset. In this case, all duplicates had the same coordinates and station name with the exception of the last one, which is probably the airport station. Above each series, we show the original filenames in ECA&D

Even though duplicates at distances larger than 200 km are possible, they are relatively rare. We found 94.1% of the duplicates to be within 20 km of each other and 99.6% within 100 km.

Figure 4 shows that the largest fraction of duplicates is found over the Eurasian continent. This is mostly caused by within-source duplicates in ECA&D, but also by subsets of ISTI (in particular for Russia, Japan and South-East Asia) that are already represented in GHCN-Daily and/or ECA&D.

In most cases, the period covered by two duplicate series is not exactly the same. It is then important not to remove entire series but only the parts that are duplicated. While we removed full duplicates (i.e. when all years are duplicates) from the dataset, we did not remove data from the partial duplicates; instead, we added a flag to mark duplicated data. Only the longest of the duplicate series in terms of available daily observations (our 'best' duplicate) is neither removed nor flagged. Figure 5 shows an example for multiple duplicates. Note that we do not perform any blending of series.

It is unlikely for two highly correlated series from different stations to be considered duplicates by our algorithm. This is also demonstrated by the fact that we hardly find duplicates in the United States, where station density is the highest but data come almost exclusively from GHCN-Daily. On the other hand, parallel records (i.e. different thermometers at the same station) can have differences that are low enough to instigate a duplicate flag (e.g., Brandsma and Van der Meulen, 2008). Moreover, if the homogenized version of a series is provided alongside the original series, the algorithm is usually able to flag only part of the series, since homogeneity adjustments are typically in the order of 1 K (Brohan *et al*., 2006; Lawrimore *et al*., 2011).

2.3 | Data quality and reporting resolution

All series underwent the set of automatic quality tests described in Durre *et al*. (2010). These include checks on basic integrity, outliers, and internal, temporal and spatial consistency.

There are 14 different types of quality flags assigned by the algorithm, but a certain observation does not get more than one flag (the one coming from the first test that the observation failed).

We also estimated the reporting resolution for each year of each series by looking at the frequency of the decimal figures in each month and taking the coarser of the 12 monthly estimations. Possible conversions from Fahrenheit and Rèaumur

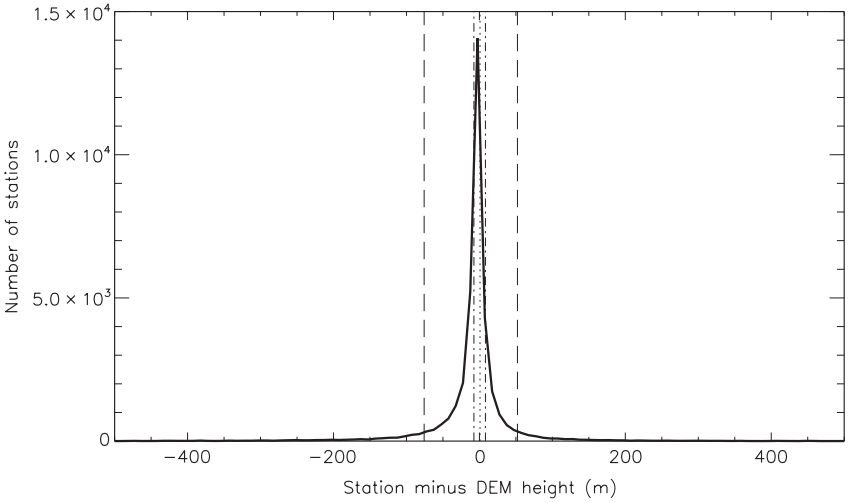

FIGURE 6 Distribution of differences between the elevations provided by the data sources and those estimated from the digital elevation model. The vertical lines indicate the 5th, 25th, 50th, 75th and 95th percentiles

scales are taken into account. To estimate the resolution, we required a minimum of 20 values in a month and of 200 values in a year.

2.4 | Metadata quality

The geographical coordinates of the stations are also affected by errors and must be subjected to quality tests. We applied a semi-automatic procedure in which we first compared the elevation of each station to the digital elevation model (DEM) produced by the Shuttle Radar Topography Mission (Farr and Kobrick, 2000; Rodríguez *et al*., 2005) at 1 km resolution. We then inspected manually those coordinates where the reported elevation differs from the DEM by more than a certain threshold, defined from the spatial variability of the topography around the station (the higher the variability, the higher the threshold) to take into account the low precision of some coordinates, which can be reported with a resolution as coarse as 0.1 degrees. We detected additional wrong coordinates by using a land/sea mask.

We corrected 128 erroneous sets of coordinates. The most common mistakes that could be detected are wrong signs of longitude next to the Greenwich meridian (e.g. 0.5 instead of −0.5), transcription errors in one digit of latitude or longitude (e.g. 30.3 instead of 33.3), and false zero elevations.

Figure 6 shows the distribution of the differences between station and DEM-derived elevations. In about half of the cases, the absolute difference is lower than 10 m, while about 90% of the differences fall between −80 and +55 m, indicating that for a large majority of stations, the elevation provided by the source is sufficiently accurate for most applications.

For about 2% of the stations (730), the elevation was not provided. We used the DEM to estimate these missing elevations (for Antarctica, we used the official elevation of the respective research stations instead).

2.5 | Homogeneity

We tested the temporal homogeneity of each series that had at least 10 years of data (23,347 series) by combining three statistical methods of breakpoint detection, described in Caussinus and Mestre (2004), Toreti *et al*. (2012) and Wang (2008), respectively. For the latter, we adopted a *p*-value of 0.05 to define significant breakpoints, while the other two methods employ the minimization of a penalized likelihood function given a maximum number of detectable breakpoints (set to be proportional to the length of the analysed series). The combination of different methods was shown to improve detection performance in previous studies (Kuglitsch *et al*., 2012; Trewin, 2018). In the present paper, however, both the individual tests and the combination of their results are fully automated.

We applied the three methods on difference series (candidate minus reference) constructed using three different temporal aggregations (annual means, April-to-September means and October-to-March means) and four variables (T_{max}: maximum temperature; T_{min}: minimum temperature; $T_{mean} = (T_{max}+T_{min})/2$; DTR $= T_{max}-T_{min}$). We used the penalized maximal *F* test of Wang (2008) for one additional absolute test (i.e. without reference series) on monthly anomalies, again on three temporal aggregations (all months, April-to-September months only and October-to-March months only) and four variables and with a *p*-value of 0.05. This test is employed when the available reference series are considered insufficient (see Section 2.5.1).

Altogether, we used up to 48 different combinations of method, temporal aggregation and variable, resulting in 48 'sets' of breakpoints (where a set is the output of a breakpoint detection test) with annual resolution (i.e. we provide only the year of the breakpoint). Two additional sets, which we call for simplicity 'breakpoints from metadata', are based on information derived from certain characteristics of the maximum and minimum temperature series: in particular,

we set breakpoints when there are changes in the reporting resolution, large gaps and changes in the data source. We did not make use of more specific metadata such as dates of relocations because these are currently not available for global datasets. These kinds of metadata, however, are often available for national datasets and constituted an additional validation tool for our method (see Section 4.1 and Kuglitsch *et al*., 2012).

For the relative tests (i.e. those using reference series), the algorithm looks for eight well-correlated ($r \geq 0.6$) reference stations located within 1,000 km of the candidate station. It also requires a minimum of 120 months of data in common with the candidate series. The statistical homogeneity tests were not applied to series with <120 months of data (a month with more than five missing days is considered missing). When more than eight references are found, priority is given to those that are geographically closer, provided that they have data in at least 80% of the period covered by the candidate series. Note that the selection is performed independently for each variable, but not for each temporal aggregation (first differences of annual means are used to calculate the correlation).

To assign the breakpoints, we used the pairwise comparison approach (Caussinus and Mestre, 2004): a breakpoint is assigned to a certain year if it is found in at least three difference series, using a tolerance of ± 1 years. If fewer than three reference series are available, then only the absolute test is performed (5,147 series affected).

With this approach, using too many reference series would result in over-detection, because the reference series are usually not homogeneous. Our choice of a maximum of eight reference series is based on tests with the synthetic dataset produced by Venema *et al*. (2012), which showed that using more than eight series increased the probability of false detections (see also Section 2.5.2).

2.5.1 | Detectability index, merged breakpoints and likelihood index

We calculated a 'detectability' index to evaluate, for each year of the candidate series, the potential performance of the relative tests. It is defined as the sum of the Pearson correlation coefficients of those reference series that have more than 50% of data available within a window of ± 5 years from the target year. For example, the index for the year 1950 is calculated from those reference series that have at least 5.5 years of data in the period 1945–1955. If fewer than three reference series are available, then the detectability index is defined as zero. Its maximum theoretical value is 8, because no more than eight reference series are selected (in reality, the correlation of the reference series is always lower than 1).

A large detectability index indicates many well-correlated reference series available and thus ideal working conditions for the relative tests. We use the results of the absolute test only in those periods when the detectability index is lower than 4. The absolute test, which has a lower power of detection than the relative tests (Wang, 2008; Venema *et al*., 2012), is thus intended as a backup test for when a relative test is hardly possible. Long series (>50 years) for which only the absolute test could be performed are found in tropical islands (Sri Lanka, Indonesia), in large deserts (Sahara) and even in station-rich areas if only an early period is covered. Also, most stations in Antarctica do not have sufficient reference series for the relative tests.

We finally added together the 50 sets of breakpoints (36 sets from the relative tests, 12 from the absolute test and two from metadata) to obtain a number of detections for each year, which is an indication of the probability of a breakpoint. We defined the most likely position of the breakpoints from the local maxima of the detections, and a 'likelihood' index for each breakpoint from the sum of the detections in the 3-year window centred on the most likely position of the breakpoint. In this 'merged' set, the breakpoints are assumed to affect all variables at the same time. Similar merged sets can be produced for each variable and each temporal aggregation using a smaller number of sets. Note that the original sets can have breakpoints in consecutive years (these are typically duplicates caused by disagreement among the reference series on the position of the breakpoint), while the merged set cannot; for this reason, the likelihood index can assume a value that is larger than the number of sets that contributed to the breakpoint.

2.5.2 | Performance of the breakpoint detection

To assess its performance, we applied our breakpoint detection algorithm to a benchmark of synthetic daily temperature series developed within the International Surface Temperature Initiative (ISTI) framework (Willett *et al*., 2014; Killick, 2016). The synthetic series are contaminated by inhomogeneities having the same statistical properties of those found in real temperature series, with the difference that the position of each breakpoint is known a priori.

The benchmark comprises four subsets representing different climatic regions in the contiguous United Stated (Wyoming, the Northeast, the Southeast and the Southwest). Each of the regions is reproduced in three or four parallel 'worlds' that represent different choices in the simulation of the breakpoints and in the station density (Table 2). Three different types of breakpoints are simulated: shelter change, station relocation and urbanization.

Since the benchmark was created for daily mean temperature only, we can only test our algorithm on that variable. In other words, we use 13 sets of breakpoints (nine relative, three absolute and one from metadata) instead of

TABLE 2 Different scenarios (or 'worlds') simulated in the ISTI benchmark (Killick, 2016)

World	Description	Regions	Inhomogeneities	Station density	Autocorrelation
1	Real world	All	All	Low	Low
2	Uniform station density	All	All	High	Low
3	No urbanization	All	Shelter changes and relocations	High	Low
4	Temporal smoothing	Wyoming	All	Low	High

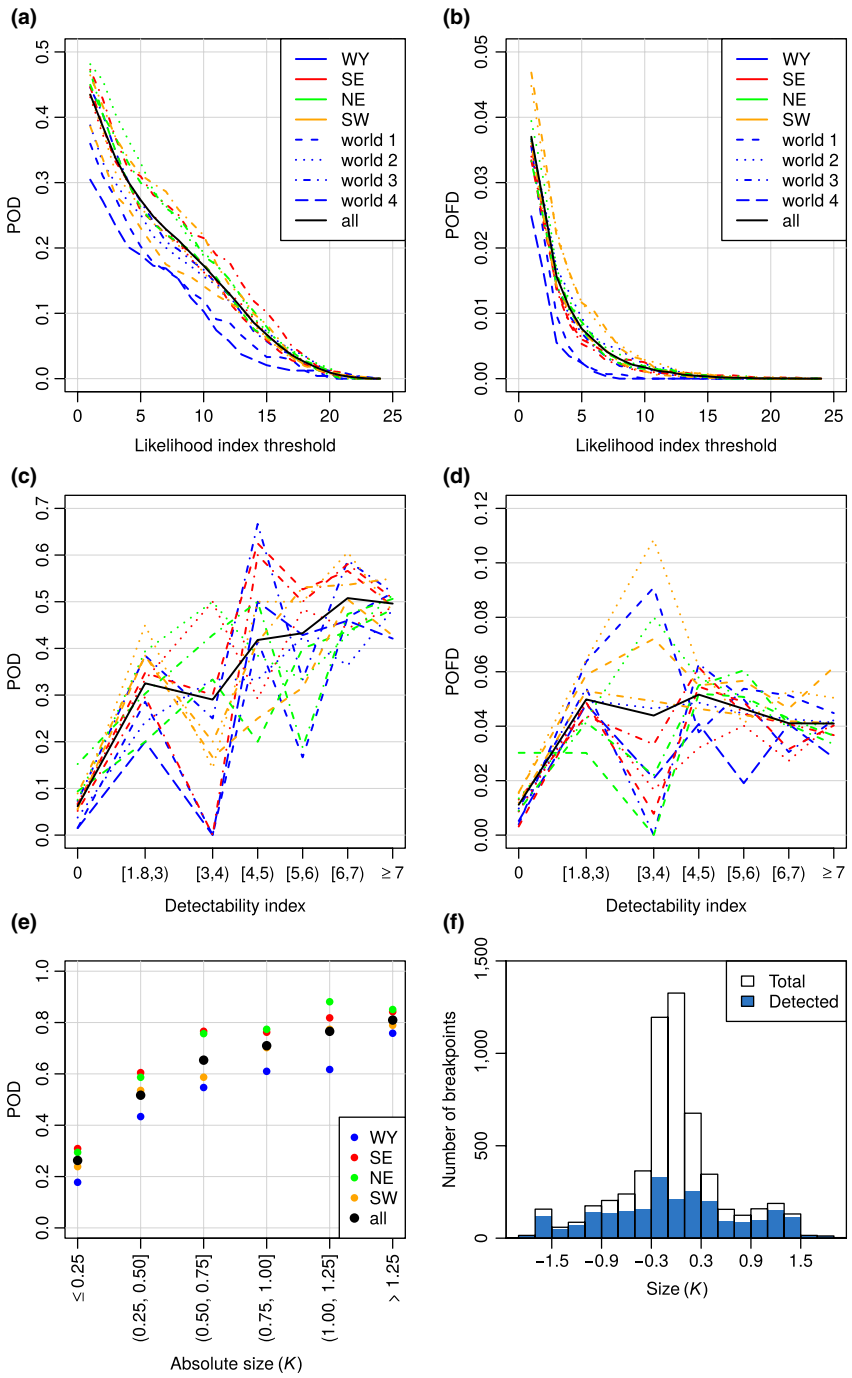

FIGURE 7 Hit rate (a) and false alarms rate (b) of the EUSTACE breakpoint detection algorithm (merged breakpoints) for the ISTI benchmark as a function of the chosen likelihood threshold. In (c) and (d), the same is shown as a function of different intervals of the detectability index. In (e), the hit rate is shown as a function of the size of the breakpoints. The distribution of the sizes is shown in (f); the blue part of the bars represents the fraction detected by the EUSTACE algorithm

TABLE 3 Length of mean homogeneous period (in years) for the relative tests (number of station years with detectability index greater or equal to 4 divided by the number of breakpoints). In parentheses, the mean homogeneous period for the absolute test (number of station years with detectability index lower than 4 divided by the number of breakpoints). For the calculation, breakpoints in consecutive years are merged into a single breakpoint

Variable	Annual	October to March	April to September	All
T_{max}	24.3 (93.4)	31.9 (231.2)	25.3 (136.1)	15.6 (62.9)
T_{min}	22.9 (69.1)	30.0 (172.9)	24.7 (101.7)	15.5 (47.6)
T_{mean}	26.2 (101.4)	34.0 (261.6)	27.8 (151.0)	16.9 (69.7)
DTR	21.7 (42.5)	28.7 (79.4)	23.8 (63.6)	14.8 (27.4)

TABLE 4 Mean homogeneous period (in years) when considering merged breakpoints (number of station years divided by the number of breakpoints) for different thresholds of the likelihood index (LI) and different intervals of the detectability index for T_{mean} (DI)

	LI ≥ 1	LI ≥ 2	LI ≥ 3	LI ≥ 5	LI ≥ 10	LI ≥ 20	Test type
DI = 0	18.4	23.6	48.8	113.9	455.0	–	Absolute
0 < DI < 4	7.4	9.7	13.6	22.5	61.1	287.6	Both
4 ≤ DI < 5	7.4	9.0	12.1	18.3	40.3	155.1	Relative
5 ≤ DI < 6	7.5	8.9	11.5	16.6	34.3	111.2	Relative
6 ≤ DI < 7	7.6	8.7	10.8	14.8	28.1	81.7	Relative
DI ≥ 7	7.4	8.4	10.0	13.2	23.7	62.6	Relative

50. Moreover, the breakpoints from metadata include only the large data gaps, because changes in the reporting resolution are not simulated.

To quantify the ability of the algorithm to detect actual breakpoints, we use the probability of detection (or hit rate):

$$POD = \frac{h}{h+m}, \quad (1)$$

where h is the number of hits (when an actual breakpoint is detected, with a tolerance of ± 1 year) and m is the number of misses (when an actual breakpoint is not detected). The probability of false detections (or false alarms rate), on the other hand, gives the frequency of false detections:

$$POFD = \frac{f}{n-h-m}, \quad (2)$$

where f is the number of false alarms (i.e. breakpoints that are given by the detection algorithm but that are not within 1 year of an actual breakpoint) and n is the number of years in the candidate series.

We allow for an error of 1 year in the position of the breakpoints because of the annual resolution of our detection; in particular, breakpoints occurring in the early part of the year are more likely assigned to the previous year (we define the year of the breakpoint as the year that ends a homogeneous sub-period).

Because an increase in the *POD* usually means an increase in the *POFD*, it is always necessary to find a compromise between the two. The advantage of defining a likelihood index for each breakpoint is that one can set a threshold to achieve the desired compromise, without performing a new detection. Setting a high minimum likelihood index allows the user to avoid false detections, but only large breakpoints can be detected; accepting breakpoints with any likelihood index maximizes the hit rate, at the price of a large number of false detections. This is illustrated by the upper panels in Figure 7, where *POD* and *POFD* obtained from the ISTI benchmark are shown for different likelihood index thresholds (here, the absolute values of the likelihood index are not representative of those provided in the EUSTACE dataset because of the reduced sets of breakpoints; see Section 2.5.2). While the increase of the *POD* with a decreasing threshold is nearly linear, the increase of the *POFD* is exponential. The lowest hit rate is found for the Wyoming subset, where the higher inter-annual variability of the continental climate decreases the signal-to-noise ratio.

In the middle panels of Figure 7, we show the impact of the detectability index on the performance of the detection. The *POD* increases on average in an approximately linear way with the detectability index; therefore, this index is indeed a good proxy for the probability of detecting a breakpoint. The index, on the other hand, does not show an influence on the *POFD* (except when only the absolute test is performed), supporting our choice of a maximum of eight reference series. The results for the individual subsets are very noisy for small values of the index because of the limited size of the samples, sometimes smaller than 100 station years (the relatively high station density in the benchmark implies that most of the series have high detectability index).

The algorithm performs best with breakpoints related to station relocations (*POD* 49%) and to shelter changes (43%), that is to abrupt changes in the station's set-up. Breakpoints caused

by urbanization have a significantly lower *POD* (32%), because they represent inhomogeneities that develop over a period of several years (usually in the form of a monotonic increase of temperature). Even if detected, such breakpoints are likely to be assigned to a much later point than the actual onset of the inhomogeneity, resulting in a miss by our definition. In fact, the results using world 3 show in general better detection scores because this world does not have urbanization. It should be noted here that the concept of an urbanization breakpoint as being one that develops over several years is a construct of the synthetic dataset – while some urbanization breakpoints in the real world may indeed take this form, others may be manifested as step changes associated with a specific development (e.g. a new building close to the observation site).

As shown by the bottom panels of Figure 7, the reason why the hit rate remains far from 100% is that the majority of the breakpoints are very small (the median of the sizes in the benchmark ranges from 0.1 K in the world 2 of the Southeast to 0.4 K in the world 4 of Wyoming). For sizes larger than 1 K, the hit rate approaches 80% (this value would probably be higher if four variables were used). Hence, about 20% of large breakpoints remain undetected, mostly because they are located close to the beginning or the end of a series, or close to another breakpoint (due to the annual resolution, we cannot detect breakpoints in two consecutive years). A lack of reference series is another reason for misses of large breakpoints, as shown in panel (c) of the figure.

2.5.3 | Statistics of breakpoints

Table 3 shows the length of mean homogeneous periods for each variable and season. For comparison, we also show the length of mean homogeneous periods for the absolute test.

Our results are qualitatively in agreement with those obtained with the Pairwise Homogeneity Assessment algorithm (Thorne *et al.*, 2016); in particular, we find more frequent breakpoints (i.e. shorter homogeneous periods) for *DTR* than for the other variables. The seasonal results are clearly biased towards the Northern Hemisphere, where the October–March semester has fewer breakpoints because of the higher interannual variability. We find considerably more breakpoints by combining the three temporal aggregations than using annual means alone.

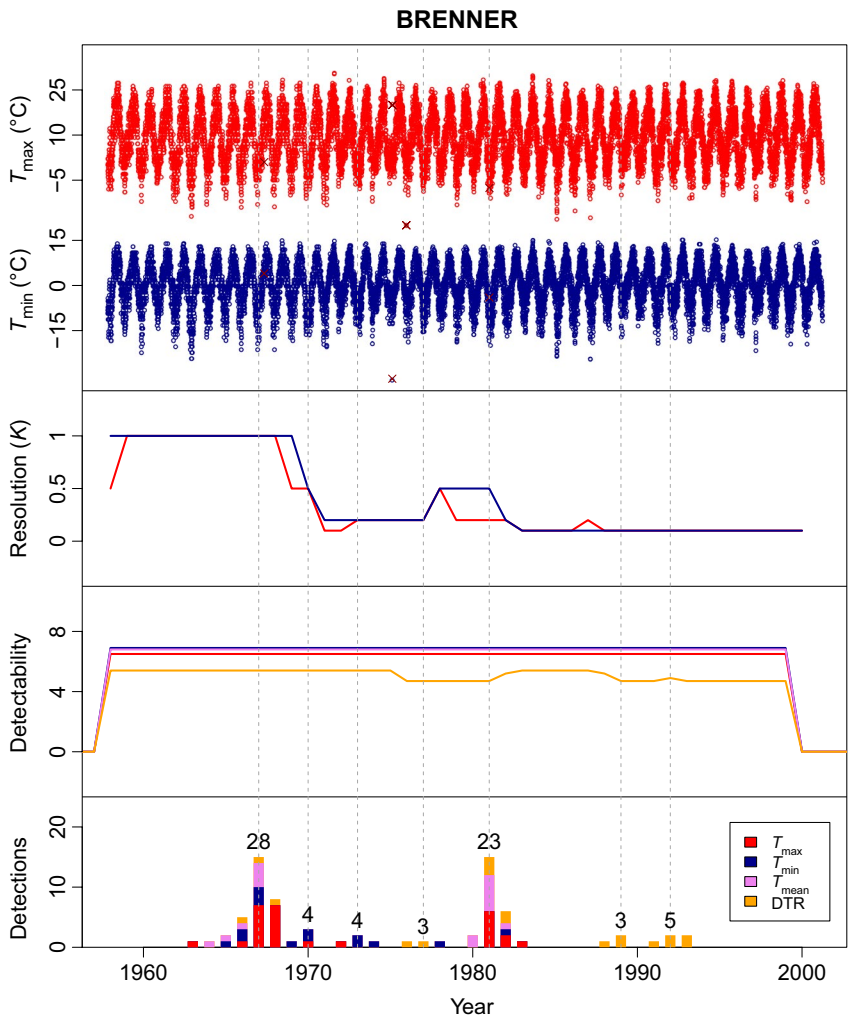

FIGURE 8 Overview of the information provided for the station of Brenner, Austria (latitude 47.0° N, longitude 11.5° E, elevation 1,450 m; source: ECA&D-TX_SOUID103851). From top to bottom: raw daily T_{max} series, raw daily T_{min} series, reporting resolution, annual detectability index and number of detections. Crossed values in the raw data series indicate observations flagged by the quality control. Vertical dashed lines indicate the position of the merged breakpoints, and the numbers on top of the bars indicate their respective likelihood index

The absolute test, as expected, has a very low detection rate. Interestingly, it can detect about one-third more breakpoints for T_{min} than for T_{max}. This difference is very consistent across the different temporal aggregations and is not found in the relative tests. It might be an indication that the sizes of the breakpoints are generally larger for T_{min} in the early years (when the absolute test is used more frequently).

Table 4 shows the length of mean homogeneous periods for the merged breakpoints for different thresholds of likelihood index and various intervals of the detectability index. The merged breakpoints are by definition more frequent than the breakpoints found for the individual variables; on average, they are to be found every 8.8 years. With ideal reference series (detectability index of 7 or larger), the mean homogeneous period reaches 7.4 years, with the most 'significant' breakpoints (likelihood index of 20 or larger) found every 62.6 years.

The results indicate that the impact of the detectability index on the hit rate of the relative tests is small for low thresholds of the likelihood index, implying good consistency of the breakpoint detection across series, but it becomes more and more important for higher thresholds. Therefore, a meaningful selection of breakpoints based on a likelihood index threshold requires that the detectability index does not vary too much within the analysed data.

3 | DATASET LOCATION AND FORMAT

The EUSTACE global land station daily air temperature dataset is archived at the Centre for Environmental Data Analysis (CEDA) and is stored in annual NetCDF files (one additional NetCDF file provides the information on breakpoints and reporting resolution). A user guide is also provided. The data are offered under a non-commercial licence.

4 | DATASET USE

4.1 | Examples

Figures 8–10 show an overview of the information available to the user for three selected stations, each giving an example of the issues that can be found in the data.

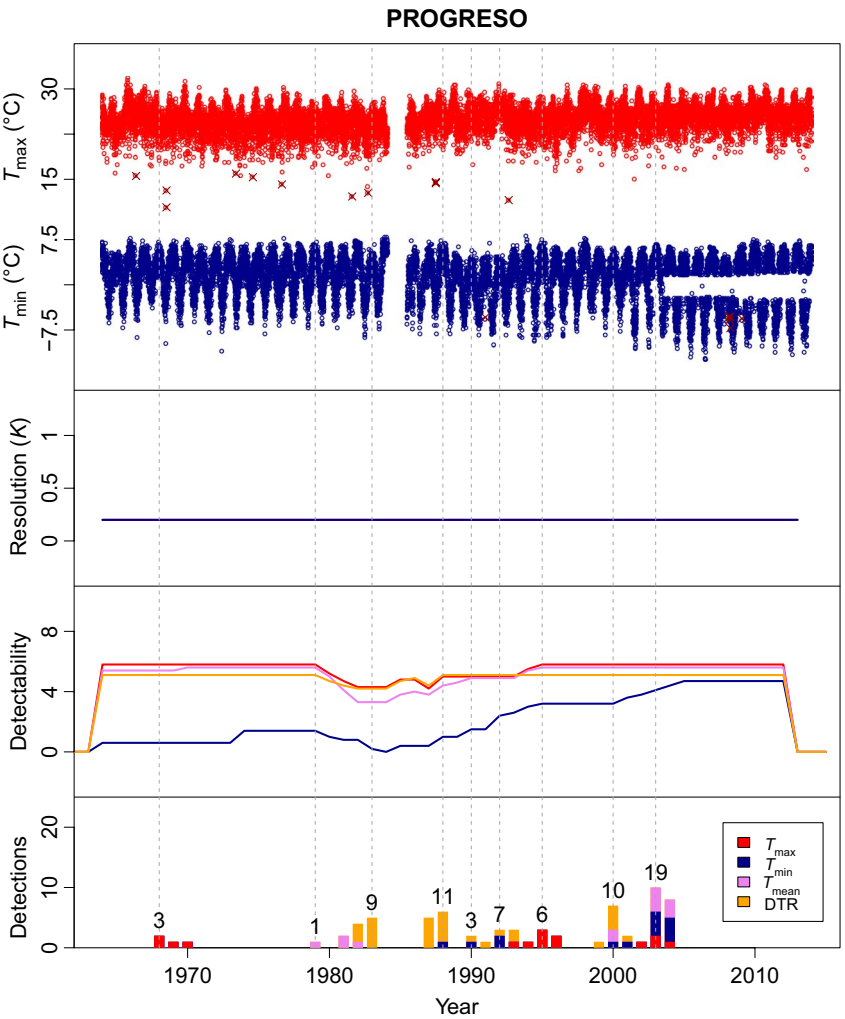

FIGURE 9 Same as Figure 8 for the station of Progreso, Peru (latitude 14.671806° S, longitude 70.367775° W, elevation 3,925 m; source: DECADE-PEPU030614)

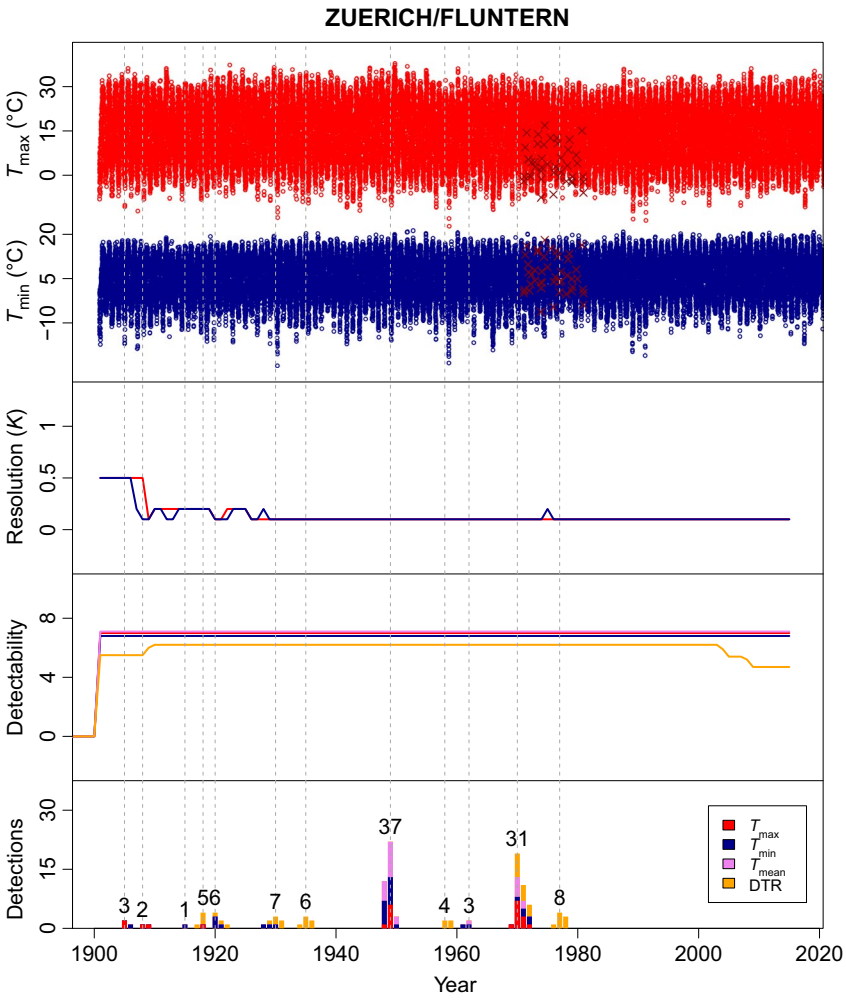

FIGURE 10 Same as Figure 8 for the station of Zurich/Fluntern, Switzerland (latitude 47.383053° N, longitude 8.566667° E, elevation 555 m; source: ECA&D-TX_SOUID100758)

The first example (Brenner, Austria) is for a station in central Europe with about 40 years of data and nearly optimal reference series, represented by a detectability index close to 8 over the whole period covered by the series. The detectability index is slightly lower for *DTR* because spatial correlation of this variable is usually lower than that of the other variables. However, the detectability index does not take into account that breakpoints in *DTR* have in general a larger amplitude (Zhang *et al*., 2011; Thorne *et al*., 2016).

The data of this station have two major breakpoints in 1967 and 1981, highlighted by their high likelihood index (28 and 23, respectively). The first large breakpoint is detected in all variables, while the second one seems to significantly affect only T_{max} (the only detection for T_{min} is caused by a change in the reporting resolution). The other five breakpoints are found in a much lower number of sets, predominantly in the *DTR* sets. A similar analysis of the sets can also be done by isolating the temporal aggregations instead of the variables (not shown).

Note that the reporting resolution until 1969 was of 1 K and only after 1981 did it reach the modern standard of 0.1 K; this is an important piece of information when analysing trends in extremes, since changes in resolution can introduce inhomogeneities in many commonly used indices (see Rhines *et al*., 2015).

The second example (Progreso, Peru) is for a station on the Andean Plateau. This region has a particularly low spatial correlation for T_{min} (Hunziker *et al*., 2018), which negatively affects the detectability index for that variable. Nevertheless, the main breakpoint in T_{min} (2003) is correctly detected. This inhomogeneity is caused by a new observer who was not trained in the use of the minimum thermometer, and it could be easily corrected if properly detected and attributed (see Hunziker *et al*., 2017, for more details).

For the third example (Zurich/Fluntern, Switzerland), we have metadata on the history of the station (Kuglitsch *et al*., 2012): the station was relocated in 1949 and a new screen was installed in 1971. These two inhomogeneities are prominent in the detections plot with likelihood index of 37 and 31, respectively. The quality flags between 1971 and 1980 are all caused by the temporal consistency test: in that decade, the Swiss NWS used a different time window to calculate T_{max}, which caused some days to have T_{max} lower than T_{min} of the previous day. This kind of information is usually not provided

by the data sources; therefore, quality control and breakpoint detection are often the only instruments available to ensure that daily observations are consistently defined.

4.2 | Case study: Impact of breakpoints on trends in South America

The likelihood index provides a metric that allows ranking of the breakpoints according to their significance. As shown in Section 4.1, breakpoints with high likelihood index usually correspond to the events in the station history that have the largest impact on data homogeneity. For many applications, breakpoints with low likelihood can be neglected, resulting in longer, 'quasi-homogeneous' sub-series.

To illustrate this in more detail, we calculated trends in warm days, that is those days when T_{max} is above the 90th percentile following the definition for 'TX90p' by the Expert Team on Climate Change Detection and Indices (ETCCDI) (see Zhang *et al*., 2011). We focus on South America, where the data availability in the EUSTACE dataset has improved the most with respect to other global datasets.

Figure 11 shows the mean detectability index for each T_{max} series over the period 1981–2010. The detection is not possible when a series has <10 years of data (see Section 2.5); in that case, the detectability index is zero, as it is when <3 reference series are found. In most of the central and eastern parts of the continent, station density is high enough for the relative tests to be applied to most of the series that have enough data, resulting in high detectability index. In the north-western part, only the absolute test could be applied; here, trends should be analysed only after a detailed inspection of the data.

Figure 12 shows the trends over the period 1981–2010 for all series that have a mean detectability index >4 (equivalent to five reference series with a correlation coefficient of 0.8) and for the quasi-homogeneous series among them defined by three different thresholds of the likelihood index. The lower the threshold, the larger the number of series that are considered inhomogeneous. We require at least 80% of the period to be covered by data, and we exclude observations with a reporting resolution coarser than 0.2 K.

When homogeneity is not taken into account, a few outliers in the trends are scattered over the region. Nevertheless, a meridional gradient is already visible (weaker trends in the southern part of the continent). After excluding the series that have at least one breakpoint with a likelihood index of 20 or greater, the trends become spatially more coherent because most outliers caused by inhomogeneous series are removed. Lower thresholds further increase this coherency and guarantee a better consistency of the breakpoint detection among the series (see Table 4), at the cost of reducing the spatial coverage. Less than 20% of the series are homogeneous when applying a threshold of 5 for the likelihood index. Even by a threshold of 10, none of the numerous series on the Andean

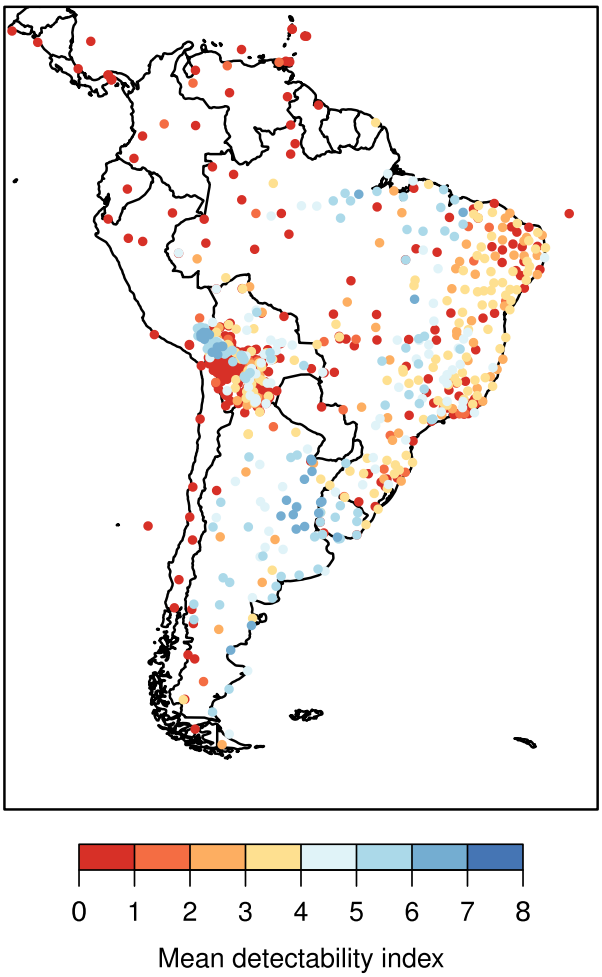

FIGURE 11 Mean detectability index for T_{max} in South America over the period 1981–2010. Series with larger index are plotted on the foreground

Plateau is homogeneous. Hunziker *et al*. (2018) found that data quality in that region is particularly poor and estimated one breakpoint per decade on average using a semi-automatic detection software.

5 | SUMMARY

The EUSTACE project has built on previous initiatives to assemble a large global collection of daily maximum and minimum air temperature series from land stations for the period 1850–2015. The data series were selected to ensure uniqueness and consistency, despite the intrinsic heterogeneity caused by the highly fragmented management of the station networks.

Data quality was assessed through a set of automatic routines that assign 14 different flags. This does not guarantee that all erroneous data were detected and still requires some expertise in the interpretation of the flags. An estimation of the reporting resolution of the observations is also provided,

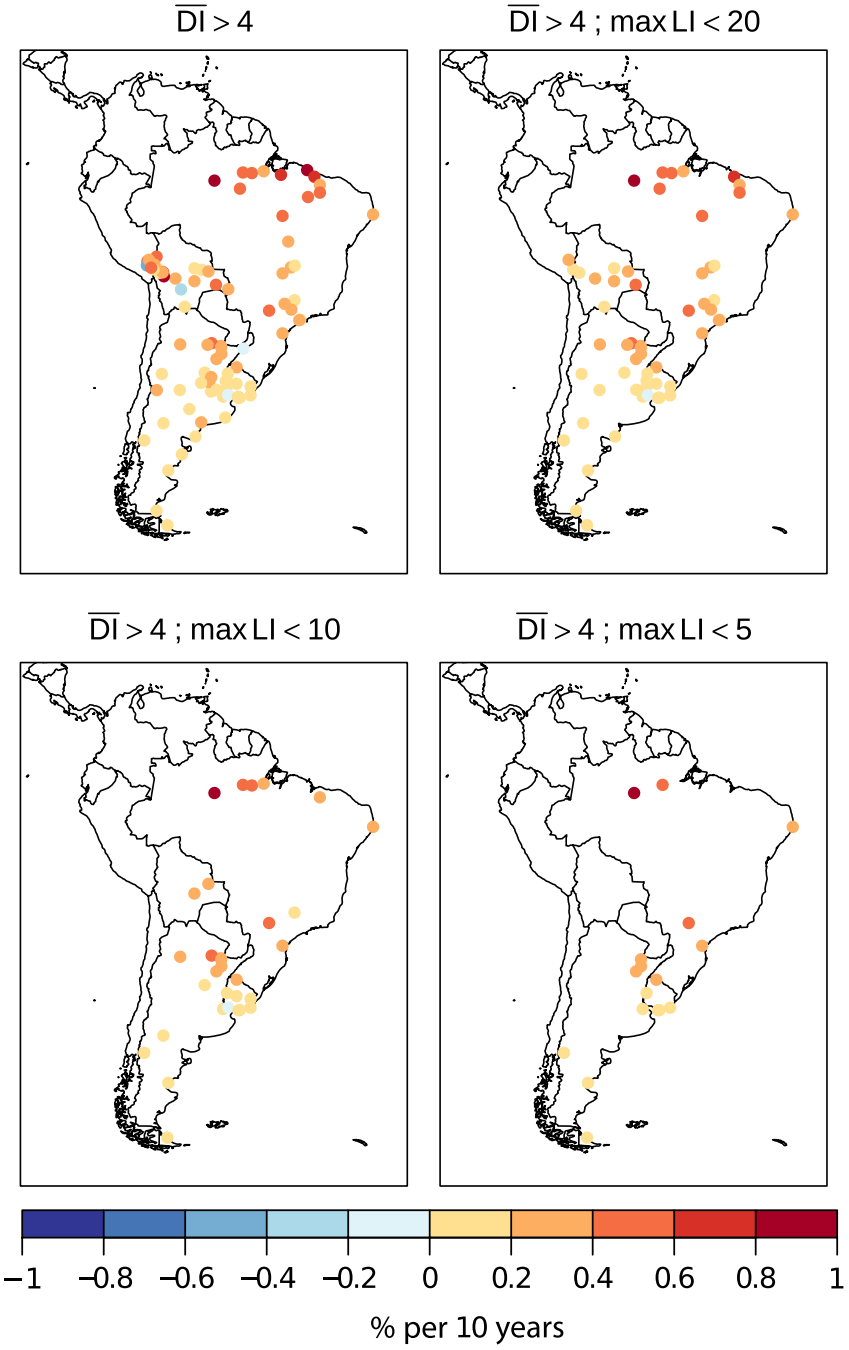

FIGURE 12 Trends in the number of warm days (TX90p) in South America over the period 1981–2010 in all T_{max} series with mean detectability index (DI) larger than 4 and in those that do not have merged breakpoints with a likelihood index (LI) equal to or larger than 20, 10 or 5

which can facilitate further processing of the data needed before calculating indices based on percentiles. Coordinates were tested against a digital elevation model and a land/sea mask and large discrepancies were inspected manually, leading in many cases to corrections or to an estimation of missing elevations.

A new algorithm was developed to assess temporal homogeneity of the data series, by combining the results of four breakpoint detection methods. Instead of providing a single set of breakpoints for each station, up to 50 different sets are provided representing different combinations of methods, variables and seasons, together with a recommended merged set. This gives flexibility to the users, who can employ the sets that best fit their specific needs. Even though the magnitude of the detected inhomogeneity was not estimated, the agreement between the sets provides an alternative metric to define the relevance of each breakpoint. An index representing the probability of detection for each year of each series, based on the availability of reference series, is also provided.

The performance of the breakpoint detection algorithm was assessed by using a state-of-the-art dataset of synthetic temperature series. It was shown that large breakpoints can be successfully detected in about 80% of the cases when

using only one variable. One drawback of the algorithm is the annual resolution of the breakpoints, which does not allow us to distinguish between breakpoints occurring in consecutive years.

About 3% of the data series presented here did not have an open licence, meaning that the authors are not allowed to provide them directly to the public. These data (and their quality flags) are replaced by missing values in the public version of the dataset. In many cases, these series can be obtained free of charge from the original data providers (inquiries can be addressed to the relevant sources listed in Table 1). All the remaining information, such as the breakpoints, is available without limitations.

Updates to the EUSTACE land station dataset are not planned. Data for the latest years can be obtained from the underlying sources (Table 1; note that the station identifiers adopted in the EUSTACE dataset are the same used in the sources). One must be aware, however, that new inhomogeneities would be likely introduced.

ACKNOWLEDGEMENTS

EUSTACE has received funding from the European Union's Horizon 2020 Programme for Research and Innovation, under Grant Agreement No. 640171. We thank Renate Auchmann, Andrea Toreti, Yang Feng and Matthew Menne for providing fundamental code modules used for the breakpoint detection and the quality control. We also thank Rachel Killick for her assistance with the daily temperature benchmark and Nick Rayner for her comments on an earlier version of the manuscript. The trends of daily temperature indices were calculated using the R package *climdex.pcic* provided by the Pacific Climate Impacts Consortium. Data were provided by the NOAA National Centers for Environmental Information, the Royal Netherlands Meteorological Institute (KNMI), the NWS of Argentina (Servicio Meteorológico Nacional), the ERA-CLIM and DECADE projects and the International Surface Temperature Initiative (www.surfacetemperatures.org).

OPEN PRACTICES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.5285/7925ded722d743fa8259a93acc7073f2 Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

REFERENCES

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, 389–400. Available at: https://doi.org/10.1002/joc.1524

Brohan, P., Kennedy, J.J., Harris, I., Tett, S.F. and Jones, P.D. (2006). Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. *Journal of Geophysical Research: Atmospheres*, 111, 12106. Available at: https://doi.org/10.1029/2005JD006548

Brönnimann, S. and Wintzer, J. (2018). Use imprint of society and history on climate data to inform climate services. *Nature*, 554, 423. Available at: https://doi.org/10.1038/d41586-018-02201-z

Brugnara, Y., Auchmann, R., Brönnimann, S., Bozzo, A., Berro, D.C. and Mercalli, L. (2016). Trends of mean and extreme temperature indices since 1874 at low-elevation sites in the Southern Alps. *Journal of Geophysical Research: Atmospheres*, 121, 3304–3325. Available at: https://doi.org/10.1002/2015JD024582

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

Dunn, R.J., Willett, K.M., Thorne, P.W., Woolley, E.V., Durre, I., Dai, A. *et al*. (2012). HadISD: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, 8, 1649–1679. Available at: https://doi.org/10.5194/cp-8-1649-2012

Dunn, R., Willett, K., Morice, C. and Parker, D. (2014). Pairwise homogeneity assessment of HadISD. *Climate of the Past*, 10, 1501. Available at: https://doi.org/10.5194/cp-10-1501-2014

Durre, I., Menne, M.J., Gleason, B.E., Houston, T.G. and Vose, R.S. (2010). Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49, 1615–1633. Available at: https://doi.org/10.1175/2010JAMC2375.1

Farr, T.G. and Kobrick, M. (2000). Shuttle Radar Topography Mission produces a wealth of data. *Eos, Transactions American Geophysical Union*, 81, 583–585. Available at: https://doi.org/10.1029/EO081i048p00583

Hunziker, S., Gubler, S., Calle, J., Moreno, I., Andrade, M., Velarde, F. *et al*. (2017). Identifying, attributing, and overcoming common data quality issues of manned station observations. *International Journal of Climatology*, 37, 4131–4145. Available at: https://doi.org/10.1002/joc.5037

Hunziker, S., Brönnimann, S., Calle, J., Moreno, I., Andrade, M., Ticona, L. *et al*. (2018). Effects of undetected data quality issues on climatological analyses. *Climate of the Past*, 14, 1–20. Available at: https://doi.org/10.5194/cp-14-1-2018

Killick, R. (2016) Benchmarking the performance of homogenisation algorithms on daily temperature data. PhD thesis, University of Exeter, UK.

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

Kuglitsch, F., Auchmann, R., Bleisch, R., Brönnimann, S., Martius, O. and Stewart, M. (2012). Break detection of annual Swiss temperature series. *Journal of Geophysical Research*, 117. Available at: https://doi.org/10.1029/2012jd017729

Lawrimore, J.H., Menne, M.J., Gleason, B.E., Williams, C.N., Wuertz, D.B., Vose, R.S. *et al*. (2011). An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3.

Journal of Geophysical Research: Atmospheres, 116. Available at: https://doi.org/10.1029/2011jd016187

Menne, M.J. and Williams, C.N. Jr (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22, 1700–1717. Available at: https://doi.org/10.1175/2008JCLI2263.1

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29, 897–910. Available at: https://doi.org/10.1175/JTECH-D-11-00103.1

Rennie, J.J., Lawrimore, J.H., Gleason, B.E., Thorne, P.W., Morice, C.P., Menne, M.J. *et al*. (2014). The International Surface Temperature Initiative global land surface databank: monthly temperature data release description and methods. *Geoscience Data Journal*, 1, 75–102. Available at: https://doi.org/10.1002/gdj3.8

Rhines, A., Tingley, M.P., McKinnon, K.A. and Huybers, P. (2015). Decoding the precision of historical temperature observations. *Quarterly Journal of the Royal Meteorological Society*, 141, 2923–2933. Available at: https://doi.org/10.1002/qj.2612

Rodríguez, E., Morris, C., Belz, J., Chapin, E., Martin, J., Daffer, W. *et al*. (2005). An assessment of the SRTM topographic products (Technical Report JPL D-31639). Pasadena, California: Jet Propulsion Laboratory.

Squintu, A.A., van der Schrier, G., Brugnara, Y. and Klein Tank, A. (2019). Homogenization of daily temperature series in the European Climate Assessment & Dataset. *International Journal of Climatology*, 39, 1243–1261. Available at: https://doi.org/10.1002/joc.5874

Stickler, A., Brönnimann, S., Valente, M.A., Bethke, J., Sterin, A., Jourdain, S. *et al*. (2014). ERA-CLIM: historical surface and upper-air data for future reanalyses. *Bulletin of the American Meteorological Society*, 95, 1419–1430. Available at: https://doi.org/10.1175/BAMS-D-13-00147.1

Thorne, P., Menne, M., Williams, C., Rennie, J., Lawrimore, J., Vose, R. *et al*. (2016). Reassessing changes in diurnal temperature range: a new dataset and characterization of data biases. *Journal of Geophysical Research: Atmospheres*, 121, 5115–5137. Available at: https://doi.org/10.1002/2015JD024583

Thorne, P.W., Allan, R.J., Ashcroft, L., Brohan, P., Dunn, R.H., Menne, M.J. *et al*. (2017). Toward an integrated set of surface meteorological observations for climate science and applications. *Bulletin of the American Meteorological Society*, 98, 2689–2702. Available at: https://doi.org/10.1175/BAMS-D-16-0165.1

Toreti, A., Kuglitsch, F.G., Xoplaki, E. and Luterbacher, J. (2012). A novel approach for the detection of inhomogeneities affecting climate time series. *Journal of Applied Meteorology and Climatology*, 51, 317–326. Available at: https://doi.org/10.1175/JAMC-D-10-05033.1

Trewin, B. (2018) The Australian Climate Observations Reference Network - surface air temperature (ACORN-SAT) version 2, Bureau of Meteorology Research Report 32, Bureau of Meteorology, Australia. Available at: <http://www.bom.gov.au/research/publications/researchreports/BRR-032.pdf> [Accessed 15 May 2019].

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J., Domonkos, P. *et al*. (2012). Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8, 89–115. Available at: https://doi.org/10.5194/cp-8-89-2012

Wang, X.L. (2008). Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or F test. *Journal of Applied Meteorology and Climatology*, 47, 2423–2444. Available at: https://doi.org/10.1175/2008JAMC1741.1

Willett, K., Williams, C., Jolliffe, I., Lund, R., Alexander, L., Brönnimann, S. *et al*. (2014). A framework for bench marking of homogenisation algorithm performance on the global scale. *Geoscientific Instrumentation, Methods and Data Systems*, 3, 187–200. Available at: https://doi.org/10.5194/gi-3-187-2014

Zhang, X., Alexander, L., Hegerl, G.C., Jones, P., Tank, A.K., Peterson, T.C. *et al*. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, 2, 851–870. Available at: https://doi.org/10.1002/wcc.147

How to cite this article: Brugnara Y, Good E, Squintu AA, van der Schrier G, Brönnimann S. The EUSTACE global land station daily air temperature dataset. *Geosci Data J*. 2019;6:189–204. https://doi.org/10.1002/gdj3.81