Environmental Modelling & Software 89 (2017) 190–195

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

An R package for daily precipitation climate series reconstruction

Roberto Serrano-Notivoli [a, b, c, *], Martín de Luis [a, c], Santiago Beguería [b]

[a] Department of Geography and Regional Planning, University of Zaragoza, Zaragoza, Spain
[b] Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas (EEAD-CSIC), Zaragoza, Spain
[c] Environmental Sciences Institute (IUCA), University of Zaragoza, Zaragoza, Spain

ARTICLE INFO

Article history:
Received 28 June 2016
Received in revised form
5 September 2016
Accepted 5 November 2016
Available online 13 January 2017

Keywords:
reddPrec
Daily precipitation
Quality control
Missing values
Grid

ABSTRACT

Daily precipitation datasets are usually large, bulky and hard to handle, but they are of key importance in many environmental studies. We developed a tool to create custom datasets from observed daily precipitation records. Reference values (RV) are computed for each day and location using multivariate logistic regression with altitude, latitude and longitude as covariates. The operations were compiled in an Open Source R package called `reddPrec`. The `reddPrec` package consists of a set of functions used to: i) apply a comprehensive quality control over original daily precipitation datasets, flagging suspect data based on five predefined criteria; ii) fill missing values in original data series by estimating precipitation values using the 10 nearest observations for each day; and iii) create new series and gridded datasets in locations where no data were recorded.

1. Introduction

The use of climate variables in any kind of environmental research requires quality-controlled, serially complete and, often, spatially dense datasets. These datasets are used to assess most of the key aspects of climate change, such as temporal trends in mean values and variability, or extreme events. Precipitation is one of the main variables under scrutiny in climate change studies, since there are theoretical reasons to expect an intensification of the global water cycle (and hence, precipitation) related to global warming (Min et al., 2011; Trenberth, 2011; Coumou and Rahmstorf, 2012). Also, extreme precipitation is expected to increase over most of the mid-latitude landmasses by the end of this century (IPCC, 2013). However, statements about temporal trends in mean and extreme precipitation are qualified by the IPCC as having a medium confidence because of two main reasons: i) the lack of data over specific regions and ii) at regional scales, precipitation predictions are hampered by observational uncertainties.

Despite existing global precipitation databases (Haylock et al., 2008; Hofstra et al., 2009; Yatagai et al., 2012; Menne et al., 2012; Schamm et al., 2014) and tools that organise and show this information (Alder and Hostetler, 2015), these data are often not adapted to the needs of regional or local studies, for which additional data may exist. Researchers are thus often faced with the need to create their own quality-controlled, serially complete and spatially dense databases. Presently, several software programs are able to create complete and reliable precipitation datasets, but very few are committed to the daily scale. *HOMER*, for instance (Mestre et al., 2013), is a software package with a synthesis of the best methods in monthly homogenisation and quality control. This work was adapted to daily data in *SPLIDHOM* (Mestre et al., 2011) but it did not include methods to reconstruct missing data. On the other hand, *ProClimDB* (Štěpánek, 2008), is able to apply homogenisation procedures and finally fill missing values with the created reference series through the process, but despite being free software, it relies on a proprietary license and does not have its source code released. *CLIMDEX* (Alexander et al., 2011; WMO-ETCCDI, 2013) is dedicated to creating gridded land-based global datasets of indices representing the more extreme aspects of climate using daily data of temperature and precipitation; however, it does not make reconstructions, so it relies on previously controlled and reconstructed data. *RClimDex* is a version of this program in R language that has detailed quality control called *EXTRAQC* (Aguilar and Prohom, 2011) over daily data. This quality control allows for checking internal coherence, duplicate dates, rounding problems and outliers (inter alia), but it neither fills missing values nor makes complete reconstructions. A few works have described stochastic

* Corresponding author. Department of Geography and Regional Planning, Pedro Cerbuna 12, 50009, Zaragoza, Spain.
E-mail address: rs@unizar.es (R. Serrano-Notivoli).

http://dx.doi.org/10.1016/j.envsoft.2016.11.005
1364-8152/

methods to estimate precipitation at ungauged values (Burton et al., 2013; Kretzschmar et al., 2014; Mehrotra et al., 2015), but they neglect prior quality control of data.

This paper presents `reddPrec`, an R package focused on daily precipitation reconstruction (cran.r-project.org/web/packages/reddPrec). Users are able to obtain serially complete precipitation datasets, estimate new data at ungauged locations and/or create regular grids of daily precipitation based on original data containing missing values or even large data gaps. The remainder of this article is organised as follows. First we give a short introduction to the methodological procedures (Section 2) and provide some information about the internal operation of the functions (Section 3) in the R language. Section 4 describes how each function works in a logical sequence from quality control to gridding. These three functions are applied to the exemplar data that can be found attached to the R package. Finally, we present future developments and conclusions (Section 5).

The analyses described here are based on the exemplar dataset included with the `reddPrec` package and are completely reproducible. The code to generate the figures is included as Supplementary material.

2. Method basics

Most quality control and gap-filling methods rely on the creation of so-called reference series, which are created from the data of neighbouring climate observatories. The creation of such reference series requires long data series with few missing records and a substantial time overlap with the candidate series (i.e. the data series we are interested in reconstructing). Other valuable data that may exist in nearby observatories but that do not fulfil these requirements need to be discarded, which is a sub-optimal use of available information. In contrast, the methodology embedded in the `reddPrec` package creates daily reference values (RV) using all the data recorded at the nearest stations for each target day. Multivariate logistic regression (MLR) is used to compute these RV based on the data of the 10 nearest neighbours (NNS), and geographic and topographic variables as covariates. This method makes an optimal use of all available information, does not depend on the length of the precipitation series, and preserves the local variability of precipitation distribution.

The process of RV creation is used to: 1) apply the quality control of original data; 2) fill the missing values in data series and 3) create gridded datasets considering each point as a blank station.

2.1. Computation of reference values (RV)

RV computation is based on a set of two predicted values: i) a binomial prediction (BP) of the probability of the occurrence of a wet day; and ii) a magnitude prediction (MP) of precipitation.

BP uses the 10 NNS codified as a binomial variable (observed wet or dry day) to compute the probability of the occurrence of precipitation on day i and location l, $BP_{i,l}$:

$$BP_{i,l} = \beta_{0i,l} + \beta_{1i,l} alt_l + \beta_{2i,l} lat_l + \beta_{3i,l} lon_l + \varepsilon_{1i,l} \quad (1)$$

where $\beta_{n^3_{0i,l}}$ are regression coefficients, alt_l, lat_l and lon_l are the altitude, latitude and longitude of the location of interest respectively, and $\varepsilon_{1_{i,l}}$ is the error term. This model is implemented in `R` through `glm()` using a binomial family.

MP uses the observed precipitation magnitudes through a quasi-binomial approach:

$$MP_{i,l} = \beta_{4i,l} + \beta_{5i,l} alt_l + \beta_{6i,l} lat_l + \beta_{7i,l} lon_l + \varepsilon_{2i,l} \quad (2)$$

where $\beta_{n^7_{4i,l}}$ are regression coefficients and $\varepsilon_{2_{i,l}}$ is the error associated with the estimation of the precipitation magnitude.

The final RV is determined by combining MP and BP, using a threshold value of $BP_{i,l} \geq 0.5$ to determine a wet day:

$$RV = \begin{cases} MP & \forall\ BP \geq 0.5 \\ 0 & \forall\ BP < 0.5 \end{cases} \quad (3)$$

In addition to the estimated RV, the method records the corresponding errors $\varepsilon_{2_{i,l}}$ as a measure of uncertainty.

RV are the result of the combination of BP and MP computed with MLR. In the first stages of the method definition, different approaches were tested: we found that the flexibility of other processes, which can be valid for other variables such as temperature, produces an overestimation in the case of precipitation. To avoid this, we used adaptive asymptotes in MLR depending on the data for each location and day.

2.2. Reconstruction process

In a first stage, these RV were used to develop a quality control test, based on the comparison of them with the original data, in order to detect and remove suspect data. Five criteria were defined to detect suspect wet or dry situations (using previous BP) and outliers (using RV). Section 4.2 gives further explanations on quality control.

The gaps in original values were filled to obtain serially complete data series. Once the original dataset was cleaned of anomalous data, RV were computed again to replace the missing data. These new estimations were multiplied by a correction coefficient based in the ratio between the monthly means of daily precipitation of all observed values and the same calculation of predicted data. This correction let preserve the peculiarities of the original series and avoid including inhomogeneities by maintaining the original structure of data series.

The process of gridding was the same as the used to fill the missing values in original series. At this stage, the filled data series were used to compute RV in the new locations using their altitude, latitude and altitude values. The precipitation magnitudes at each location were estimated using the same 10 NNS each day, so there was not needed a subsequent correction.

3. Software design

The precipitation reconstruction method was developed as an R package because of its Open Source characteristics. Implementing this complex model in a widespread programming language allows us to reach a wider audience, and has two main benefits: 1) inexperienced users do not need specific knowledge in climate reconstruction to obtain serially complete datasets; and 2) advanced users are able to explore the source code to learn, modify and improve the procedures therein, and provide useful feedback to the maintainers of the code.

The package `snowfall` was used to embed parallelisation (Knaus et al., 2009), allowing an optimum use of current multiprocessor computers. However, we recommend making several tests with different numbers of processor cores in a subset of the data before running it over the entire dataset. The time elapsed to complete a task can vary greatly, even taking more time when executing the same tasks with more CPUs. The trade-off between computing and data transfer operations may cause a non-linear relationship between the number of CPUs used and the total

computational time; often, an optimum is found.

4. The reddPrec package

Application of the functions in the package allows one to obtain i) a dataset free of anomalous data; ii) a serially complete precipitation dataset; iii) estimation of data series at ungauged locations; and iv) grids of daily precipitation series covering a specified area. To achieve these objectives, the package includes three functions that need be run sequentially (although they can be run separately with the proper input data):

1) `qcPrec()` applies quality control to original data by flagging and removing suspect data not corresponding to the precipitation distribution of each day;
2) `gapFilling()` fills the missing values in each data series from the previously cleaned dataset. This cleaned dataset is dependent on the quality control criteria used to detect and remove suspect data;
3) `gridPcp()` creates new data series in any set of locations based on their latitude, longitude and altitude.

4.1. Input data requirements

The `reddPrec` package does not have a specific function to import data; the input variables are different depending on the function that we want to run:

`qcPrec()` function needs two objects: i) A matrix with real observations of daily precipitation in tenths of millimetre (1/10 mm), with daily observations in rows and different stations in columns. Days with missing values (i.e. no record) are codified as `NA`. ii) A data-frame with location information, with four columns corresponding to the identifier of the code of the meteorological station (ID) and their longitude, latitude and altitude (X, Y, ALT) in metres. For each single day, a minimum number of 11 stations where precipitation values were recorded is needed to calculate a reference value and complete the quality control process.

gapFilling() function fills with estimated values the cleaned dataset obtained from the `qcPrec()` function.

gridPcp() uses the filled dataset obtained from the `gapFilling()` function (or any other daily precipitation dataset), the station information and a data-frame with the identifier of each point of the grid (ID), their coordinates (X, Y) in metres, and the altitude of each one (ALT). The dataset used as input may still contain `NA` values, depending on the parameters used to fill the station precipitation. Nevertheless, using a complete dataset without missing values is recommended, because each point of the grid would use the same stations each day to estimate precipitation.

Additionally, the package includes a daily precipitation dataset as an example to test the operation of the functions. The dataset has been random created for this purpose and contains the location information of 48 stations with their corresponding daily precipitation in 1/10 mm.

4.2. Quality control of daily observations

The quality control made by `qcPrec()` is based on a comparison of the recorded values at each location and day with their corresponding reference values (RV) built with their 10 nearest stations (NNS), as explained in Section 2.1.

Once the data have been loaded, we set the initial and the end date of the quality control process. The initial date has to be the initial day recorded in the dataset. We can use parallel processing and set the number of core processors to use based on our hardware configuration.

A threshold parameter (`thres`) is integrated to set a maximum distance (in km) to the search of the 10 NNS. If it is not possible to find 10 NNS within that distance, the RV cannot be computed, and the original value will not be checked for quality. If the threshold is set as `NA` no distance limit will be applied.

When computing RV, if all NNS observations in a day are zero, no suspect data are flagged. Otherwise, five criteria have been specified to determinate whether it is necessary to flag suspect data. The function applies them sequentially:

QC.1) **Suspect data**: Observed value is over zero and all their 10 NNS are zero.
QC.2) **Suspect zero**: Observed value is zero and all its 10 NNS are over zero.
QC.3) **Suspect outlier**: The magnitude of the observed value is 10 times higher or lower than that predicted by its 10 NNS.
QC.4) **Suspect wet**: Observed value is zero, wet probability is over 0.99 and predicted magnitude is over 5 mm.
QC.5) **Suspect dry**: Observed value is over 5 mm, dry probability is over 0.99 and predicted magnitude is under 0.1 mm.

When `qcPrec()` ends the quality control of all stations, it starts again but using as NNS the 10 closest stations with unflagged values. This process is iterated since no more observations are flagged. The cleaned dataset is written as an `.RData` file containing the cleaned dataset and the station information. If we `set printmeta` as `TRUE`, the removed values are recorded in the subfolder `./meta`, with one file per day. These files contain information about the date, the identifier of the station, the data removed and a codification of the removing criteria (1: Suspect data; 2: Suspect zero; 3: Suspect outlier; 4: Suspect wet; and 5: Suspect dry).

These five criteria were defined to generic situations, but they could not work well in specific situations or type of climates, when other kind of criteria could work better. However, the thresholds can be changed, especially in QC.3, QC.4 and QC.5 to be more flexible or stricter.

4.3. Filling missing values

The process of filling the missing values implies computing new RV based on the depurated dataset resulting from `qcPrec()`. RV are computed with `gapFilling()` for all days and locations, with and without original observations:

The function produces a `Filled.RData` file that contains a matrix with the serially completed stations data and a `./days/` subfolder with one file per day. These files have seven columns indicating: 1) the identifier (ID) of all stations (with recorded data in that day or not); 2) the observed value (obs); 3) the binomial prediction (predb) corresponding to the probability of the occurrence of a wet day; 4) the raw magnitude prediction of precipitation (pred1); 5) the predicted RV (pred2); and 7) the standard error of magnitude prediction (err). In addition, a standardised RV (pred3) (column 6) represents the final corrected values. The reconstruction process does not use the same stations for all days because it selects the 10 available NNS at each moment. If we use the raw estimations (i.e. the outputs of each day and location model) as a final data series, it will have a non-time-homogeneous data series. The correction is calculated by multiplying pred2 by the ratio between obs and pred2 to obtain RV with same mean as that of the observed series. This is applied monthly; that is, computing the ratio of observations and predictions in January, then in February, etc., until December. To ensure a solid correction, we recommend

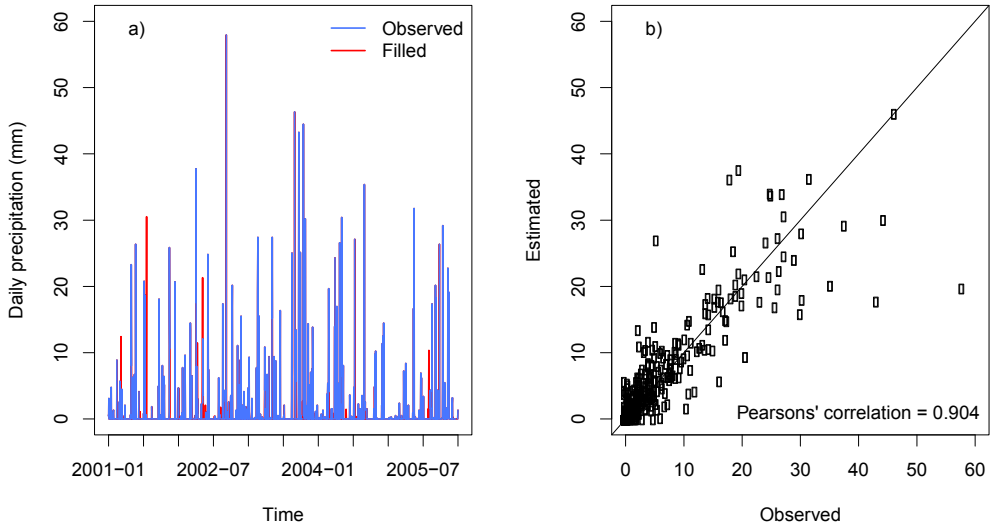

Fig. 1. Reconstruction of a data series (sts_21) with an 8.93% of original missing values. a) Final series with original cleaned data (blue) and estimated values (red) for missing days; b) Comparison between observations and estimations for days with precipitation record. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

using data series with at least 10 years of original values.

The result of this function is a gap-filled matrix of daily precipitation data (Fig. 1a). However, if days with fewer than 11 observations exist in the original dataset, or if a very restrictive distance threshold limiting the possibilities of finding 10 NNS is imposed, a final dataset including missing values could be obtained.

The result of the gap filling process includes the standard error associated with the estimations (Fig. 2). Confidence limits can be constructed from the standard errors, and they can be propagated to further calculations based on the filled data.

4.4. Gridding

The `gridPcp()` function allows new daily gridded datasets to be created. The spatial resolution of the grid will determine the computational time required to complete the estimation process for each day in the dataset. In this case, only the locations of the grid (`points`) are filled, and the original stations (`sts`) are used as the nearest stations to estimate precipitation values. There is no distance limit of searching nearest stations for each grid point; this is due to the need to obtain a completely filled precipitation dataset. With fewer than 10 observations in a day, the grid cannot be computed.

The output of this function is a directory named `./gridded/`, where one file per day is written. These files have the identifier of all the grid points (ID), the estimated precipitation values for each (pred), and the standard error (err) associated with the estimation. The result is not a continuous surface of precipitation, but is rather

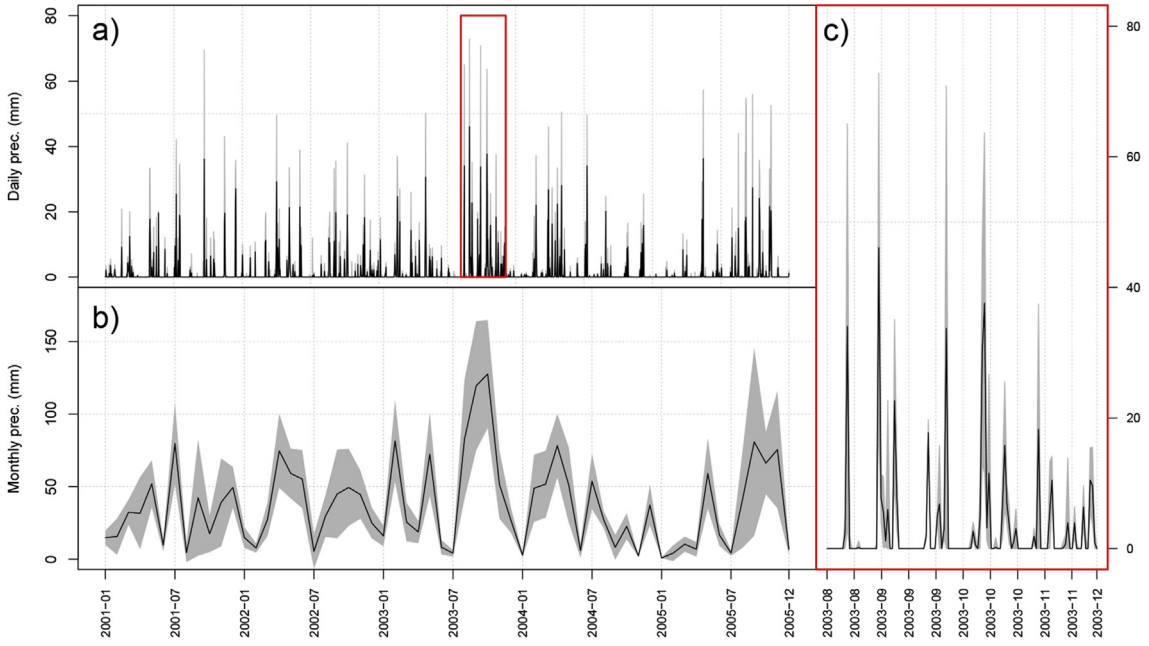

Fig. 2. Estimated values for a data series (sts_21). a) Daily values (black line) and their standard errors (grey lines); b) Aggregated daily values at monthly scale (black line) and their corresponding aggregated standard errors (grey shaded areas); c) Detail of daily estimations (black line) and their standard errors (grey lines).

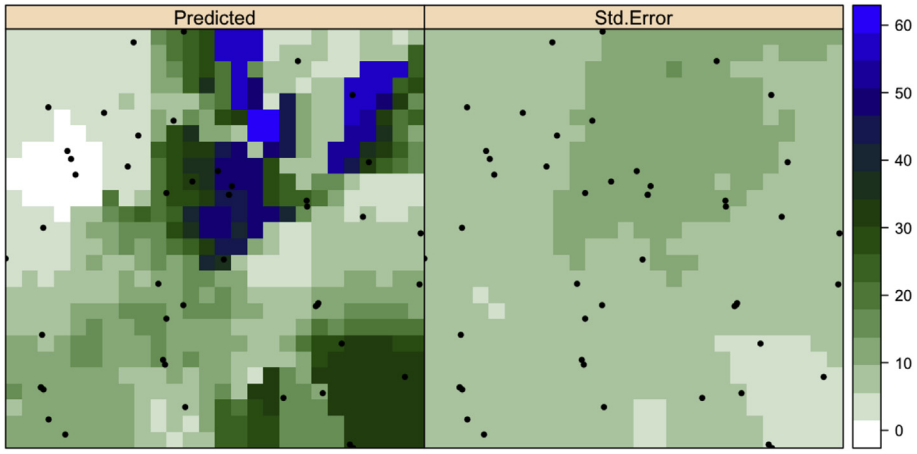

Fig. 3. Precipitation estimation over a gridded set of coordinates for a specific day and its standard error (absolute errors are related to the magnitude of estimations and relative errors are highly related to the irregularity of precipitation in each location). Dots represent the observatories.

the result of a statistical inference process that estimates daily precipitation in specific locations represented by the grid points. To plot the results for any day, it is necessary to join the ID column to the ID of the matrix containing the points' coordinates used as input (Fig. 3). The standard error spatial distribution resulting from the gridding process usually shows higher errors where high values have been estimated. However, in relative terms (i.e. Std. Error/Predictions), the errors are higher where precipitation is more irregular, globally low and with high frequency of extreme events.

5. Future developments, limitations and conclusions

The `reddPrec` R package contains a set of functions to reconstruct daily precipitation datasets, including quality-control, gap-filling and estimation at ungauged locations. However, one of the main limitations is its applicability to areas where there is a low number of observatories, or they are far from each other. Since the quality of the reconstruction, and especially the creation of gridded datasets, is strongly affected by this lack of original information, the uncertainty of predictions could be spread (as noted in Beguería et al., 2015), and final users are not able to minimise the effect of this issue. This uncertainty is measured by the standard error values provided for each precipitation estimation. These standard errors could be used in a bias-correction post-processing stage. With respect to quality control, we set five different criteria to detect suspect values in the original dataset. However, such criteria could not be suitable for every climate region; different thresholds would be necessary in some contexts. Our method is optimised for precipitation data with a daily time resolution. It should not be difficult to adapt it to work with data at finer (sub-daily) or coarser (weekly, monthly) resolutions, although we have not tested this. Such limitations, and others that can be suggested by users, will be implemented in future versions of the package `qcPrec()`.

Acknowledgements

This study was supported by research projects CGL2012-31668, CGL2015-69985-R, CGL2011-24185 and CGL2014-52135-C3-1-R, financed by the Spanish Ministerio de Economía y Competitividad (MINECO) and FEDER-ERDF funds. We were financially supported by the Government of Aragón through the 'Programme of research groups' (groups 'H38, Clima, Cambio Global y Sistemas Naturales' and 'E68, Geomorfología y Cambio Global').

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2016.11.005.

References

Aguilar, E., Prohom, M., 2011. EXTRAQC. Centre for Climate Change and Servei Meteorològic de Catalunya, Spain. www.c3.urv.cat/data.html (Accessed 18 March 2016).

Alexander, L., Donat, M., Takayama, Y., Yang, H., 2011. The CLIMDEX Project: Creation of Long-term Global Gridded Products for the Analysis of Temperature and Precipitation Extremes. WCRP OSC conference, Denver, CO, USA. 24–28 October 2011.

Alder, J.R., Hostetler, S.W., 2015. Web based visualization of large climate data sets. Environ. Model. Softw. 68, 175–180. http://dx.doi.org/10.1016/j.envsoft.2015.02.016.

Beguería, S., Vicente-Serrano, S.M., Tomás-Burguera, M., Maneta, M., 2015. Bias in the variance of gridded datasets leads to misleading conclusions about changes in climate variability. Int. J. Climatol. http://dx.doi.org/10.1002/joc.4561 early view.

Burton, A., Glenis, V., Jones, M.R., Kilsby, C.G., 2013. Models of daily rainfall cross-correlation for the United Kingdom. Environ. Model. Softw. 49, 22–33. http://dx.doi.org/10.1016/j.envsoft.2013.06.001.

Coumou, D., Rahmstorf, S., 2012. A decade of weather extremes. Nat. Clim. Change 2 (7), 491–496. http://dx.doi.org/10.1038/nclimate1452.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. J. Geophys. Res. Atmos. 113 (20) http://dx.doi.org/10.1029/2008JD010201 art. no. D20119.

Hofstra, N., Haylock, M., New, M., Jones, P.D., 2009. Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. J. Geophys. Res. 114, D21101. http://dx.doi.org/10.1029/2009JD011799.

IPCC, 2013. Summary for policymakers. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Knaus, J., Porzelius, C., Binder, H., Schwarzer, G., 2009. Easier parallel computing in R with snowfall and sfCluster. R J. 1/1, 54–59.

Kretzschmar, A., Tych, W., Chappell, N.A., 2014. Reversing hydrology: estimation of sub-hourly rainfall time-series from streamflow. Environ. Model. Softw. 60, 290–301. http://dx.doi.org/10.1016/j.envsoft.2014.06.017.

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. J. Atmos. Ocean. Technol. 29, 897–910. http://dx.doi.org/10.1175/JTECH-D-11-00103.1.

Mehrotra, R., Li, J., Westra, s., Sharma, A., 2015. A programming tool to generate multi-site daily rainfall using a two-stage semi parametric model. Environ. Model. Softw. 63, 230–239. http://dx.doi.org/10.1016/j.envsoft.2014.10.016.

Mestre, O., Gruber, C., Prieur, C., Caussinus, H., Jourdain, S., 2011. SPLIDHOM: a method for homogenization of daily temperature observations. J. Appl. Meteorol. Climatol. 50 (11), 2343–2358. http://dx.doi.org/10.1175/2011JAMC2641.1.

Mestre, O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, E., Böhm, R., Aguilar, E., Guijarro, J., Vertachnik, G., Klancar, M., Bubuisson, B., Stepanek, P., 2013. HOMER: HOMogenisation softwarE in R- methods and applications.

Időjárás 117, 47–67.
Min, S., Zhang, X., Zwiers, F., Hegerl, G., 2011. Human contribution to more-intense precipitation extremes. Nature 470, 378–381. http://dx.doi.org/10.1038/nature09763.
Schamm, K., Ziese, M., Becker, A., Finger, P., Meyer-Christoffer, A., Schneider, U., Schröder, M., Stender, P., 2014. Global gridded precipitation over land: a description of the new GPCC First Guess Daily product. Earth Syst. Sci. Data 6, 49–60. http://dx.doi.org/10.5194/essd-6-49-2014.
Štěpánek, P., 2008. ProClimDB – Software for Processing Climatological Datasets. CHMI, regional office Brno. In: www.climahom.eu/ProcData.html (Accessed 18 March 2016).
Trenberth, K., 2011. Changes in precipitation with climate change. Clim. Res. 47, 123–138. http://dx.doi.org/10.2254/cr00953.
WMO-ETCCDI, 2013. CLIMDEX Software. The World Meteorological Organisation (WMO) Expert Team on Climate Change (ETCCDI), the Australian Research Council's (ARC). Linkage Project LP100200690. www.climdex.org (Accessed 18 March 2016).
Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N., Kitoh, A., 2012. APHRODITE: constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. Bull. Am. Meteorol. Soc. http://dx.doi.org/10.1175/BAMS-D-11-00122.1.