Freely available daily hydrometeorological data from Czechia: further insights

Ondrej Ledvinka[1,*], and *Matej* Jedlicka[1]

[1]Czech Hydrometeorological Institute, Hydrology Database and Water Budget Department, Na Sabatce 2050/17, 143 06 Prague 412, Czechia

Abstract. Quite recently, freely available mean daily discharge series representing the territory of Czechia were complemented by ten daily series of nine climate variables, spanning usually from 1961 to the present. Their length thus allows the analyses similar to those conducted approximately a year ago in relation to the long series of discharge. Besides this possibility, the current paper goes further and shows how these long series (including discharge) can be used in order to assess the presence of the so-called Hurst phenomenon. Using a wavelet-based approach, several important differences in the wavelet spectra and values of the Hurst exponent (as well as the uncertainty in their estimation) were found when focusing on discharge, air temperature and precipitation. Furthermore, using the stationarity and unit root tests, it was revealed that, unlike precipitation, air temperature and discharge series may be characterized by long-memory processes in many cases. Finally, as the paper is devoted mainly to students, a short R script is presented in Appendix that makes it easier to work with the online files offered by Czech climatologists.

1 Introduction

In 2017, the academia and other researchers or scientists from around the world were informed about the existence of freely available mean daily discharge series from Czechia [1]. The author of that paper briefly demonstrated how the data set could be used especially in relation to the study of possible influences of climate change (CC) on water resources monitored by the Czech Hydrometeorological Institute (CHMI). At the end of that paper, it was foreseen that 'there would be more free data in the near future', which finally came true. Notably, the climatologists from the CHMI followed the idea of their colleagues from the Hydrology Division and made various daily time series recorded at ten selected stations available online. Besides climatologists worldwide, also hydrologists could clearly benefit from this decision.

From the perspective of all who are interested in time series analysis in the field of hydrometeorology, a relatively important aspect was intentionally postponed to the next work. On the one hand, the stationarity of the offered discharge series was tested, but on the other, values of the Hurst exponent (i.e. a key indicator of long-range dependence,

* Corresponding author: ledvinka@chmi.cz

otherwise called long-term persistence, LTP) were not reported in [1]. This exponent is closely related to the fractional differencing parameter *d* of the models characterizing stochastic processes typical of LTP (i.e. so-called long-memory processes; see e.g. [2–4]). It is well known that different geophysical time series (hydrometeorological ones inclusive) reveal different values of the Hurst exponent, usually greater than 0.5 (for further details see [5–7]), which means that investigators should seriously consider using such models that are, moreover, capable of capturing the hyperbolic decay of the autocorrelation function, as described in [8]. The issue of misunderstanding the underlying processes can also lead to a large number of falsely detected (deterministic) trends since the presence of LTP results in the occurrence of trend-like patterns, among other components. Even the highly recognized nonparametric trend tests, widely used by hydrologists dealing with CC, have problems to distinguish properly between deterministic trends and those occurring by chance under the so-called scaling hypothesis. Therefore, also some modifications of these tests came into being exactly in hydrology to overcome this drawback (e.g. [9, 10]).

In this current paper, we thus go further and look again at the same discharge series in order to find out what values of the Hurst exponent may be expected in relation to the rivers in Czechia. In addition, we investigate the selected time series of air temperature and precipitation from the newly offered data set comprising climatological stations. Finally, due to the fact that the paper is primarily dedicated to students, a short R statistical software [11] script is presented in Appendix that should make it easier to work with the atypical structure of online climatological data.

2 Data and their sources

For the hydrological data sets composed of discharge, the source was the same as in [1] (i.e. the website of the Czech National Committee for Hydrology, CNCH; http://cnvh.cz/). The only difference was that the new data representing the year 2016 were added to the files in spring 2017. Since the analysis regarding the Hurst phenomenon requires the longest series available, we incorporated the newest data in our assessment as well. Thus, although the series of discharge could have different starting points, the end point was always 31 December 2016. Having prior information about the homogeneity and the missing values within the series, we proceeded with five selected water-gauging stations that were subjected to the stationarity and unit root tests in [1] (i.e. stations 091000, 151000, 240000, 294000 and 421500). In all cases, the series length was over 95 years, which satisfies the conditions applying to the study of the Hurst phenomenon.

In the case of climate, there are data from ten additional stations at http://portal.chmi.cz/historicka-data/pocasi/denni-data/. However, the observations at two of them started in the 1970s, which is not sufficient for the analysis intended. Table 1 lists the remaining eight climatological stations that were finally taken into account here. At the time of writing, their time series spanned from 1 January 1961 to 31 December 2016. In other words, there were 56-year daily series of nine climate variables in total. For the purpose of our study, we selected only two variables: mean daily air temperature given in °C and daily amounts of precipitation given in mm.

Table 1. Relation between the IDs (referred to in the text) and names of eight climatological stations considered in the study. More detailed information on these (and two additional) stations can be acquired at http://portal.chmi.cz/historicka-data/pocasi/denni-data/.

B2BTUR01	**L2PRIM01**	**O1LYSA01**	**O1MOSN01**	**P1PRUZ01**	**P3PRIB01**	**U1MILE01**	**U2LIBC01**
Brno Tuřany	Přimda	Lysá hora	Mošnov	Praha Ruzyně	Přibyslav	Milešovka	Liberec

3 Statistical tools for data processing

For the sake of brevity, we skipped the basic exploratory data analysis (EDA) as well as the homogeneity check. Regarding the discharge series, we believed that the addition of new data from 2016 did not change much their behaviour. In the case of air temperature and precipitation, we assumed that the tests of homogeneity were carried out by climatologists before offering them online. For the next analyses, the series should be deseasonalized, as emphasized, for example, in [1, 12, 13]. Observing suspicious patterns in wavelet spectra after the wavelet-based deseasonalization suggested by [14] (not shown), we decided rather to employ classical deseasonalization outlined in [12, 15, 16].

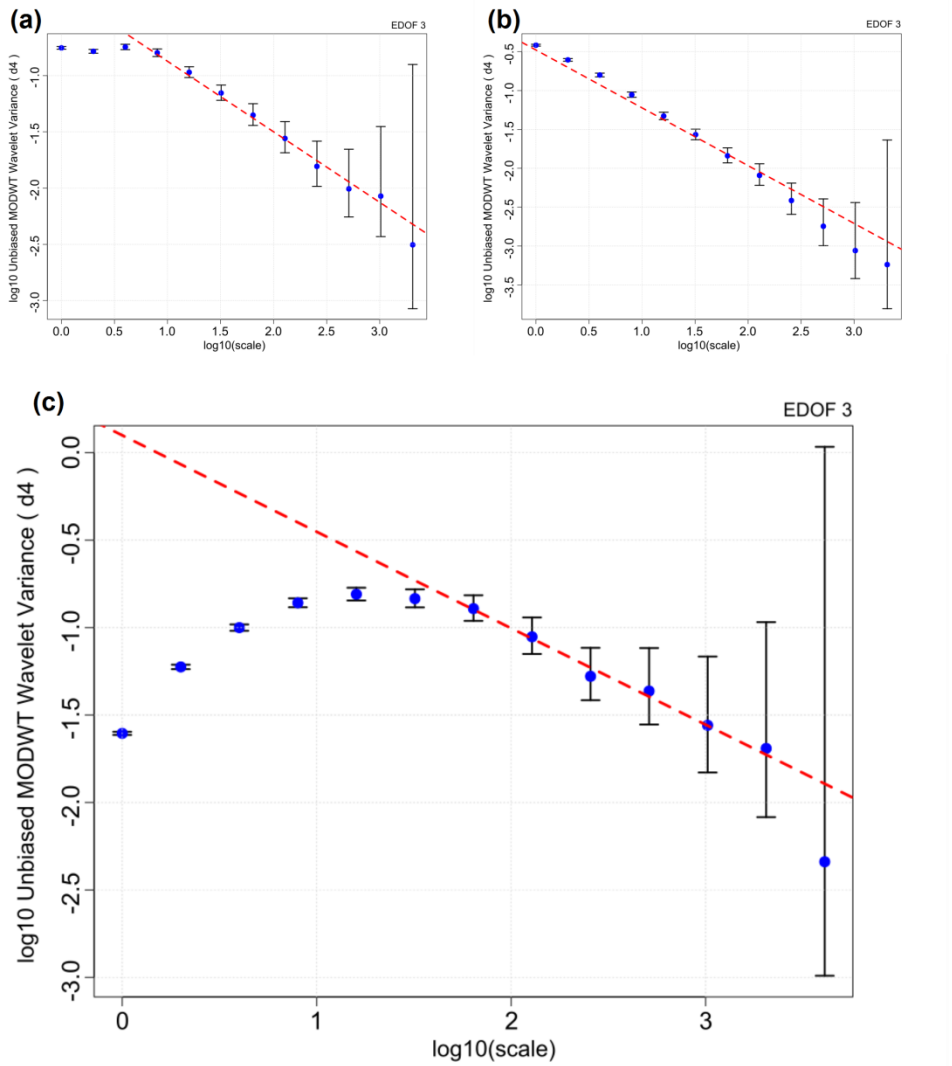

Fig. 1. Wavelet spectra depicted as double (decimal) logarithmic plots of wavelet variance against scale for: (a) air temperature and (b) precipitation at climatological station U2LIBC01, and (c) discharge at water-gauging station 421500. The Daubechies filter of length 4 (d4) and periodic boundary conditions were used. The range of vertical lines (i.e. 95% confidence intervals of variance) increases with scale because of decreasing equivalent degrees of freedom (EDOF). The dashed red lines represent the fitted WLS regression models (see Section 3.2 or [17] for details).

3.1 Testing for stationarity

As the deseasonalization procedure was different from that in [1], the stationarity and unit root tests were applied to both climate variables and, once again, to discharge. The combination of the Kwiatkowski–Phillips–Schmidt–Shin [18] and the Phillips–Perron [19] tests was employed in the same way as in [1, 12, 20–22]. For this purpose, the 'tseries' R package [23] was very useful.

3.2 Hurst exponent estimation

There are many approaches through which one can estimate the value of the Hurst exponent H. As, for instance, [12] or [24] show, they can be based on time domain, frequency domain or a mixture of both such as the wavelet domain. Although some estimators were developed exactly in hydrology as well (see e.g. [25, 26]) and they are generally considered very good [27], in our study, the technique making use of the so-called maximal overlap discrete wavelet transform (MODWT; [17, 28, 29]) was utilized because it should be insensitive to the presence of deterministic trends [12], which is beneficial in the cases when different components of time series may interact and mimic each other (deterministic and stochastic trends inclusive).

The method builds on the estimation of wavelet variance at the so-called dyadic scales (i.e. integer powers of the number two from $2^0 = 1$ to the maximum allowed by the length of data), which results in the estimates of spectra depicted, for instance, in Fig. 1. The Hurst phenomenon usually starts after some scale, which is clearly visible for daily time series. If one fits a regression line to the decreasing range of variance logarithms (knowing the logarithms of scale as the explanatory variable), the Hurst exponent can be obtained as $H = \beta/2 + 1$ where β is the slope of the regression line [16]. During the estimation of β, the method of weighted least squares (WLS) is used because of the decreasing number of the so-called equivalent degrees of freedom (EDOF) with increasing scales. The whole procedure of estimating β via MODWT, using either matrix notation or explicit expression, is described in [17]. Note that there are many types of wavelets possible when constructing wavelet spectra. Inspired by the work [12], we applied the Daubechies filter of length 4 (d4) and did not investigate other types of wavelets. Several functions of the R package 'wmtsa' [30] were incorporated in our own code for fitting the WLS regression lines.

3.3 Evaluation of uncertainty in the Hurst exponent estimation

Also the uncertainty in the estimation of the Hurst exponent was studied here. This was performed using the maximum entropy bootstrap (MEB) implemented in the R package 'meboot' [31, 32]. MEB properly preserves the serial dependence in the original geophysical series when resampling [33, 34]. Since the technique is computationally demanding, especially as regards long daily series, we produced ensembles of only 200 replicates from which the empirical quantiles corresponding to 95% confidence intervals were derived according to [32].

4 Results and discussion

This time, the tests devoted to stationarity revealed that fractional differencing would be better for the modelling of discharge at stations 151000 and 294000. Such a finding is somewhat contradictory to [1], which may be explained by different deseasonalization used here. Fractional differencing would also be valuable for all series of air temperature, apart from stations O1LYSA01 and U2LIBC01. It seems that the percentage of temperature

series with LTP is higher in Czechia than in Italy (cf. [12]). The reason should be studied in the future. For precipitation, on the contrary, no LTP is evidenced by the tests, which is in accordance with [15, 22].

Looking at Fig. 1, where only examples are given but the same pattern can be seen for groups with regard to variables, it is more than clear that the wavelet spectra have their specific courses. In the case of discharge, the wavelet variance rises to the scales approximately equal to 2^5–2^6 days (i.e. months) and then declines. The WLS regression lines were therefore fitted to the range between 2^6 and 2^{10} days. The larger scales were not included due to a higher risk of sampling errors. The resulting estimates of the Hurst exponent can be seen in Fig. 2. They are greater that those depicted in Fig. 3 that shows the values of the Hurst exponent for air temperature and precipitation. Also, the uncertainty in the estimation is larger for discharge (at least compared to temperature), which may be caused by the fact that, in the procedure of estimation regarding climate variables, more values of wavelet variance were included when fitting the regression lines (i.e. scales between 2^3 and 2^9 days for temperature and between 2^0 and 2^9 days for precipitation). This was because the wavelet spectra decay earlier here. Notably, while the spectra of precipitation decay almost immediately, the spectra associated with air temperature after 2^2–2^3 days. Interestingly, as regards air temperature, the wavelet variance is somewhat stable during the days before the Hurst phenomenon starts (see Fig. 1). Moreover, without exception, air temperature reveals higher Hurst exponents than precipitation (Fig. 3).

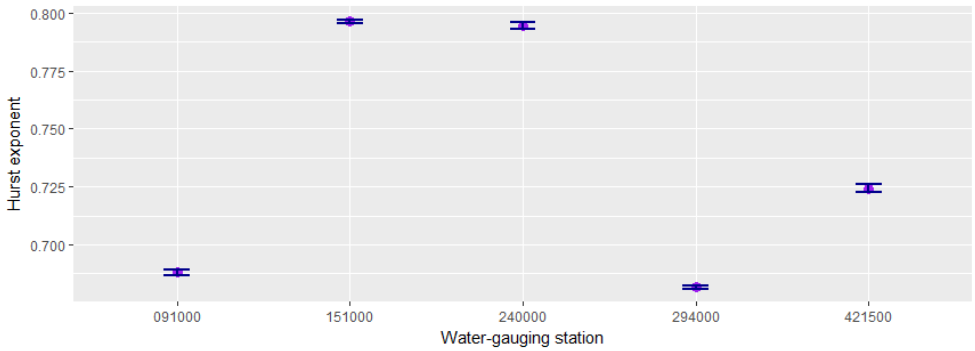

Fig. 2. Hurst exponent wavelet-based estimates in the case of five selected water-gauging stations and discharge as the only variable. Error bars indicate the 95% confidence interval obtained via MEB (see Section 3.3). Assessed period varied according to the series lengths (see Section 2 or [1]).

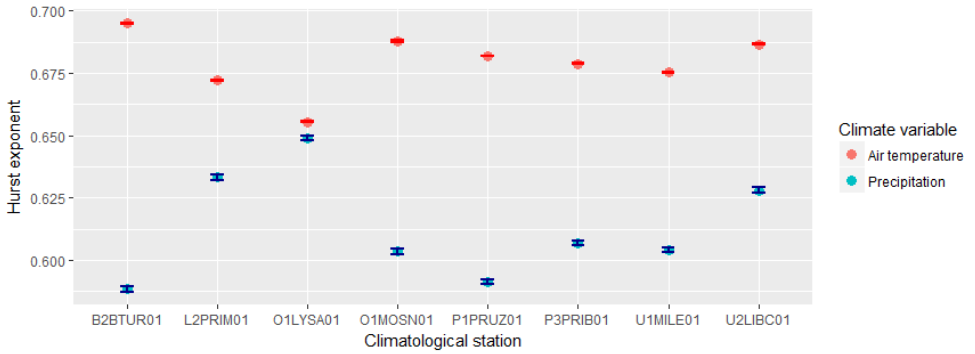

Fig. 3. Hurst exponent wavelet-based estimates in the case of eight selected climatological stations and air temperature and precipitation as variables. Error bars indicate the 95% confidence interval obtained via MEB (see Section 3.3). Assessed period was always 1961–2016.

5 Conclusion and recommendations

In this study, we analyzed selected daily series of discharge, air temperature and precipitation from the territory of Czechia that are currently offered for free. Especially, the length of the series allowed the assessment of the presence of long-term persistence by means of the estimation of the Hurst exponent and stationarity testing. When estimating Hurst exponents, we recommend the use of wavelet analysis that enables one to look at the properties of time series at different scales more thoroughly, which is relevant mainly for the daily time step exhibiting various patterns in wavelet spectra also for different hydrometeorological variables.

Among other insights, it was found that while many air temperature and discharge time series models would benefit from the inclusion of the fractional differencing parameter, the series of precipitation may be modelled without it. Another finding relates to the technique of deseasonalization. Namely, the wavelet-based method should be inspected further. On the other hand, the properties of different types of wavelets were not investigated. Since there were several studies conducted regarding LTP and hydrometeorological series from different parts of Czechia, it is advisable to collect and compare the findings therein and perform a review elucidating also the spatial distribution of LTP in relation to different types of data.

We greatly acknowledge the decision of the CHMI climatologists who made their selected daily time series available online.

Appendix: CHMI climatological data and the R environment

The climatological data offered online by the CHMI climatologists are in a somewhat atypical form resulting from the download from the CLIDATA database. In fact, the files are created in well-known MS Excel, but the data are presented in matrices whose first two columns give information about years and months of observations. Days, however, are provided as column names whose number is always 31, irrespective of the lengths of individual months (not to mention the leap years and the issue of 29 February). Transforming such matrices manually in MS Excel prior to the next time series analyses would be cumbersome and there would certainly be a high risk of mistakes. As we decided to use R statistical software [11] completely for all the computations herein, we needed to load the files into its environment (and/or create more suitable files for future work).

Suppose, for instance, that our goal was to create text files that would bring the data in the so-called ‘long form’ that is undoubtedly much better for time series analysts. There should also be individual files for different stations and variables created. We assume here that the readers have already set the working directory where the original files were placed and that they know how to install additional packages. First, it is necessary to know that the MS Excel files are natively foreign to R. According to our experience, the package ‘readxl’ [35] is currently best in loading these files, able also to look into their particular sheets (probably with different variables). In our example, the variable chosen was precipitation. The script/function for creating files that R is familiar with may be as follows:

```
#create a function entitled ‘xls2txt.files’
xls2txt.files<-function(sheet=8,start="1961-01-01",end="2016-12-31") {
#the function has three arguments specifying the sheet (i.e. precipitation here)
#and the desired time period for a control table (serving also for sorting purposes later)
    library(readxl)                 #load the necessary two packages [35,36]
    library(data.table)
```

```
    for (i in dir()) {                  #we have more files, therefore the 'for loop'
#load (sequentially for each 'i') the matrices from desired sheet as an object called 'tab'
        tab<-read_excel(i,sheet=sheet,skip=3)
#transform to an object of class 'data.table' that can be 'melted'
        tab<-as.data.table(tab)
        tab<-melt(tab,id.vars=1:2)
#get rid of the unwanted periods in the column 'variable' with the indices of days
        tab$variable<-sub("\\.","",tab$variable)
#since there is a problem with creating 'Date' objects using non-existing days, transform to
#a 'data.frame' object
        tab<-as.data.frame(tab)
#and add a new column 'date' created from other three columns
        tab$date<-as.Date(paste(tab[,1],tab[,2],tab[,3],sep="-"))
#create a vector with real dates spanning from the desired beginning to the desired end
        check<-seq(as.Date(start),as.Date(end), "day")
#transform this vector to a 'data.frame' object having only one column 'date'
        check<-data.frame(date=check)
#merge these tables, preserving only the real dates
        result<-merge(check,tab,all.x=T)
#add a column with the station ID
        result$id<-sub(".xls","",i)
#finally, write the table 'result' to a desired text file
        write.table(result[,c(6,2:5)],paste(sub(".xls","",i),".txt",sep=""),col.name=F,row.name
=F,quote=F,sep=",")
    }
}
```

The above block of code can be copied and directly pasted to the R Console, which adds a new function called 'xls2txt.files()'. The arguments of the function default to the values that are thought of as the most frequent. However, they can be changed to any sheet containing data and any time period. The produced text files can be loaded into R as usual. Moreover, they better allow the mapping of missing values, as done for instance in [1] in the case of discharge series.

References

1. O. Ledvinka, E3S Web Conf. **17**, 00051 (2017)
2. J. Beran, Wiley Interdiscip. Rev. Comput. Stat. **2**, 26 (2010)
3. J. Beran, *Statistics for Long-Memory Processes* (Chapman & Hall, New York, 1994)
4. J. Beran, Y. Feng, S. Gosh, R. Kulik, *Long-Memory Processes: Probabilistic Properties and Statistical Methods* (Springer, Berlin, 2013)
5. A. Montanari, in *Theory and Applications of Long-Range Dependence*, edited by P. Doukhan, G. Oppenheim, M.S. Taqqu (Birkhäuser, Boston, Massachusetts, 461–472, 2003)
6. M.N. Khaliq, T.B.M.J. Ouarda, P. Gachon, L. Sushama, Water Resour. Res. **44**, W08436 (2008)
7. M.N. Khaliq, L. Sushama, in *Hydrologic Time Sereries Analysis: Theory and Practice* (Springer Netherlands, Dordrecht, 201–221, 2012)
8. A.I. McLeod, J. Time Ser. Anal. **19**, 473 (1998)

9. K.H. Hamed, J. Hydrol. **349**, 350 (2008)
10. E. Ehsanzadeh, K. Adamowski, Hydrol. Process. **24**, 970 (2010)
11. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2017)
12. S. Fatichi, S.M. Barbosa, E. Caporali, M.E. Silva, J. Geophys. Res. **114**, D18121 (2009)
13. A. Montanari, R. Rosso, M.S. Taqqu, Water Resour. Res. **33**, 1035 (1997)
14. E. Szolgayová, J. Arlt, G. Blöschl, J. Szolgay, J. Hydrol. Hydromech. **62**, 24 (2014)
15. J.W. Kantelhardt, E. Koscielny-Bunde, D. Rybski, P. Braun, A. Bunde, S. Havlin, J. Geophys. Res. **111**, D01106 (2006)
16. O. Ledvinka, in *Hydrogy of a Small Basin 2014*, edited by K. Brych, M. Tesař (Czech Hydrometeorological Institute, Praha, 287–295, 2014)
17. D.B. Percival, A.T. Walden, *Wavelet Methods for Time Series Analysis* (Cambridge University Press, Cambridge, New York, 2000)
18. D. Kwiatkowski, P.C.B. Phillips, P. Schmidt, Y. Shin, J. Econom. **54**, 159 (1992)
19. P.C.B. Phillips, P. Perron, Biometrika **75**, 335 (1988)
20. S.M. Barbosa, M.E. Silva, M.J. Fernandes, in *Nonlinear Time Series Analysis in the Geosciences: Applications in Climatology, Geodynamics and Solar-Terrestrial Physics*, edited by R.V. Donner, S.M. Barbosa (Springer-Verlag, Berlin, 157–173, 2008)
21. O. Ledvinka, Proc. Int. Assoc. Hydrol. Sci. **366**, 188 (2015)
22. O. Ledvinka, Acta Hydrol. Slovaca **16**, 199 (2015)
23. A. Trapletti, K. Hornik, *tseries: Time Series Analysis and Computational Finance* (2017)
24. M.S. Taqqu, V. Teverovsky, W. Willinger, Fractals **3**, 785 (1995)
25. D. Koutsoyiannis, Hydrol. Sci. J. **48**, 3 (2003)
26. H. Tyralis, D. Koutsoyiannis, Stoch. Environ. Res. Risk Assess. **25**, 21 (2011)
27. X. Navarro, F. Porée, A. Beuchée, G. Carrault, Digit. Signal Process. **23**, 1610 (2013)
28. D.B. Percival, Biometrika **82**, 619 (1995)
29. D.B. Percival, in *Nonlinear Time Series Analysis in the Geosciences: Applications in Climatology, Geodynamics and Solar-Terrestrial Physics*, edited by R.V. Donner, S.M. Barbosa (Springer-Verlag, Berlin, 61–79, 2008)
30. W. Constantine, D. Percival, *wmtsa: Wavelet Methods for Time Series Analysis* (2017)
31. H.D. Vinod, J. Asian Econ. **17**, 955 (2006)
32. H.D. Vinod, J. López-de-Lacalle, J. Stat. Softw. **29**, 1 (2009)
33. S.M. Barbosa, J. Clim. **24**, 2516 (2011)
34. S.M. Barbosa, M.G. Scotto, A.M. Alonso, Nat. Hazards Earth Syst. Sci. **11**, 3227 (2011)
35. H. Wickham, J. Bryan, *readxl: Read Excel Files* (2017)
36. M. Dowle, A. Srinivasan, *data.table: Extension of `data.frame`* (2017)