EUSTACE

Horizon 2020

H2020-EO-2014 New ideas for Earth-relevant Space Applications

EUSTACE

(Grant Agreement 640171)

EU Surface Temperature for All Corners of Earth

Deliverable D1.7

Global LSAT dataset with discontinuities identified where possible

Deliverable Title	*Global LSAT dataset with discontinuities identified where possible*	
Brief Description	*Three break detection methods will be applied to the global GHCN-Daily data, which will allow assessment of regions for which the break detection methodology works. This assessment will be reported to WP2 and WP3. GHCD-Daily data with break points identified where feasible, together with an estimate of uncertainty will be delivered (Task 1.2).*	
WP number	1	
Lead Beneficiary	*University of Bern*	
Contributors	*Yuri Brugnara, University of Bern* *Renate Auchmann, University of Bern* *Stefan Brönnimann, University of Bern*	
Creation Date Version Number Version Date	8 February 2017 1 8 February 2017	
Deliverable Due Date Actual Delivery Date	31 December 2016 Data delivered to project 14 December 2016, report submitted to portal 6 March 2017	
Nature of the Deliverable		*R - Report*
		DEM – Demonstrator, Pilot, Prototype
		DEC – Dissemination, Exploitation or Communication
	X	*O - Other*
Dissemination Level/ Audience	X	*PU - Public*
		CO - Confidential, only for members of the consortium, including the Commission services

Version	**Date**	**Modified by**	**Comments**
1	08/02/17	Y. Brugnara	First draft
2	01/03/17	Y. Brugnara	First revision
3	30/04/18	Y. Brugnara	Updates to the files format and the break detection output

EUSTACE

Contents

1. Executive Summary

The aim of this deliverable is to provide information on the homogeneity of all available station-based daily air temperature series (period 1850-2015). In detecting non-climatic breaks in more stations than just those available in GHCN-D, we have exceeded our expectations. This involved considerable pre-processing of the data, which will be only briefly described in this report. The analysis was extended to monthly data to address the problem for later activities of EUSTACE of large areas having few or no observations available with daily resolution, particularly before the 1950s.

The station data will be made publicly available once product verification steps have been undertaken.

2. Project Objectives

With this deliverable, the project has contributed to the achievement of the following objectives (DOA, Section B1.1):

No.	Objective	Yes	No
1	Intensively develop the hitherto immature use of Earth Observation estimates of Earth's surface **skin** temperature to enable new Climate Data Records of the surface **air** temperature Essential Climate Variable (ECV) to be created, for all locations over all surfaces of Earth (i.e. land, ocean, ice and lakes), for every day since 1850. EUSTACE will achieve this by: combining information estimated from multiple satellites with surface air temperature measurements made *in situ* and **creating complete analyses** of surface air temperature, through the application of novel statistical in-filling methods.	X	
2	Integrate these new daily surface air temperature Climate Data Records into a range of applications in Earth System Science and Climate Services and research, amongst others. EUSTACE will achieve this via the active and continuous engagement of trail-blazer users, and the provision of products through already-existing user community data portals and service mechanisms, in standard formats.		X
3	Undertake and report detailed research into the relationships between surface skin temperature estimated from Earth Observation satellite measurements and surface air temperature observed *in situ* by conventional measurements, over all surfaces of the Earth, including the polar regions. This is likely to provide information useful for refining coupling in Earth system models.		X

4	Create a sustainable, automated system at an appropriate level of maturity for the potential production of the products beyond the lifetime of the project. To enable this, EUSTACE will also identify Earth Observation and conventional data streams that could be used to update the surface air temperature Climate Data Records in the future, including those from Sentinel missions.		X
5	Extensively validate the new surface air temperature Climate Data Records against independent, surface-based reference data, sourced by the project for this purpose.	X	
6	Develop and report new, consistent, validated estimates of uncertainty both in already-existing Earth Observation surface skin temperature estimates and in the new surface air temperature Climate Data Records, at all locations and times across the Earth's surface.		X
7	Develop links with related activities within Europe and beyond to help to ensure the execution of a joined-up work programme, the Copernicus Services and to enable the provision of requirements for the future surface skin temperature and surface air temperature observing system.		X
8	Other – not directly linked to one of the above objectives	X	

3. Detailed Report

Our primary aim was to create a global data set of daily station measurements of maximum and minimum temperature where non-climatic discontinuities arising, for example, from station moves were identified. In some regions, few daily data are available, so in order to constrain the later EUSTACE analyses in those regions, monthly station series (of which there are many more) were also run through the "break detection" process. Section 3.1 summarises the processing of the daily data and Section 3.2 covers the monthly data.

3.1 Daily Data

The data referred to here are daily maximum and minimum temperatures (hereafter Tmax and Tmin). How these were exactly measured, and how a day is defined, is often unknown. When any information on measurement method and/or day definition was available, it has been retained in the final output.

3.1.1 Data Sources

We used seven data sources:

1. GHCN-Daily version 3.22 (Menne et al, 2012)
2. ISTI version 1.00 (Rennie et al., 2014)
3. ECA&D (non-blended, updated October 2016) (Klein Tank et al., 2002)
4. ERA-CLIM and ERA-CLIM 2 projects (Stickler et al., 2014)
5. DECADE project (http://www.geography.unibe.ch/research/climatology_group/research_projects/decade/index_eng.html)
6. Homogenised series from Brugnara et al. (2016)
7. Data provided to EUSTACE by the national weather service of Argentina (SMN)

Not all data are freely re-distributable. For instance, this concerns part of the DECADE dataset (owned by SENAMHI Peru and SENAMHI Bolivia), whose policy is not completely clear yet, as this project is still under way. With its end in late 2017, data policy constraints are expected to relax (DECADE data should be publicly released then). Also part of ECA&D and the data from SMN are not re-distributable.

3.1.2 Pre-processing

The following steps were necessary before a breakpoint detection algorithm could be applied:

- all data were converted into a common format (ASCII);
- duplicate data were removed; and
- quality control was applied (Durre et al., 2010).

In addition, data resolution was estimated for each year of each series, since data are often converted and/or rounded (even multiple times) before being released by their provider. For example, temperature in most north-American stations was measured with a resolution of 1°F. This has to be taken into account when using the data as it has an impact on uncertainty.

3.1.3 Breakpoint detection

A hybrid R/Fortran-based software was developed specifically for EUSTACE to apply three independent statistical tests (see Kuglitsch et al., 2012; also EUSTACE Deliverable 1.4 and references therein) in a fully automatic way.

Each test gives a list of years with potential discontinuities (or breakpoints), based on the analysis of standardized differences with up to eight reference series from nearby stations. The tests were applied separately to Tmax, Tmin, Tmean=(Tmax+Tmin)/2, and DTR=Tmax-Tmin, averaged into three temporal aggregations (annual means, October-March means, and April-September means). After discussion with colleagues from EUSTACE WP2 (dataset construction), a finer temporal resolution of the breakpoints (e.g., monthly) was considered not useful.

The combination of statistical tests, variables, and temporal aggregations results in up to 36 sets of breakpoints for each station, which are merged by taking local maxima in the number of detections as the most probable positions of the merged breakpoints.

Each merged breakpoint is provided with a "likelihood" index, indicating how many sets contributed to it. The higher the index, the more confident we are on the breakpoint. Each year of each series was assigned a "detectability" index (from 0 to 8), defined as the sum of the correlation coefficients of the reference series available in that year; this index is proportional to the probability of the detection (i.e., the probability that a real breakpoint is detected). If the number of available reference stations is lower than three, the detection by means of reference series is considered not possible (in this case the index is zero).

An additional test (Wang, 2008) was performed for any series with a low detection score in any year, i.e. those series with few well-correlated neighbours. This is an absolute test, i.e. without reference series, hence with a higher false alarm rate. Breakpoints detected with this test are marked by a specific flag in the final product to indicate their higher uncertainty.

A third type of breakpoint was obtained from metadata (when available) or from certain features of the data: a change of data source, of reporting resolution, or a large data gap were all considered potential breakpoints. These breakpoints also have a specific flag.

The breakpoints from the absolute test, from changes in source or resolution, and from gaps, all contribute to the final merged set of breakpoints (i.e., they are additional sets on top of the 36 "regular" sets).

3.1.4 Post-processing

The entire data set, together with the information on breakpoints and data resolution, was converted into the NetCDF format previously agreed within EUSTACE. This includes 17 different flag values for data quality, 40 for variable definition, and 7 for data source (see Appendix A for details). A total of 130'950 merged breakpoints were detected in 35'375 stations. Figure 1 shows the geographical distribution of the stations.

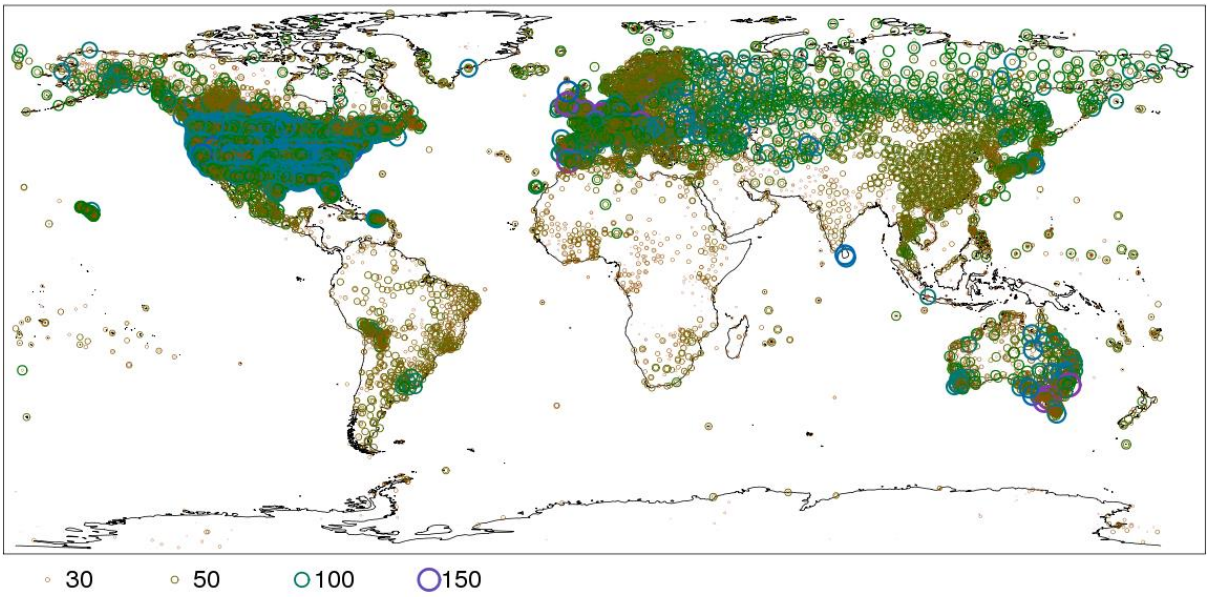

Figure 1. Station distribution and length of Tmax series (in years; size and colour – ranging from brown to purple – of the circles are proportional to length) for daily data in the EUSTACE data set.

3.2 Monthly Data

The data referred to here are the monthly averages of Tmax and Tmin, and a not-well-defined monthly mean temperature. The latter covers longer periods, particularly in the 19th century. Note that often the monthly averages of Tmax and Tmin are available for stations for which daily data are not available.

3.2.1 Data source

Monthly data has a unique source, the ISTI merged dataset version 1.1 (stage 3, Rennie et al., 2014).

3.2.2 Pre-processing

Since there is only one source, which previously underwent a merging process (Rennie et al., 2014), the only pre-processing needed was the quality control. For monthly data, we applied basic automatic quality tests (deviations from climatology, series of identical values, internal consistency).

3.2.3 Breakpoint detection

The procedure for the breakpoint detection is identical to that applied to daily data, except that mean temperature is not always calculated from Tmax and Tmin. The points where two series are known to have been merged into one (Rennie et al., 2014) are considered breakpoints and flagged accordingly.

3.2.4 Post-processing

The format of the final product is analogous to that of the daily data. However, for monthly data, only two flag values are used for data quality ("ok" and "suspect") and there is no

information given on definitions and resolution. A total of 151’999 merged breakpoints were detected over 26'070 stations. Figure 2 shows the geographical distribution of the stations.

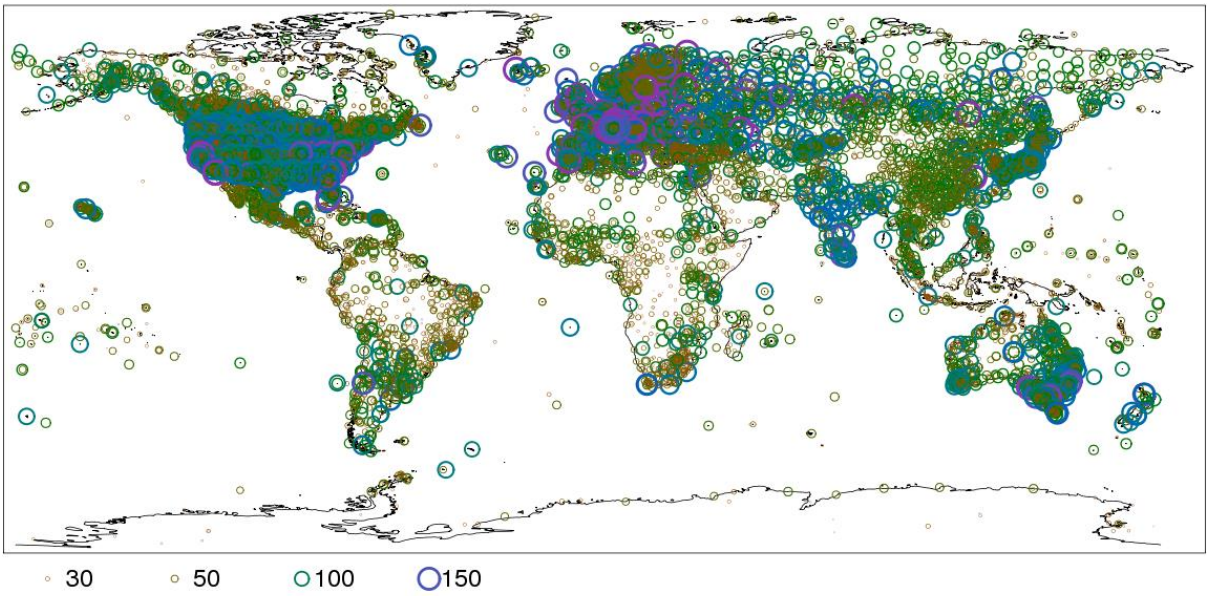

Figure 2. Station distribution and length of monthly mean temperature series (in years).

3.3 Data provision

The station data will be made publicly available via the CEDA archive once product verification steps have been undertaken.

References

Brugnara, Y., Auchmann, R., Brönnimann, S., Bozzo, A., Berro, D. C., & 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(7), 3304-3325.

Brugnara, Y., Squintu, A., and van der Schrier, G (2016). Report on Homogenised daily Land Surface Air Temperature. EUSTACE Deliverable 1.4

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., & Vose, R. S. (2010). Comprehensive automated quality assurance of daily surface observations. Journal of Applied Meteorology and Climatology, 49(8), 1615-1633.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An overview of the global historical climatology network-daily database. Journal of Atmospheric and Oceanic Technology, 29(7), 897-910.

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., 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(12), 1441-1453.

Kuglitsch, F. G., Auchmann, R., Bleisch, R., Brönnimann, S., Martius, O., & Stewart, M. (2012). Break detection of annual Swiss temperature series. Journal of Geophysical Research: Atmospheres, 117(D13).

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(2), 75-102.

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(9), 1419-1430.

Wang, X. L. (2008). Penalized maximal F test for detecting undocumented mean shift without trend change. Journal of Atmospheric and Oceanic Technology, 25(3), 368-384.

Appendix A: file contents

Detailed information on variables and flags for data quality and data source in the EUSTACE global daily station data set. Daily temperature data (observations and quality flags) are provided in one file per year, while "status" data (information on homogeneity and resolution) are contained in a single file.

A.1 Temperature files

DIMENSIONS

time	days since 01-01-1850
station	station progressive number (starting from 1, identical for every year)
name_strlen	length of station name (max 50 characters)
code_strlen	length of station code (max 50 characters)

VARIABLES

station_name (station, name_strlen)	Name of the station (not necessarily unique)
station_code (station, code_strlen)	Official station ID from source (unique)
latitude (station)	Latitude in degrees North
longitude (station)	Longitude in degrees East
elevation (station)	Elevation above mean sea level in m
data_source (station)	Source ID: 0. GHCN-Daily version 3.22 1. ISTI version 1.00 2. ECA&D (non-blended, updated October 2016) 3. ERA-CLIM and ERA-CLIM 2 projects 4. DECADE project 5. Homogenised series from Brugnara et al. (2016) 6. Data provided by national weather service of Argentina (SMN)
data_policy (station)	Is the data re-distributable? (0=Yes, 1=No)
tasmax (time, station)	Daily Tmax in Kelvin
tasmin (time, station)	Daily Tmin in Kelvin

tasmax_qc (time, station)	QC flag, the following options are possible: **0 (ok):** Did not fail any quality assurance test **1-14 (capital letters):** QC flags as defined in GHCN-D (see Durre et al., 2010): D = failed duplicate check G = failed gap check I = failed internal consistency check K = failed streak/frequent-value check L = failed check on length of multiday period M = failed megaconsistency check N = failed naught check O = failed climatological outlier check R = failed lagged range check S = failed spatial consistency check T = failed temporal consistency check W = temperature too warm for snow X = failed bounds check Z = flagged as a result of an official Datzilla investigation **15 (GSOD):** Observation obtained from the Global Summary Of the Day (often unreliable) **16 (duplicate):** Observation already included in another (longer) series; used for partial duplicates
tasmin_qc (time, station)	See tasmax_qc

tasmax_definition (time, station)	Flag for how Tmax is defined, the following options are possible: **0-14 (TX--):** Definition codes from ECA&D: TX1: Maximum temperature unknown interval TX2: Maximum temperature 18-18 UT TX3: Maximum temperature 0-0 UT TX5: Maximum temperature morning previous day 06, 07, 08 until morning today (shifted 1 day back by ECA staff) TX6: Maximum temperature morning today 06, 07, 08 until morning next day TX7: Maximum temperature between 06 and 18 UT today TX8: Maximum temperature 21-21 CET TX9: Maximum temperature morning previous day 09 h GMT until morning today (shifted 1 day back by ECA staff) TX10: Maximum temperature from 2130 previous day until 2130 CET TX11: Maximum temperature morning today 9 UTC until morning next day TX12: Maximum temperature 19-19 UTC TX13: Maximum temperature within 00-24, 12-12 or 06-06 TX14: Maximum temperature 0-0 LT based on hourly intervals TX15: Maximum temperature 17-17 CET TX16: Maximum temperature 22-22 or 23-23 UT **15-20 (ISTI_---):** Definition codes from ISTI: ISTI_101: Daily value original ISTI_102: Daily value calculated from main standard synoptic observations (00, 06, 12, 18 UTC) ISTI_103: Daily value calculated from main and intermediate synoptic observations (00, 03, 06, 09, 12, 15, 18, 21 UTC) ISTI_104: Daily value calculated from other sub-daily observations (at least 3 obs available) ISTI_105: Daily value calculated from other sub-daily observations (at least 20 obs available) ISTI_999: Missing/Unknown/Not Applicable **-128 (missing value):** No information available on definition

tasmin_definition (time, station)	Flag for how Tmin is defined, the following options are possible: **0-14 (TN--):** Definition codes from ECA&D: TN1: Minimum temperature unknown interval TN2: Minimum temperature 18-18 UT TN3: Minimum temperature 0-0 UT TN5: Minimum temperature morning previous day 06, 07, 08 until morning day TN6: Minimum temperature between 18 UT previous day and 06 UT today TN8: Minimum temperature 21-21 CET TN9: Minimum temperature morning previous day 09 h GMT until morning today TN10: Minimum temperature from 2130 previous day until 2130 CET TN11: Minimum temperature 19-19 UTC TN12: Minimum temperature within 00-24, 12-12 or 18-18 TN13: Minimum temperature 0-0 LT based on hourly intervals TN14: Minimum temperature 17-17 CET TN15: Minimum temperature 22-22 or 23-23 UT **15-20 (ISTI_---):** Definition codes from ISTI: ISTI_101: Daily value original ISTI_102: Daily value calculated from main standard synoptic observations (00, 06, 12, 18 UTC) ISTI_103: Daily value calculated from main and intermediate synoptic observations (00, 03, 06, 09, 12, 15, 18, 21 UTC) ISTI_104: Daily value calculated from other sub-daily observations (at least 3 obs available) ISTI_105: Daily value calculated from other sub-daily observations (at least 20 obs available) ISTI_999: Missing/Unknown/Not Applicable **-128 (missing value):** No information available on definition

A.2 Status file

DIMENSIONS

detection_time	time dimension (annual resolution) for the detection score (days since 01-01-1850 of the first day of the year)
uncertainty_time	time dimension (annual resolution) for the resolution info (days since 01-01-1850 of the first day of the year)
station	station progressive number (starting from 1, identical to the temperature files)
name_strlen	length of station name (max 50 characters)
code_strlen	length of station code (max 50 characters)
tas_break	breakpoint progressive number for Tmean (starting from 1)
tasmax_break	breakpoint progressive number for Tmax (starting from 1)
tasmin_break	breakpoint progressive number for Tmin (starting from 1)
tasdtr_break	breakpoint progressive number for DTR (starting from 1)
merged_break	merged breakpoint progressive number (starting from 1)

VARIABLES

station_name (station, name_strlen)	Name of the station (not necessarily unique)
station_code (station, code_strlen)	Official station ID from source (unique)
latitude (station)	Latitude in degrees North
longitude (station)	Longitude in degrees East
elevation (station)	Elevation above mean sea level in m
data_source (station)	Source ID: as above
tasmax_uncertainty_maximum (uncertainty_time, station)	Worst resolution of Tmax observations within a year
tasmin_uncertainty_maximum (uncertainty_time, station)	Worst resolution of Tmin observations within a year
tas_break_time (tas_break)*	Year in which a breakpoint in Tmean was detected (days since 01-01-1850 of the first day of the year)
tas_break_station (tas_break)*	Station at which a breakpoint in Tmean was detected (station number)
tas_detectability (detection_time, station)*	Detectability index for Tmean
tas_break_type (tas_break)*	Type of breakpoint for Tmean (0=from the relative tests, 1=from the absolute test, 2=from metadata)
tas_break_season (tas_break)*	Temporal aggregation used to detect a breakpoint for Tmean (0=annual means, 1=Oct-Mar, 2=Apr-Sep)
tas_break_count (tas_break)*	Number of relative tests that detected a breakpoint for Tmean (1-3)
merged_break_time (merged_break)	Year of a merged breakpoint (days since 01-01-1850 of the first day of the year)
merged_break_station (merged_break)	Station at which a merged breakpoint was detected (station number)
merged_break_likelihood (merged_break)	Likelihood index for a merged breakpoint
detection_feasibility (station)	Feasibility of detection at a certain station (0=not possible, 1=only absolute test, 2=all tests)

* These variables are repeated for tasdtr (DTR), tasmax (Tmax) and tasmin (Tmin).