RHtests_dlyPrcp User Manual

By

Xiaolan L. Wang and Yang Feng

Climate Research Division
Atmospheric Science and Technology Directorate
Science and Technology Branch, Environment Canada
Toronto, Ontario, Canada

Published online at
http://etccdi.pacificclimate.org/software.shtml
14 August 2013

Table of contents

1. Introduction

The RHtests_dlyPrcp software package is similar to the RHtestsV3 and RHtestsV4 packages, except that it is specifically designed for homogenization of daily precipitation data time series. It is based on the transPMFred algorithm (Wang et al. 2010), which integrates a data adaptive Box-Cox transformation procedure into the PMFred algorithm (Wang 2008a). The PMFred algorithm is based on the penalized maximal *F* (PMF) test (Wang 2008b) that is embedded in a recursive testing algorithm (Wang 2008a), and is used in the case "without a reference series" in the RHtestsV3 and RHtestsV4 packages. The Box-Cox transformation is necessary, because daily precipitation amounts are not normally distributed. Since daily precipitation is highly variable both spatially and temporally (it could be raining in this side of the street, but not the other side), it is hardly possible to find a suitable reference series (except in the case of parallel measurements). Thus, this software does not use any reference series. Since daily precipitation is not a continuous process, discontinuities in the occurrence frequency of precipitation might exist and should be dealt with first to avoid complicating the homogenization of daily precipitation data time series. Please refer to Section 6 of Wang et al. (2010) for more details on how to deal with frequency discontinuities.

This simple manual is to provide a quick reference to the usage of the functions included in the RHtests_dlyPrcp package (and also to the usage of the equivalent FORTRAN functions, which are available by sending a request in English to Xiaolan.Wang@ec.gc.ca). Users are assumed to have the general knowledge of R (how to start and end an R session and how to call an R function).

2. Input data format for the RHtests_dlyPrcp

- The RHtests_dlyPrcp functions handle daily precipitation series. Each input data series should be stored in a separate file (e.g., a file named Example.dat), in which the first three columns are the dates (calendar year YYYY, month MM, and day DD) of observations, and the fourth column, the observed data values (or missing value code). For example,

```
(Daily series)
…
1947 12 8 8.8
1947 12 9 17.6
1947 12 10 2.9
1947 12 11 0
1947 12 12 0
1947 12 13 0
1947 12 14 0
1947 12 15 -999.9
1947 12 16 -999.9
1947 12 17 2.9
1947 12 18 5.9
1947 12 19 0
1947 12 20 0
1947 12 21 2.9
1947 12 22 0
…                                               …
```

The dates of input data **must be consecutive** and in the calendar order. Otherwise, the program will exit with an error message containing the first date on which the data error occurs. For example, the four rows from 15-16 December 1947 in the daily series above must be included in the input data file. They should not be deleted because they are missing values. December 18, 1947 should not occur before December 17, 1947, etc.

3. How to use the RHtests_dlyPrcp functions

The RHtests_dlyPrcp software package provides three functions for detecting, and adjusting for, artificial shifts in daily precipitation data series without using a reference series.

First of all, enter *source (“RHtests_dlyPrcp.r”)* at the R prompt (“>”) to load the RHtests_dlyPrcp functions to R.

Briefly, the steps to follow are: (see Sections 3.1 or 3.2 below for the details)

1) Call function *FindU* with an appropriate list of input parameters (see Sections 3.1 or 3.2 below).
2) Go to Step 5) if you don’t have metadata. Otherwise, call *FindUD* with an appropriate list of input parameters (see Sections 3.1 or 3.2 below).
3) Modify the resulting *_mCs.txt file (the list of changepoints identified so far, which is in the data directory, i.e., the directory in which the data series being tested resides), if necessary, to incorporate metadata information in the results. (Here, the * stands for a user specified prefix for the name of the output files).
4) Call function *StepSize* with an appropriate list of input parameters (see Section 3.1 or 3.2 below) to assess the significance and magnitude of the retained changepoints.
5) Analyze the latest version of the *_mCs.txt file and delete the smallest shift if it is statistically, or subjectively determined to be, not significant. Then, call function *StepSize* again to re-assess the significance of the remaining changepoints. Repeat this procedure (5) until each and every changepoint in the list is determined to be significant.

Specifically, the general procedure should be:

(1) Call the *FindU* function, to detect all changepoints that could be significant at the nominal level even without metadata support (these are called Type-1 changepoints). If there is no significant changepoint identified so far, the time series being tested can be declared to be homogeneous; and there is no need to go further in testing this series.

(2) **Go to (5) if there is no metadata available, or if one wants to detect only those changepoints that are significant even without metadata support, i.e., Type-1 changepoints.**
Otherwise, call the *FindUD* function. The resulting additional changepoints are called Type-0 changepoints (these changepoints could be significant **only if** they

are supported by reliable metadata. This step is meant to help narrow down the metadata investigation, which should focus on the periods encompassing these Type-0 changepoints).

(3) Investigate the available metadata to see whether or not anything happened **at or near** the identified changepoint times/dates that could have caused the shifts. Retain only those Type-0 changepoints that are actually supported by metadata, along with all the Type-1 changepoints as identified in (1). The date of a changepoint may be changed to the documented date of change as obtained from the metadata if one is confident about the cause of the change; also change the changepoint type from 1 to 0 if a Type-1 changepoint turns out to have reliable metadata support.

(4) Call function *StepSize* to assess the significance and magnitude of the remaining changepoints (listed in the latest version of the *_mCs.txt file).

(5) Analyze the latest version of the *_mCs.txt file and delete the least significant changepoint if it is determined to be not significant at the nominal level. Then, call function *StepSize* (or *StepSize.wRef*) to re-assess the significance of the remaining changepoints. Repeat this procedure (5) until all the retained changepoints are significant.

In the GUI mode (Section 3.1), the final output files reside in the output subdirectory of the data directory (i.e., where the data series being tested resides); both the data and output directory path are also shown in the GUI window. In the command line mode (Section 3.2) user gets to specify the directory for storing the output files by including the output directory path in the output parameter string (see Section 3.2).

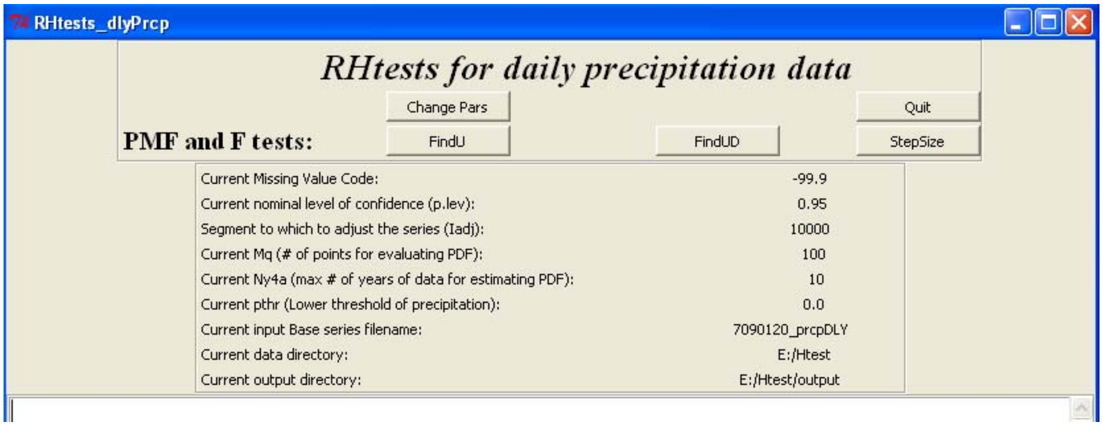
3.1 The graphical user interface (GUI) mode

The *FindU*, *FindUD*, and *StepSize* functions are based on the transPMFred algorithm (Wang et al. 2010), which allows the time series being tested to have a linear trend throughout the whole period of the data record (i.e., no shift in the trend component; see Wang 2003), with linear trend and lag-1 autocorrelation of the base series being estimated in tandem through iterative procedures, while accounting for all the identified mean-shifts. No reference series will be used in any of these functions. Please refer to section 3.2 below for more details about these three functions.

In this simple graphical user interface (GUI) mode, the prefix of the input data filename is used as the prefix for the names of the output files. For example, if Example.dat is the input data filename, the output files will be named Example_*.*.

Specifically, the procedure is as follows:

(1) To start the GUI session, enter StartGUI() after entering *source (“RHtests_dlyPrcp.r”)* at the R prompt. The following window shall appear.

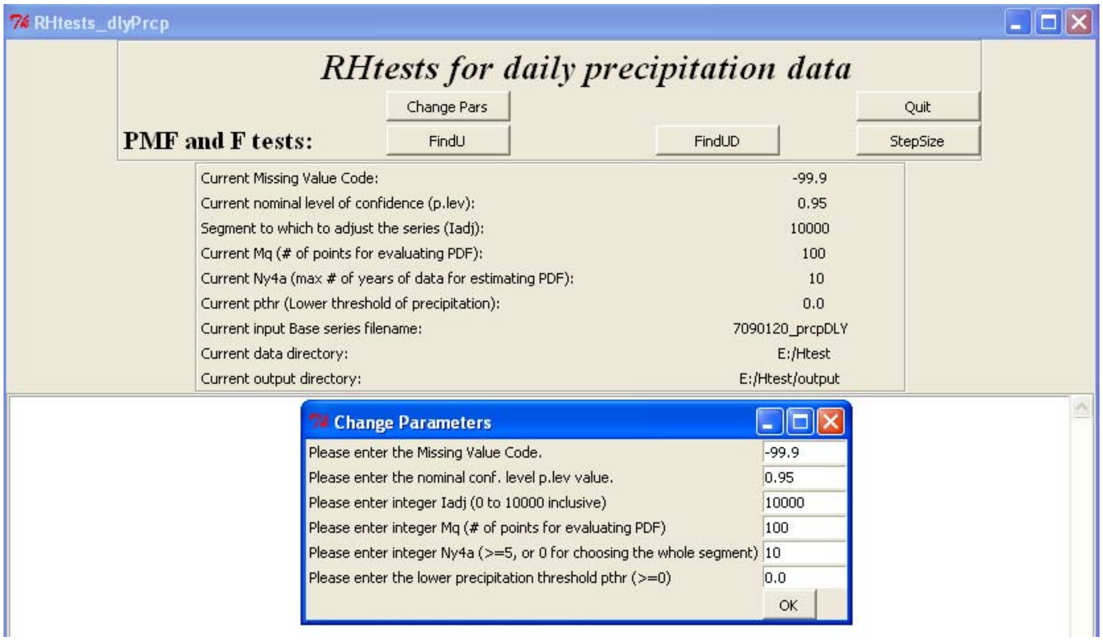


RHtests for daily precipitation data

PMF and F tests:

Current Missing Value Code:	-99.9
Current nominal level of confidence (p.lev):	0.95
Segment to which to adjust the series (Iadj):	10000
Current Mq (# of points for evaluating PDF):	100
Current Ny4a (max # of years of data for estimating PDF):	10
Current pthr (Lower threshold of precipitation):	0.0
Current input Base series filename:	7090120_prcpDLY
Current data directory:	E:/Htest
Current output directory:	E:/Htest/output

(2) Click the ChangePars button to set the following parameter values: (a) the missing value code used in the data series to be tested, e.g., "-99.9" in the window below (note that the code entered here **must be exactly the same** as used in the data; e.g., "-99." and "-99.0" are different; one can not enter "-99." instead of "-99.0" when "-99.0" is used in the input data series; it will produce erroneous results); (b) the nominal level of confidence at which to conduct the test; (c) the base segment (to which to adjust the series); (d) the number of points (Mq) for which the empirical probability distribution function (PDF) are to be estimated for use in deriving the QM-adjustments (Wang et a. 2010); (e) the maximum number of years of data immediately before or after a changepoint to be used to estimate the PDF (Ny4a = 0 for choosing the whole segment); and (f) the lower threshold of precipitation (any value below this threshold will be excluded 0 during the test). The default values used are: *p.lev=0.95, Iadj=10000, Mq=12, Ny4a=0, pthr=0.0*. Then, click the OK button to accept the parameter values shown in the window.



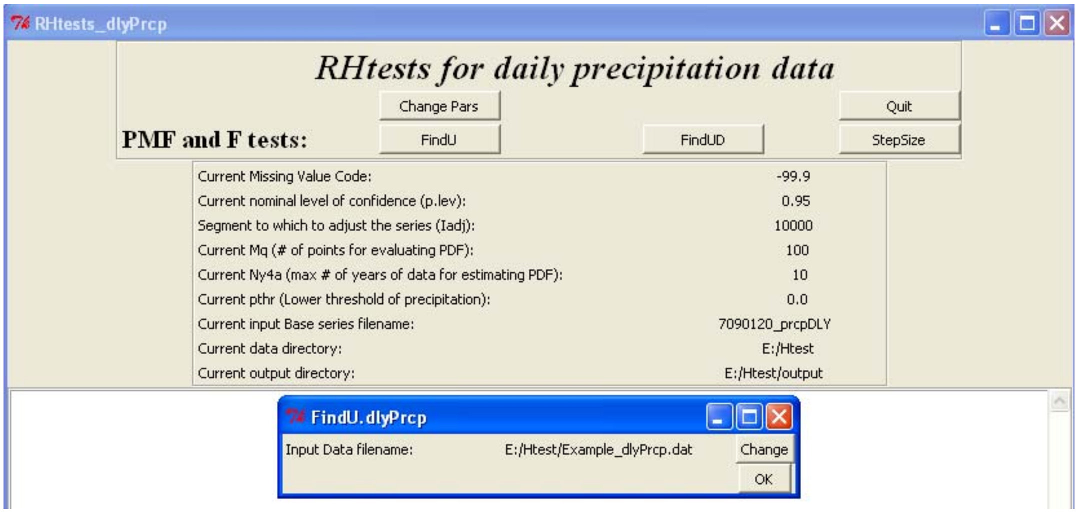
RHtests for daily precipitation data

PMF and F tests:

Current Missing Value Code:	-99.9
Current nominal level of confidence (p.lev):	0.95
Segment to which to adjust the series (Iadj):	10000
Current Mq (# of points for evaluating PDF):	100
Current Ny4a (max # of years of data for estimating PDF):	10
Current pthr (Lower threshold of precipitation):	0.0
Current input Base series filename:	7090120_prcpDLY
Current data directory:	E:/Htest
Current output directory:	E:/Htest/output

Please enter the Missing Value Code.	-99.9
Please enter the nominal conf. level p.lev value.	0.95
Please enter integer Iadj (0 to 10000 inclusive)	10000
Please enter integer Mq (# of points for evaluating PDF)	100
Please enter integer Ny4a (>=5, or 0 for choosing the whole segment)	10
Please enter the lower precipitation threshold pthr (>=0)	0.0

(4) Click the FindU button to open a window, select the data series (say Example1.dat) to be tested and click the Open button to execute the transPMFred test (Wang et al. 2010). This will produce the following files in the output directory: Example1_1Cs.txt, Example1_Ustat.txt, Example1_U.dat, and Example1_U.pdf (see section 3.2 for description of the content of these files). A copy of the first file is also stored in file Example1_mCs.txt in the output directory, which lists all changepoints that could be significant at the nominal level even without metadata support (i.e., Type-1 changepoints).



Current Missing Value Code: -99.9
Current nominal level of confidence (p.lev): 0.95
Segment to which to adjust the series (Iadj): 10000
Current Mq (# of points for evaluating PDF): 100
Current Ny4a (max # of years of data for estimating PDF): 10
Current pthr (Lower threshold of precipitation): 0.0
Current input Base series filename: 7090120_prcpDLY
Current data directory: E:/Htest
Current output directory: E:/Htest/output

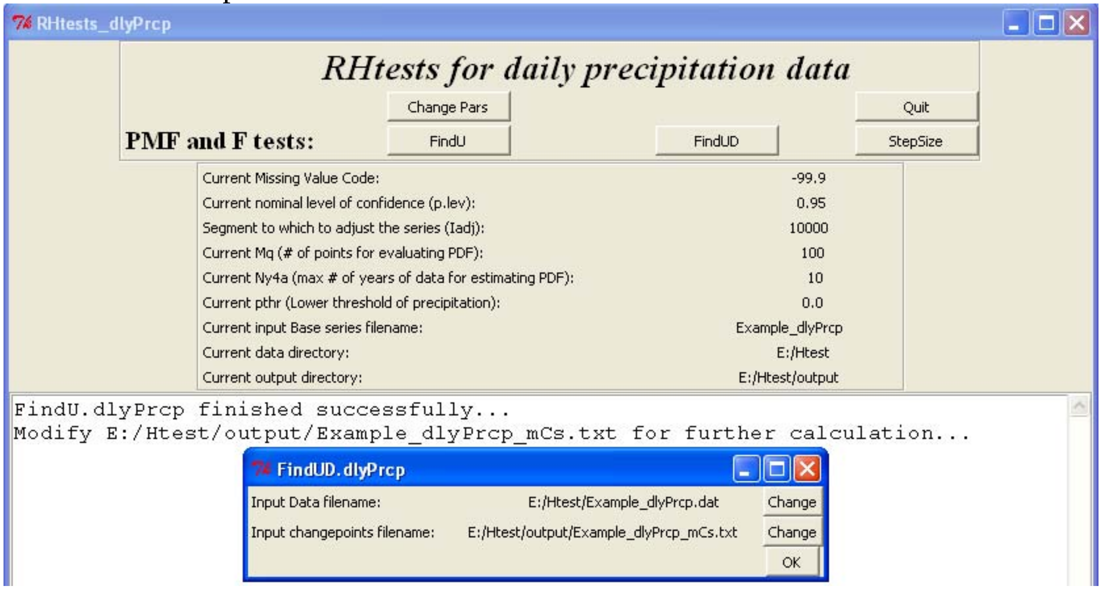
FindU.dlyPrcp
Input Data filename: E:/Htest/Example_dlyPrcp.dat

An example of the *1Cs.txt file looks like this:

```
2 changepoints in Series …Example_prcpDLY.dat…
1 Yes   19350927 (    1.0000-    1.0000) 0.950   59.2034 (   16.4042-    18.3580)
1 ?     19870327 (    0.9999-    0.9999) 0.950   16.8418 (   16.2018-    18.1121)
```

Here, the first column (the1's) is an index indicating these are Type-1 changepoints. The second column indicates whether or not the changepoint is statistically significant for the changepoint type given in the first column; they are "Yes" or "?" in the *_1Cs.txt file, but in other *Cs.txt files they could be the following: (1) "Yes" (significant); (2) "No" (not significant for the changepoint type given in the first column); (3) "?" (may or may not be significant for the type given in the first column), and (4) "YifD" (significant if it is documented, i.e., supported by reliable metadata). The third column lists the changepoint dates YYYYMMDD, e.g., 19350927 denotes 27 September 1935. The numbers in the fourth column (in parentheses) are the 95% confidence interval of the p-value, which is estimated assuming the changepoint is documented. The nominal p-value (confidence level) is given in the fifth column. The last three columns are the values of the test statistic $PF_{\max}$ and the 95% confidence interval of the $PF_{\max}$ percentiles that correspond to the nominal confidence level, respectively. A copy of the file OutFile_1Cs.txt is stored in file OutFile_mCs.txt in the output directory for possible modifications later (so that an original copy is kept unchanged).

(5) **If you know all the documented changes that could cause a mean-shift, add these changepoints in the file** Example1_Ref1_mCs.txt **or** Example_mCs.txt **if they are not already there, and go to procedure (7) below. If you do not have metadata, or if you only want to detect Type-1 changepoints, also go to procedure (7) below.** Otherwise, click the FindUD.wRef button (or FindUD in case of not using a reference series) to identify all Type-0 changepoints (i.e., changepoints that could be significant ***only if*** they are supported by metadata) for the chosen input data series. The window below will appear for you to choose or confirm the input files to run the FindUD function.



Current Missing Value Code:	-99.9
Current nominal level of confidence (p.lev):	0.95
Segment to which to adjust the series (Iadj):	10000
Current Mq (# of points for evaluating PDF):	100
Current Ny4a (max # of years of data for estimating PDF):	10
Current pthr (Lower threshold of precipitation):	0.0
Current input Base series filename:	Example_dlyPrcp
Current data directory:	E:/Htest
Current output directory:	E:/Htest/output

```
FindU.dlyPrcp finished successfully...
Modify E:/Htest/output/Example_dlyPrcp_mCs.txt for further calculation...
```

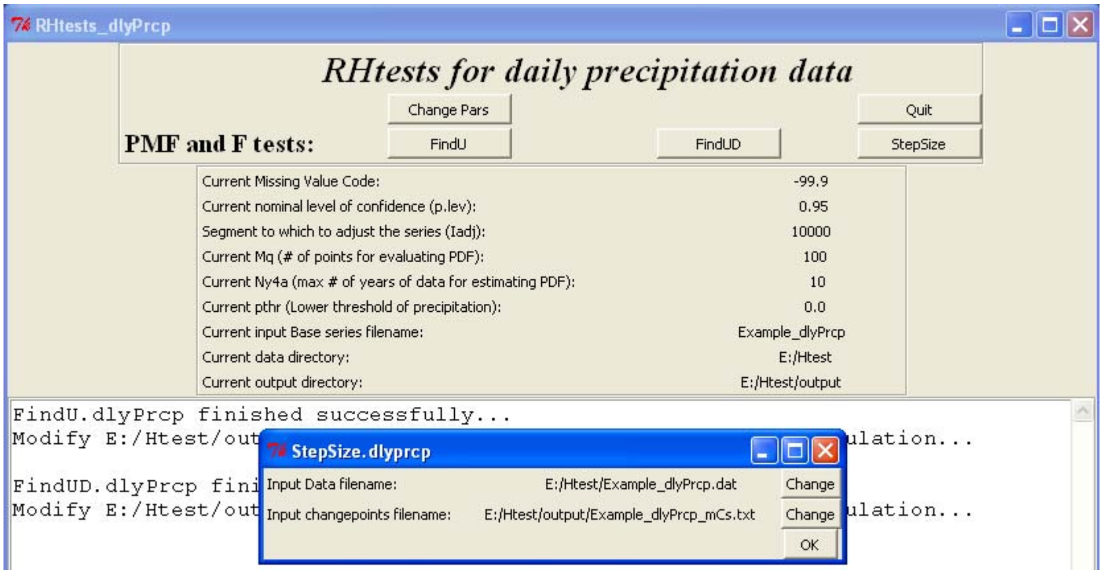
FindUD.dlyPrcp

Input Data filename: E:/Htest/Example_dlyPrcp.dat

Input changepoints filename: E:/Htest/output/Example_dlyPrcp_mCs.txt

Four files will be produced in the output directory, e.g., Example1_pCs.txt, Example1_UDstat.txt, Example1_UD.dat, and Example1_UD.pdf by calling the FindUD function with Example1.dat as the input data series (see Section 3.2 for description of the content of these files). A copy of Example1_pCs.txt is also stored in file Example1_mCs.txt in the output directory (so that the previous version of this file, if exists, is updated).

(6) Investigate metadata and delete from the Example1_mCs.txt file in the output directory all Type-0 changepoints that are not supported by metadata. Click the StepSize to re-assess the significance/magnitudes of the remaining changepoints, which will produce the following files: Example1_fCs.txt, Example1_Fstat.txt, Example1_F.dat, Example1_F.pdf, and an updated Example1_mCs.txt in the output directory (see Section 3.2 for description of the content of these files). Please check the input file names to ensure they are what you want to use here.



Current Missing Value Code:	-99.9
Current nominal level of confidence (p.lev):	0.95
Segment to which to adjust the series (Iadj):	10000
Current Mq (# of points for evaluating PDF):	100
Current Ny4a (max # of years of data for estimating PDF):	10
Current pthr (Lower threshold of precipitation):	0.0
Current input Base series filename:	Example_dlyPrcp
Current data directory:	E:/Htest
Current output directory:	E:/Htest/output

```
FindU.dlyPrcp finished successfully...
Modify E:/Htest/out                                   ulation...

FindUD.dlyPrcp fini
Modify E:/Htest/out                                   ulation...
```

Input Data filename: E:/Htest/Example_dlyPrcp.dat

Input changepoints filename: E:/Htest/output/Example_dlyPrcp_mCs.txt

(7) Analyze the results obtained so far to determine if the smallest shift is significant [see (F5) in section 3.2 for the details of how to do so]. If it is determined to be not significant, delete it from the file Example1_mCs.txt in the output directory and click the StepSize button to re-assess the significance and magnitudes of the remaining changepoints, which will update (or produce) the following files with the new estimates: Example1_fCs.txt, Example1_Fstat.txt, Example1_F.dat, Example1_F.pdf, and Example1_mCs.txt in the output directory.

(8) Repeat the procedure (7) above, until each and every changepoint retained in the file Example1_mCs.txt is determined to be significant (no more deletions will be done). The following four final output files are in the output directory: a) Example1_fCs.txt, which lists the changepoints identified and their significance and statistics; b) Example1_Fstat.txt, which stores the estimated mean-shift sizes and a copy of the content in Example1_fCs.txt; c) Example1_F.dat, which stores the mean-adjusted base series in its fifth column, the QM-adjusted base series in its ninth column, and the original base series in its third column; and d) Example1_F.pdf. The Example1_F.pdf file stores five plots: (i) segments of the original dailyP>pthr series for the short periods surrounding each changepoint; (ii) segments of the Box-Cox transformed dailyP>pthr series and the estimated mean-shifts and linear trend for the short periods surrounding each changepoint; (iii) the original dailyP>pthr series for the whole period and the estimated mean-shifts and linear trend; (iv) the QM adjusted dailyP>pthr series for the whole period; and (v) the IBC-adjusted dailyP>pthr series for the whole period (Wang et al. 2010). Please see Section 3.2 for description of the content of these output files.

In addition to the functions with GUI above, the RHtests_dlyPrcp also provides three functions for detecting abrupt changes (mean-shifts) in daily precipitation time series without a graphical user interface. One should click the Quit button first and then call these functions at the R prompt (see Section 3.2 below for the details).

3.2 The command line mode

In this mode, the five detailed procedures are:

(**F1**) Call function *FindU.dlyPrcp* to identify all Type-1 changepoints in the InSeries by entering the following at the R prompt:

```
FindU.dlyPrcp(InSeries="C:/inputdata/InFile.csv", MissingValueCode="-999.0",
  ,pthr=0.0, p.lev=0.95, Iadj=10000, Mq=10, Ny4a=0, output="C:/results/OutFile")
```

Here, the C:/inputdata/ is the data directory path and the InFile.csv is the name of the file containing the data series to be tested; while C:/results/ is a user specified output directory path and the OutFile is a user selected prefix for the name of the files to store the results; -999.0 is the missing value code that is used in the input data file InFile.csv; *p.lev* is a pre-set (nominal) level of confidence at which the test is to be conducted (choose from one of these: 0.75, 0.80, 0.90, 0.95, 0.99, and 0.9999); *pthr* is the lower precipitation threshold (daily precipitation below this threshold will be excluded in the test); *Iadj* is an integer value corresponding to the segment to which the series is to be adjusted (referred to as the base segment), with *Iadj*=10000 corresponding to adjusting the series to the last segment; *Mq* is the number of points (categories) for which the empirical probability distribution function (PDF) are to be estimated, and *Ny4a* is the maximum number of years of data immediately before or after a changepoint to be used to estimate the PDF (*Ny4a=0* for choosing the whole segment). One can set *Mq* to any integer between 1 and 100 inclusive, or set *Mq*=0 if this number is to be determined automatically by the function (the function re-sets *Mq* to 1 if 0 is selected eventually or to 100 if a larger number is selected or given). The default values used are: *p.lev=0.95, Iadj=10000, Mq=12, Ny4a=0, pthr=0.0*. Note that the MissingValueCode entered here **must be exactly the same** as used in the data; e.g., one cannot enter "-999." instead of "-999.0" when "-999.0" is used in the input data series; otherwise it will produce erroneous results. Also, note that character strings should be included in double quotation marks, as shown above. After a successful call, this function produces the following five files in the output directory:

- OutFile_1Cs.txt (and OutFile_mCs.txt): The first number in the first line of this file is the number of changepoints identified in the series being tested. If this number is $N_c > 0$, the subsequent N_c lines list the dates and statistics of these N_c changepoints. For example, it looks like this for a case of $N_c = 2$:

  ```
  2 changepoints in Series ⋯Example_prcpDLY.dat⋯
  1 Yes   19350927 (    1.0000-     1.0000) 0.950   59.2034 (    16.4042-    18.3580)
  1 ?     19870327 (    0.9999-     0.9999) 0.950   16.8418 (    16.2018-    18.1121)
  ```

  The first column (the1's) is an index indicating these are Type-1 changepoints (also indicated by the "1Cs" in the filename). The second column indicates whether or not the changepoint is statistically significant for the changepoint type given in the first column; all of them are "Yes" in this *_1Cs.txt file, but in other *Cs.txt files they could be the following: (1) "Yes" (significant); (2)

"No" (not significant for the changepoint type given in the first column); (3) "?" (may or may not be significant for the type given in the first column), and (4) "YifD" (significant if it is documented, i.e., supported by reliable metadata). The third column lists the changepoint dates YYYYMMDD, e.g., 19350927 denotes 27 September 1935. The numbers in the fourth column (in parentheses) are the 95% confidence interval of the p-value, which is estimated assuming the changepoint is documented (thus this value is very high for a significant Type-1 changepoint). The nominal p-value (confidence level) is given in the fifth column. The last three columns are the $PF_{\max}$ statistics and the 95% confidence interval of the $PF_{\max}$ percentiles that correspond to the nominal confidence level, respectively. A copy of the file OutFile_1Cs.txt is stored in file OutFile_mCs.txt in the output directory for possible modifications later (so that an original copy is kept unchanged).

- OutFile_Ustat.txt: In addition to all the results stored in the OutFile_1Cs.txt file, this output file contains the parameter estimates of the $(N_c + 1)$-phase regression model fit, including the sizes of the mean-shifts identified, the linear trend and lag-1 autocorrelation of the series being tested.

- OutFile_U.dat: This file contains the dates of observation (2nd column), the original daily precipitation series (3rd column), the estimated linear trend and mean-shifts of the daily precipitation series (4th column), the QM-adjusted daily precipitation series (5th column), the mean-adjusted daily precipitation series (6th column), the estimated linear trend and mean-shifts of the QM-adjusted daily precipitation series (7th column), the Box-Cox transformed original daily precipitation series (8th column), the estimated linear trend and mean shifts of the Box-Cox transformed original daily precipitation series (9th column), the mean-adjusted Box-Cox transformed daily precipitation series (10th column), the estimated linear trend of the mean-adjusted Box-Cox transformed series (11th column), the different between the QM-adjusted series and the original series (12th column=column5-column3) and the difference between the mean-adjusted series and the original series (13th =column6-column3).

- OutFile_U.pdf: This file stores five plots: (i) segments of the original dailyP>pthr series for the short periods surrounding each changepoint; (ii) segments of the Box-Cox transformed "*dailyP > pthr*" series and the estimated mean-shifts and linear trend for the short periods surrounding each changepoint; (iii) the original "*dailyP > pthr*" series for the whole period and the estimated mean-shifts and linear trend; (iv) the QM-adjusted "*dailyP > pthr*" series for the whole period; and (v) the IBC-adjusted "*dailyP > pthr*" series for the whole period (Wang et al. 2010).

If there is no significant changepoint identified, the time series being tested can be declared to be homogeneous; and no need to go further in testing this series.

(F2) **If you know all the documented changes that could cause a shift, add these changepoints in the file** Example_mCs.txt **if they are not already there, and go to (F4) now. If there is no metadata available, or if you want to detect only those changepoints that are significant even without metadata support (i.e., Type-1 changepoints), also go to (F4) now**. Otherwise, call function *FindUD* to identify all Type-0 changepoints in the series, in the presence of all the Type-1 changepoints listed in file OutFile_1Cs.txt, by entering the following at the R prompt:

```
FindUD(InSeries="C:/inputdata/InFile.csv", MissingValueCode="-999.0", pthr=0.0,
        p.lev=0.95, Iadj=10000, Mq=10, Ny4a=0, InCs="C:/results/OutFile_1Cs.txt",
         output="C:/results/OutFile")
```

Here, the OutFile_1Cs.txt file contains all the Type-1 changepoints identified by calling *FindU* in (F1) above, and all the other files are the same as in (F1). Here, a successful call also produces five files: OutFile_pCs.txt and OutFile_mCs.txt, OutFile_UDstat.txt, OutFile_UD.pdf, and OutFile_UD.dat. The contents of these files are similar to the relevant files in (F1), except that the changepoints that are now modeled are those listed in the OutFile_pCs.txt or OutFile_mCs.txt file, which contains all the Type-1 changepoints listed in OutFile_1Cs.txt, **plus** all Type-0 changepoints. The OutFile_mCs.txt file is now a copy of OutFile_pCs.txt for possible modifications later.

(**F3**) As mentioned earlier, the Type-0 changepoints **could be** statistically significant at the pre-set level of significance **only if** they are supported by reliable metadata. Also, some of the Type-1 changepoints identified could have metadata support as well, and the exact dates of change could be slightly different from the dates that have been identified statistically. Thus, one should now investigate available metadata, focusing around the dates of all the changepoints (Type-1 or Type-0) listed in the OutFile_mCs.txt file. **Keep only those Type-0 changepoints that are supported by metadata, along with all Type-1 changepoints**. Modify the statistically identified dates of changepoints to the documented dates of change (obtained from highly reliable metadata) if necessary. For example, the original OutFile_mCs.txt is as follows:

```
38 changepoints in Series …Example_prcpDLY.dat…
0 Yes   19151212 (    1.0000-    1.0000) 0.950   19.8021 (   13.0273-   14.3260)
0 YifD  19200813 (    0.9993-    0.9993) 0.950   11.4365 (   13.0616-   14.3676)
0 Yes   19230720 (    1.0000-    1.0000) 0.950   19.2023 (   12.8899-   14.1601)
0 YifD  19250715 (    0.9804-    0.9804) 0.950    5.5348 (   12.4966-   13.7064)
0 Yes   19261125 (    1.0000-    1.0000) 0.950   31.9058 (   12.1369-   13.2909)
0 Yes   19270619 (    1.0000-    1.0000) 0.950   27.1131 (   11.9673-   13.0945)
0 Yes   19271013 (    1.0000-    1.0000) 0.950   43.1756 (   11.8784-   12.9925)
0 Yes   19290225 (    0.9999-    0.9999) 0.950   15.0093 (   12.7626-   14.0073)
0 YifD  19310503 (    0.9966-    0.9966) 0.950    8.9943 (   13.0085-   14.3032)
1 No    19350927 (    0.9904-    0.9904) 0.950    7.2408 (   12.9625-   14.2476)
0 YifD  19361010 (    0.9760-    0.9760) 0.950    5.4677 (   13.0185-   14.3153)
0 YifD  19420506 (    0.9967-    0.9967) 0.950    8.7573 (   13.1034-   14.4182)
0 Yes   19451113 (    1.0000-    1.0000) 0.950   64.8952 (   12.8477-   14.1094)
0 Yes   19460109 (    1.0000-    1.0000) 0.950   40.4886 (   12.8275-   14.0852)
0 Yes   19490113 (    1.0000-    1.0000) 0.950   38.4868 (   13.0578-   14.3629)
```

```
0 YifD  19540326 (    0.9997-    0.9997) 0.950   13.0602 (   12.9548-   14.2382)
0 YifD  19540812 (    0.9984-    0.9984) 0.950   11.3299 (   11.5751-   12.6502)
0 Yes   19540919 (    1.0000-    1.0000) 0.950   31.6919 (    7.1814-    7.7339)
0 Yes   19541015 (    1.0000-    1.0000) 0.950   19.2973 (   12.9661-   14.2519)
0 Yes   19590809 (    0.9999-    0.9999) 0.950   14.5873 (   13.1303-   14.4507)
0 Yes   19640519 (    1.0000-    1.0000) 0.950   18.9390 (   12.9371-   14.2167)
0 YifD  19640728 (    0.9986-    0.9986) 0.950    9.4018 (   12.9365-   14.2160)
0 No    19680805 (    0.7590-    0.7590) 0.950    1.6107 (   13.2751-   14.6260)
0 YifD  19760611 (    0.9992-    0.9992) 0.950   11.2777 (   13.5402-   14.9472)
0 YifD  19851108 (    0.9977-    0.9977) 0.950    8.2166 (   13.2075-   14.5442)
0 Yes   19860306 (    1.0000-    1.0000) 0.950   19.0308 (   12.3039-   13.4843)
0 YifD  19861014 (    0.9673-    0.9673) 0.950    4.7821 (   12.2611-   13.4347)
0 YifD  19870106 (    0.9948-    0.9948) 0.950    8.4427 (   11.6850-   12.7735)
1 Yes   19870327 (    1.0000-    1.0000) 0.950   44.4813 (   11.9157-   13.0345)
0 Yes   19870821 (    0.9999-    0.9999) 0.950   15.9799 (   12.4156-   13.6137)
0 YifD  19880425 (    0.9969-    0.9969) 0.950    9.8995 (   12.3188-   13.5016)
0 YifD  19880731 (    0.9955-    0.9955) 0.950    9.2208 (   11.1945-   12.2133)
0 Yes   19880824 (    0.9989-    0.9989) 0.950   13.1230 (   10.8759-   11.8399)
0 YifD  19881001 (    0.9729-    0.9729) 0.950    5.3150 (   12.1200-   13.2713)
0 Yes   19890413 (    1.0000-    1.0000) 0.950   21.1044 (   12.2441-   13.4151)
0 YifD  19890805 (    0.9909-    0.9909) 0.950    6.2282 (   13.1516-   14.4765)
0 YifD  19971004 (    0.9998-    0.9998) 0.950   13.3885 (   13.1510-   14.4759)
0 Yes   19980521 (    0.9996-    0.9996) 0.950   14.4661 (   12.3300-   13.5145)
```

If it is determined after metadata investigation that there are documented causes for one shift, and that the exact dates of these shifts are October 1945, one should modify the OutFile_mCs.txt file to (the modified numbers are shown in bold):

```
6 changepoints in Series …Example_prcpDLY.dat…
1 No    19350927 (    0.9904-    0.9904) 0.950    7.2408 (   12.9625-   14.2476)
0 Yes   19451113 (    1.0000-    1.0000) 0.950   64.8952 (   12.8477-   14.1094)
0 Yes   19460109 (    1.0000-    1.0000) 0.950   40.4886 (   12.8275-   14.0852)
0 Yes   19490113 (    1.0000-    1.0000) 0.950   38.4868 (   13.0578-   14.3629)
0 Yes   19590809 (    0.9999-    0.9999) 0.950   14.5873 (   13.1303-   14.4507)
1 Yes   19870327 (    1.0000-    1.0000) 0.950   44.4813 (   11.9157-   13.0345)
```

[**Please do not change the format of the first three columns,** which are to be read as input later with a format that is equivalent to **format(i1, a4, i10)** in FORTRAN]

Note that it is possible that metadata support is not found for some of the Type-0 changepoints identified (e.g., in the example above, only four Type-0 changepoints has metadata support); in this case, all the un-supported Type-0 changepoints should be deleted from the list. It could also happen that no modification to the OutFile_mCs.txt is necessary (neither in the number nor in the dates of the changepoints; so the OutFile_pCs.txt and OutFile_mCs.txt files are still identical); in this case the procedure (F4) below can be skipped.

(**F4**) Call function *StepSize* to re-estimate the significance and magnitude of the changepoints listed in OutFile_mCs.txt, e.g., enter at the R prompt the following:

```
StepSize(InSeries="C:/inputdata/InFile.csv", MissingValueCode="-999.0", pthr=0.0,
                  p.lev=0.95, Iadj=10000, Mq=10, Ny4a=0,
                  InCs="C:/results/OutFile_mCs.txt", output="C:/results/OutFile")
```

which will produce the following five files in the output directory as a result:

- OutFile_fCs.txt, which is similar to the input file OutFile_mCs.txt above, except that it contains the new estimates of significance/statistics of the changepoints listed in the input file OutFile_mCs.txt. It looks like this:

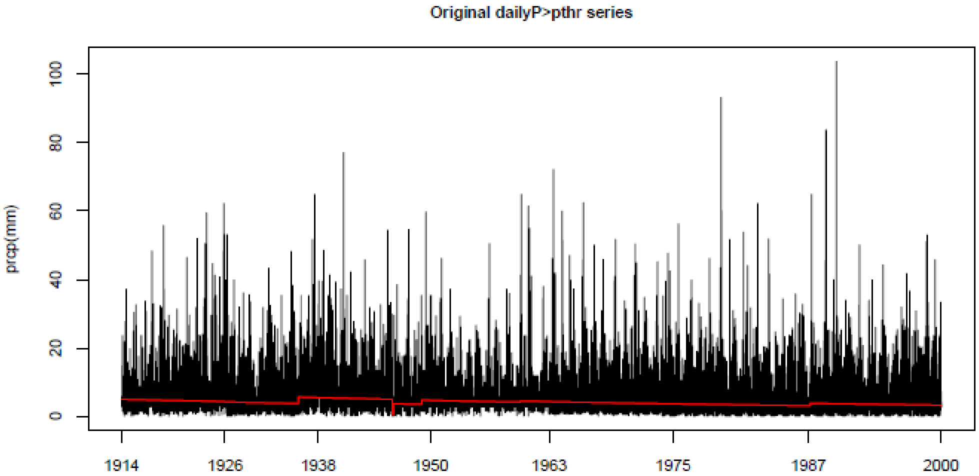
```
6 changepoints in Series ⋯Example_prcpDLY.dat⋯
1 Yes   19350927 (    1.0000-    1.0000) 0.950   48.0830 (   14.4499-   15.9967)
0 Yes   19451113 (    1.0000-    1.0000) 0.950   39.7475 (   13.6787-   15.0576)
0 Yes   19460109 (    1.0000-    1.0000) 0.950   32.3047 (   13.3338-   14.6374)
0 Yes   19490113 (    1.0000-    1.0000) 0.950   33.8784 (   13.8193-   15.2287)
0 No    19590809 (    0.9242-    0.9242) 0.950    3.5663 (   14.9664-   16.6256)
1 Yes   19870327 (    1.0000-    1.0000) 0.950   16.8990 (   15.1114-   16.8022)
```

  A copy of OutFile_fCs.txt is also stored as the OutFile_mCs.txt file (i.e., its input version is updated with the new estimates of significance/statistics) for further analysis.

- OutFile_Fstat.txt, which is similar to the OutFile_Ustat.txt or OutFile_UDstat.txt file above, except that the changepoints that are accounted for here are those that are listed in OutFile_mCs.txt.

- OutFile_F.dat, which is similar to the OutFile_U.dat or OutFile_UD.dat file above, except that the changepoints that are accounted for here are those that are listed in OutFile_mCs.txt.

- OutFile_F.pdf, which is similar to the OutFile_U.pdf or OutFile_UD.pdf above, except that the changepoints that are accounted for here are those that are listed in OutFile_mCs.txt. For the example above, it looks like this:



(**F5**) Now, one needs to analyze the results, to determine whether or not the smallest shift among all the shifts/changepoints is still significant (the magnitudes of shifts are included in the OutFile_Fstat.txt or OutFile_Ustat.txt file). To this end, one needs to compare the p-value (if it is Type-0) or $PF_{\max}$ statistic (if it is Type-1) of the smallest shift with the corresponding 95% uncertainty range. This smallest shift can be determined to be significant if its p-value or the $PF_{\max}$ statistic is larger than the corresponding upper bound, and to be not significant if it is smaller than the lower bound. However, if the p-value or the $PF_{\max}$ statistic lies within the corresponding 95% uncertainty range, one **has to determine subjectively** whether or not to take this changepoint as significant (viewing the plot in OutFile_F.pdf or OutFile_U.pdf could help here); this is due to the uncertainty inherent in the estimate of the unknown lag-1 autocorrelation of the series (see Wang 2008a).

If the smallest shift is determined to be not significant (for example, the last changepoint above is determined to be not significant), delete it from file OutFile_mCs.txt and call function *StepSize* again with the new modified list of changepoints, e.g., with this list:

```
5 changepoints in Series …Example_prcpDLY.dat…
1 Yes   19350927 (    1.0000-    1.0000) 0.950   48.0830 (   14.4499-   15.9967)
0 Yes   19451113 (    1.0000-    1.0000) 0.950   39.7475 (   13.6787-   15.0576)
0 Yes   19460109 (    1.0000-    1.0000) 0.950   32.3047 (   13.3338-   14.6374)
0 Yes   19490113 (    1.0000-    1.0000) 0.950   33.8784 (   13.8193-   15.2287)
1 Yes   19870327 (    1.0000-    1.0000) 0.950   16.8990 (   15.1114-   16.8022)
```

One should repeat this re-assessment procedure (i.e. repeat calling function *StepSize*) until **each and every** changepoint listed in OutFile_fCs.txt or OutFile_mCs.txt is determined to be significant. For example, if the 5th changepoint above (now the smallest shift among the six) is also determined to be not significant, one should delete it and call function *StepSize* again with the remaining three changepoints, which would produce the following new estimates in the OutFile_fCs.txt:

```
5 changepoints in Series …Example_prcpDLY.dat…
1 Yes   19350927 (    1.0000-    1.0000) 0.950   48.0830 (   14.4499-   15.9967)
0 Yes   19451113 (    1.0000-    1.0000) 0.950   39.7475 (   13.6787-   15.0576)
0 Yes   19460109 (    1.0000-    1.0000) 0.950   32.3047 (   13.3338-   14.6374)
0 Yes   19490113 (    1.0000-    1.0000) 0.950   33.8784 (   13.8193-   15.2287)
1 Yes   19870327 (    1.0000-    1.0000) 0.950   16.8990 (   15.1114-   16.8022)
```

Here, all these five changepoints are significant even without metadata support, because each of the corresponding $PF_{\max}$ statistics (column 5 above) is larger than the upper bound of its percentile that corresponds to the nominal level (the last number in each line). Thus, the results obtained from the last call to function *StepSize* are the final results for the series being tested.

References

Wang, X. L., H. Chen, Y. Wu, Y. Feng, and Q. Pu, 2010: New techniques for detection and adjustment of shifts in daily precipitation data series. *J. Appl. Meteor. Climatol.* **49** (No. 12), 2416-2436. DOI: 10.1175/2010JAMC2376.1

Wang, X. L., 2008a: Accounting for autocorrelation in detecting mean-shifts in climate data series using the penalized maximal *t* or *F* test. *J. Appl. Meteor. Climatol.*, 47, 2423-2444.

Wang, X. L., 2008b: Penalized maximal F-test for detecting undocumented mean-shifts without trend-change. *J. Atmos. Oceanic Tech.*, **25** (No. 3), 368-384. DOI:10.1175/2007/JTECHA982.1.

Wang, X. L., 2003: Comments on “Detection of Undocumented Changepoints: A Revision of the Two-Phase Regression Model”. *J. Climate*, **16**, 3383-3385.