DOI:10.28974/idojaras.2024.2.2

IDŐJÁRÁS

Quarterly Journal of the HungaroMet Hungarian Meteorological Service
Vol. 128, No. 2, April – June, 2024, pp. 155–170

Operational homogenization of daily climate series in Spain: experiences with different variables

Belinda Lorenzo*, **José A. Guijarro**, **Andrés Chazarra**, **César Rodríguez-Ballesteros**, **José V. Moreno**, **Ramiro Romero-Fresneda**, **Maite Huarte**, and **Ana Morata**

Spanish State Meteorological Agency (AEMET)
Leonardo Prieto Castro 8, 28071 Madrid, Spain

** Corresponding author E-mail: blorenzom@aemet.es*

(Manuscript received in final form February 29, 2024)

Abstract— Calculation of the new climatological standard normals for the period 1991–2020 was a motivation to carry out the homogenization of the required climatic variables in the Spanish Meteorological Agency (AEMET).

The national observation network has undergone changes along its history that often introduce non-climatic interferences to the series. On the other hand, for the calculation of various parameters and climatic indices, it is essential to have complete daily series. With this in mind, homogenization of daily series of precipitation, maximum and minimum temperatures, sunshine hours, relative humidity, station level pressure, mean wind speed, and maximum wind gust was carried out.

This paper shows how the homogenization process was performed, covering the period 1975–2020 with carefully selected daily data sets from the national climatological database. The homogenization software Climatol v.4.0 was used for this process, and derived variables such as average temperature, sea level pressure, and vapor pressure were calculated from their related homogenized series.

The peculiarities and issues of each variable are explored and, finally, the homogenization results were used to readily calculate the 1991–2020 climatological standard normals with the dedicated software CLINO_tool v.1.5.

Key-words: standard normals, homogenization, Climatol, CLINO_tool, climatic variables, daily data

1. Introduction

Calculation of the new climatological standard normals for the period 1991–2020 (*WMO*, 2017) was a motivation to carry out the homogenization of all the daily series of the required climatic variables in the Spanish State Meteorological Agency (AEMET).

Having high-quality series which allow objective monitoring of the climatic behavior in a region is fundamental. Climatological series are subjected to changes of different nature. Some examples could be: instrumentation, location, changes in environment conditions near the station, different measurement methods, etc.

All these changes produce alterations in the data series which have no climatological or meteorological origin. Thus, in order to study possible changes in the climate or watch its behavior, time series long enough and without non-climatological alterations are necessary. These are homogenized series (*WMO*, 2020).

The land observation network of AEMET is made up of stations with specialist observers, volunteer observers, and automatic stations. The stations with professional workers are well distributed throughout the country, and usually do observations of many different variables. The stations with volunteers, complements the network throughout the entire territory. This selfless collaboration is very valuable. It provides measurements from areas that would otherwise lack information. Mostly, these stations measure precipitation and many also get temperature data. In recent times, the number of these collaborating stations has been reduced, but the number of automatic stations has increased. The automatic stations provide data for other variables in addition to temperature and precipitation.

All the data collected through these different methods is stored in the national climatological database.

This study focuses on the calculation of homogenized series of different climate variables.

To make this possible, the series were homogenized using a pre-release of the R package *Climatol* version 4 (*Guijarro*, 2023b). Given the computational requirements, the daily processing was computed in AEMET's HPC Cirrus.

Throughout the entire study period, the available data has different nature of origin but complements each other to perform the homogenization. Climatol responds well to short series, such as those obtained from recent automatic stations. Furthermore, series that no longer exist but that existed throughout the period can help to homogenize series that currently exist. All this variety of data series is very valuable for understanding the behavior of the variables in an area throughout the entire period considered.

When it was necessary to obtain the normals for the last reference period 1981–2010, the homogenized series of precipitation and temperature in that period were calculated (*Botey et al.*,2013; *Chazarra et al.*, 2018). On that occasion, monthly data series were homogenized using an earlier version of homogenization software Climatol (*Guijarro*, 2014). These previous works were taken as a base and reference, being expanded with the calculation of homogenization of daily series and were included more variables such as: sunshine duration, relative humidity, station level pressure, mean wind speed, and maximum wind gust.

Working with daily series offers several advantages. For example, straightforward calculation of daily parameters and climatic indices, such as number of days over certain thresholds. However, it also brings certain disadvantages such as their high variability and the huge calculation capacity that is required.

When homogenized series were obtained, the derived variables, such as average temperature, sea level pressure, and vapor pressure were calculated from their related homogenized series.

Finally, once all the relevant series have been obtained, the 1991–2020 climatological standard normals were calculated automatically and written in the required format to be sent to WMO by means of the CLINO_tool (v.1.5) (*Guijarro*, 2023b).

2. Procedure

2.1. Data, regions and period selection

This study is carried out covering the period 1975-2020. Although the purpose was to obtain the 1991–2020 standard climatological normals, the period 1975–2020 was chosen to allow the recalculation of the 1981–2010 normals, adding 6 years backwards to improve the detection of inhomogeneities in the early 80s.

A large amount of daily data for the period 1975–2020 are available in the AEMET Climatological Database, especially when it comes to precipitation and temperature. For the other variables, considerably less daily data are stored.

This is because the collaborator stations mainly measure precipitation and to a lesser extent temperature, as it was mentioned above. Their data is of a daily nature. The remaining variables are measured at the main stations of the official AEMET network. Their number has increased in recent times due to the introduction of automatic stations that provide ten-minute data. This type of data has been stored in the climatological database since last years of the 2000s. In the main network stations, the nature of the data can be hourly, for example if it is manual, or ten-minute if it is automatic. Therefore, there are three types of data according to their measurement: daily, hourly, and ten-minute. All of them are computed to obtain the corresponding daily data and stored in the climatological database.

In a first step, only data series with a minimum of 5 years of daily data are accepted. Except in the case of precipitation, where only series with a minimum of 10 years are taken into consideration, due to their high number and the irregularity of their data.

When two or more variables have a functional dependency, those observed have been selected (extreme temperatures, relative humidity, station level pressure), leaving the others (mean temperature, water vapor partial pressure, sea level pressure) for a posterior calculation.

For precipitation and temperature, the study area is divided into 12 regions: mainland Spain (divided in turn into ten regions with similar climatic conditions and coinciding with the main hydrographic basins), the Balearic Islands, and the Canary Islands, due to their spatial and climatic differentiation.

For the rest of variables except precipitation and temperature, the study area is divided into three regions: mainland Spain (including autonomous cities of Ceuta and Melilla), the Balearic Islands, and the Canary Islands, due to their spatial and climatic differentiation. Mainland is divided into fewer regions mainly due to lack of data that makes its performing difficult.

2.2. General steps using Climatol

To perform the homogenization of the daily series of all the variables, a pre-release of the R package Climatol v4.0 was used. This software is very flexible, it allows to homogenize different climatic variables and in different time scales (*Guijarro*, 2023b).

Daily data have a high variability over time that makes it difficult to detect changes in their mean. The basic process with Climatol consists of obtaining monthly series from the daily data and homogenizing them. The daily series are divided with the breakpoints obtained in this monthly homogenization. Finally, all the series can be reconstructed from their homogeneous sub-periods in a final stage by estimating all their missing data.

The schematic general steps used in this work with Climatol are:

1. Preparation of the series in the required input format. To do this, the function *csv2climatol* from the Climatol package is used, set to each variable.
2. Only for precipitation: replace trace values by zeros and dissaggregate values accumulated during a few days. The distribution of accumulated is calculated with the *homogen* function using the *cumc* = -4 parameter.
3. First exploratory analysis of the daily series to control their quality (*homogen* function).
4. Obtain the monthly series from the daily data. The daily data is grouped into monthly series with the idea of obtaining the breakpoints. The *dd2m* function is used.

5. Homogenization of the monthly aggregates. The monthly homogenization is carried out to obtain the breakpoints, using the *homogen* function, with the appropriate parameters in each case (*Table 1*).
6. Careful manual review of these results by service specialists, identifying bad quality issues, deleting anomalous periods or whole series where needed, and repeating step 5.
7. Daily homogenization. Taking advantage of the breakpoints offered by the previous steps, it proceeds to the homogenization of the daily data. The parameter metad = TRUE is used in the *homogen* function.
8. Calculation of the daily series of variables considered derived from the corresponding homogenized ones.

Finally, with the homogenized series, the CLINO files required with the climatological normals for the standard period 1991–2020 can be straightforward obtained. CLINO_tool v.1.5 software is used.

In the homogenizations carried out, the following parameters (Table 1) have been used in the *homogen* function, following the recommendations of the documentation (*Guijarro*, 2023a):

- *dz.max* sets the thresholds for rejecting anomalous data or warning about suspicious data,
- *inht* is the inhomogeneity threshold, that is, the value of the homogeneity test above which the series will be split,
- *std* is the type of normalization applied to the data (e.g., default: std=3 means subtract the mean and divide by the standard deviation; with std=2 the data will be divided by its mean value; std=1 only centers the data by subtracting its mean value),
- *nref* is the maximum number of nearby data to use to estimate those of the problem series,
- *vmin* and *vmax* serve to limit the possible values that the data can take, and
- *gp* is the graphics parameter e.g., gp = 4 indicates moving annual sums instead of average values in final graphics.

Table 1. Parameters used in *homogen* function

	monthly		daily		both			
	dz.max	***inht***	***dz.max***	***nref***	***std***	***vmin***	***vmax***	***gp***
Precipitation	15	20	25	1	2	-	-	4
Temperature	6	-	15	-	-	-	-	-
Sunshine hours	6	-	12	-	1	0	-	-
Relative humidity	6	-	15	-	-	0	100	-
Station level pressure	6	-	15	-	-	0	-	-
Mean wind speed and max. wind gust	6	-	15	-	2	-	-	-

The peculiarities of the process for each variable are described below.

2.3. Precipitation

Firstly, the input data is selected. From the information stored in the climatological database, those stations with precipitation series of at least 10 years of daily data are selected resulting in a total of 6293 initial stations distributed throughout the entire territory (*Fig. 1*). Due to the huge amount of data, to facilitate its computation, the territory is divided into the 12 regions mentioned above.

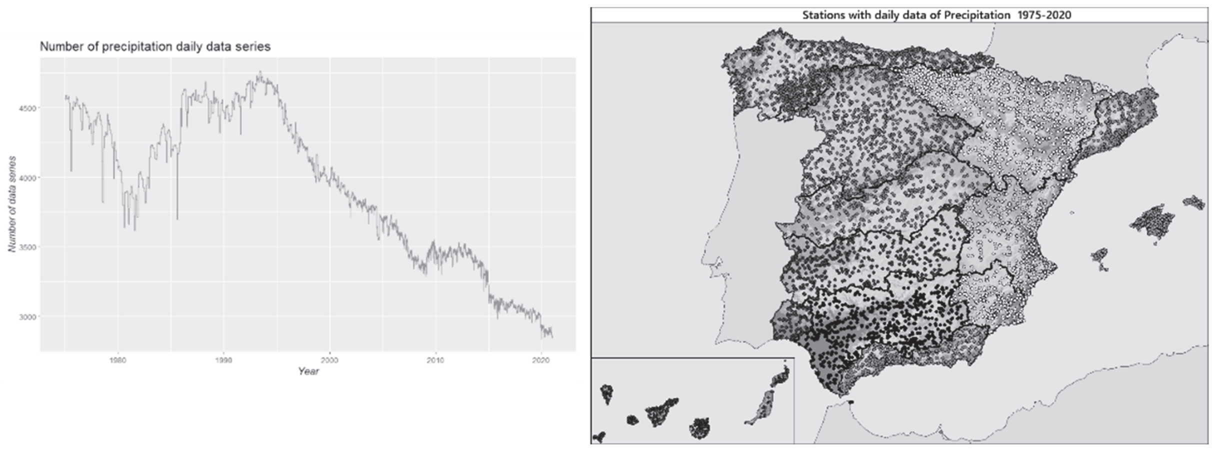

*Fig. 1***.** Number of daily data series for precipitation (left) and distribution of precipitation stations (right).

In this variable, when the precipitation is negligible (<0.1 mm), the data is stored in the database with a value of -3. This values are changed to 0, to avoid small errors in aggregate data.

When the measurement of the amount of precipitation of the day cannot be carried out, it is marked as "accumulated". On the day when the observation can already be made, the total precipitation accumulated in the previous days is recorded. This situation occurs mostly in manual measurements, mainly in collaborator stations. But this can also happen with automatic stations, for example, when snow freezes in rain gauge without a heater. Before proceeding with homogenization, these accumulations must be distributed over the days without measurements. Climatol version 4.0 allows this process to be done using the *cumc* = -4 parameter in the *homogen* function.

The data is then grouped into monthly values, using the *dd2m* function from the Climatol package. It must be taken into account that in this case the monthly values are totals and not average values, so to indicate this, the value of the *valm* parameter must be changed to 1.

With the parameters indicated (*Table 1*), the monthly homogenization is made.

Subsequently, the results are carefully manually reviewed by specialist personnel. Among other criteria used, this review is also based on metadata when possible. As a result of this review, 30 series are rejected throughout the entire territory. To avoid the influences of these rejected stations on nearby stations, the entire process is redone.

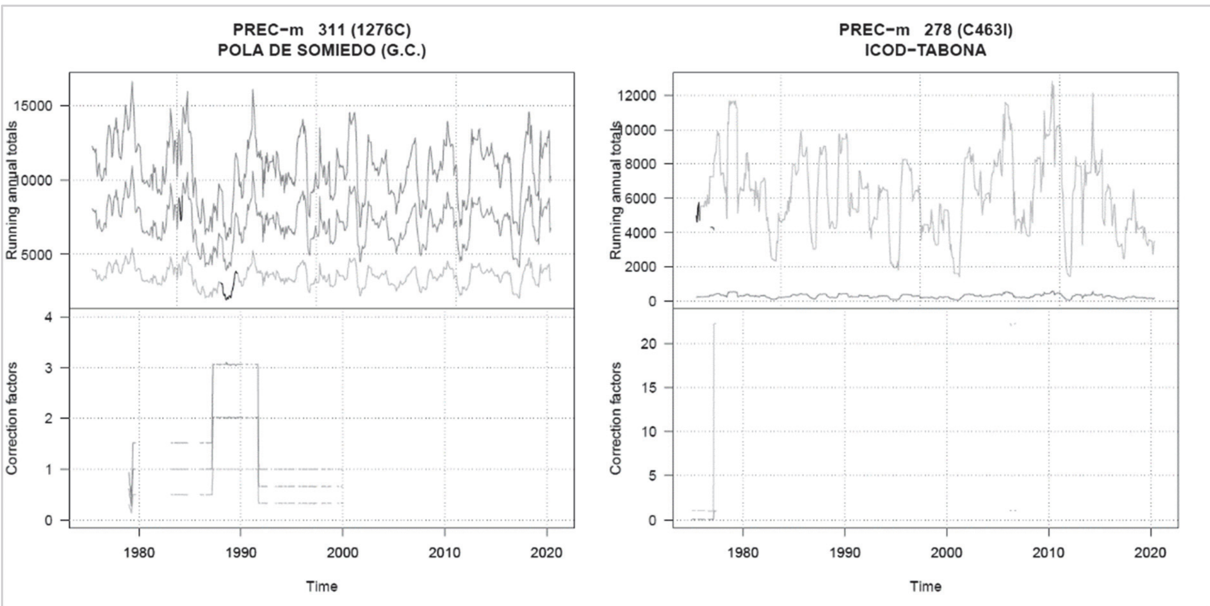

Fig. 2. Examples of series rejected after monthly homogenization.

An example of this is shown in *Fig. 2*. The running annual totals of the reconstructed monthly precipitation series of two rejected stations are shown. In this case, upper graphs show the running annual totals of the reconstructed monthly precipitation series (parameter *gp*=4, *Table 1*). The original series are drawn darker. The other lines represent the reconstructed series from the different homogeneous fragments. Lower graphs show the correction factors applied to the original data in the reconstruction of the series.

In both figures it can be seen that there are few data, especially in the one on the right. Both series are rejected due to their lack of quality data, and none of the reconstructed series offers conclusive or representative information.

The breakpoints obtained in the monthly homogenization are used to calculate the daily homogenization.

2.4. Maximum and minimum temperature

Regarding temperature, the homogenized variables are the maximum and minimum daily temperature. The daily series of average temperature from the homogenized series of maximum and minimum temperatures are obtained in a derivative way. In this case, from the series stored in the climatological database, those with a minimum of five years of data are selected resulting in a total of 3704 selected stations for the entire territory, which is divided into 12 regions (*Fig. 3*).

Fig. 3. Number of daily data series for temperature and distribution of temperature stations.

The entire process described above is carried out, but in this case the distribution of accumulated values is no longer necessary, and the grouping into monthly values is done by calculating daily averages.

After homogenizing the grouped monthly data, results are manually reviewed, and 34 series are rejected. The whole process is recalculated to obtain the homogenized daily series.

2.5. Sunshine duration

In this case, a total of 238 series are selected throughout the territory, which is divided into 3 different climatic regions (*Fig. 4*) corresponding to mainland Spain, the Balearic Islands, and the Canary Islands. The increase in the number of stations observed at the beginning of the 2010s is due to the installation of automatic measurement equipment.

After the monthly homogenization, a total of 12 series are rejected after manual review, and all process is redone. After the daily homogenization, the total hours of sunshine can exceed the theoretical number of hours that a certain place should have. So to fix this if necessary, an auxiliary function *fix.sunshine* is used.

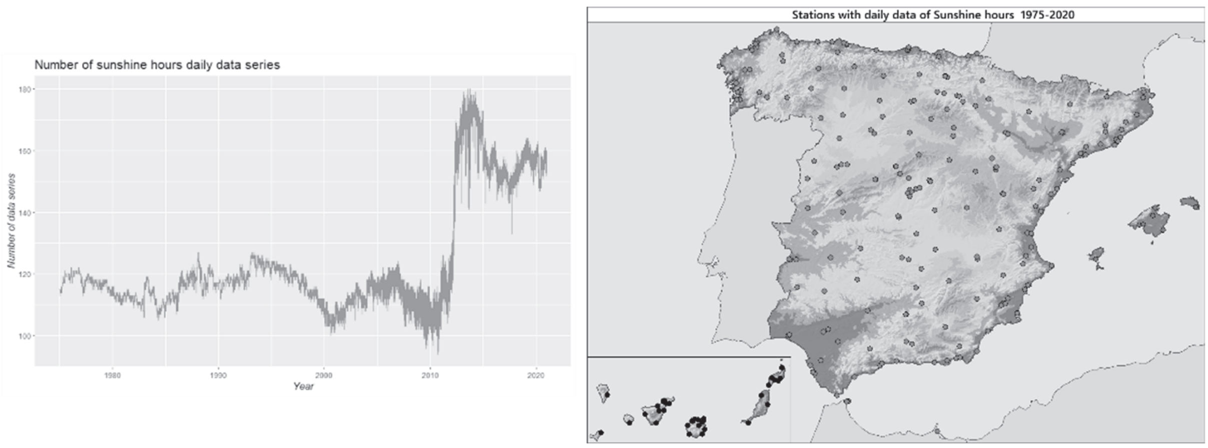

Fig. 4. Number of daily data series and distribution of stations for sunshine hours.

2.6. Relative humidity

Vapor pressure is the main variable required for the calculation of the climatic normal. However, it is a variable calculated from other observed variables. Considering the consistency with the observation, it was decided to homogenize the relative humidity. Vapor pressure is obtained then from the resulting data series and the corresponding average temperature series.

As at least 5 years of daily data are required, we have a total of 914 data series available for calculation. Processing is divided into three climatic regions

(*Fig. 5*) corresponding to mainland Spain, the Balearic Islands, and the Canary Islands.

In this variable, there is also a considerable increase in the number of available stations around 2010. This is attributed to the installation of a large number of automatic stations during that period. Nevertheless, the daily mean value of this variable is calculated using data from 07, 13, and 18 UTC hours to maintain consistency with historical data. Before 1996, only information at these specified hours was recorded.

After the monthly homogenization, about 15 stations are eliminated from the calculation.

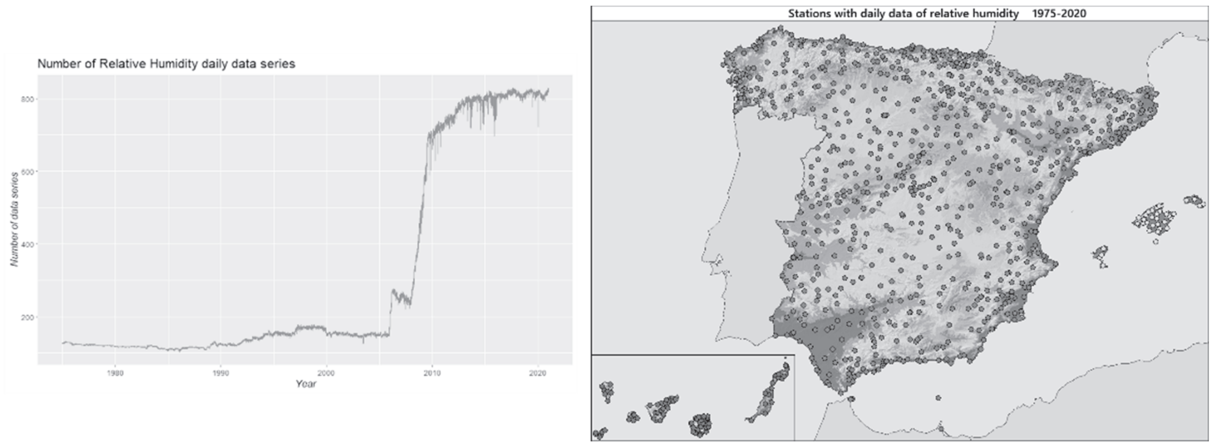

Fig. 5. Number of daily data series for relative humidity and distribution of its stations.

2.7. Station level pressure

In this case, the main variable required in the climatological normals is the pressure at sea level. However, it is a variable derived from the pressure at the station level, which is the one observed. For consistency with the observation/record method, it was decided to homogenize the observed variable and from the results, to obtain the homogenized daily series of pressure at sea level.

Similar to what happens with relative humidity, the daily mean value of this variable is obtained from the values at 07, 13, and 18 UTC hours.

After discarding those series with less than 5 years of data, a total of 274 stations are available (*Fig. 6*), of which 13 are rejected after evaluating their quality.

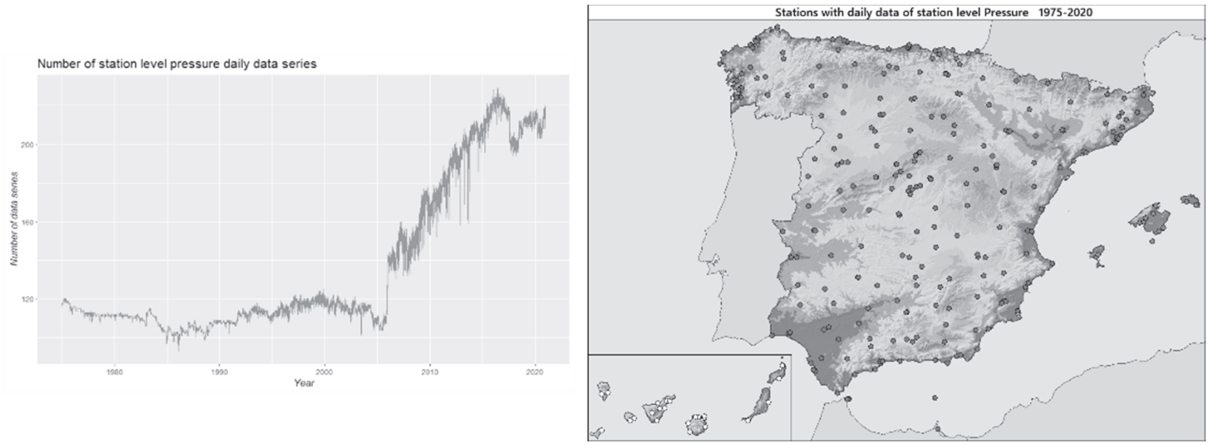

Fig. 6. Number of daily data series for station level pressure and distribution of its stations.

2.8. Mean wind speed and maximum wind gust

The maximum wind gust is treated as a secondary variable to report normal climatological values. Taking advantage of this calculation, the average wind speed has also been homogenized.

Similar to the previous variables, the daily mean value of this variable continues to be calculated from the data at 07, 13, and 18 UTC hours, even though the majority of measurement stations are automatic.

In both variables, after rejecting those series with less than five years of daily data, near 900 series are gathered (*Fig. 7*), of which 11 are rejected after a quality review after the monthly homogenization.

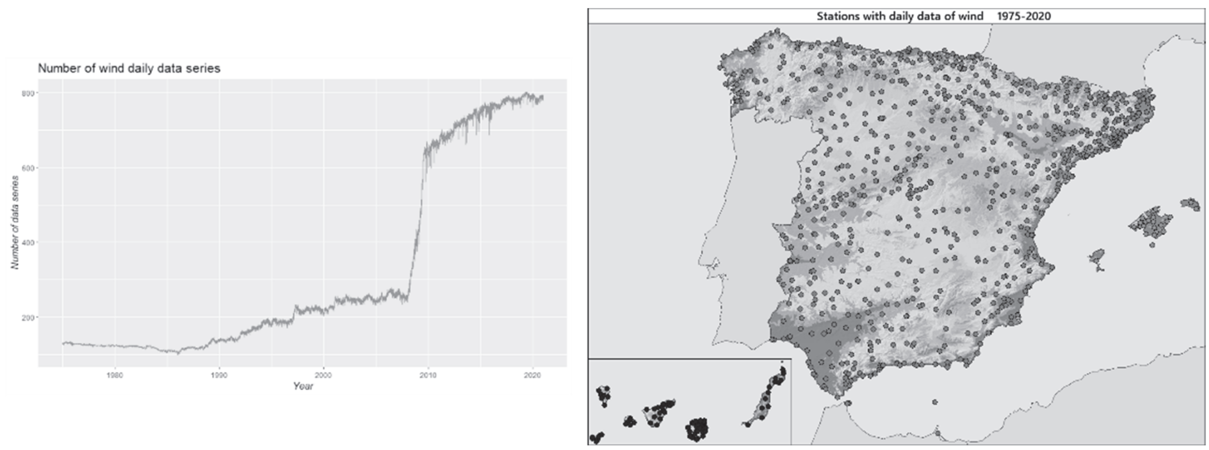

Fig. 7. Number of daily data series for wind and distribution of its stations.

2.9. Derivate values

As previously mentioned, the series of derived variables are calculated from the homogenized daily series of the variables from which they are derived.

The average temperature is calculated from the series of daily maximum and minimum temperature. The vapor pressure is obtained from the series of relative humidity and average temperature. Likewise, the pressure at sea level is derived from the pressure at the station level.

2.10. Normal values

Once all the relevant series have been obtained, the CLINO files can be calculated straightforward using the software CLINO_tool v.1.5. This CLINO files are requested with a specific structure. With CLINO_tool, these files are automatically generated from the Climatol output after the daily homogenizations (*Fig. 8*).

```
World Meteorological Organization Climate Normals for 1991-2020,,,,,,,,,,,,,,,,
Single Station Data Sheet For All Climatological Surface Parameters,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
Station Header Record,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
Country_Name,Spain,,,,,,,,,,,,,,,
Station_Name,A CORUÑA,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
WMO_Number,Latitude,Longitude,Station_Height,,,,,,,,,,,,,
8001,43|21|57|N,008|25|17|W,58.000000,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
WMO Integrated Global Observing System (WIGOS) Station Identifier (if available)
0-20000-0-08001,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
Principal Climatological Surface Parameters,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
Parameter_Code,Parameter_Name,Units,,,,,,,,,,,,,,
1,Precipitation_Total,mm,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,
WMO_Number,Parameter_Code,Calculation_Name,Calculation_Code,January,February,March,April
,May,June,July,August,September,October,November,December,Annual
8001,1,Sum,4,120.8,88.5,87.6,86.9,66.3,45.7,31.2,40.7,57.0,120.7,149.9,122.3,1017.6
```

Fig. 8. Example of the CLINO file heading.

It must be taken into account that the CLINO_tool only calculates normals from series having the required minimum of 80% of data in the reference period 1991–2020.

3. Results

The main results are the homogenized data series of the variables involved. Some particular cases about the relationship between changes involved in stations and breaks detected are discussed below.

3.1. Summary of breaks

Table 2 shows a summary of number of series and breaks detected by variable.

Table 2. Summary of the breaks by variable

Climatological variable	Number of series	Number of break-points	Average break-points/series
Precipitation	6293	1759	0.28
Maximum temperature	3704	5630	1.52
Minimum temperature	3704	5577	1.51
Sunshine hours	238	126	0.52
Relative humidity	914	601	0.66
Station level air pressure	274	353	1.29
Mean wind speed	900	586	0.65
Maximum wind gust	883	495	0.56

It highlights the low value of average breaks-points/series for precipitation. High number of series and irregularity of their data are observed. Fewer jumps in the average are detected. Moreover, the review of the test histograms and anomaly graphs suggested a lowering of the default value of parameter *inht*, which is 25, to 20 (*Table 1*). Variability is inherent to this variable. It is of interest to increase the value of the parameter *dz.max,* so that fewer anomalous data points are rejected.

The same does not happen with temperature. After observing the anomaly graphs, the threshold value *inht* = 25, is appropriate for monthly values of temperature. Although in a little conservative way, we are trying not to detect false jumps in the average at the cost of letting the minor ones pass.

3.2. Change in location

Fig. 9 shows a graph with the running annual means of hours of sunshine at the Tenerife Sur airport located in the Canary Islands. In this case, a break was detected for November 2014. According the stored metadata, it has been verified that in November 2014, the location station at the airport was changed, thus giving a possible explanation for this break. In this case, this break produces a split in two series, where the last period undergoes an additional correction of more than 0.5 hours in the data prior to the break.

We can find another example in *Fig. 10*, that shows a graph with the running annual totals of precipitation at Llinars del Vallès (mainland). Here the Climatol has detected a break in 1995. According to the registered metadata it is described, that in the middle of the decade there was a change of location from a garden to an interior terrace at 250 m apart.

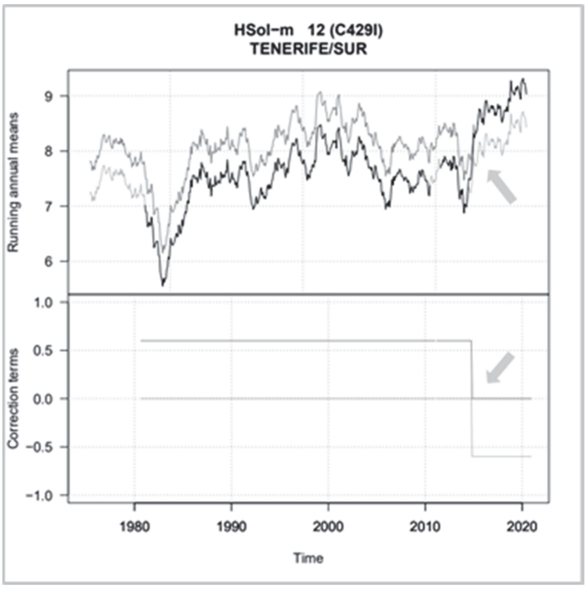

Fig 10. Running annual means for sunshine in Tenerife Sur with a break in November 2014.

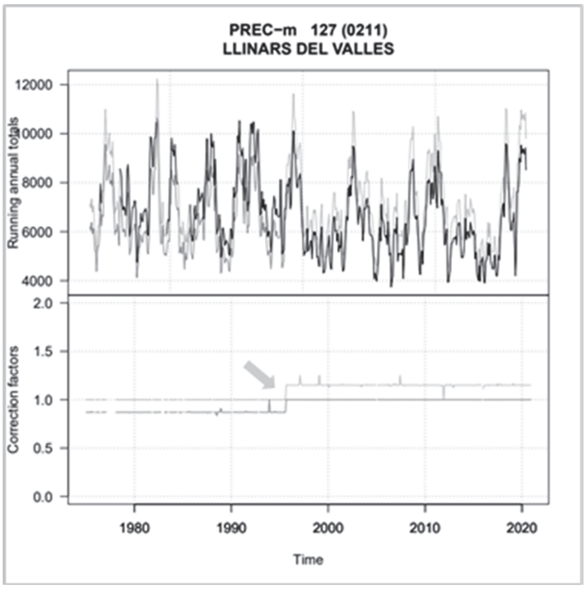

Fig. 9. Running annual totals for precipitation with a break in 1995.

3.3. Change in shelter

Fig. 11 shows a graph with the running annual means of maximum temperature at Hervás (mainland) with a break detected in 2018. It is a THIES-type automatic station that started operating in 2008. According to the records, in 2018, a change of shelter was carried out with a change of height.

3.4. Change in instrumentation

Fig. 12 shows how a change of instrumentation could affect the measurements of wind. In this case, a break was detected in the late 2000s. It is a SEAC-type automatic station that at the end of the decade was replaced by a THIES-type station. On this occasion, the series prior to the break see their values clearly increased.

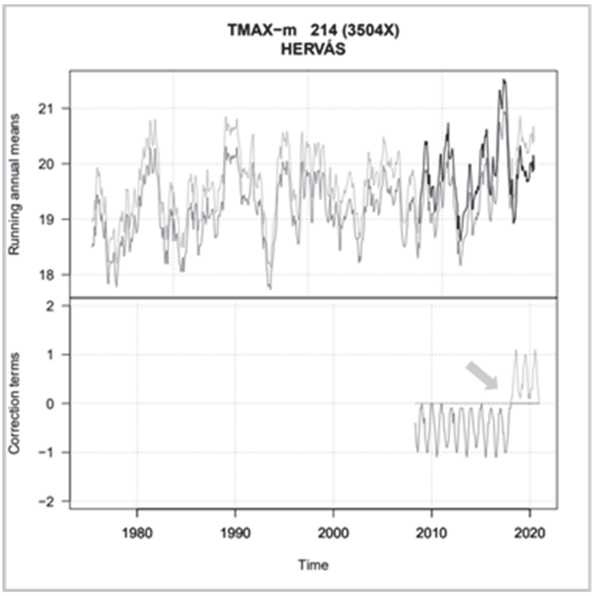

Fig. 11. Running annual means for maximum temperature with a break in 2018.

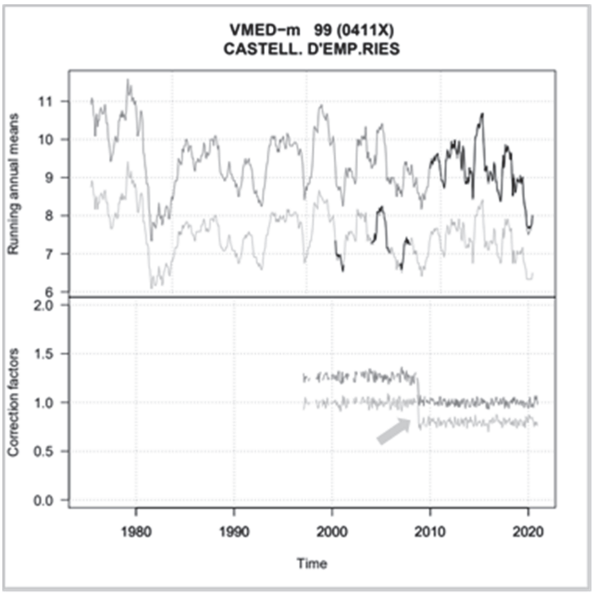

Fig. 12. Running annual means for mean wind speed.

3.5. Change in calculation/measurement methods

The following example (*Fig. 13*) shows a graph with the running annual means for the station level pressure. In this case, a break was detected in 1996. According to metadata, after 1996, the barometric reference was changed to reduce the pressure calculation. There is no further recorded metadata that could explain the other break detected in 2008.

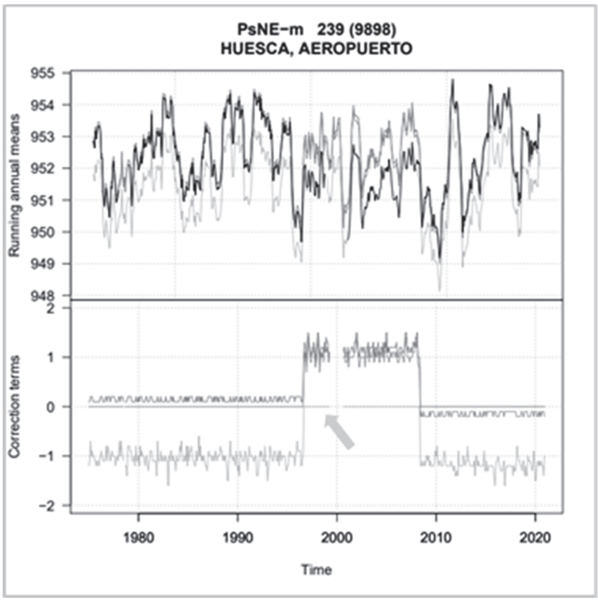

Fig. 13. Running annual means for station pressure.

4. Conclusions

Due to the large amount of data and simultaneously running threads, the calculation involves a computationally heavy load. It became necessary to resort to processing in the homogenization process HPC, especially thinking about the calculation time.

Climatol has responded well to detecting jumps whose origin appears to be related to non-climatological changes recorded in the metadata.

Obtaining the normal values from the homogenized series using CLINO_tool has been straightforward.

As a continuation, the maps corresponding to the climatological normals for the period 1991–2020 will be obtained.

An interesting perspective for the future could be obtaining daily data grids and maps of variables such as precipitation and temperature from the homogenized data. They could be compared with non-homogenized grids but validated by automatic methods.

References

Botey, R., Guijarro, J.A. and *Jiménez, A.*, 2013: Valores normales de precipitación mensual 1981-2010. Agencia Estatal de Meteorología.

Chazarra, A., Flórez, E., Peraza, B., Tohá, T., Lorenzo, B., Criado, E., Moreno, J.V., Romero, R. and *Botey, R.*, 2018. Mapas Climáticos de España (1981–2010) y ETo (1996–2016); AEMET.

Guijarro, J.A., 2014: User’s guide of the *climatol* R Package (version 2.2).

Guijarro, J.A., 2023a: CLINO tool. https://climatol.eu/pub/CLINO_tool.tgz

Guijarro, J.A., 2023b: User’s guide of the *climatol* R Package (version 4). https://climatol.eu/climatol4-en.pdf

WMO, 2017: Guidelines on the Calculation of Climate Normals. WMO-No. 1203. https://library.wmo.int/idurl/4/55797

WMO, 2020: Guidelines on Homogenization. WMO-No. 1245, 54 pp, Geneva, Switzerland.