Journal of Atmospheric and Solar-Terrestrial Physics 128 (2015) 70–83

Contents lists available at ScienceDirect

Journal of Atmospheric and Solar-Terrestrial Physics

journal homepage: www.elsevier.com/locate/jastp

Extraction of the global absolute temperature for Northern Hemisphere using a set of 6190 meteorological stations from 1800 to 2013

Demetris T. Christopoulos

National and Kapodistrian University of Athens, Department of Economics, Greece

ARTICLE INFO

Article history:
Received 11 October 2014
Received in revised form
23 March 2015
Accepted 24 March 2015
Available online 28 March 2015

Keywords:
Absolute temperature
Northern Hemisphere
Valid station
Data quality
Seasonal bias
Extreme values distribution
Missing records
Big data analysis

ABSTRACT

Starting from a set of 6190 meteorological stations we are choosing 6130 of them and only for Northern Hemisphere we are computing average values for absolute annual *Mean*, *Minimum*, *Q*1, *Median*, *Q*3, *Maximum* temperature plus their standard deviations for years 1800–2013, while we use 4887 stations and 389 467 rows of complete yearly data. The data quality and the seasonal bias indices are defined and used in order to evaluate our dataset. After the year 1969 the data quality is monotonically decreasing while the seasonal bias is positive in most of the cases. An Extreme Value Distribution estimation is performed for minimum and maximum values, giving some upper bounds for both of them and indicating a big magnitude for temperature changes. Finally suggestions for improving the quality of meteorological data are presented.

1. Data selection and description

We are using data provided by the *Met Office Hadley Centre* and more specifically the *CRUTEM4 dataset*, see Met Office Hadley Centre Observations Datasets and Jones et al. (2012), while the format description is in Met Office Hadley Centre Observations Datasets. We are interested for the first 8 rows of every file, those with Number, Name, Country, Lat, Long, Height, Start year and End year. We exclude from final data all those stations where there exists errors in coordinates, because either they have neither a latitude (-99) nor a longitude entry (-199). We also exclude 4 files because their content is only a few years of records and cannot contribute to any proper analysis. All such excluded stations are given in Table 4. By applying those filters we reduce the number of accepted stations to 6167 from the initially given number of 6190. We want now to keep from each station only the full monthly yearly data, so if a file-station for some year has a -99.0 entry for just one month, then we exclude the relevant year from that file. After such a procedure we found that 37 stations had not even one full year to contribute, so we excluded them and ended at a number of 6130 files for further analysis. The first excluded station in such a way was id #156150 (Mussalah Top, Bulgaria) and we present it in Table 5 where it is obvious that the station is unacceptable due to the many monthly missing values per year.

A pie chart of the remaining 6130 meteorological stations is presented in Fig. 1, while Fig. 2 is the spatial density and by a simple inspection we observe that its size is analogous to the industrial and population density, at least for western countries. This could possibly be a reason for generating a primary bias at the records, since the industry and population grow the temperature close those places will also increase, due to direct thermal pollution.

2. Annual average values for yearly temperature measures as a function of time and number of valid stations

Let us define formally the concept of *valid station* for a year that we study temperature of Earth's atmosphere.

Definition 2.1. Let us consider a year Y and a set of meteorological stations S. From the subset S_Y of stations that have records for Y we extract the subset V_Y of stations that have full monthly records, from January to December, for the same year Y and name it as the set of valid stations for year Y, while the relevant number of valid stations is $N_v = |V_Y|$.

Example 2.1. For $Y=2013$ we find a set of $N_v = |V_{2013}| = 895$ valid

E-mail address: dchristop@econ.uoa.gr

http://dx.doi.org/10.1016/j.jastp.2015.03.009
1364-6826/

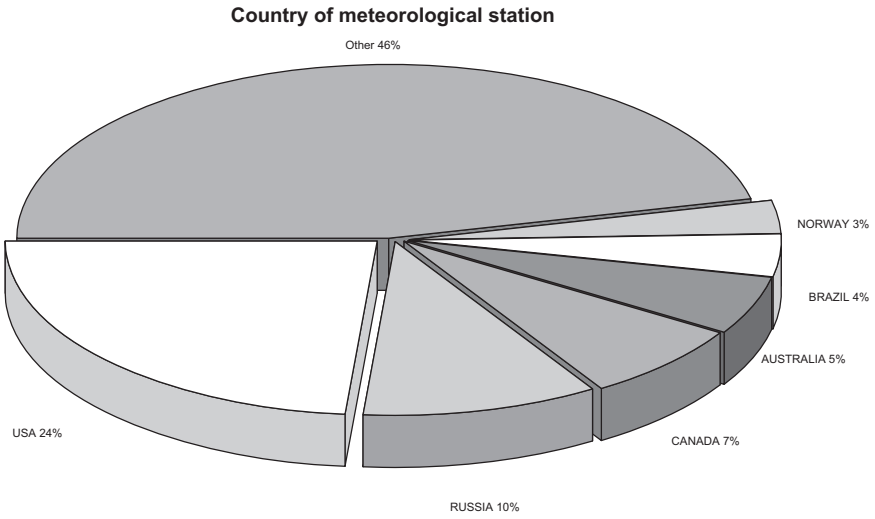

Fig. 1. Pie chart of meteorological stations.

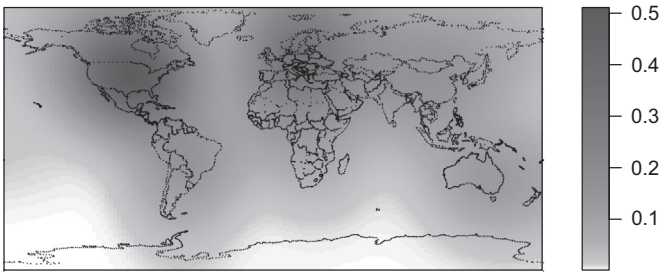

Fig. 2. Spatial density of 6130 meteorological stations for Hadley Centre CRUTEM4 dataset.

Mean, *Minimum*, Q_1, *Median*, Q_3 and *Maximum* temperature, for every year Y since 1800–2013, while at the same time we have kept a record for the number $N_v = |V_Y|$ of valid stations that were used. A sample of the rows for our data frame is presented at Table 6 [1] where we see that station *Belgaum* of *India*, although had $nys = 50$ yearly records, we have used only the 40 valid entries for its contribution to our computations. There were also 10 years (1952, 1953, etc.) with non-full monthly records, so, while working to compute means and other measures, we will not use that station's records for 1952, 1953, etc. After doing the massive work we found an interesting result, presented in Fig. 4, where we observe that as long as we are leaving from the year 1969, when

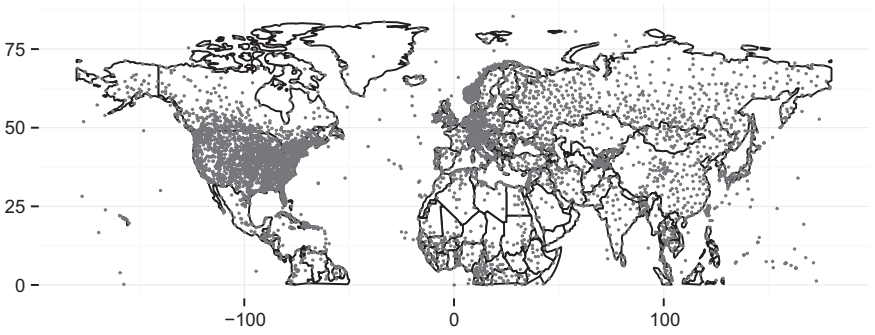

Fig. 3. Northern Hemisphere with 4887 stations selected.

stations from a total of $N_S = |S_{2013}| = 1210$ available stations.

After collecting from positive latitude in the Northern Hemisphere we end up with a set of 4887 station files, see Fig. 3, with a grand total of 446 477 yearly row entries. The number of full monthly rows is $N_v = 389\,467$, which is also the total number of valid stations for all available years of observations. Now for that subset we can compute average values and standard deviations without the need to remove Non-Available (*NA*) entries. By using *R* – R Core Team – and its wide functionality we create a data frame of the above row size and put at every line all the relevant information about the station record for each working year. So, we finally have a huge number of yearly observations for more than three centuries (1701–2014) and all the Northern Earth. Using this final data frame we can easily compute all kinds of descriptive statistics that we desire, part of them is in Table 7, what we have computed, for Northern Hemisphere only, the average values of the annual

we had the maximum of 4247 valid stations, we are always getting a lower number of them. It is extremely important that after the year 2004, when the annual mean temperature seems to explode, at the same time the number of valid stations seems to collapse, leading to a small value of just 895 for the year 2013 which is approximately the same number as the year 1891, see the bold rows of Table 7. The correlations between number of valid stations N_v and average values of annual *Mean*, *Median*, Q_1 and Q_3 temperature for [2004, 2013] are highly negative, see Table 1. It is a well known result that *Mean* (μ) is affected from 'outliers', so it is not a representative measure of the central tendency. In many research disciplines, like Economics, it is not used any more (for example they will present only *median income* of households, not

[1] Keep in mind that our dataset has eastern longitudes marked as negatives, contrary to the usual convention.