Indecis
Sectorial Climate Services

Integrated approach for the **d**evelopment across **E**urope of user oriented **c**limate **i**ndicators for GFCS high-priority **s**ectors: Agriculture, disaster risk reduction, energy, health, water and tourism

Work Package 3

Deliverable 3.3

Release of the INDECIS Homogenization Suite and Manual

José A. Guijarro (AEMET)

December 28nd, 2019

This report arises from the Project INDECIS which is part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR), with co-funding by the European Union's Horizon 2020 research and innovation programme.

Table of Contents

Summary

This report summarizes the procedures followed to homogenize the ECA&D daily series needed for the INDECIS project to supply high quality and complete databases to the other teams responsible of building series of the climate indices focused to the main economic sectors. The first part is devoted to explain the encountered difficulties, followed by an explanation of the final approach used to obtain the desired results. Some discussion and recommendations are given at the end to serve as guide for future updates of the homogeneous databases.

1. Introduction

This Work Package aims at producing a quality controlled and homogenized daily dataset of the essential climatic variables, with any missing data filled in, to allow other teams of the INDECIS project to obtain series of climate indices focused on the main economical sectors.

Several benchmark datasets were developed in a previous task of this Work Package (see deliverable D3.2) to allow testing homogenization procedures for daily data in order to choose the most appropriate for every variable of interest. However, only ACMANT v.4 and Climatol v.3 were tested with all the benchmarks, and Climatol was finally selected due to the availability of its source code, which allow to apply any needed modification to the software to successfully homogenize all the ECA&D daily series.

2. Data and previous quality controls

The raw daily data were extracted from the European Climate Assessment and Dataset (ECA&D; Klein Tank et al., 2002), and consisted in the following ten variables (with expression of their units):

1. CC : Cloud Cover (oktas)
2. FG : Wind Speed (0.1 m s^{-1})
3. HU : Relative Humidity (1 %)
4. PP : Sea Level Pressure (0.1 hPa)
5. RR : Precipitation Amount (0.1 mm)
6. SD : Snow Depth (cm)
7. SS : Sunshine (0.1 hours)
8. TG : Mean Temperature (0.1 °C)
9. TN : Minimum Temperature (0.1 °C)
10. TX : Maximum Temperature (0.1 °C)

This series were in the first place subjected to the quality controls of the INQC software (see deliverable D3.1 for further details). This software applies various quality controls, assigning different flags to the data depending on the probability of being erroneous. In our case, only data with flags 1 (error) or 2 (most likely error) were rejected to avoid the deletion of real extreme data, which may be important for their potential impact on economic activities.

However, this conservative approach allowed some erroneous data to remain in the series, which were then detected by Climatol (outliers or short periods of zeros or other abnormal data).

3. First homogenization attempts

Climatol works best on dense station networks providing fairly correlated series, even with a high proportion of missing data. However, due to the high number of data contained in long daily series, the homogenization of more than a few hundred series may take a very long time (many days or even weeks, when many inhomogeneities are found). In these cases, it is advisable to divide the area into subareas with a more reduced number of series, and the Climatol package provides a function to split the input files into "rectangular" areas defined by a set of parallels and meridians provided by the user. Nevertheless, the lower the number of series in the area, the higher the probability of an abnormal termination of the process due to the simultaneous lack of observations in one or more time steps.

This procedure was attempted with the ECA&D series, whose geographic domain was split into 229 3x6° lat-lon areas, but due to the extremely heterogeneous density of stations (Figure 1), they often contained either too few or too many series (from 5 to 2136 for daily precipitation). In the first case, the homogenization failed due to simultaneous lack of data at some time steps, while the latter case resulted in unacceptable long computing times. Therefore, this approach was found unpractical.

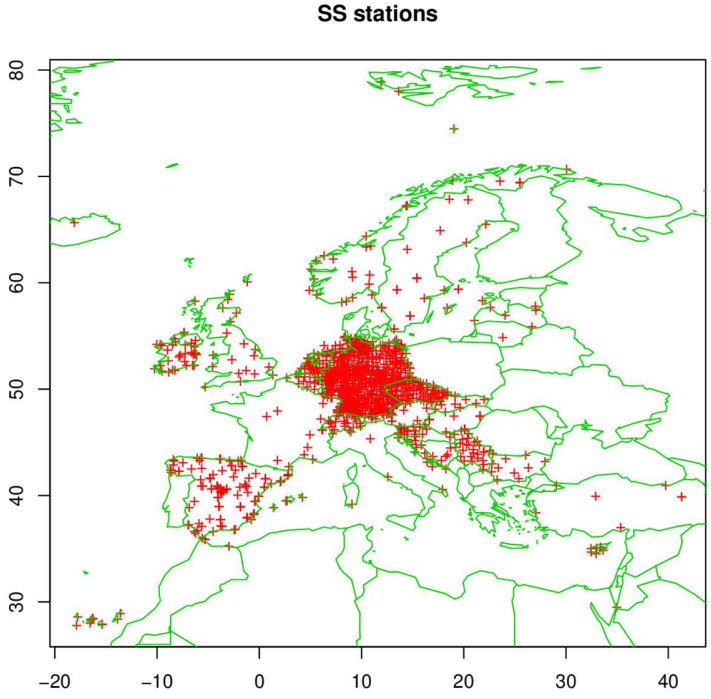

Fig.1: Sunshine duration stations available at ECA&D.

Reanalysis series were used to solve the problem of frequently finding time steps completely void of data. From the various possibilities available, the 20CR (v2c) was chosen because it provides all the needed variables at daily resolution and their series start in 1851, hence providing reference data from the beginning of the period stored at ECA&D (1900). The drawback is that it ends in 2014, and

unfortunately many ECA&D series are not updated in the last years, resulting in the persistence of the simultaneous lack of data condition. The addition of series from the ERA5 reanalysis to cover the last years was also tried, but obtaining its daily series is much more difficult because they are stored with a very high spatial and temporal (hourly) resolution.

4. Final homogenization strategy

In view of the difficulties encountered so far, the decision was to homogenize the series one by one, using the closest 20CR series as reference, for the period 1950-2014, and to add the data available in the last 2015-2018 years in their raw state. In this way, the computing time of the process is very much reduced (from many weeks to five days in a standard personal computer), while the resulting series are still suitable for being gridded into new E-OBS datasets (although missing data in the last period will not be filled in). From the 10 variables initially considered, snow depth (SD) was skipped because its homogenization was found terribly difficult in the previous benchmarking study (see deliverable D3.2), and average temperature (TG) was also left aside for its redundancy with maximum (TX) and minimum (TN) temperature.

Table 1 shows the total number of series provided and the series homogenized after requiring them to have a minimum of 10 years of observation in 1950-2014. Some series appeared to be repeated (they had the same station identifier, but different providers), but as their data were not coincident, all their versions were kept by adding a distinctive letter to their identifier. The number of break-points detected and corrected is of the same order of magnitude as the number of series, except in the case of precipitation, whose high variability makes more difficult the detection of changes in their means. Very few outliers were corrected, mainly in surface pressure and precipitation.

Table 1: Number of series available, homogenized and repeated, and number of break-points and outliers corrected during the homogenization process.

	CC	FG	HU	PP	RR	SD	SS	TG	TN	TX
Total series	2147	1492	2395	1684	15962	8619	1252	5920	6442	6287
>=10 years	1672	1232	1961	1293	13440	--	1006	--	5163	5012
Repeated	6	18	6	42	756	7	12	698	744	726
Homogenized	1672	1232	1961	1293	13439	--	1005	--	5163	5012
Break-points	1548	2697	2928	2482	1449	--	519	--	3439	1969
Outliers	0	1	0	176	115	--	3	--	0	0

5. Conclusions and recommendations

Due to the large number of series and their temporal and spatial heterogeneity, the approach used here to homogenize the ECA&D daily series of the different variables needed by the INDECIS project seems the only feasible to be applied periodically in order to update the homogenized dataset.

However, future improvements are needed to achieve the homogenization and missing data filling of the whole period. The easiest way would be to use a reanalysis which begins early enough as to provide references to the whole period of the ECA&D data and at the same time is periodically updated at the same path as new data are incoming. A new version of the 20CR reanalysis is available which ends en 2015, but it is foreseeable that it will not reach the timely update rythm of the data to be homogenized. Another possibility is to use another reanalysis to act as reference of the lasts years, although mixing two different reanalysis as references may have a negative impact on the homogenization procedure that should be investigated. Anyway, as the homogeneity of the last data of the series is more easily detected after a few years, when more data are incorporated, the only clear drawback of lacking suitable references for the last part of the period is the inability to estimate replacements for all missing data.

The homogenization procedure explained in this report can be reproduced with the help of the homogenization suite and manual gathered in the archive homogenECAD.tgz.

6. Acknowledgements

Project INDECIS is part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co-funding by the European Union (Grant 690462).

ECA&D must also be acknowledged, since its data are crucial for this project.

Support for the Twentieth Century Reanalysis Project version 2c dataset is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER), and by the National Oceanic and Atmospheric Administration Climate Program Office

7. References

20CR (last accessed December 2019): https://psl.noaa.gov/data/20thC_Rean/

ACMANT v.4 (last accessed December 2019): https://github.com/dpeterfree/ACMANT

Climatol v.3.1 (last accessed December 2019): https://CRAN.R-project.org/package=climatol

Klein Tank, A.M.G. and Coauthors (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453.