Environment and Climate Change Canada | Environnement et Changement climatique Canada

Canada

Methods and software for climate data homogenization

The Fifth Symposium on Monitoring and Detection of Regional Climate Change
Yinchuan, China

Xiaolan Wang
Climate Research Division
Science and Technology Branch

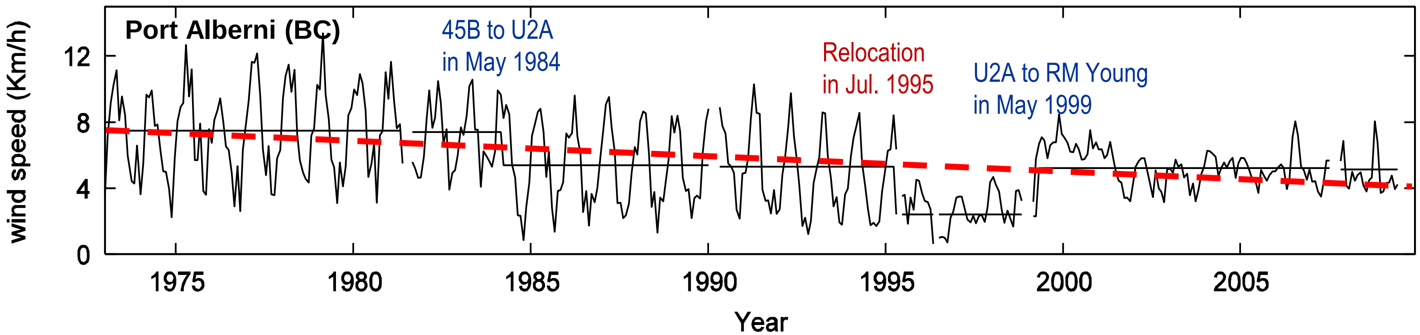
26-28 August 2018

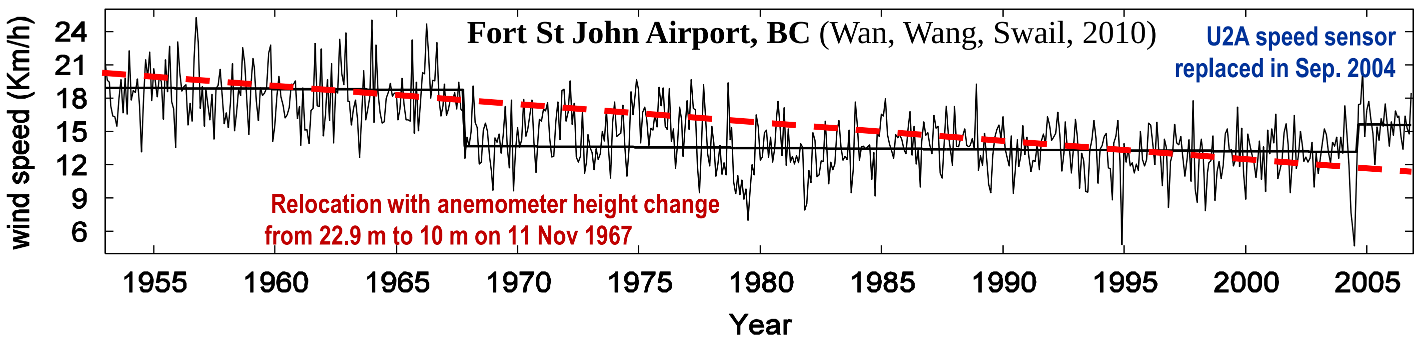
Outline

- Introduction – why climate data homogenization is inevitable
- Gaussian data – homogeneity testing methods and software
- A few simple methods for dealing with some non-Gaussian data, such as monthly precipitation
- Daily precipitation data – homogeneity testing method and software (or daily wind speed)
- Distributional shift detection method
- Sparse data series – homogenization method
- Adjustment methods for diminishing inhomogeneities
- Precautionary notes
- An upcoming book on climate data homogenization and trend analysis

Climate data homogenization is inevitable, because

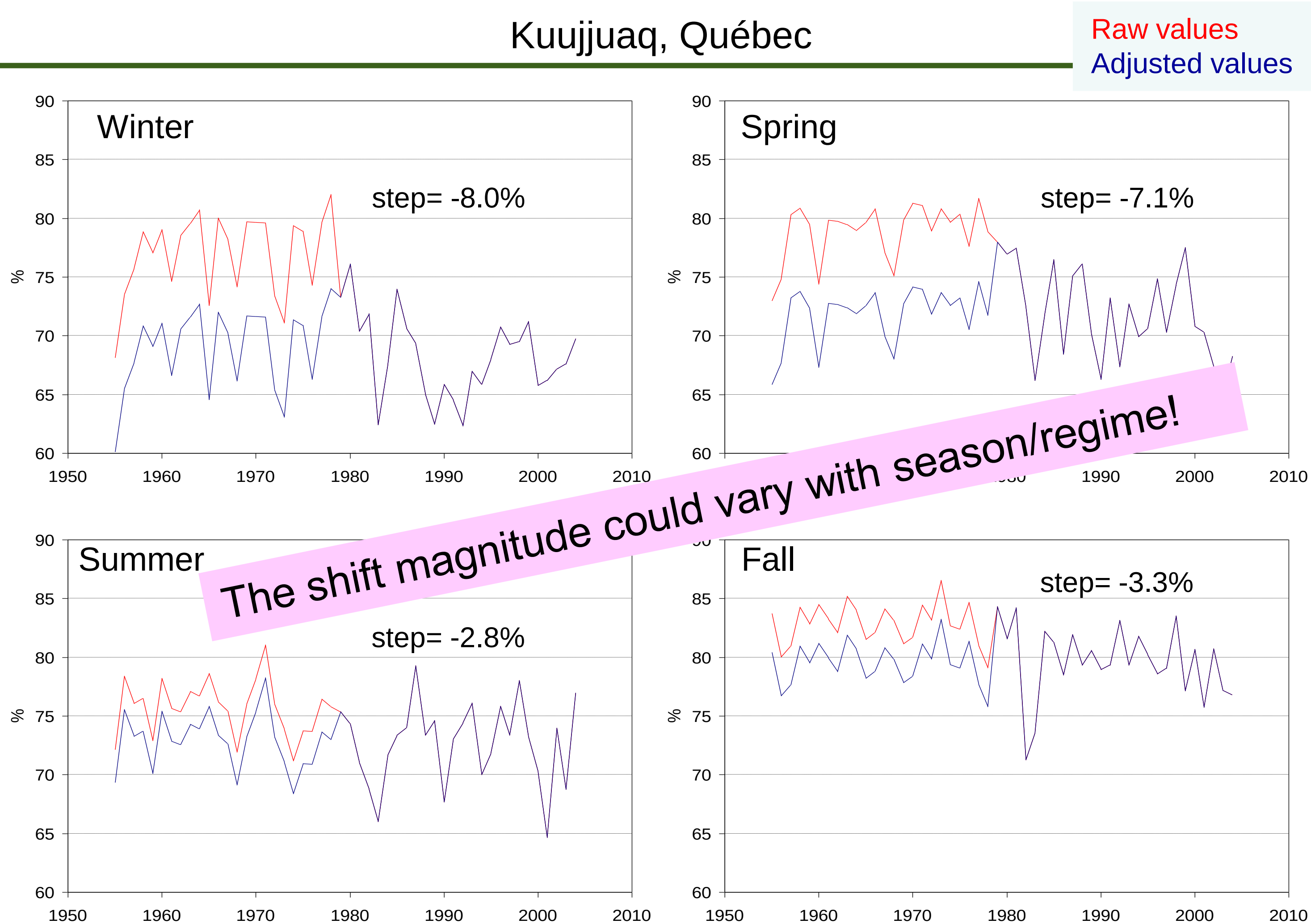
changes in observing technology and observing environment is inevitable, e.g.,





Without data homogenization, the trend would be largely biased!

Relative humidity – discontinuity due to introduction of **Dewcel** in June 1978



See Vincent et al. (2007), Surface temperature and humidity trends in Canada for 1953-2005. *J. Clim.*, **20**, 5100-5113.

Data homogenization involves two major steps:

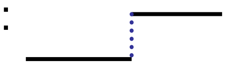
1. Identify inhomogeneities in the data series being analyzed
2. Adjust the data series to diminish the identified artificial inhomogeneities

Different types of data and changepoints need different identification methods

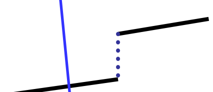
- Metadata is crucial for the identification of documented changepoints of all types; one only needs a statistical test to determine the statistical significance of such changepoints.

- But metadata is usually incomplete or unavailable. For undocumented shifts, most existing methods focus on identifying mean-shifts in Gaussian data series, e.g.,

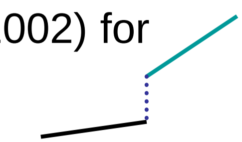
  SNHT (Alexandersson 1986) and **PMT test** (Wang et al. 2007) for mean-shifts in a difference (base-minus-reference) data series with no trend:



  PMF test (Wang 2008) for mean-shifts in a data series with a constant trend throughout the record period:



  4-parameter two-phase regression, TPR4 test (Lund and Reeves 2002) for mean-shifts that might be accompanied with a trend change:



RHtests software package uses these tests

❖ The unique features of the RHtests packages include:

1) This and the RHtests_dlyPrcp package (to be presented later) are the only existing data homogenization software that allows users to test both documented and undocumented changepoints.

2) The RHtestsV5 allows users to make Quantile-Matching (QM) adjustments to daily or subdaily (up to hourly) data series for the changepoints already identified in the corresponding annual or monthly data series.

3) The lag-1 autocorrelation in the data series being tested is accounted for, which greatly minimizes the false alarm rate;

4) The annual cycle, lag-1 autocorrelation, linear trend of the base series (when no reference is used), and all identified shifts are modelled simultaneously.

5) All can be done in graphical user interface (GUI) mode. Both the mean-adjusted and QM-adjusted data series, along with plots of the series and the resulting regression fit are provided in the graphical output.

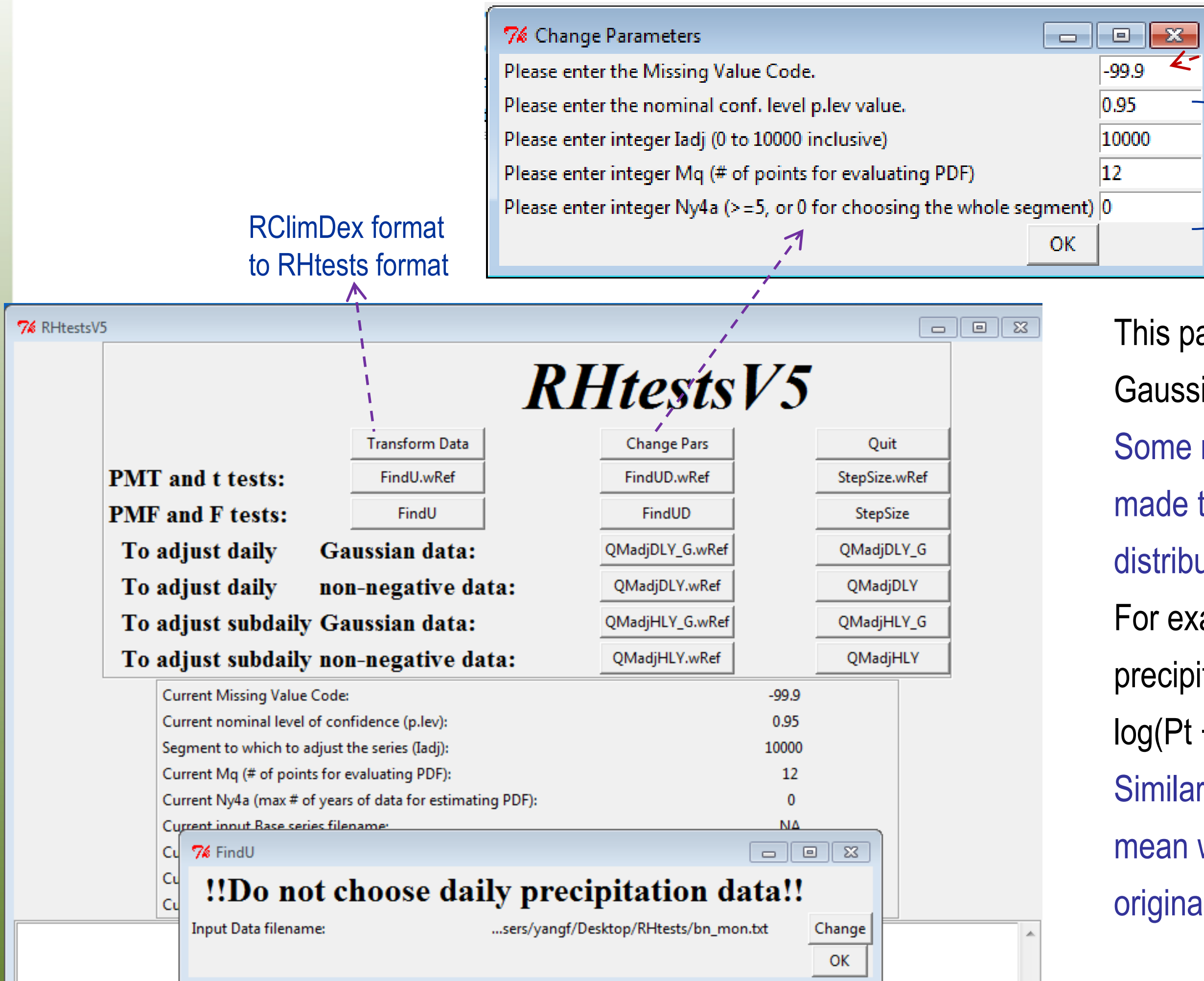
6) Users can also use the **Change Pars** button

i) choose the segment to which to adjust the base series (e.g., to the most accurate or latest segment);

ii) choose to use the whole or part of the segments before and after a shift to estimate the QM-adjustments;

iii) choose the level of significance at which to conduct the tests

The GUI mode of the RHtests package:



RClimDex format to RHtests format

Must be changed to the missing value code in your data

You can change these default values if you wish

This package is for homogenization of Gaussian data series.

Some non-Gaussian data can easily be made to approximate a Gaussian distribution.

For example, for monthly or annual total precipitation Pt: test log(Pt) or log(Pt + 0.1), instead of testing Pt.

Similarly, test deseasonalized monthly mean wind speed series rather than the original monthly mean wind speed series

will talk about how to use the other functions later

Daily precipitation – homogenization method and software:

Precipitation is typically not normally distributed; daily precipitation is not a continuous variable!

- Log transformation is often sufficient for monthly/annual total precipitation (Prcp) data series

→ recommend use the RHtests functions to test a log-transformed monthly/annual Prcp series

- Homogenization of daily precipitation data is much more challenging, and yet much needed for characterizing extremes

Log-transformation is often not good enough; a data-adaptive transformation procedure is needed.

- Integrate a Box-Cox transformation in the PMFred algorithm, developing the transPMFred algorithm & RHtests_dlyPrcp package for homogenization of daily precipitation data series

→ alleviates the limitation of the assumption of normal distribution in the RHtests package

Box-Cox transformation: $X_i = h(Y_i;\lambda) = \begin{cases} (Y_i^{\lambda}-1)/\lambda, & \lambda \neq 0 \\ \log Y_i, & \lambda = 0 \end{cases}$ where $Y_i > 0\,(i = 1,2,\ldots,N)$ is a series of non-zero daily precipitation amounts

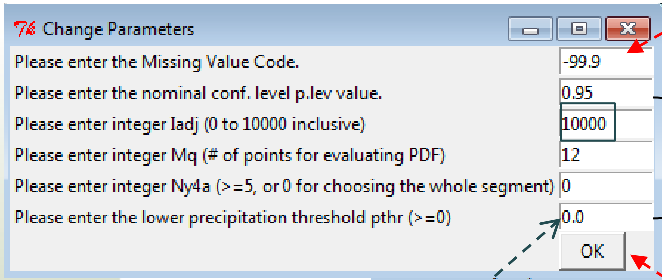
Y_i can be other positive values, e.g., non-zero wind speeds

The gist of the transPMFred algorithm:

- For a set of trial λ values, use the PMFred algorithm to test each transformed series X_i
- Use a profile log-likelihood statistic to find the best λ for the series being tested

A data-adaptive transformation, because different λ values (transformations) may be chosen for different series

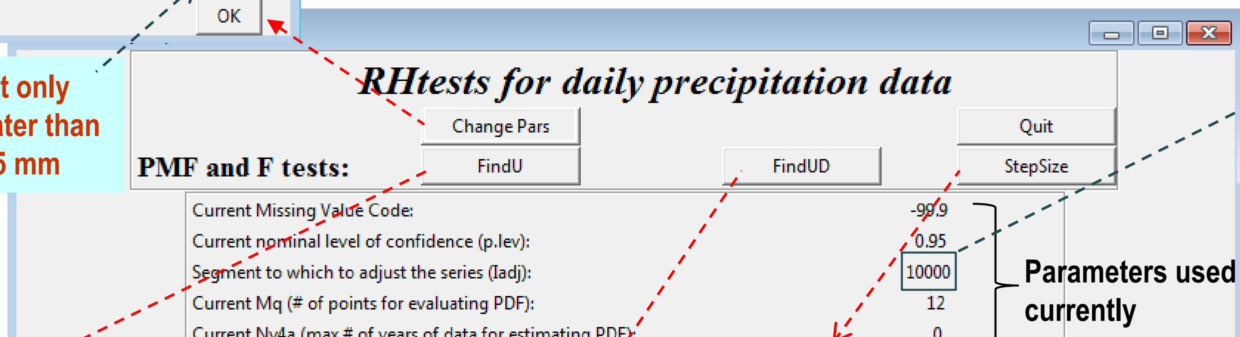
The software RHtests_dlyPrcp consists of three functions; all of them available in GUI mode:



must be changed to the code used for missing values in your data!

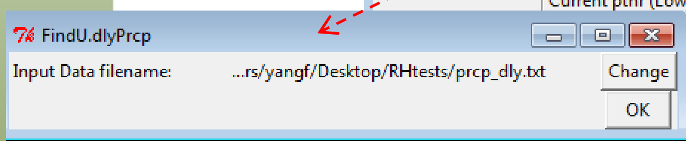
You can change these default values to the values you want to use. Namely, you can choose (1) the significance level to conduct the test, (2) the segment to which to adjust the series, (3) the number of categories/points you want to use to estimate the probability distribution, (4) to use all or part of the data in a segment to estimate the QM adjustments

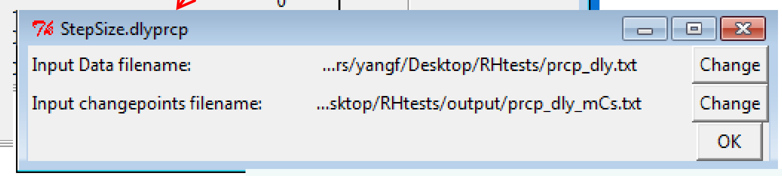
You can also choose to test only precip. values that are greater than a chosen threshold, say 0.5 mm



to adjust the data to the latest seg. (better to adjust to the highest seg.)

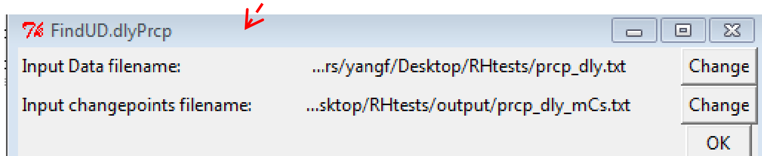
Parameters used currently





Click **FindU** button to choose the precip. series to be tested. Then, click Ok to run the test. This will find significant Type-1 (unknown) changepoints, i.e., those that are significant even without metadata support

Click **StepSize** button to re-estimate the size and significance of shifts after you make any change in the list of changepoints, for example, add a documented shift, or change the date to a nearby documented date of change, or decide not to adjust a statistically detected shift…

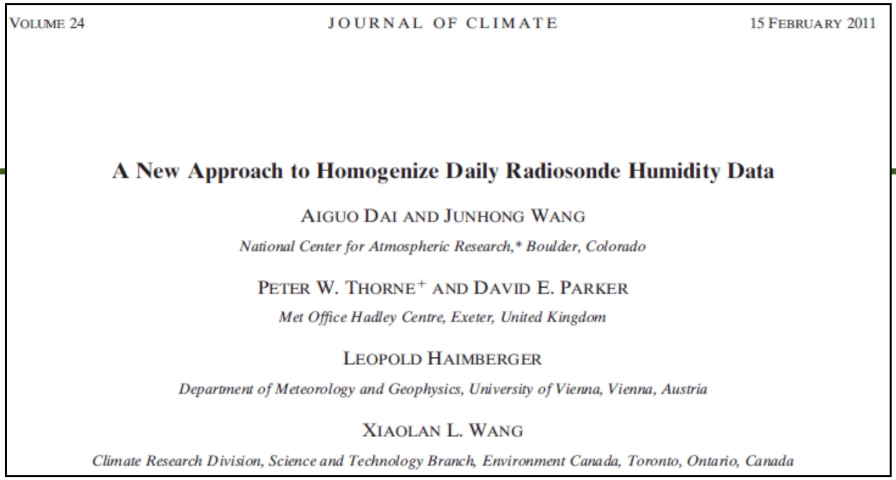


Click **FindUD** button to find potential Type-0 changepoints, namely, those that are significant only if they are supported by metadata. Skip this step if you don't have metadata or only want to focus on Type-1 shifts

Canada

Distributional change detection method

Dai et al., 2011, J. Clim., 24, 965-990:


VOLUME 24 JOURNAL OF CLIMATE 15 FEBRUARY 2011

A New Approach to Homogenize Daily Radiosonde Humidity Data

AIGUO DAI AND JUNHONG WANG
National Center for Atmospheric Research, Boulder, Colorado*

PETER W. THORNE+ AND DAVID E. PARKER
Met Office Hadley Centre, Exeter, United Kingdom

LEOPOLD HAIMBERGER
Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

XIAOLAN L. WANG
Climate Research Division, Science and Technology Branch, Environment Canada, Toronto, Ontario, Canada

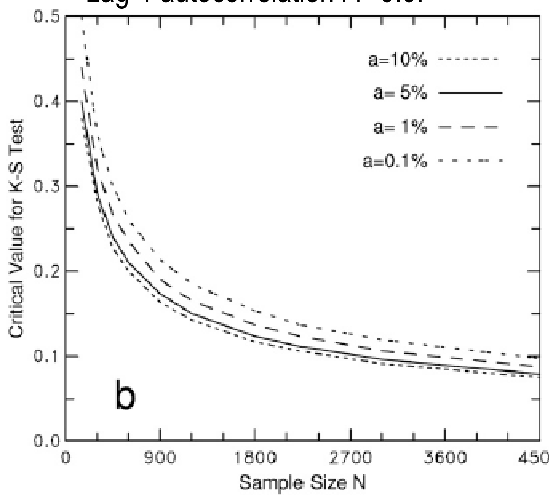
Developed a variant of the Kolmogorov-Smirnov (K-S) test - a test for differences in two distributions

Due to the need to search for the most probable changepoints, the standard K-S test is not suitable for detecting unknown changepoints but can be applied to test significance of documented changes.

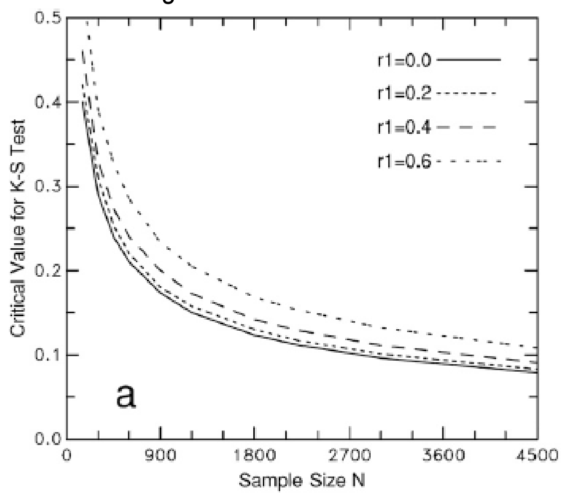
For a given level of significance, the critical value of our K-S test is much larger than that for the standard K-S test, and was estimated by Monte Carlo simulations:

We used this variant of the K-S test in combination with the RHtests to homogenize the global DPD data

Lag-1 autocorrelation r1=0.0:



For significance level a = 5%:



Sparse data series – homogenization method

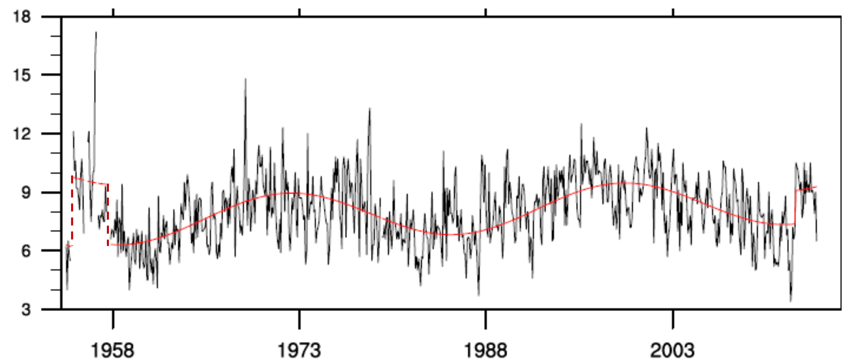
(i.e., data series from data sparse areas and/or periods)

Wen, Wang and Wong (2011a & 2011b; Wen's Ph.D. Thesis) developed

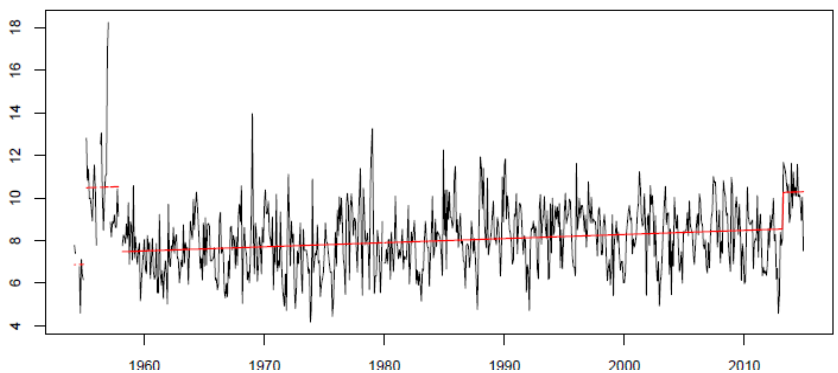
- An Overlapped Grouping Periodogram Test for Detecting Multiple Hidden Periodicities in Mixed Spectra, and
- A hybrid-domain approach for modeling climate data time series, which includes a two - phase competition procedure to address the confounding issue between modeling periodic variations and mean shifts
- This is not yet included in the RHtests package; it will be eventually.
- One can also use other spectrum analysis methods in combination with a changepoint detection algorithm
- The **gist** is to use spectrum analysis to estimate and set aside the low-frequency oscillation while testing the series for mean shifts, with an iterative procedure to find the best fit to the data series in question

An example of application to a monthly mean wind speed series:

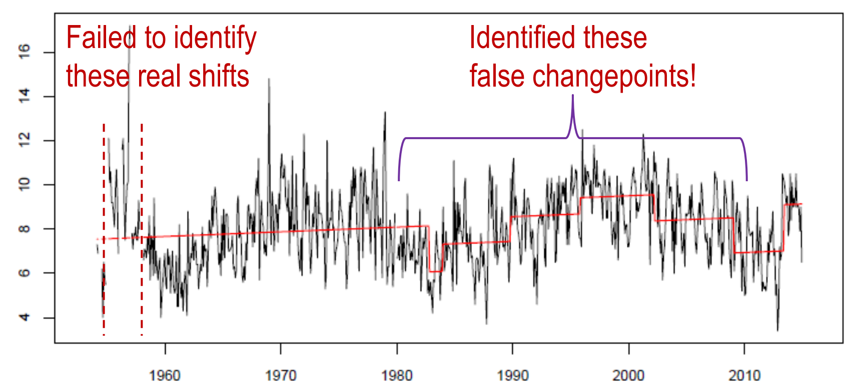
This series has a 313-month (~26-year) cycle and three shifts:



Detection result for the original minus 33-month cycle series:

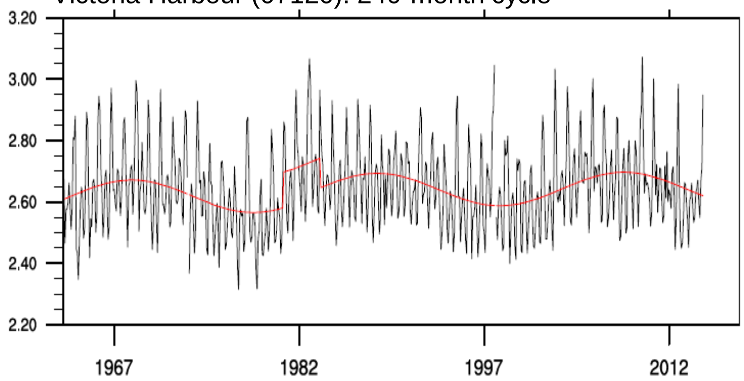


Detection result for the original series (without a reference):

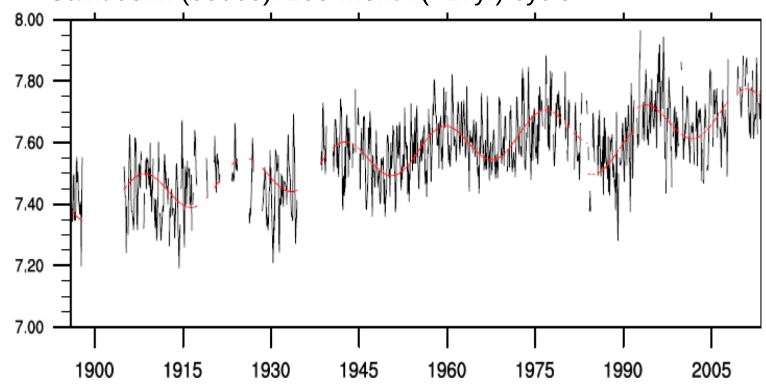


More examples – series of monthly means of maximum water level (from tide gauges):

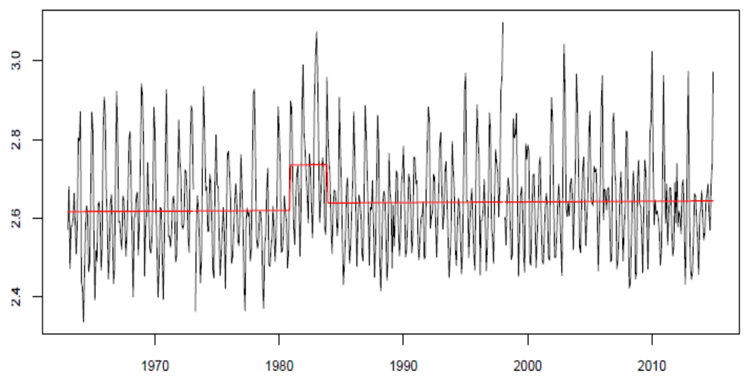
Victoria Harbour (07120): 240-month cycle



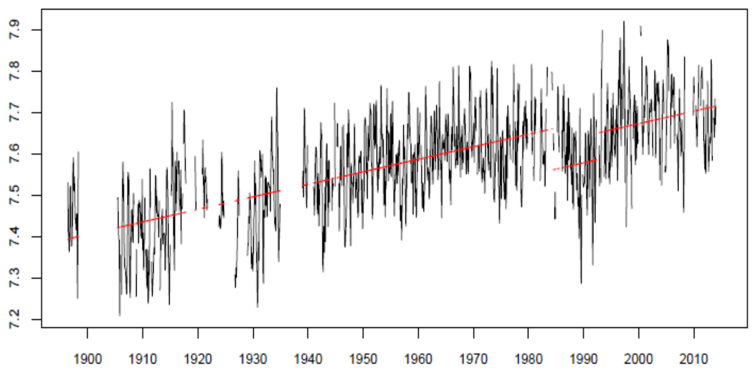
Saint John (00065): 205-month (~17yr) cycle



Series with the 240-month cycle being set aside:



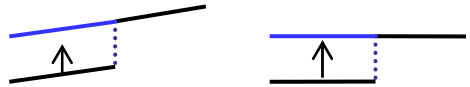
Series with the 205-month cycle being set aside:

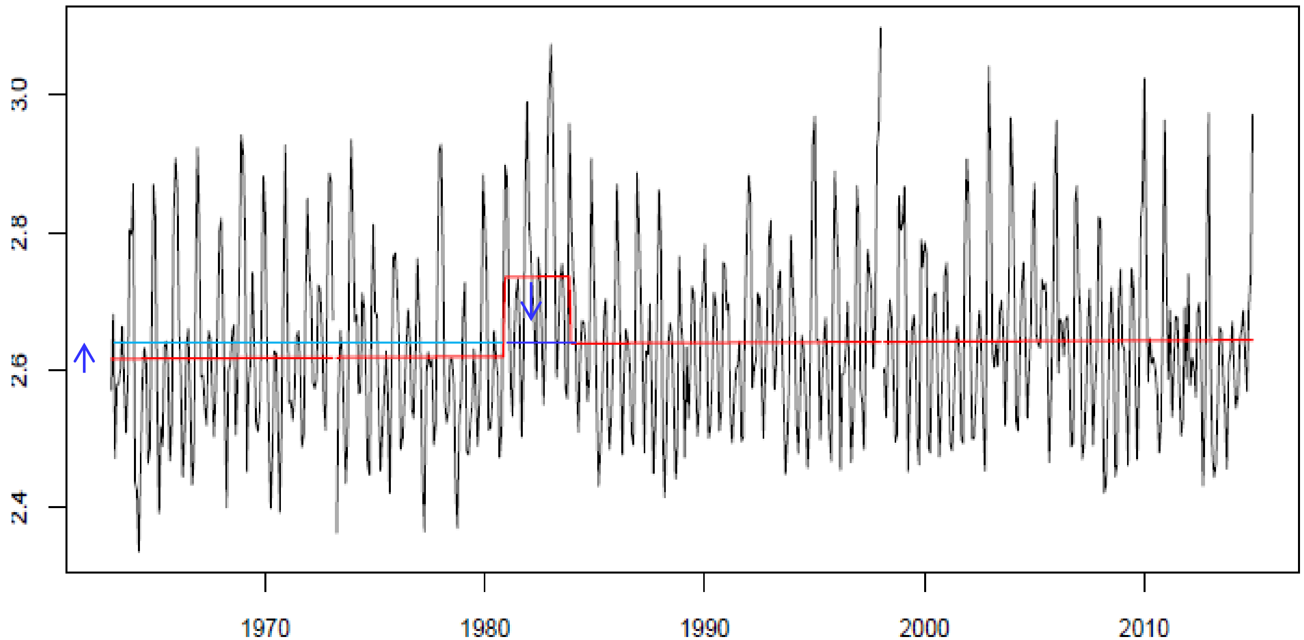


Adjustment methods for diminishing inhomogeneities

- Mean adjustment methods
- Distributional adjustment methods

(i) Mean adjustment methods:





could be based on the base-minus-Ref series or base-Ref regression residual series.

When no Ref:
could base on a multi-phase regression fit to the base series.

Mean adjustment is acceptable if the shift is only in the mean with no seasonality & no variance change, or if the homogenized data is only for use to estimate trend in the series, but not for any other applications especially not for studies of extremes

(ii) Distribution adjustment methods include

Percentile Adjustment method (Trewin and Trevitt, 1996)

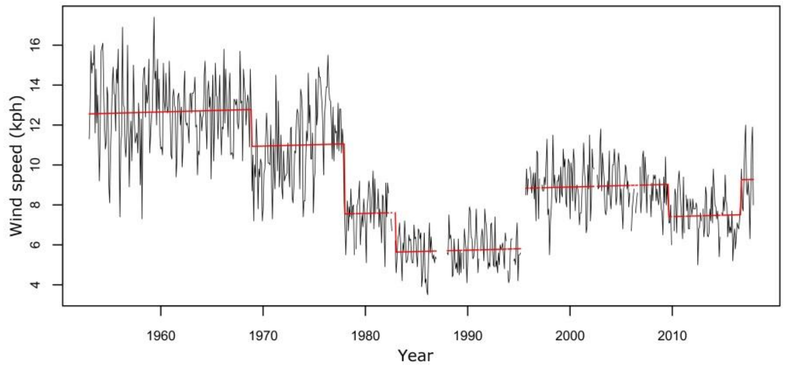
HOM adjustment method (Della-Marta and Wanner, 2006)

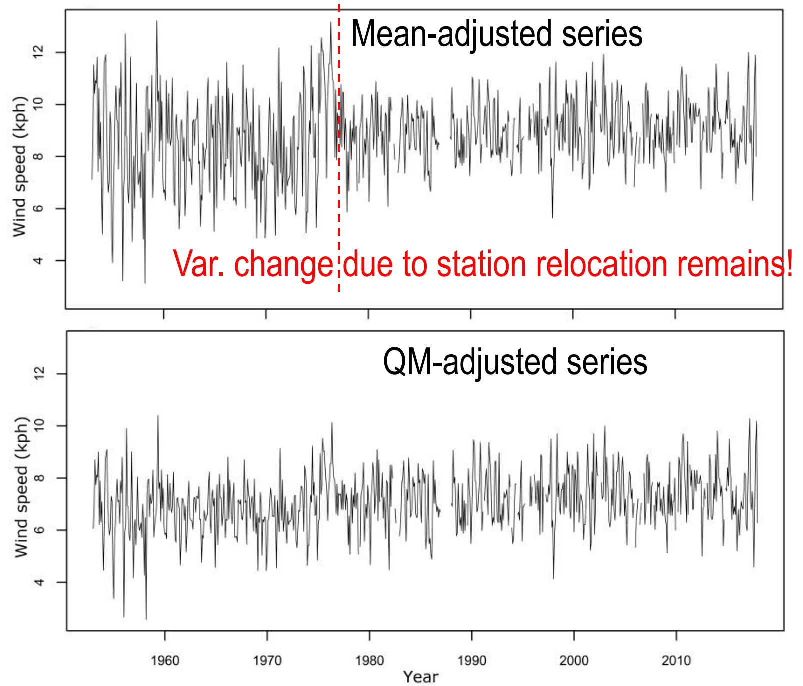
SPLIDHOM (spline daily homogenization) method (Mestre et al. 2011)

Adjustments are based on regression residuals; weak relationship → failure

Quantile Matching (QM) adjustment method (Wang et al. 2010, Vincent et al. 2012)

Adjustment based on the base-minus-reference series when using a reference or de-trended base series when not using a reference





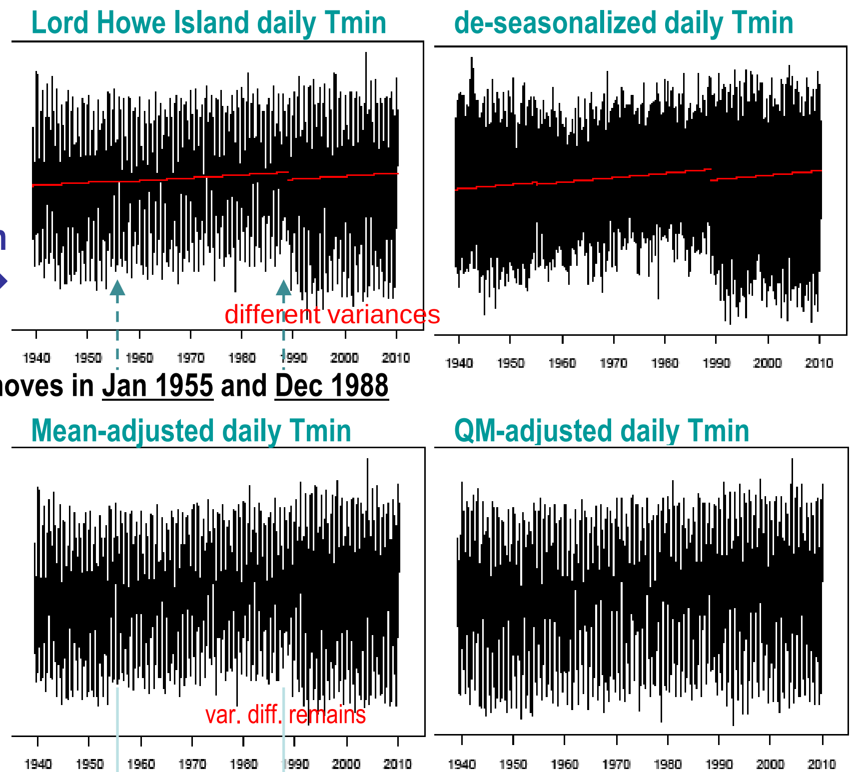
- regime dependent shifts
- seasonality of shifts, e.g., ...

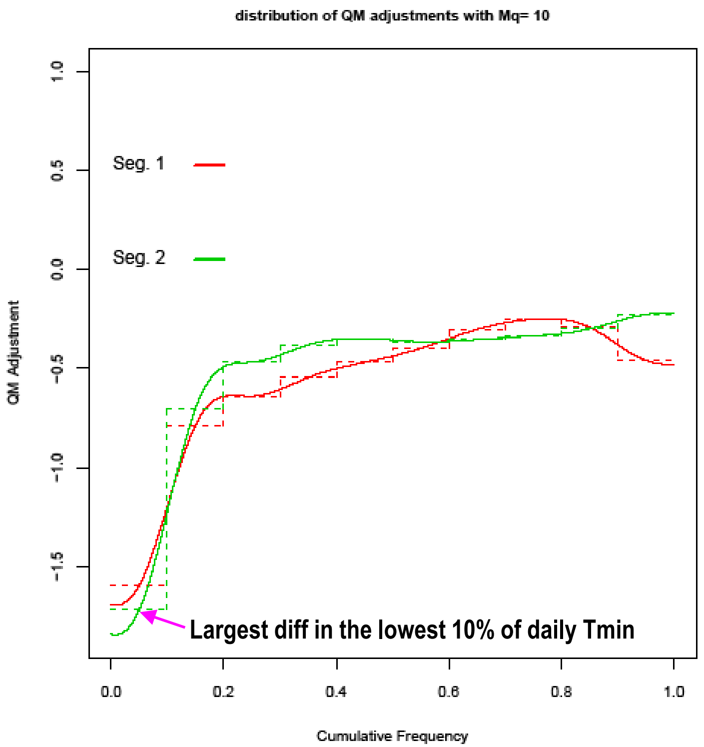
QM adjustment method is for adjusting quantile-dependent shifts,

i.e. shifts that affect not only the mean, but also the entire distribution of the data.

Site moves at an Australian station → quantile-dependent shifts:

Larger effects on cold extremes →





Gist of QM adjustments – to match the distributions of different segments of the base-minus-reference series or the de-trended base series, i.e., to diminish differences in the distribution caused by non-climatic factors.

to preserve in the QM-adjusted series the linear trend estimated from a multi-phase regression fit
- important not to remove the natural trend!

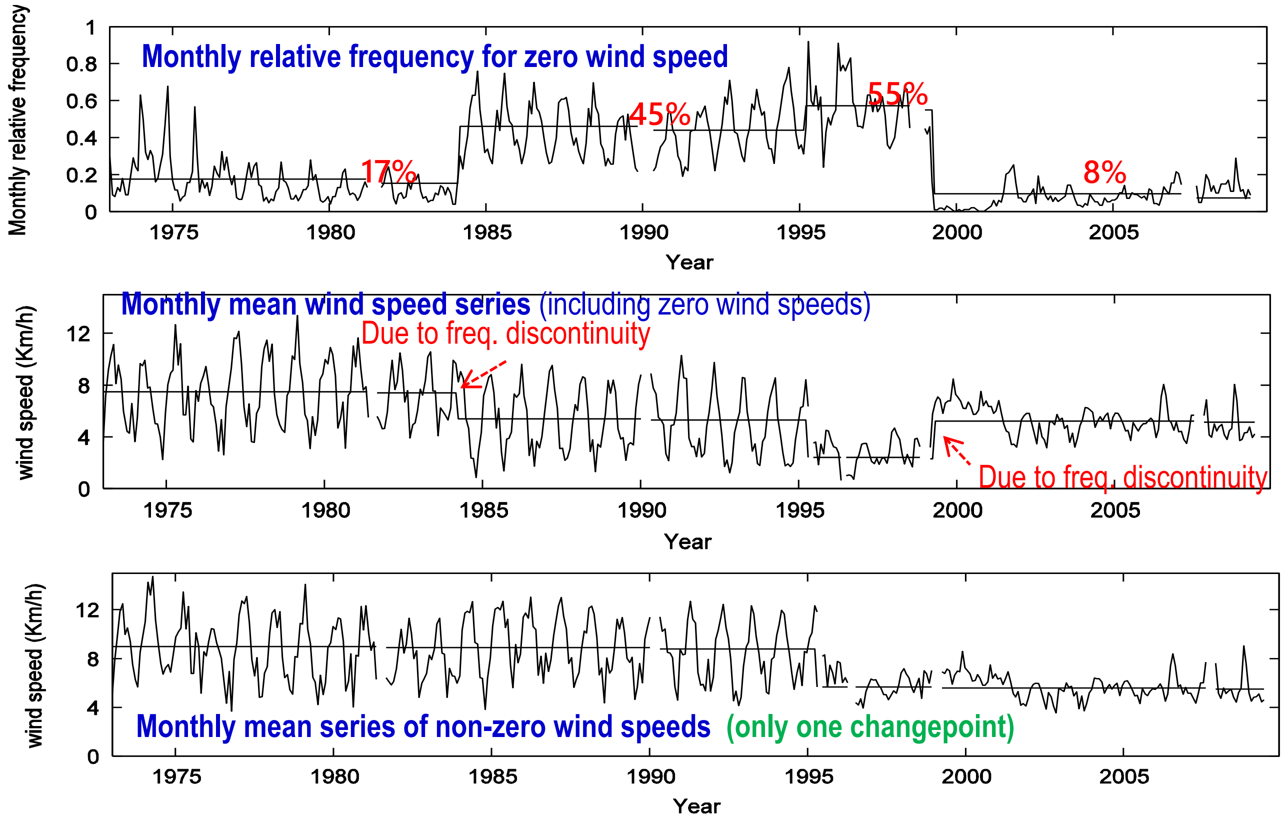
Caveat of quantile/percentile matching adjustments for non-continuous data

(e.g., daily or subdaily precipitation or wind speed data…)

Quantile/percentile matching algorithms would work only if there is no frequency inhomogeneity, because they line up the adjustments by empirical frequency, implicitly assuming homogeneous frequencies.

→ **they should be used after all freq. discontinuities have been diminished! Otherwise, freq. dismatch!**

An example of frequency discontinuity:

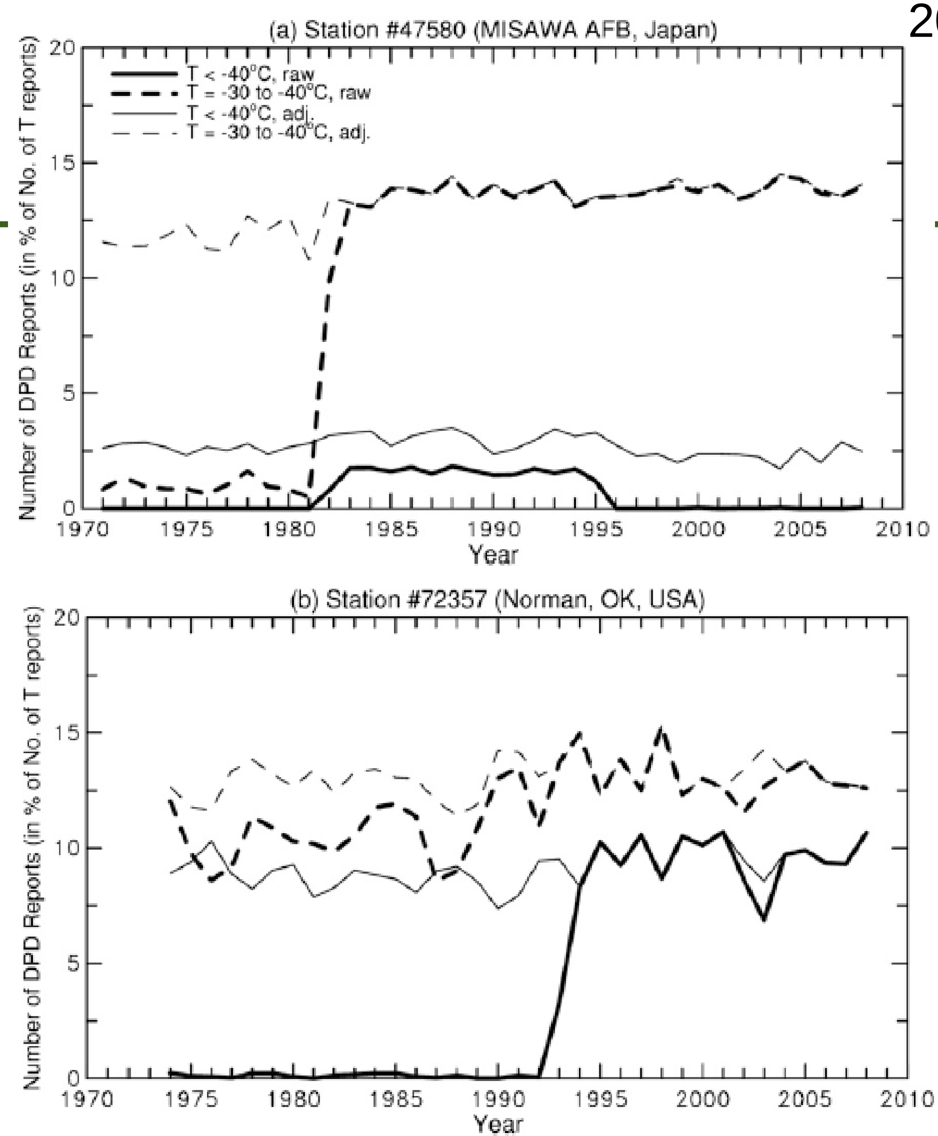


Another example of frequency discontinuity:

The early radiosonde hygrometer are considered unreliable under extreme cold conditions.

So it was a standard practice until ~1993 for U.S. (and other) stations to report humidity as missing when T < 40˚C:

Dai et al., 2011 (J Clim., 24, 965-990) used the relationship between temperature/dewpoint temperature and vapor pressure to estimate the missing DPD reports under the cold conditions.



Time series of annual Dew Point Depression (DPD) reports, expressed in percentage of the total temperature (T) reports

Precautionary notes

- Climate data homogenization is inevitable and should be done with extra caution and using all available metadata. Using any method or software in a fully automatic procedure could lead to very bad, misleading results!

- The data properties should be considered when doing climate data homogenization:
  Normally distributed? Autocorrelated? Continuous or non-continuous? Frequency discontinuities? …

- Whenever possible, first use physical-based relationship to adjust known problems, such as using wind profile to adjust for anemometer height changes (Wan et al. 2010, J.Clim, 23, 1209-1225), or use the relationship between temperature/dewpoint temperature and vapor pressure to estimate the missing humidity values (Dai et al., 2011).

- Compare the maps of trend estimated using the raw and homogenized data. The homogenized data should show a trend pattern of better spatial consistency. Also, visualize the time series along with the fit with changepoints and compare it with the neighbor station series to help make the final decision.

- The entire data homogenization procedure should be well documented. This document (could be a journal publication) should be stored/published at the same place as the resulting homogenized data.

- Homogenized data is more homogenous than the raw data but is <u>not necessarily closer</u> to the truth, although often it is.

The upcoming book:

Changepoint Detection, Data Homogenization, and Trend Analysis in Climate Research

By Xiaolan L. Wang and Francis W. Zwiers

To be published by Cambridge University Press

501 – Proprietor Agreement

ACADEMIC PROPRIETOR AGREEMENT (AUTHORED BOOK)

CAMBRIDGE UNIVERSITY PRESS

EFFECTIVE DATE: 27 of April, 2018

BETWEEN:

1 **The Chancellor, Masters, and Scholars of the University of Cambridge acting through its department:**
Cambridge University Press
University Printing House
Shaftesbury Road
Cambridge CB2 8BS
UK
('Cambridge')

AND:

2 **Her Majesty the Queen in Right of Canada, as represented by the Minister of the Environment, c/o Intellectual Property Office, 200 boulevard Sacré Coeur, Gatineau, QC, K1A OH3, Canada (the 'Proprietor')**

(each a 'Party' and, together, the 'Parties')

FOR a work co-authored by **Xiaolan Wang ('the Author')** whose is employed by the Proprietor. Said work is provisionally entitled

Changepoint Detection, Data Homogenization, and Trend Analysis in Climate Research

(the 'Work')

BACKRGOUND:

A. The Parties wish to produce and publish the Work co-authored by the Author and Francis Zwiers;
B. The Author is an employee of the Proprietor and, as such, this Agreement reflects the ownership by Her Majesty the Queen in Right of Canada of any copyright in the contributions to the Work created by the Author hereunder; and
C. Cambridge has agreed terms in respect of Francis Zwiers contributions to the Work under a separate agreement with Francis Zwiers (the 'Co-Author').

Cambridge and the Proprietor hereby accept and agree to the terms of this Agreement, which incorporates the following attached Sections:

SECTION I:	Specific terms and conditions
SECTION II:	Standard terms and conditions

SIGNED: ____________________
for and on behalf of the Chancellor, Masters, and Scholars of the University of Cambridge acting through its department, Cambridge University Press

SIGNED: ____________________
for and on behalf of the Proprietor, by Claude Bélisle
Director- Procurement Servicesand Contracting and Intellectual Property Office
Corporate Services and Finance Branch

This book will have the following 12 chapters:

1. Introduction
2. Climate data quality control procedures
3. Related statistical concepts, definitions, and models
4. Regression based mean-shift models for climate data time series
5. Other changepoint detection/testing methods
6. Non-negative climate data time series with changepoints
7. Homogenization of satellite climate data records
8. Daily or sub-daily data time series with changepoints
9. Discrete-valued climate data time series with changepoints (*e.g., cloudiness data – categorical data*)
10. Data homogenization methods
11. Practical aspects of climate data homogenization
12. Statistical methods for trend analysis in climate research

The Wang and Swail (2001, *J. Clim, 14, 2204-2221*) trend analysis method

- a Mann-Kendall (Sen-Theil) trend estimator and test that accounts for the effect of lag-1 autocorrelation.

This method has been found to perform best in comparison with other trend calculation methods, especially for short time series - **IPCC AR5, page 2SM-12** :

6. **Wang and Swail (2001) iterative method (WS2001).** A method of trend calculation iterating between computing Sen–Theil trend slope for time series prewhitened as in equation (2.SM.13), computing data residuals of the original time series with regards to the line with this new slope, estimating $\hat{\rho}$ from these residuals (as in Equations (2.SM.10) to (2.SM.12)), prewhitening the original time series using this $\hat{\rho}$ value, etc. Zhang and Zwiers (2004) compared this method with other approaches, including Maximum Likelihood for linear trends with AR(1) error, and found it to perform best, especially for short time series.

We have been and are providing our R and FORTRAN codes for anyone to apply this method.

IPCC AR5,

Table 2.SM.3 | Trends (degrees Celsius per decade) and 90% confidence intervals for HadCRUT4 global mean annual time series for periods 1901–2011, 1901–1950 and 1951–2011 calculated by methods described in the Supplementary Material. Effective sample size N_r and lagged by one time step correlation coefficient for residuals $\hat{\rho}$ are given for methods that compute them. Note differences in the width of confidence intervals between methods that assume independence of data deviations from the straight line (OLS and Sen–Theil methods) and those that allow AR(1) dependence in the data (all other methods). Two of these methods use non-parametric trend estimation (Sen–Theil and WS2001).

Method	1901–2011			1901–1950			1951–2011		
	Trend	N_r	$\hat{\rho}$	Trend	N_r	$\hat{\rho}$	Trend	N_r	$\hat{\rho}$
OLS (Ordinary LS)	0.075 ± 0.006			0.107 ± 0.016			0.107 ± 0.015		
S2008 (Santer et al. 2008)	0.075 ± 0.013	28	0.599	0.107 ± 0.026	21	0.407	0.107 ± 0.028	21	0.494
GLS (Generalized LS)	0.073 ± 0.012		0.599	0.100 ± 0.023		0.407	0.104 ± 0.025		0.494
Prewhitening	0.077 ± 0.013		0.594	0.113 ± 0.022		0.362	0.111 ± 0.026		0.488
Sen–Theil	0.075 (–0.006, +0.007)			0.113 (–0.019, +0.019)			0.109 (–0.017, +0.019)		
WS2001	0.079 (–0.014, +0.012)		0.596	0.114 (–0.026, +0.023)		0.352	0.110 (–0.028, +0.029)		0.487

Thank you very much for listening!

Questions?