CoRain: a free and open source software for rain series comparison

CoRain: a free and open source software for rain series comparison

D. Guenzi[a]*, F. Acquaotta[ab], D. Garzena[a], S. Fratianni[ab]

[a] Università degli Studi di Torino, dipartimento di Scienze della Terra, via Valperga Caluso 35 – 10125 Torino (Italy)

[b] Centro interdipartimentale sui rischi naturali in ambiente montano e collinare NatRisk, via Leonardo da Vinci 44 – 10095 Grugliasco (TO, Italy)

*Corresponding author. Tel.: +39 011 670 5172, e-mail: diego.guenzi@unito.it

Abstract

A good climatic analysis requires accurate and homogeneous daily precipitation series; unluckily, inhomogeneity is frequently found and have to be considered, especially when it is due to non-climatic parameters. CoRain is a free and open source software written in R language that could greatly help analyzing inhomogeneity caused by rainfall measuring instruments. CoRain compares two parallel rain series (with an overlapping period) and tries to highlight overestimations and underestimations due to rain gauges in a specific condition, so that the user can consider it for future analysis. CoRain offers many information on the two analyzed series, starting with cleaning input data, comparing them and classifying rainy days by severity. CoRain is a cross-platform software, easily adaptable to different needs, that takes in input a single text file with daily information of the two rain series and outputs tables (in CSV format) and plots (as PNG images) that help in the interpretation of the data. Use of the program is very simple: the execution can be either interactive or non-interactive. CoRain code has been tested on different rain series in the Piedmont region (northwestern Italy), showing its importance in identifying climate variations and instrumentation errors.

Keywords

Open source software, R project, precipitation series comparison, rain gauges network, parallel measurements, quality control

Published on Earth Science Informatics (http://link.springer.com/article/10.1007/s12145-017-0301-y and http://rdcu.be/qEsU)

1. Introduction

Studying and analyzing extreme rain events, dry and wet periods or trends and return times can help in planning and containing the effects of the climate change (Acquaotta et al. 2015, Terzago et al. 2010, 2012, 2013, Zandonadi et al. 2016). Availability of daily precipitation series is a necessary but not sufficient condition to make a good climatic analysis and to better understand extreme events (Mekis and Vincent 2011, Venema et al. 2013). To make accurate climatic analysis we have to use homogeneous daily series of good quality (Acquaotta et al. 2009, Aguilar et al. 2003, Parker 1994, Peterson et al. 1998). In this perspective several international projects are created, aimed at the promotion, recovery and exchange of meteorological series of high quality as the MEDARE initiative - MEditerranean DAta REscue (WMO 2012), an international project born under the auspice of the World Meteorological Organization, with the main objective of developing, consolidate and progress climate data and metadata rescue activities across the Greater Mediterranean Region. Moreover, the importance of the meteorological data can be seen into many international dataset (ECAD - European Climate Assessment and Data, GCOS - Global Climate Observing System, GHCN - Global Historical Climatology Network…) where long instrumental climate records are available. These datasets are essential since they are the basis for assessing century-scale trends and can be used in the validation of climate models as well as detection and attribution of climate change at regional scale. The value of these datasets, however, depends strongly on the homogeneity of the time series. In fact, once climate change became an issue of central importance, some skepticism arose about the results of data analysis work, which frequently indicated sharp and determined changes in regional climates. It is now well recognized that variations in many long term time series are not only caused by changes in weather and climate, but also by changes in the positioning of the stations, changes in instruments, formulae used to calculate means, observing practices and station environment (Göktürk et al. 2008, Heino 1994, Karl and Williams 1987). In addition, inhomogeneity in rain gauges precipitation measurements can be caused also by changes in wind-induced undercatchment, wetting losses (water adhering to the surface of the inner walls) and evaporation losses (Bodtmann and Ruthroff 1976, Sevruk and Zahlavova 1994, Sevruk et al. 2009); changes in instrument geometry, in the neighboring environment and in the methods of recording can also cause inhomogeneity due to undercatchment. The Commission for Instruments and Methods of Observation (CIMO) of the World Meteorological Organization (WMO) has recognized the need to conduct a series of comparisons of instruments in order to highlight and classify these discontinuities in precipitation recordings (Goodison et al. 1998, Lanza and Vuerich 2009, Sevruk and Klemm 1989). CoRain software has the objective of highlighting and classifying dissimilarities in daily precipitation among different instruments for rain measuring, in addition to the analysis of the inhomogeneity caused by rainfall measuring instruments. It compares the *candidate series* (e.g.: a rain series coming

from an old instrument) with the *reference series* (e.g.: the one recorded from a more recent and efficient instrumentation), evaluating mean errors between the two series and the overestimation or underestimation of a particular instrument in a specific condition (Baciu et al. 2005, Boroneant et al. 2006). The information acquired with this program can improve the understanding of inhomogeneity and continuity of the series also allowing, in some cases, the correction of discontinuities. The development of efficient comparison method is very important for detection and correction of inhomogeneity because an accurate comparison of series can increase both the significance and the power of the correction factor estimated by homogeneity test. CoRain software has been written in R language (R Development Core Team 2011), is open source and freely available online[1] under GNU GPL v3 license (Free Software Foundation 2007).

2. Methodology

CoRain software uses an innovative analysis approach combining a set of well-known statistical tools and works in three steps (Acquaotta et al. 2016): (i) statistical analysis, (ii) comparison between the series and (iii) comparison between precipitation classes.

The statistical information calculated on each series are: the minimum value of precipitation, the 1st quantile, the median, the mean, the 3rd quantile, the maximum value, the number of missing values, the total number of values and the results of Shapiro-Wilk test (Acquaotta et al. 2009, 2016, Giaccone et al. 2015, Isotta et al. 2013); in addition to daily values, monthly values are also calculated for each series.

The second step is the comparison between the parallel series, using the overlapping period. In order to be able to make a direct comparison only between the recorded daily rain series, any values that were missing in one series were also set to be missing in its counterpart, to avoid modifying the series. Then the values lower than 1 mm were dropped (Wang et al. 2010). A set of statistical tests were carried out to show the differences or the similarities between the series. The Student's T test allowed identifying if the series have the same mean, the Wilcoxon rank test was used to establish if the series have the same median and the Kolmogorov-Smirnov test was used to see if the series have the same distribution. In addition, CoRain also applies the Kruskal-Wallis test, computes the root mean square error (RMSE) and the correlation using Spearman correlation coefficient. A p=5% significance level was used for all the tests. In order to identify the months or seasons with the greatest differences, the percentage relative errors (Lanza and Stagi 2012) are also calculated from the monthly precipitation data. On this new variable, a statistical analysis is carried out, calculating the median, the 1st and the 3rd quantiles and the trends.

[1] https://github.com/UniToDSTGruppoClima/CoRain

The third step allows identifying if the two rain gauges recorded the same precipitation events. The daily rain events were classified in five classes, *weak*, *mean*, *heavy*, *very heavy* and *extreme*. After that, the software estimates the number of common events included in these ranges and the maximum Systematic Error of the Precipitation measurement (SEP) is calculated for these events (Sevruk and Klemm 1989, Sevruk et al. 2009, WMO-CIMO 2008).

3. Program requirements

Since R is an interpreted language, CoRain requires it to be executed, resulting in easy portability on different types of platforms and operating systems. It has been successfully tested under R versions 3.2.2 (Fire Safety) and 3.3.1 (Bug in Your Hair), both under Windows and Linux, but it is likely to run even on later versions. To work, CoRain requires some packages installed in the R environment that are *class*, *zoo*, *hydroGOF*, *xts*, *hydroTSM*, *zyp*, *Kendall* and their dependencies. After their installation, the program can be run either interactively (i.e.: from RStudio) or in a non-interactive way (i.e.: from Rscript). If the execution is interactive, CoRain asks for an input text file; in the other case, name and path of that file are hardcoded inside the first part of the source code (see Online Resource 1) and the user has to modify them according to the environment.

4. Input file

The input of the program is a text file formatted with five TAB-separated columns. Two of the five columns are the rain series that are assumed to have already passed an external quality check, to highlight and remove incorrect values such as daily precipitation lower than 0 mm (Gonzalez-Rouco et al. 2001), for example. The first row of that file contains the headers of the columns (column names) while the first three columns contain year, month and day of the two series. Column four contains the values in mm/day for the candidate rain series and column five the values of the reference rain series; missing values are allowed if explicitly expressed as *NA*. Series in input file must start 1st of January and end on 31st December, possibly having at least 5 years of data (Vincent and Mekis 2009). Attached with the program it is possible to find some examples of input files that could be used to execute CoRain and practice with it (see Online Resources from 2 to 7).

5. Program features and implementation

The program works in three consecutive phases, following what is described in paragraph 2: the first one is the cleaning of input data and computation of statistical analysis; the second one is the comparison between the two series and the

last one is the comparison between the different precipitation classes. A graphical overview of these steps could be seen in Fig. 1.

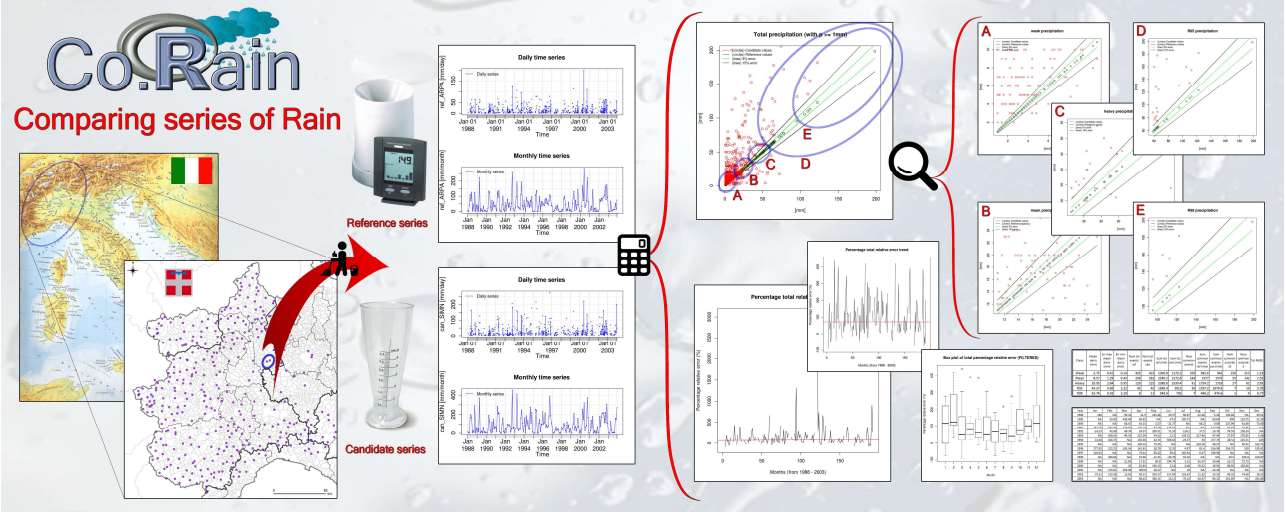

Fig. 1*: overview of the main steps and features of CoRain program - cleaning of data, comparison of rain series and organization in five classes*

5.1. Cleaning of input data and statistical analysis

First, a statistical analysis is carried out on the two series, the candidate one (column 4 of the input file) and the reference one (column 5). After this, the program removes all values with daily rain < 1mm, setting those values to *NA*, and then puts the same missing values on both series (where there is a *NA* in a series, it puts a *NA* on the other). The next part is the application of the previous statistical analysis on the new series obtained after the cleaning process. In this stage, in interactive mode, the program shows to the user a summary of the statistical analysis, asking if the user wants to proceed or, in case the cleaning process has been too aggressive (i.e.: has removed too many values), to stop the execution to review the input file. If the program is in non-interactive mode or the user agrees in going over, CoRain proceeds directly with the next step. The results of the statistical analysis are also reported in the CSV file *0_statistics_input_file.csv* (see Table 1), both for the original input data and for the cleaned one (that could be also verified in the *1_cleaned_input_file.txt*). In the two CSV files called *2_<series_name>_daily_avalibillity.csv* (one for the candidate series and one for the reference one) there are written the monthly number of days used in this analysis (see Table 2) while, in the other two CSV files named *2_<series_name>_monthly_rain.csv* (one for the candidate series and one for the reference one) it is possible to find the monthly amount of rain of the two series (see Table 3). In addition, in the two PNG files called *3_<series_name>_day_month_rain.png* (always one for the candidate series and one for the reference one) are reported the plots of daily and monthly precipitation series (see Fig. 2).

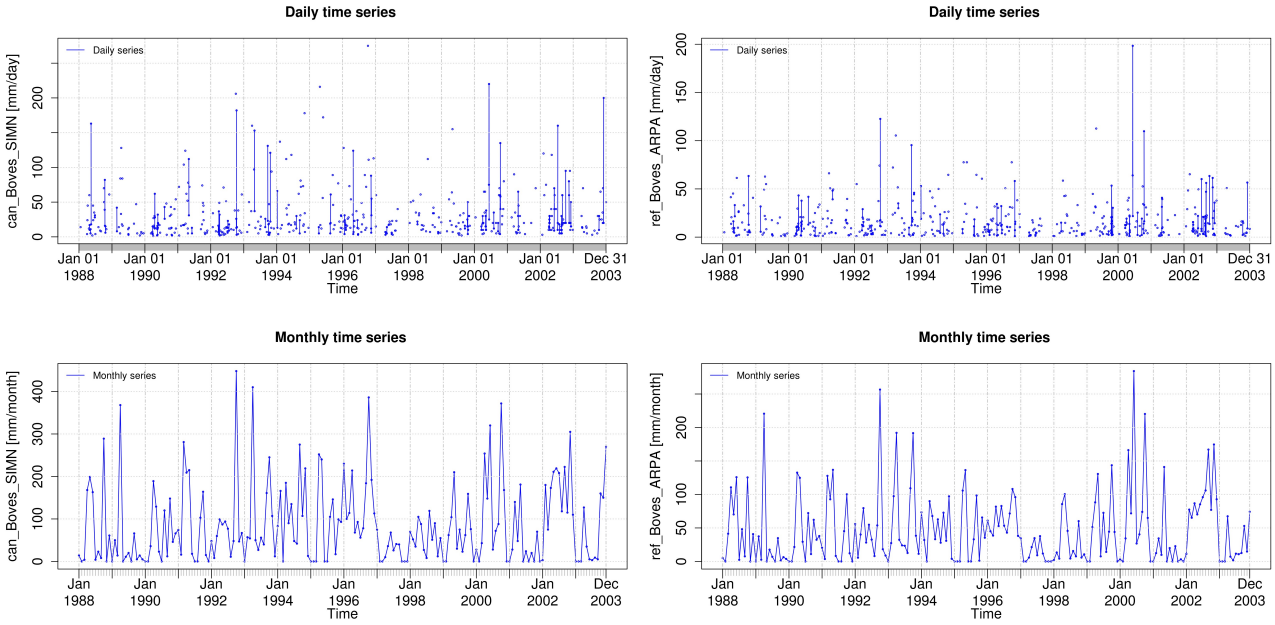

***Fig. 2**: plots of daily and monthly precipitation, both for candidate (left) and reference (right) series, taken from the output of the processing of input file example2.txt (Boves stations, NW Italy - see Online Resource 3)*

5.2. Comparison between the cleaned series

In this second phase of the program, a statistical comparison analysis between the two cleaned series is carried out, applying different statistical methods described in paragraph 2. The results of the comparison analysis are shown in the CSV file *4_statistics_between_daily_series.csv* (see Table 4) where, for every test, CoRain reports the statistical values and the p value. After this, the maximum error boundaries (equal to ± 15% of daily value of the reference series) and minimum error boundaries (equal to ± 5% of daily value of the reference series) are calculated (Sevruk and Klemm 1989, Sevruk et al. 2009, WMO-CIMO 2008). These boundaries, currently, are hardcoded and can be changed only by editing the source code of the program; in the next release of the software we planned a parameterization of these values, adding the possibility of changing them easily. A scatter plot named *5_total_precipitation_plot_with_15_5.png* is created with daily values from both series and with the addition of previously defined error boundaries. In this scatter plot, red

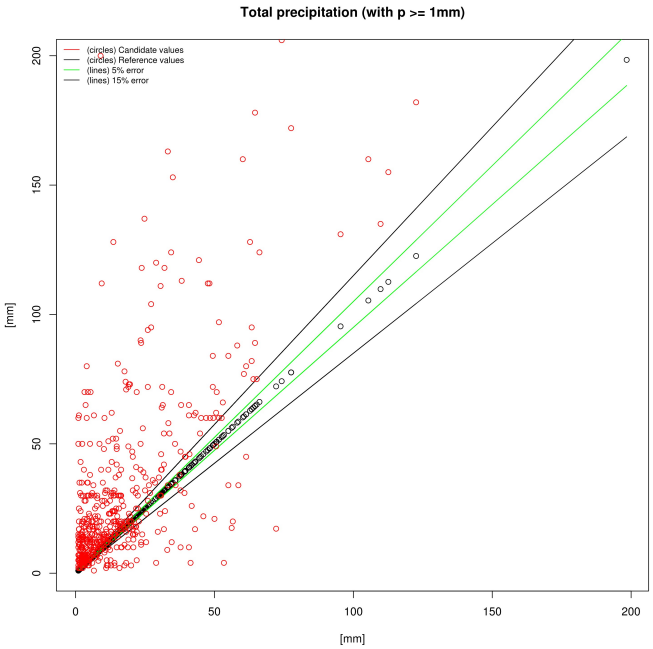

***Fig. 3**: precipitation daily values from candidate series (red circles) and reference series (black circles), with error boundaries of 5% (green line) and 15% (black line), taken from the output of the processing of input file example2.txt (Boves stations, NW Italy - see Online Resource 3)*

circles represents candidate values, black ones are reference values and the green and black lines represents the boundaries of 5% end 15% error, respectively (see Fig. 3). After this, the monthly percentage relative error is calculated for the two cleaned series (Lanza and Stagi 2012) and is reported both numerically in the CSV file *6_percentage_relative_error.csv* (see Table 5) and using a box plot called *7_percentage_relative_error_boxplot.png* (see Fig. 4). On the percentage relative error series, CoRain also computes the non-parametric trend using the Theil-Sen approach. The Mann-Kendall test for the trend is then run on the resulting time series to compute the level of significance (Sen 1968, Toreti and Desiato 2008, Zhang et al. 2000). The results of the trend are plotted in *8_percentage_relative_error_trend.png* (see Fig. 5) and also written in the CSV file *8_stats_percentage_relative_error_trend.csv* where CoRain reports the intercept, the trend over the total period (delta of whole period analysis), the tau of Mann-Kendall test and its p value (see Table 6). After this analysis, if some values of percentage relative error are outside a specific range (according to Acquaotta et al. 2016, defined in ±500% by default but easily configurable inside the source code), CoRain removes them and start a re-computation only of the values inside the accepted range, writing new results in *6_filtered_percentage_relative_error.csv*, *7_filtered_percentage_relative_error_boxplot.png*, *8_filtered_percentage_relative_error_trend.png* and *8_filtered_stats_percentage_relative_error_trend.csv*, with the same information described previously.

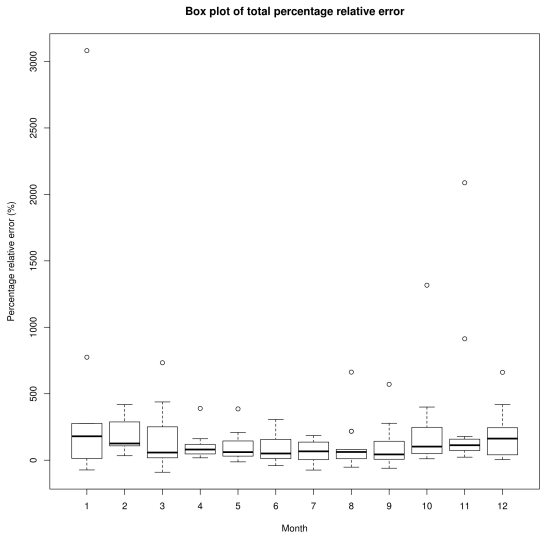

***Fig. 4**: box plot of monthly-aggregated percentage relative error, taken from the output of the processing of input file example2.txt (Boves stations, NW Italy - see Online Resource 3)*

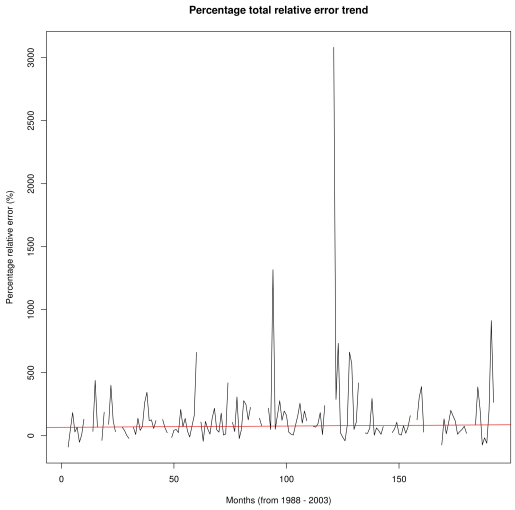

***Fig. 5**: trend of percentage relative error for every month, taken from the output of the processing of input file example2.txt (Boves stations, NW Italy - see Online Resource 3)*

5.3. Comparison between five classes of precipitation events

In this final step, CoRain calculates the quantiles of the reference series to identify the thresholds for every class. Every rainy day is classified as *weak, mean, heavy, very heavy* (R95) or *extreme* (R99). For each class CoRain returns a scatter

plot of daily data, using the same logic and graphical conventions presented in the previous scatter plot called *5_total_precipitation_plot_with_15_5.png*. The five plots are class-specific and are called *9_events_weak.png*, *9_events_mean.png*, *9_events_heavy.png*, *9_events_R95.png* and *9_events_R99.png* (see Fig. 6); in these scatter plots are also reported the error boundaries, ± 15% of daily value and ± 5% of daily value. For each class, in the CSV file *10_class_events_and_RMSE.csv*, CoRain reports statistics like the thresholds, the mean value of maximum and minimum errors, number of events for reference and candidate series, precipitation amount for reference and candidate series, number of events recorded in the same day, precipitation amount for reference and candidate series recorded in the same day, number of events outside the range for maximum and minimum error and RMSE (see Table 7). To better explain all inputs, outputs and parameters used by CoRain software, we have summarized them in Table 8.

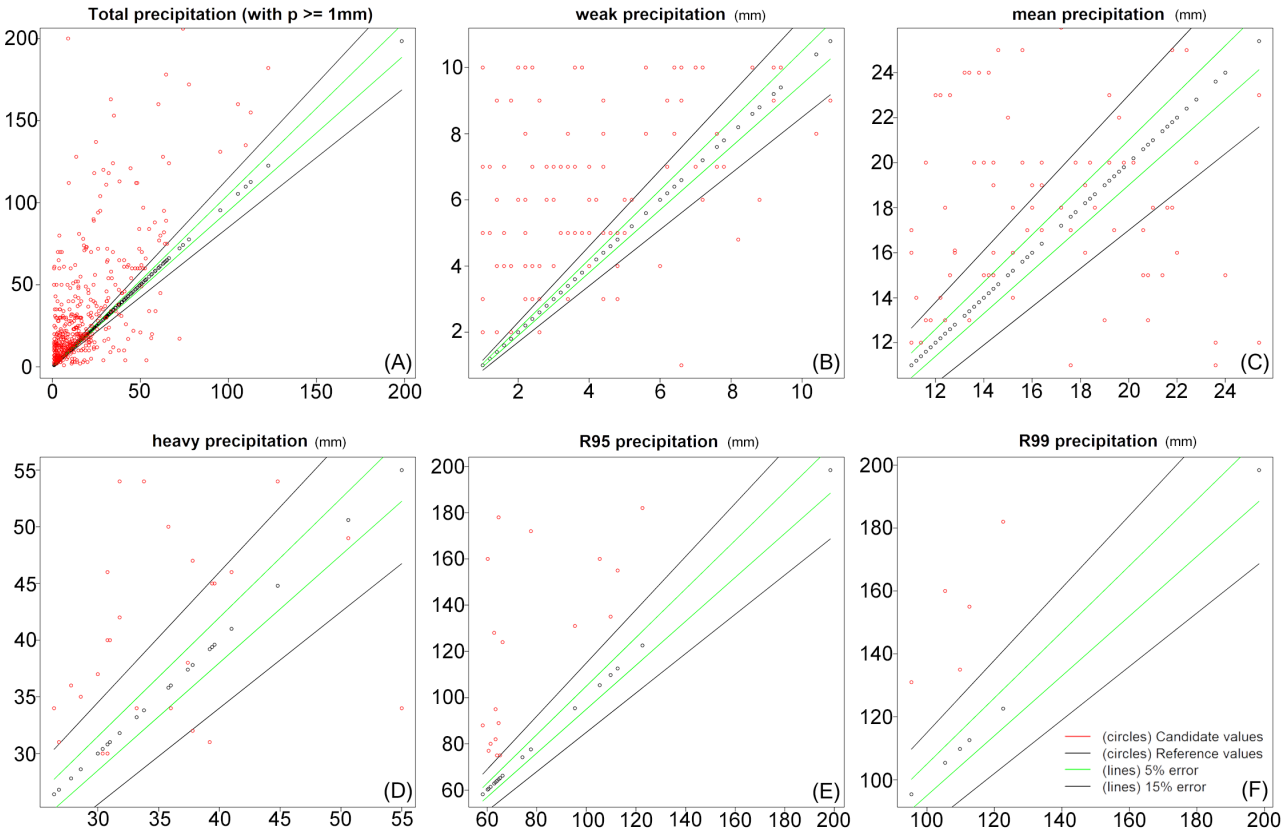

Fig. 6*: class organization of all rainy days (A): weak precipitation (B), mean precipitation (C), heavy precipitation (D), very heavy precipitation also called R95 (E) and extreme precipitation also called R99 (F); plots are taken from the output of the processing of input file example2.txt (Boves stations, NW Italy - see Online Resource 3)*

6. CoRain application: a case study

Here we are going to present a case study where we applied CoRain software using all its features, described above. Boves is an Italian town located in Piedmont (northwestern Italy) with two neighboring weather stations, one manned and one automatic. The manned weather station is the older one; it started to record the rain in 1913 and it was closed in

2003. The station was located at an elevation of 590 m ASL. The automatic weather station has started to record in 1988 and is currently operative. Its elevation is 575m ASL and the distance from the manned weather station is equal to 1240 m. The two weather stations have an overlapping period of 16 years, resulting a perfect example for the CoRain application. Moreover, having 16 years of overlapping data between a manned station and an automatic one is a very rare and interesting case. Most nations have just 1 or 2 overlapping years of data (e.g.: Canada - Milewska and Hogg 2002; Romania - Baciu et al. 2005; Spain - Gilabert 2016, Brunet et al. 2006; Norway - Forland et al. 1998), making the comparison between stations less accurate. These information can be used to highlight discontinuities among rain monitoring networks data, to enhance the following homogenization tests correction. In this way, for example, it is possible to analyze non-climatic parameters that could alter real trends of meteorological series.

The first step of CoRain has highlighted that the two series have an equal number of missing values (around 2% of the data) but, after the cleaning process of the daily values, only 10% of data (approx.) can be utilized (see Table 1). The statistical analysis shows great differences between the two series: for example, the mean, maximum value and quantiles are very different. The differences are confirmed by the statistical tests in step two (see Table 4). Kolmogorov-Smirnov test does not show similar probability distributions between the pairs of stations. The Student's T test and the Wilcoxon rank test does not highlight the same mean and median. The RMSE is also very high for these stations and the rank correlations is equal to 0.59. The monthly mean percentage relative error is equal to 10% (approx.); the largest differences were recorded in the winter months, where the mean monthly percentage relative error was 47% for December, 35% for January and 54% for February. The trend calculated on percentage relative error does not highlight a systematic long-term change in the quality between reference and candidate series (see Table 6). The analysis of precipitation events into classes shows in detail the differences between reference and candidate series (see Table 7): except for the *weak* class, the candidate series shows a greater total number of events. On average, the candidate series measures 36 events more. The greater difference is recorded for the *very heavy* events, with 59 events more than the candidate series, followed by the *mean* class with 30 additional events. Only for the *weak* class the reference series records a greater number of events (115 more). According to the characteristics of the area, the results obtained through the use of CoRain for these two series has shown major differences in the registrations of rain gauges data, thus indicating that the series cannot be joined without the application of a daily homogenization tests.

7. Conclusions and future work

Parallel measurements analysis is a critical step before performing climate analysis in order to identify non-climatic changes in climate records. This is especially true when working with precipitation, where the relative statistical

homogenization is hampered by low cross-correlations between stations. The Parallel Observations Science Team is working on a large database with parallel station measurements of all the essential climatic elements in order to be able to study the characteristics of non-climatic changes in large sets as a function of the local climate (POST 2015). Following these ideas, this study describes CoRain, a free and open source R software used to compare different rain series. CoRain has been tested on a large amount of series, mainly in the northwestern part of Italy and has shown ease of use, efficiency and quickness, requiring only a basic knowledge of R language. Being an open source software is very important for this program, since it could be easily modified and improved directly from the community of users, fitting most needs of climatologists and researchers. In this sense, users are encouraged to report bugs, feature requests or changes they make to the code, either directly to the authors or by using GitHub platform. In the future, if the software becomes popular, it could also be re-factored as an R package and submitted to CRAN for easier and broader adoption. Furthermore, we are working to introduce new features in CoRain for the comparison of other meteorological parameters such as the temperature and we are investigating the comparison of CoRain results with other recent studies that empirically described the correlation of nearby spatial rain measurements (e.g.: Guenzi et al. 2016, Peleg et al. 2013).

Appendix A. Supplementary material

Supplementary data and the program source code associated with this article can be found both in the online version of the article at http://link.springer.com/article/10.1007/s12145-017-0301-y and at https://github.com/UniToDSTGruppoClima/CoRain

References

Acquaotta F., Fratianni S., Cassardo C. and Cremonini R. (2009) On the continuity and climatic variability of meteorological stations in Torino, Asti, Vercelli and Oropa. Meteorology and Atmospheric Physics, 103:279–287 doi: 10.1007/s00703-008-0333-4

Acquaotta F., Fratianni S. and Garzena D. (2015) Temperature changes in the North-Western Italian Alps from 1961 to 2010. Theoretical and Applied Climatology, 122:619–634 doi: 10.1007/s00704-014-1316-7

Acquaotta F., Fratianni S. and Venema V. (2016) Assessment of parallel precipitation measurements networks in Piedmont, Italy. International Journal of Climatology, doi: 10.1002/joc.4606

Aguilar E., Auer I., Brunet M., Peterson T.C. and Wieringa J. (2003) Guidelines on climate metadata and homogenization. WMO-TD No. 1186, WCDMP No. 53, World Meteorological Organization, Geneva (Switzerland) 52 pp.

Baciu M., Copaciu V., Breza T., Cheval S., and Pescaru I.V. (2005) Preliminary results obtained following the intercomparison of the meteorological parameters provided by automatic and classical stations in Romania. In WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation (TECO-2005), Bucharest

Bodtmann W. and Ruthroff C. (1976) The measurement of 1 min rain rates from weighing raingage recording. Journal of Applied Meteorology 15:1160–1166

Boroneant C., Baciu M. and Orzan A. (2006) On the statistical parameters calculated for the essential climatological variables during 2-years of parallel observations with automatic and classical stations in Romania. In 5th Seminar on Homogenization and Data Quality in the Climatological Databases, Budapest

Brunet M., Saladié O., Jones P., Sigrò J., Aguilar E., Moberg A., Lister D., Walther A., Lòpez D. and Almarza C. (2006) The development of a new daily adjusted temperature dataset for Spain (1850 – 2003). International Journal of Climatology 26:1777–1802 doi:10.1002/joc.1338

Forland E. J., Alexandersson H., Dahlstrom B., Drebs A., Frich P., Hanssen-Bauer I., Heino R., Helminen J., Jhonsson T., Nordli P. O., Palsdottir T., Smith T., Tuomenvirta H., Tveito O. E. and Vedin H. (1998) REWARD: relating extreme weather to atmospheric circulation using a regionalized dataset, final report (1996–1998). DNMI Report 17/98 KLIMA

Free Software Foundation (2007) GNU General Public License (GPL v3). Available online at http://www.gnu.org/licenses/gpl.html Accessed on 19 June 2016

Giaccone E., Colombo N., Acquaotta F., Paro L. and Fratianni S. (2015) Climate variations in a high altitude Alpine basin and their effects on a glacial environment (Italian Western Alps). Atmosfera 28:117–128

Gilabert A. (2016) Assessment of the bias introduced by the automatisation of climate record combining climatological and meteorological. Ph.D Thesis, Geography Department Centre for Climate Change, p 209

Göktürk O. M., Bozkurt D., Şen O. L. and Karaca M. (2008) Quality control and homogeneity of Turkish precipitation data. Hydrological Processes, 22 (16):3210–3218 doi: 10.1002/hyp.6915

Gonzalez-Rouco J. F., Jimenez J. F., Quesada V. and Valero F. (2001) Quality Control and Homogeneity of Precipitation Data in the Southwest of Europe. Journal of Climate 14:964-978

Goodison B. E., Louie P. Y. T. and Yang D. (1998) WMO solid precipitation measurement intercomparison - Final report. WMO-TD No. 872, Instruments and observing methods report No. 67, World Meteorological Organization, Geneva (Switzerland) 318 pp.

Guenzi D., Fratianni S., Boraso R. and Cremonini R. (2016) CondMerg: an open source implementation in R language of conditional merging for weather radars and rain gauges observations. Earth Science Informatics, ISSN 1865-0473, doi: 10.1007/s12145-016-0278-y

Heino R. (1994) Climate in Finland during the period of meteorological observations. Finnish Meteorological Institute Contributions, 12, p 209

Isotta F., Frei C., Weilguni V., Tadic M., Lassègues P., Rudolf B., Pavan V., Cacciamani C., Antolini G., Ratto S., Munari M., Micheletti S., Bonati V., Lussana C., Ronchi C., Panettieri E., Marigo G. and Vertacnik G. (2013) The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data. International Journal of Climatology 34:1657–1675 doi: 10.1002/joc.3794

Karl T. R. and Williams C. (1987) An approach to adjusting climatological time series for discontinuous inhomogeneities. Journal of Climate and Applied Meteorology, 26: 1744–1763

Lanza L. and Stagi L. (2012) Non-parametric error distribution analysis from the laboratory calibration of various rainfall intensity gauges. Water Science and Technology 65(10):1745–1752

Lanza L. and Vuerich E. (2009) The WMO field intercomparison of rain intensity gauges. Atmospheric Research 94:534–543

Mekis E. and Vincent L. (2011) An overview of the second generation adjusted daily precipitation dataset for trend analysis in Canada. Atmosphere-Ocean 2:163–177 doi: 10.1080/07055900.2011.583910

Milewska E. and Hogg W. D. (2002) Continuity of climatological observations with automation. Atmosphere-Ocean 40(3), 333-359

Parker D.E. (1994) Effects of changing exposure of thermometers at land stations. International Journal of Climatology 14:1–31 doi: 10.1002/joc.3370140102

Peleg N., Ben-Asher M. and Morin E. (2013) Radar subpixel-scale rainfall variability and uncertainty: lessons learned from observations of a dense rain-gauge network. Hydrology and Earth System Sciences, 17:2195-2208, doi: 10.5194/hess-17-2195-2013

Peterson T.C., Easterling D.R., Karl T.R., Groisman P., Nicholls N., Plummer N., Torok S., Auer I., Boehm R., Gullett D., Vincent L., Heino R., Tuomenvirta H., Mestre O., Szentimrey T., Salingeri J., Førland E.J., Hanssen-Bauer I., Alexandersson H., Jones P. and Parker D. (1998) Homogeneity adjustments of in situ atmospheric climate data: a review. International Journal of Climatology 18:1493–1517 doi: 10.1002/(SICI) 1097-0088(19981115)18:13<1493::AID-JOC329>3.0.CO;2-T

POST (2015) Parallel Observations Science Team. Available online at http://www.surfacetemperatures.org/databank/parallel_measurements Accessed on 19 June 2016

R Development Core Team (2011) The R Project for Statistical Computing. Available online at https://www.r-project.org/ Accessed on 19 June 2016

Sen P.K. (1968) Estimates of the regression coefficient based on Kendall's Tau. Journal of the American Statistical Association 63:1379–1389

Sevruk B. and Klemm S. (1989) Catalogue of national standard precipitation gauge. WMO-TD No. 313, Instruments and observing methods report No. 39, World Meteorological Organization, Geneva (Switzerland) 24 pp.

Sevruk B. and Zahlavova L. (1994) Classification system of precipitation gauge site exposure: evaluation and application. International Journal of Climatology 14:681–689

Sevruk B., Ondras M. and Chvila B. (2009) The WMO precipitation measurement intercomparisons. Atmospheric Research 92:376–380

Terzago S., Cassardo C., Cremonini R. and Fratianni S. (2010) Snow precipitation and snow cover climatic variability for the period 1971–2009 in the SouthWestern Italian Alps: the 2008–2009 snow season case study. Water 2:773–787 doi: 10.3390/w2040773

Terzago S., Cremonini R., Cassardo C. and Fratianni S. (2012) Analysis of snow precipitation during the period 2000-09 and evaluation of a snow cover algorithm in SW Italian Alps. Geografia Fisica e Dinamica Quaternaria 35:91–99

Terzago S., Fratianni S. and Cremonini R. (2013) Winter precipitation in Western Italian Alps (1926–2010) Trends and connections with the North Atlantic/Arctic Oscillation. Meteorology and Atmospheric Physics 119:125-136 doi:10.1007/s00703-012-0231-7

Toreti A. and Desiato F. (2008) Changes in temperature extremes over Italy in the last 44 years. International Journal of Climatology 28:733–745 doi: 10.1002/ joc.1576

Venema V. K. C., Mestre O., Aguilar E., Auer I., Guijarro J. A., Domonkos P., Vertacnik G., Szentimrey T., Stepánek P., Zahradnicek P., Viarre J., Müller-Westermeier G., Lakatos M., Williams C. N., Menne M.J., Lindau R., Rasol D., Rustemeier E., Kolokythas K., Marinova T., Andresen L., Acquaotta F., Fratianni S., Cheval S., Klancar M., Brunetti M., Gruber C., Prohom Duran M., Likso T., Esteban P., Brandsma T. and Willet K. (2013) Benchmarking homogenization algorithms for monthly data. AIP Conference Proceedings,1552 8:1060-1065 doi 10.1063/1.4819690

Vincent L. and Mekis E. (2009) Discontinuities due to joining precipitation station observations in Canada. Journal of Applied Meteorology and Climatology 48:156–166 doi: 10.1175/2008JAMC2031.1

Wang X., Chen H., Wu Y., Feng Y. and Pu Q. (2010) New techniques for the detection and adjustment of shifts in daily precipitation data series. Journal of Applied Meteorology and Climatology 49:2416–2436 doi: 10.1175/2010JAMC2376.1

WMO (2012) Medare initiative. Available online at http://www.omm.urv.cat/MEDARE/ Accessed on 19 June 2016

WMO-CIMO (2008) Guide to Meteorological Instruments and Methods of Observation. WMO-No. 8, 7th edition, World Meteorological Organization, Geneva (Switzerland) 681 pp.

Zandonadi L., Acquaotta F., Fratianni S. and Zavattini J. A. (2016) Changes in precipitation extremes in Brazil (Paraná River Basin). Theoretical and applied climatology, 123:741-756

Zhang X., Vincent L. A., Hogg W. D. and Niitsoo A. (2000) Temperature and precipitation trends in Canada during the 20th century. Atmosphere-Ocean 38:395–42

Tables

Table 1: raw input data statistics (columns 2 and 3), statistics after the removal of values < 1mm (column 4 and 5) and statistics after putting the same missing values on both series (columns 6 and 7) using daily rain information about Boves stations (Piedmont, NW Italy) from 1988 to 2003 (see Online Resource 3); Shapiro-Wilk normality test is run only on dataset containing less than 5000 values.

	Raw candidate series	Raw reference series	Partially cleaned candidate series	Partially cleaned reference series	Cleaned candidate series	Cleaned reference series
Min	0	0	1	1	1	1
1st quantile	0	0	6	2.8	10	4.2
Median	0	0	14	6.8	18	11
Mean	3658	3051	25.4	13.36	31.75	17.32
3rd quantile	0	0.4	30	16.6	37	21.8
Max	310	198.4	310	198.4	275	198.4
Number of NAs	129	154	5021	4557	5329	5329
Number of values	5715	5690	823	1287	515	515
Shapiro-Wilk normality test W	NA	NA	0.62	0.68	0.7	0.72
Shapiro-Wilk normality test p-value	NA	NA	0	0	0	0

Table 2: monthly number of days used for the candidate series analysis of Boves stations (Piedmont, NW Italy) - see Online Resource 3

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1988	1	0	1	6	5	5	1	2	1	8	0	1
1989	0	2	1	7	0	1	3	0	2	1	3	1
1990	0	0	5	8	9	2	0	4	1	8	2	2
1991	2	1	5	5	3	2	0	0	3	6	2	0
1992	2	1	3	6	4	5	6	1	2	7	2	2
1993	0	2	2	3	2	2	3	2	3	9	2	1
1994	3	3	0	3	4	2	3	2	6	3	3	1
1995	0	0	0	4	3	0	0	4	3	1	2	5
1996	6	6	3	4	4	6	2	3	4	2	5	1
1997	2	0	0	2	3	5	2	3	1	0	0	0
1998	2	1	1	4	3	2	1	2	2	3	2	3
1999	0	0	2	3	3	2	6	2	3	9	3	0
2000	1	0	3	9	7	3	2	4	3	7	4	0
2001	0	1	3	4	7	0	2	0	1	0	1	0
2002	1	2	1	9	5	5	4	10	7	2	7	5
2003	0	0	0	4	3	1	1	1	1	5	4	3

Table 3: monthly amount of rain obtained from the candidate series of Boves station (Piedmont, NW Italy) - see Online Resource 3

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1988	14	0	4	168	199	163	4	23	8	289	0	61
1989	0	50	14	368	0	11	20	0	66	5	14	5
1990	0	0	36	189	129	23	0	120	12	148	46	66
1991	74	16	281	209	215	18	0	0	103	164	15	0
1992	48	8	59.8	99	87	94	77	11	48	448	47	67
1993	0	57	54.2	410	50	27	56	40	161	245	107	12
1994	83	166	0	185	90	135	48	42	275	107	219	13
1995	0	0	0	252	240	0	0	105	146	17	99	93
1996	230	100	114	214	68	93	56	78	184	386	192	113
1997	75	0	0	10	36	68	26	41	40	0	0	0
1998	70	52	35	105	88	27	8	119	51	90	12	55
1999	0	0	62	104	210	30	75	23	62	159	76	0
2000	28	0	43	254	148	320	28	72	88	371.8	168	0
2001	0	28	140	48	181	0	24	0	20	0	70	0
2002	3	180	75	173	211	219	208	118	222.5	115	305	110
2003	0	0	0	127	35	5	3	9	5	160	150	270

Table 4: results of the comparison analysis between candidate and reference series of Boves (Piedmont, NW Italy) from 1988 to 2003 (see Online Resource 3)

Test	Value
RMSE	30.77
T test t	7.75
T test df	793.69
T test p-value	0
Kolmogorov-Smirnov D	0.24
Kolmogorov-Smirnov p-value	0
Wilcoxon W	174914
Wilcoxon p-value	0
Kruskal-Wallis chi-squared	289.59
Kruskal-Wallis df	193
Kruskal-Wallis p-value	0
Spearman S	9279358.93
Spearman rho	0.59
Spearman p-value	0

Table 5: monthly percentage relative error for the two cleaned series of Boves (Piedmont, NW Italy) - see Online Resource 3

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1988	180	NA	-90.34	51.9	183.48	29.57	66.67	-52.28	5.26	130.46	NA	49.51
1989	NA	33.69	438.46	66.82	NA	-37.5	185.71	NA	89.66	400	118.75	31.58
1990	NA	NA	66.67	42.53	3.37	-21.77	NA	66.2	9.09	137.94	41.98	73.68
1991	262.75	344.44	119.87	125.22	57.16	119.51	NA	NA	127.88	63.67	22.95	NA
1992	-14.29	42.86	48.76	24.37	208.51	71.53	136.2	37.5	-10.78	74.59	158.24	661.36
1993	NA	108.03	-44.35	113.54	54.32	12.5	139.32	217.46	47.44	27.87	178.65	5.26
1994	13.08	418.75	NA	105.56	32.74	306.63	-23.57	50	277.75	247.4	125.31	225
1995	NA	NA	NA	138.19	75.95	NA	NA	218.18	48.37	1316.67	50.91	162.71
1996	277.05	122.22	195.34	161.61	28.79	12.32	4.87	81.4	156.98	256.75	100	195.81
1997	118.02	NA	NA	72.41	68.22	95.4	182.61	8.47	238.98	NA	NA	NA
1998	3081.82	288.06	733.33	22.66	-12.35	-40.79	81.82	662.82	571.05	49.5	106.9	418.87
1999	NA	NA	21.09	17.91	60.8	294.74	3.31	61.97	39.64	10.72	72.73	NA
2000	775	NA	25	52.83	106.13	12.6	4.48	78.22	18.92	68.85	158.46	NA
2001	NA	125.81	306.98	389.8	28.37	NA	20	NA	-12.28	NA	2087.5	NA
2002	-73.21	132.56	15.03	99.31	200.57	157.04	116.67	11.32	33.23	49.35	74.68	18.53
2003	NA	NA	NA	88.43	386.11	212.5	-74.14	-16.67	-60.32	201.89	913.51	263.88

Table 6: statistics of the trend computed on percentage relative error using Boves stations (Piedmont, NW Italy) from 1988 to 2003 (see Online Resource 3)

Test	Value
Trend	0.1
Intercept	66.51
Trend over total period	19.4
Mann-Kendall tau	0.01
Mann-Kendall p-value	0.82

Table 7: classes of rainy days and statistics about them; data is taken from Boves stations (Piedmont, NW Italy) from 1988 to 2003 (see Online Resource 3)

Class	Mean daily precip. (mm)	Max error mean daily (mm)	Min error mean daily (mm)	Num. total events ref. series	Num. total events can. series	Sum total prec. ref. (mm)	Sum total prec. can. (mm)	Num. common events	Sum common events ref. (mm)	Sum common events can. (mm)	Num. common events outside ±15%	Num. common events outside ±5%	Tot RMSE
Weak	3.89	0.58	0.19	255	140	1200.4	860.8	113	439.6	702.8	96	108	3.51
Mean	16.71	2.51	0.84	157	187	2669.6	3177.8	78	1303.2	1405.1	43	65	5.62
Heavy	35.32	5.3	1.77	77	103	2997.4	3873.7	26	918.4	1048	17	21	10.08
R95	82.08	12.31	4.1	26	85	2054.6	8438	22	1805.8	3103	21	22	76.62
R99	124.03	18.61	6.2	6	34	744.2	4869	6	744.2	983	5	6	42.18

Table 8: input, output and other parameters used by CoRain software

Name / Description	Type	Notes
A single txt file	INPUT (default /data/test/ example.txt)	Text file containing five TAB-separated columns (year, month, day, candidate rain series and reference rain series)
Threshold below which the precipitation is not taken into account	Hard coded parameter (default 1 mm)	In a future release of the program, this will be parameterized, specifying it in a variable called min_rain_rate inside a “Global variables” section
Err_pos_15	Hard coded parameter (default +15%)	Upper limit of the maximum error boundaries, based on daily data of the reference series. In a future release of the program, this will be parameterized, specifying it in inside a “Global variables” section
Err_neg_15	Hard coded parameter (default -15%)	Lower limit of the maximum error boundaries, based on daily data of the reference series. In a future release of the program, this will be parameterized, specifying it in inside a “Global variables” section
Err_pos_5	Hard coded parameter (default +5%)	Upper limit of the minimum error boundaries, based on daily data of the reference series. In a future release of the program, this will be parameterized, specifying it in inside a “Global variables” section
Err_neg_5	Hard coded parameter (default -5%)	Lower limit of the minimum error boundaries, based on daily data of the reference series. In a future release of the program, this will be parameterized, specifying it in inside a “Global variables” section
Threshold_percentage_relative_error	Hard coded parameter (default ±500%)	Range of accepted percentage relative errors. Customizable parameter specified inside the “Global variables” section
0_statistics_input_file.csv	OUTPUT	Main statistics on input file
1_cleaned_input_file.txt	OUTPUT	The file that is really used as input, after cleaning the reference and candidate series
2_<series_name>_daily_avalibillity.csv	OUTPUT	Number of days of available data on the series named <series_name>. This is produced two times, one for the candidate series and one for the reference one

2_<series_name>_monthly_rain.csv	OUTPUT	Total amount of rain by month on the series named <series_name>. This is produced two times, one for the candidate series and one for the reference one
3_<series_name>_day_month_rain.png	OUTPUT	Plot of daily and monthly data on the series named <series_name>. This is produced two times, one for the candidate series and one for the reference one
4_statistics_between_daily_series.csv	OUTPUT	Statistics between the two daily series
5_total_precipitation_plot_with_15_5.png	OUTPUT	Scatter plot of daily events with error boundaries
6_percentage_relative_error.csv	OUTPUT	Percentage relative error grouped by month
7_percentage_relative_error_boxplot.png	OUTPUT	Box plot of percentage relative error grouped by month
8_percentage_relative_error_trend.png	OUTPUT	Trend of percentage relative error grouped by month
8_stats_percentage_relative_error_trend.csv	OUTPUT	Statistics on the trend of percentage relative error grouped by month
6_filtered_percentage_relative_error.csv	Optional OUTPUT	Percentage relative error grouped by month after the exclusion of values that are outside of the defined range (by Threshold_percentage_relative_error)
7_filtered_percentage_relative_error_boxplot.png	Optional OUTPUT	Box plot of percentage relative error grouped by month after the exclusion of values that are outside of the defined range (by Threshold_percentage_relative_error)
8_filtered_percentage_relative_error_trend.png	Optional OUTPUT	Trend of percentage relative error grouped by month after the exclusion of values that are outside of the defined range (by Threshold_percentage_relative_error)
8_filtered_stats_percentage_relative_error_trend.csv	Optional OUTPUT	Statistics on the trend of percentage relative error grouped by month after the exclusion of values that are outside of the defined range (by Threshold_percentage_relative_error)
9_events_R99.png	OUTPUT	Scatter plot of extreme precipitation events (greater than 99 percentile)
9_events_R95.png	OUTPUT	Scatter plot of very heavy precipitation events (greater than 95 percentile)
9_events_heavy.png	OUTPUT	Scatter plot of heavy precipitation events (between 80 percentile and 95 percentile)
9_events_mean.png	OUTPUT	Scatter plot of mean precipitation events (between 50 percentile and 80 percentile)
9_events_weak.png	OUTPUT	Scatter plot of weak precipitation events (lower than 50 percentile)
10_class_event_and_RMSE.csv	OUTPUT	Summary of information on the classification of precipitation events

Supplementary material captions

Online Resource 1: file “Co.Rain.R”, containing the source code of the program discussed here

Online Resource 2: file “example1.txt”, containing both candidate and reference series used as example

Online Resource 3: file “example2.txt”, containing both candidate and reference series used as example; data is taken from Boves stations, Piedmont (NW Italy)

Online Resource 4: file “example3.txt”, containing both candidate and reference series used as example

Online Resource 5: file “example4.txt”, containing both candidate and reference series used as example

Online Resource 6: file “example5.txt”, containing both candidate and reference series used as example

Online Resource 7: file “example6.txt”, containing both candidate and reference series used as example