RClimDex (1.0)

User Manual

By

Xuebin Zhang and Feng Yang

Climate Research Branch
Environment Canada
Downsview, Ontario
Canada

September 10, 2004

Acknowledgement

TABLE OF CONTENTS

1. Introduction

ClimDex is a Microsoft Excel based program that provides an easy-to-use software package for the calculation of indices of climate extremes for monitoring and detecting climate change. It was developed by Byron Gleason at the National Climate Data Centre (NCDC) of NOAA, and has been used in CCl/CLIVAR workshops on climate indices from~~in~~ 2001.

The original objective was to port ClimDex into an environment that does not depend on a particular operating system. It was very natural to use R as our platform, since R is a free and yet very robust and powerful software for statistical analysis and graphics. It runs under both Windows and Unix environments. In 2003 it was discovered that the method used for computing percentile-based temperature indices in ClimDex and other programs resulted in inhomogeneity in the indices series. A fix to the problem requires a bootstrap procedure that makes it almost impossible to implement in an Excel environment. This has made it more urgent to develop this R based package.

RClimDex (1.0) is designed to provide a user friendly interface to compute indices of climate extremes. It computes all 27 core indices recommended by the CCl/CLIVAR Expert Team for Climate Change Detection Monitoring and Indices (ETCCDMI) as well as some other temperature and precipitation indices with user defined thresholds. The 27 core indices include almost all the indices calculated by ClimDex (Version 1.3). This version of RClimDex has been developed under R 1.84. It should run with R 1.84 or a later version.

A main objective of constructing climate extremes indices is to use for climate change monitoring and detection studies. This requires that the indices be homogenized. Data homogenization has been planned but is not implemented in this release. Current RClimDex only includes a simple data quality control procedure that was provided in ClimDex. As in ClimDex, we require that data are quality controlled before the indices can be computed. This users' manual provides step-by-step instructions on 1) The installation of R and setting up the user environment, 2) Quality control of daily climate data, 3) Calculation of the 27 core indices.

2. Installation and running of R

R is a language and environment for statistical computing and graphics. It is a GNU implementation of the S language developed by John Chambers and colleagues at Bell Laboratories (formerly AT&T, now Lucent Technologies). S-plus provides a commercial implementation of the S language.

2.1 How to install R

RClimDex requires the base package of R and graphic user interface TclTk. The installation of R involves a very simple procedure. 1) Connect to the R project website at

http://www.r-project.org, 2) Follow the links to download the most recent version of R for your computer operating system from any mirror site of CRAN.

For Microsoft Windows (95, 98, 2000, and XP), download the Windows setup program. Run that program and R will be automatically installed in your computer, with a short cut to R on your desktop. The TclTk is included in the default installation of R 1.9.0 or later versions. It may need to be installed separately if you are running an earlier version of R.

For Linux, download proper precompiled binaries and follow the instruction to install R. For other unix systems, you many need to download source code and compile it yourself.

2.2 How to run R

Under the Windows environment, double click the R icon on your desktop, or launch it through Windows "start" menu. This usually gets you into the R user interface. For some computers, you may need to first setup an environment variable called "HOME". See R for Windows FAQ (Appendix E) for details if you have any problems.

Under a unix environment, just run R to give you the R console.

Exit from R by entering *q()* in the R console under both Windows and unix. Under Windows, you may also click "File" menu and then "Exit".

3. How to use RClimDex

3.1 Loading of RClimDex

Within the R consol prompt ">", enter *source("rclimdex.r")*. This will load RClimDex into R environment. You may need to include the full path before the filename rclimdex.r. Or you may download the most recent version from ETCCDMI web site by entering *source ("http://cccma.seos.uvic.ca/ETCCDMI/RClimDex/rclimdex.r")* if your computer is connected to the internet. Under windows, RClimDex can also be loaded from drop down menu. Choose the "File" from the RGui menu, and then select "Source R code". This will bring a new pop-up window within which you can select our R source code "rclimdex.r" from the directory where the program was saved or type *http://cccma.seos.uvic.ca/ETCCDMI/RClimDex/rclimdex.r* to download the latest version from the web site .
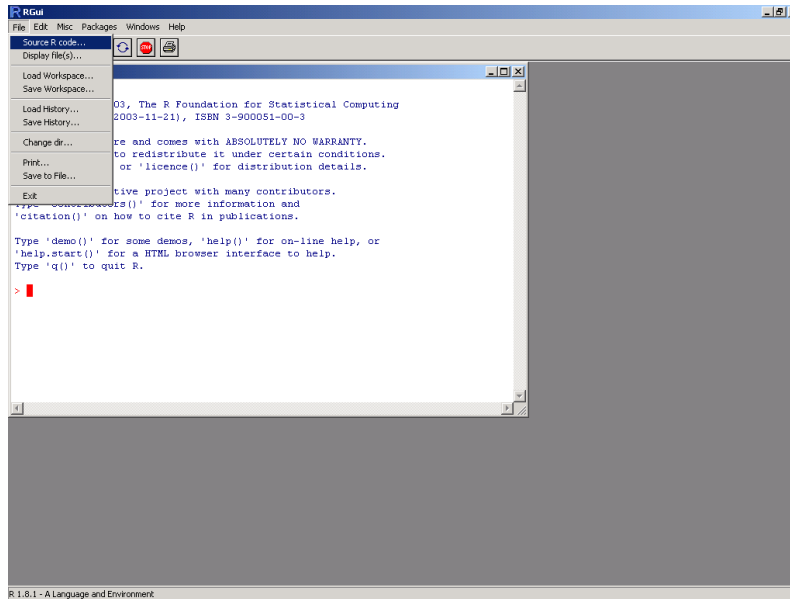
Once the source code is successfully loaded, the RClimDex main menu will appear.
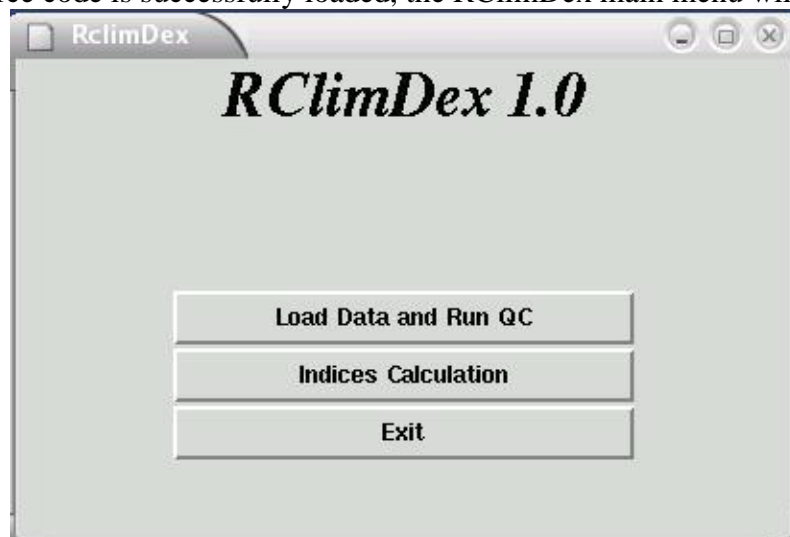
3.2. Load Data and Run QC

Data Quality Control is a prerequisite for indices calculations. The RClimDex QC performs the following procedure: 1) Replace all missing values (currently coded as -99.9) into an internal format that R recognizes (i.e. NA, not available), and 2) Replace all unreasonable values into NA. Those values include a) daily precipitation amounts less than zero and b) daily maximum temperature less than daily minimum temperature. In addition, QC also identifies outliers in daily maximum and minimum temperature. The outliers are daily values outside a region defined by the user. Currently, this region is defined as the mean plus or minus n times standard deviation of the value for the day, that is, [mean − n*std, mean+n*std]. Here std represents the standard deviation for the day and n is an input from the user and mean is computed from the climatology of the day.

Select "Load Data and Run QC" from the RClimDex Menu to open a window as shown below. This allows users to select (load) the data file from which indices are to be computed.
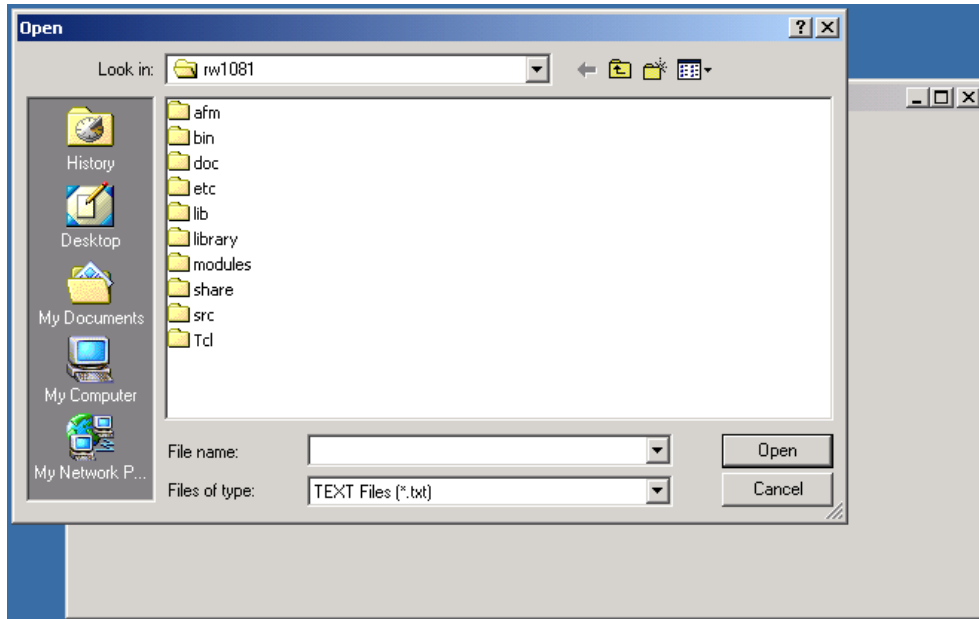
The filename should be of the form "stationname.txt". The values in the file should be of the format described in Appendix B. In this menu, we use data from a station whose data are stored in an ASCII file "21946.txt" for the purpose of demonstration. A pop-up window, as shown below, will appear once the data for station 21946 are successfully loaded.
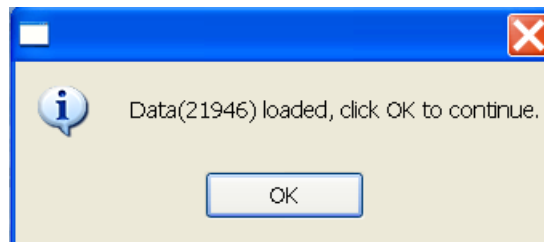
Error messages will appear in the R console if this step has not been completed successfully. This is usually caused by the wrong input data format. Please compare your format with our sample data if you see such messages.

Unreasonable values are identified automatically but identification of outliers in temperature data requires input from the user.
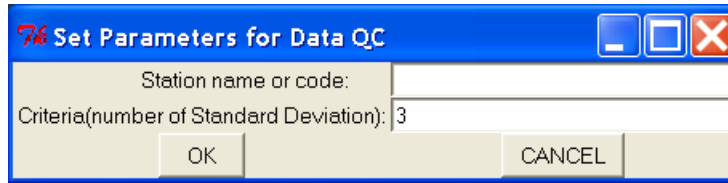
The default value for *n* is 3 (Criteria in the "Set Parameters for Data QC") window, but this number may be overwritten by the user. As a value of 3 may flag a very large number of values, users may wish to start by setting this value to 4. There is no need to fill in "Station name or code" as this parameter is for a later use. After setting the parameter, click "OK" to continue.
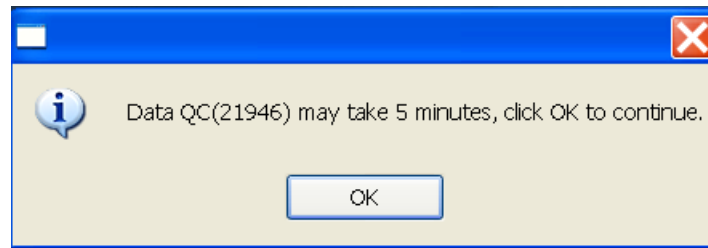
In some slower PC's, this process may take a few minutes.

Pop-up windows will appear if unreasonable values are found. For instance, when minimum daily temperature is greater than maximum daily temperature, the following message appears.
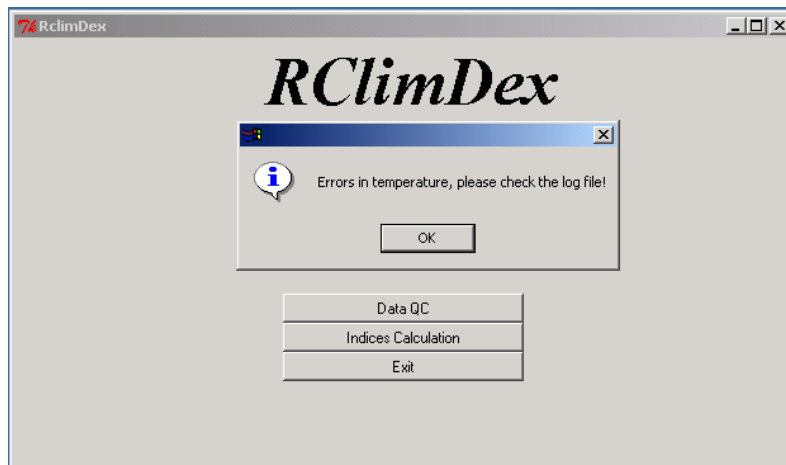
If there are any negative values (other than missing values coded as -99.9) in the daily precipitation amount, the following message will appear.

If there are outliers, the following window appears.

A pop-up window appears once the data QC is complete. At the same time, four Excel files, "21946tempQC.csv", "21946prcpQC.csv", "21946tepstdQC.csv", and "21946indcal.csv" are created in a subdirectory called *log*. The first two files contain information about unreasonable values for temperature and precipitation. The third file flags all possible outliers in daily temperature with the dates on which those outliers occur. The last file contains the QC'd data and will be u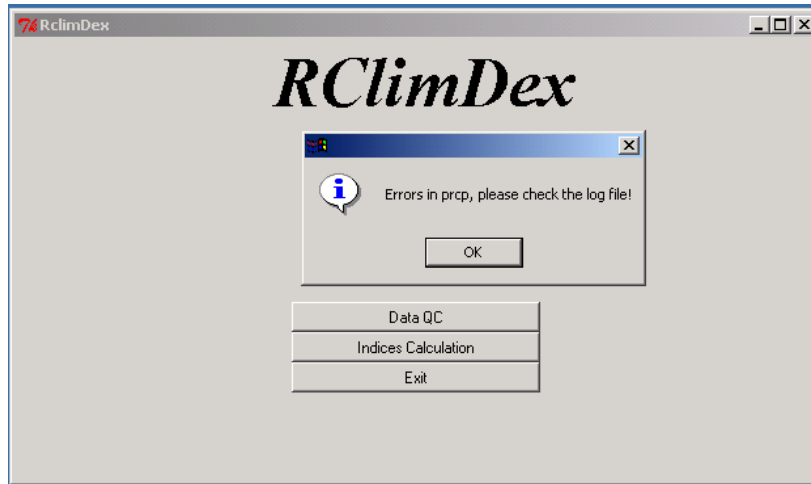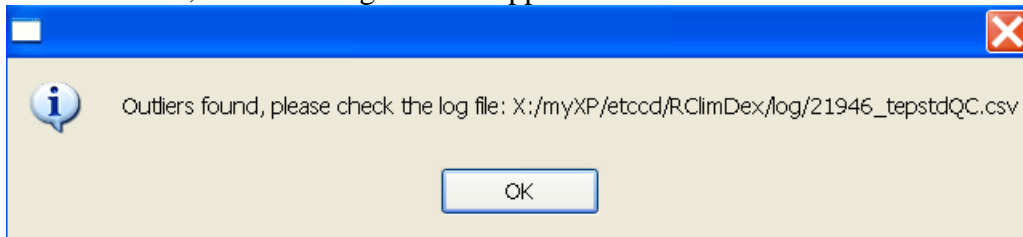sed for the indices calculation. Note that, in this file, only missing values and unreasonable values are replaced with NA, flagged possible outliers are NOT changed. For an easy visualization, 4 PDF files containing time series plots (missing values are plotted as red dots) of daily precipitation amount, daily maximum, minimum temperatures and daily temperature range are also stored in *log*.

At this point, the user may check the data in the file "21946tepstdQC.csv" to determine if any value marked as an outlier is really an outlier. The file "21946indcal.csv" can be modified using Excel under Windows and any editor under Unix if any action needs to be taken. After the completion of this step, the user may Click OK on the following window to proceed with indices calculation.
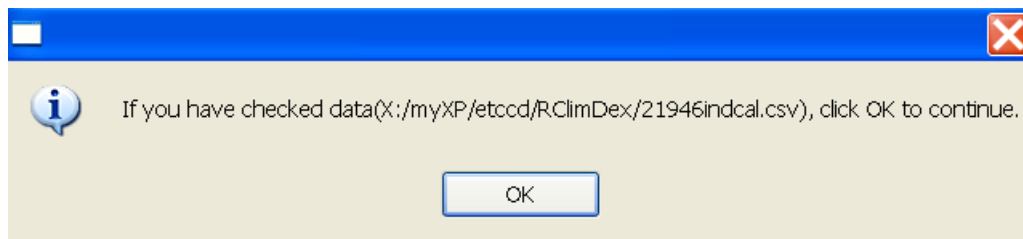
Note that, the indices are computed from the QC'd data. The original input file is not altered in any manner. So if a user chose to modify the original data file to correct some of the problematic values, the Load Data and Run QC procedure needs to be performed again on the improved data set before the changes can be reflected in the indices calculation.

3.3. Indices calculation

RClimDex is capable of computing all 27 core indices listed in Appendix A. Users may, however, compute only those indices they require.

After selecting "Indices Calculation" from the main menu, a user is asked to set up some parameters for the indices calculation. The "Set Parameter Values" window allows the user to enter the first and last years of the base period for the threshold calculation, the station latitude (Southern Hemisphere is negative) to determine in which hemisphere the station is located, a user defined daily precipitation threshold, P (in mm), to compute the number of days when daily precipitation amounts exceed this threshold (the Rnn indicator), and 4 user defined temperature thresholds. The "User defined Upper Limit of Day High" allows the calculation of the number of days when daily maximum temperature has exceeded this threshold. The "User defined Lower Limit of Day High" allows the calculation of the number of days when daily maximum temperature is below this value. The "User defined Upper Limit of Day Low" allows the calculation of the number of days when daily minimum temperature has exceeded this threshold. The "User defined Lower Limit of Day Low" allows the calculation of the number of days when daily minimum temperature is below this limit. These indices are called SUmm, FDmm, TRmm, IDmm where "mm" corresponds to user defined value. This step includes some data processing, so it will take a few seconds to finish.

Once this step is completed, a window will appear to allow the user to select their desired indices for calculation. All indices are selected by default.

Uncheck indices that are not needed, then click "OK" to perform the computation. Depending on the indices selected, this procedure may take a while.
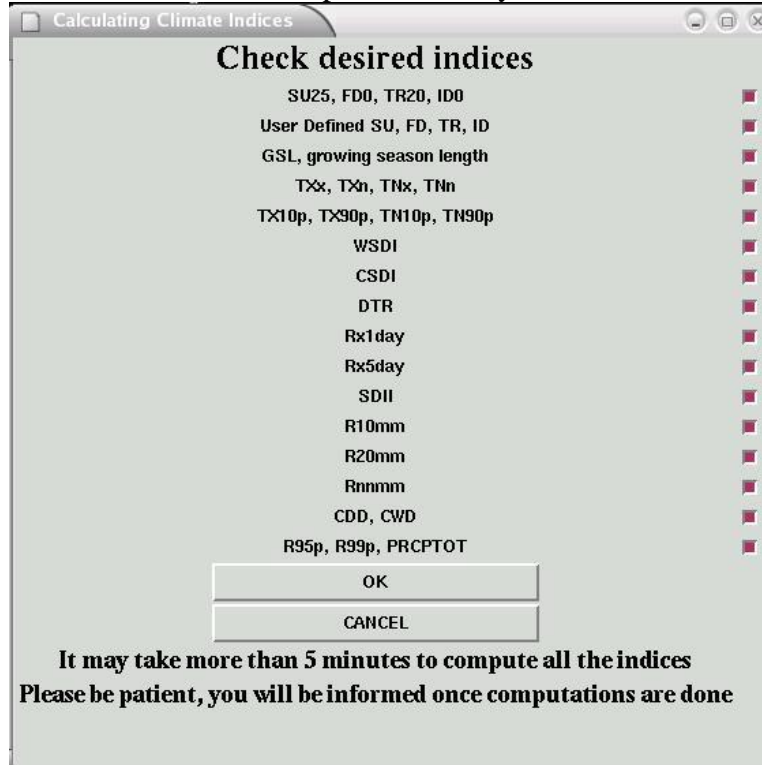
A pop-up window will appear once the selected indices are computed.
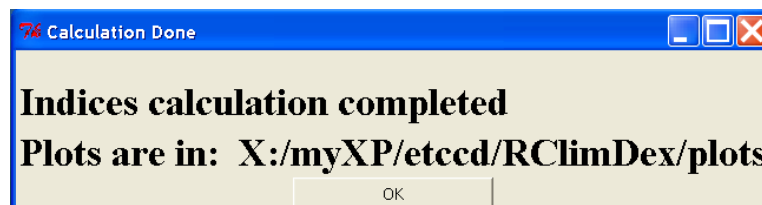
Resulting indices series are stored in a sub-directory called *indices* in Excel format. The indices files have names "21946_XXX.cvs" where XXX represents the name of the index. Data columns are separated by a comma (","). For the purpose of visualization, we plot annual series, along with trends computed by linear least square (solid line) and locally weighted linear regression (dashed line). Statistics of the linear trend fitting are displayed on the plots. These plots are stored in a sub-directory called *plots* in JPEG format. The filenames for plots follow the same rule except that "cvs" is changed to "jpg".

Select "Indices Calculation" from the main menu to compute additional indices for the same station. For additional stations, select "Data QC" and repeat the above process. Select "Exit" if all required calculations are completed.

4. Known bugs

There is a known bug in this and earlier versions of RClimDex. The program will stop running if the first year of the available data is the same as the first year of the base

period. This is caused by come computation that requires data beyond the boundary of the base period. The calculation of percentile based temperature indices is an example. One way to avoid this problem is to add an extra record for the day (with values marked as missing just before the beginning of the base period. For example, if base period is 1961-1990 and the data also starts in 1961, one may add "1960 12 31 -99.9 -99.9 -99.9" as the first line for the input data file.

5. Bug report

Please report any bugs/errors to yang.feng@ec.gc.ca with error messages and data being used for the calculation of the indices. This will be helpful in producing a better release in the near future. We would also appreciate your suggestions for further improvement.

APPENDIX A: List of ETCCDMI core Climate Indices

ID	Indicator name	Definitions	UNITS
FD0	Frost days	Annual count when TN(daily minimum)<0ºC	Days
SU25	Summer days	Annual count when TX(daily maximum)>25ºC	Days
ID0	Ice days	Annual count when TX(daily maximum)<0ºC	Days
TR20	Tropical nights	Annual count when TN(daily minimum)>20ºC	Days
GSL	Growing season Length	Annual (1st Jan to 31st Dec in NH, 1st July to 30th June in SH) count between first span of at least 6 days with TG>5ºC and first span after July 1 (January 1 in SH) of 6 days with TG<5ºC	Days
TXx	Max Tmax	Monthly maximum value of daily maximum temp	ºC
TNx	Max Tmin	Monthly maximum value of daily minimum temp	ºC
TXn	Min Tmax	Monthly minimum value of daily maximum temp	ºC
TNn	Min Tmin	Monthly minimum value of daily minimum temp	ºC
TN10p	Cool nights	Percentage of days when TN<10th percentile	Days
TX10p	Cool days	Percentage of days when TX<10th percentile	Days
TN90p	Warm nights	Percentage of days when TN>90th percentile	Days
TX90p	Warm days	Percentage of days when TX>90th percentile	Days
WSDI	Warm spell duration indicator	Annual count of days with at least 6 consecutive days when TX>90th percentile	Days
CSDI	Cold spell duration indicator	Annual count of days with at least 6 consecutive days when TN<10th percentile	Days
DTR	Diurnal temperature range	Monthly mean difference between TX and TN	ºC
RX1day	Max 1-day precipitation amount	Monthly maximum 1-day precipitation	Mm
Rx5day	Max 5-day precipitation amount	Monthly maximum consecutive 5-day precipitation	Mm
SDII	Simple daily intensity index	Annual total precipitation divided by the number of wet days (defined as PRCP>=1.0mm) in the year	Mm/day
R10	Number of heavy precipitation days	Annual count of days when PRCP>=10mm	Days
R20	Number of very heavy precipitation days	Annual count of days when PRCP>=20mm	Days
Rnn	Number of days above nn_mm	Annual count of days when PRCP>=nn_mm, nn is user defined threshold	Days
CDD	Consecutive dry days	Maximum number of consecutive days with RR<1mm	Days
CWD	Consecutive wet days	Maximum number of consecutive days with RR>=1mm	Days

R95p	Very wet days	Annual total PRCP when RR>95th percentile	Mm
R99p	Extremely wet days	Annual total PRCP when RR>99th percentile	mm
PRCPTOT	Annual total wet-day precipitation	Annual total PRCP in wet days (RR>=1mm)	mm

APPENDIX B: Input Data Format

All of the data files that are read or written are in list formatted format. The exception is the very first data file that is processed in the " Quality Control" step. This input data file has several requirements:

1.   ASCII text file
2.   Columns as following sequences: Year, Month, Day, PRCP, TMAX, TMIN. (NOTE: PRCP units = millimeters and Temperature units= degrees Celsius)
3.   The format as described above must be space delimited (e.g. each element separated by one or more spaces).
4.   For data records, missing data **must** be coded as -99.9; data records **must** be in calendar date order. Missing dates allowed.

Example data Format for the initial data file (e.g. used in the 'Quality Control' step):

```
1901  1     1       -99.9 -3.1   -6.8
1901  1     2       -99.9 -1.3   -3.6
1901  1     3       -99.9 -0.5   -7.9
1901  1     4       -99.9 -1     -9.1
1901  1     7       -99.9 -1.8   -8.4
```

APPENDIX C: Indices definitions

Definitions for indicators listed in Appendix A. For practical reasons, in this version of the software, not all indices are calculated on a monthly basis. Monthly indices are calculated if no more than 3 days are missing in a month, while annual values are calculated if no more than 15 days are missing in a year. No annual value will be calculated if any one month's data are missing. For threshold indices, a threshold is calculated if at least 70% of data are present. For spell duration indicators (marked with a *), a spell can continue into the next year and is counted against the year in which the spell ends e.g. a cold spell (CSDI) in the Northern Hemisphere beginning on 31st December 2000 and ending on 6th January 2001 is counted towards the total number of cold spells in 2001.

1. FD0

Let Tn_{ij} be the daily minimum temperature on day i in period j. Count the number of days where:

$$Tn_{ij} < 0°C$$

2. SU25

Let Tx_{ij} be the daily maximum temperature on day i period j. Count the number of days where:

$$Tx_{ij} > 25°C$$

3. ID0

Let Tx_{ij} be the daily maximum temperature on day i in period j. Count the number of days where:

$$Tx_{ij} < 0°C$$

4. TR20

Let Tn_{ij} be the daily minimum temperature on day i in period j. Count the number of days where:

$$Tn_{ij} > 20°C$$

5. GSL

Let T_{ij} be the mean temperature on day i in period j. Count the number of days between the first occurrence of at least 6 consecutive days with:

$$T_{ij} > 5^o C$$

and the first occurrence after 1st July (1st January in SH) of at least 6 consecutive days with:

$$T_{ij} < 5^o C$$

6. TXx

Let Tx_{kj} be the daily maximum temperatures in month k, period j. The maximum daily maximum temperature each month is then:-

$$TXx_{kj} = \max(Tx_{kj})$$

7. TNx

Let Tn_{kj} be the daily minimum temperatures in month k, period j. The maximum daily minimum temperature each month is then:-

$$TNx_{kj} = \max(Tn_{kj})$$

8. TXn

Let Tx_{kj} be the daily maximum temperatures in month k, period j. The minimum daily maximum temperature each month is then:-

$$TXn_{kj} = \min(Tx_{kj})$$

9. TNn

Let Tn_{kj} be the daily minimum temperatures in month k, period j. The minimum daily minimum temperature each month is then:-

$$TNn_{kj} = \min(Tn_{kj})$$

10. Tn10p

Let Tn_{ij} be the daily minimum temperature on day i in period j and let $Tn_{in}10$ be the calendar day 10th percentile centred on a 5-day window (calculated using method from Appendix D). The percentage of time is determined where:

$Tn_{ij} < Tn_{in}10$

11. Tx10p

Let Tx_{ij} be the daily maximum temperature on day i in period j and let $Tx_{in}10$ be the calendar day 10th percentile centred on a 5-day window (calculated using method from Appendix D). The percentage of time is determined where:

$Tx_{ij} < Tx_{in}10$

12. Tn90p

Let Tn_{ij} be the daily minimum temperature on day i in period j and let $Tn_{in}90$ be the calendar day 90th percentile centred on a 5-day window (calculated using method from Appendix D). The percentage of time is determined where:

$Tn_{ij} > Tn_{in}90$

13. Tx90p

Let Tx_{ij} be the daily maximum temperature on day i in period j and let $Tx_{in}90$ be the calendar day 90th percentile centred on a 5-day window (calculated using method from Appendix D). The percentage of time is determined where:

$Tx_{ij} > Tx_{in}90$

14. WSDI*

Let Tx_{ij} be the daily maximum temperature on day i in period j and let $Tx_{in}90$ be the calendar day 90th percentile centred on a 5-day window (calculated using method from Appendix D). Then the number of days per period is summed where, in intervals of at least 6 consecutive days:-

$Tx_{ij} > Tx_{in}90$

15. CSDI*

Let Tn_{ij} be the daily minimum temperature at day i in period j and let $Tx_{in}10$ be the calendar day 10th percentile centred on a 5-day window (calculated using the method from Appendix D). Then the number of days per period is summed where, in intervals of at least 6 consecutive days:-

$Tn_{ij} < Tn_{in}10$

16. DTR

Let Tx_{ij} and Tn_{ij} be the daily maximum and minimum temperature respectively on day i in period j. If I represents the number of days in j, then:

$$DTR_j = \frac{\sum_{i=1}^{I}\left(Tx_{ij} - Tn_{ij}\right)}{I}$$

17. RX1day

Let RR_{ij} be the daily precipitation amount on day i in period j. Then maximum 1-day values for period j are:

$$Rx1day_j = \max(RR_{ij})$$

18. Rx5day

Let RR_{kj} be the precipitation amount for the 5-day interval ending k, period j. Then maximum 5-day values for period j are:

$$Rx5day_j = \max(RR_{kj})$$

19. SDII

Let RR_{wj} be the daily precipitation amount on wet days, $w(RR \geq 1mm)$ in period j. If W represents number of wet days in j, then:

$$SDII_j = \frac{\sum_{w=1}^{W} RR_{wj}}{W}$$

20. R10

Let RR_{ij} be the daily precipitation amount on day i in period j. Count the number of days where:

$$RR_{ij} \geq 10mm$$

21. R20

Let RR_{ij} be the daily precipitation amount on day i in period j. Count the number of days where:

$$RR_{ij} \geq 20mm$$

22. Rnn

Let RR_{ij} be the daily precipitation amount on day i in period j. If nn represents any reasonable daily precipitation value then, count the number of days where:

$$RR_{ij} \geq nnmm$$

23. CDD*

Let RR_{ij} be the daily precipitation amount on day i in period j. Count the largest number of consecutive days where:

$$RR_{ij} < 1mm$$

24. CWD*

Let RR_{ij} be the daily precipitation amount on day i in period j. Count the largest number of consecutive days where:

$$RR_{ij} \geq 1mm$$

25. R95pTOT

Let RR_{wj} be the daily precipitation amount on a wet day $w(RR \geq 1.0mm)$ in period j and let $RR_{wn}95$ be the 95th percentile of precipitation on wet days in the 1961-1990 period. If W represents the number of wet days in the period, then:

$$R95p_j = \sum_{w=1}^{W} RR_{wj} \text{ where } RR_{wj} > RR_{wn}95$$

26. R99p

Let RR_{wj} be the daily precipitation amount on a wet day $w(RR \geq 1.0mm)$ in period j and let $RR_{wn}99$ be the 99th percentile of precipitation on wet days in the 1961-1990 period. If W represents number of wet days in the period, then:

$$R99p_j = \sum_{w=1}^{W} RR_{wj} \text{ where } RR_{wj} > RR_{wn}99$$

27. PRCPTOT

Let RR_{ij} be the daily precipitation amount on day i in period j. If I represents the number of days in j, then

$$PRCPTOT_j = \sum_{i=1}^{I} RR_{ij}$$

Appendix D : Threshold estimation and base period temperature indices calculation

Empirical quantile estimation:

The quantile of a distribution is defined as
$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}, \; 1<p<1,$$
where $F(x)$ is the distribution function. Let $\{X_{(a)}, ..., X_{(n)}\}$ denote the order statistics of $\{X_1, ..., X_n\}$ (i.e. sorted values of $\{X\}$), and let $\hat{Q}_i(p)$ denote the ith sample quantile definition. The sample quantiles can be generally written as

$$\hat{Q}_i(p) = (1-\gamma)X_{(j)} + \gamma X_{(j+1)}.$$

Hyndman and Fan (1996) suggest a formula to obtain medium un-biased estimate of the quantile by letting $j = \mathrm{int}(p*n+(1+p)/3))$ and letting $\gamma = p*n+(1+p)/3 - j$, where int(u) is the largest integer not greater than u. The empirical quantile is set to the smallest or largest value in the sample when $j<1$ or $j>n$ respectively. That is, quantile estimates corresponding to $p<1/(n+1)$ are set to the smallest value in the sample, and those corresponding to $p>n/(n+1)$ are set to the largest value in the sample.

Bootstrap procedure for the estimation of exceedance rate for the base period:

It is not possible to make an exact estimate of the thresholds due to sampling uncertainty. To provide temporally consistent estimate of exceedance rate throughout the base period and out-of-base period, we adapt the following procedure (Zhang et al. 2004) to estimate exceedance rate for the base period.

a) The 30-year base period is divided into one "out-of-base" year, the year for which exceedance is to be estimated, and a "base-period" consisting the remaining of 29 years from which the thresholds would be estimated.

b) A 30-year block of data is constructed by using the 29 year "base-period" data set and adding an additional year of data from the "base-period" (i.e., one of the years in the "base-period" is repeated). This constructed 30-year block is used to estimate thresholds.

c) The "out-of-base" year is then compared with these thresholds and the exceedance rate for the "out-of-base" year is obtained.

d) Steps (b) and (c) are repeated for an additional 28 times, by repeating each of the remaining 28 in-base years in turn to construct the 30-year block.

e) The final index for the "out-of-base" year is obtained by averaging the 29 estimates obtained from steps (b), (c) and (d).

Reference:
Hyndman, R.J., and Y. Fan, 1996: Sample quantiles in statistical packages. *The American Statistician,* **50**, 361-367.

Zhang, X., G. Hegerl, F.W. Zwiers, and J. Kenyon, 2004: Avoiding inhomogeneity in percentile-based indices of temperature extremes. *J. Climate*, submitted.

Appendix E: R for Windows FAQ

Adobe
Acrobat Docum