

Hydrological Data Screening

Detecting inhomogeneity and trends in hydrological time series due to climatic and environmental change



Raymond Venneker
UNESCO-IHE Institute for Water Education
Delft, the Netherlands

Hydrological Change

Alterations in the hydrological regime occur due to:

- Influence from changing climatic conditions
 - GHG emissions, pollution, deforestation, ...
- Influence from physical changes at the land surface
 - Land use, infrastructure, water use, ...

In order to determine the impacts on the hydrology we need data

- Continuous
- Good quality
- Long records

Characteristics of water resources data

- Lower bound of zero, negative values often not plausible
- Presence of outliers, commonly in high values
- Positive skewness, most values in the lower region
- Therefore often non-normal distribution
- Seasonal patterns are often present
- Autocorrelation, mostly positive
- Dependence on other variables, e.g. elevation, soils, land use

After: Helsel and Hirsch (2002)

Time series

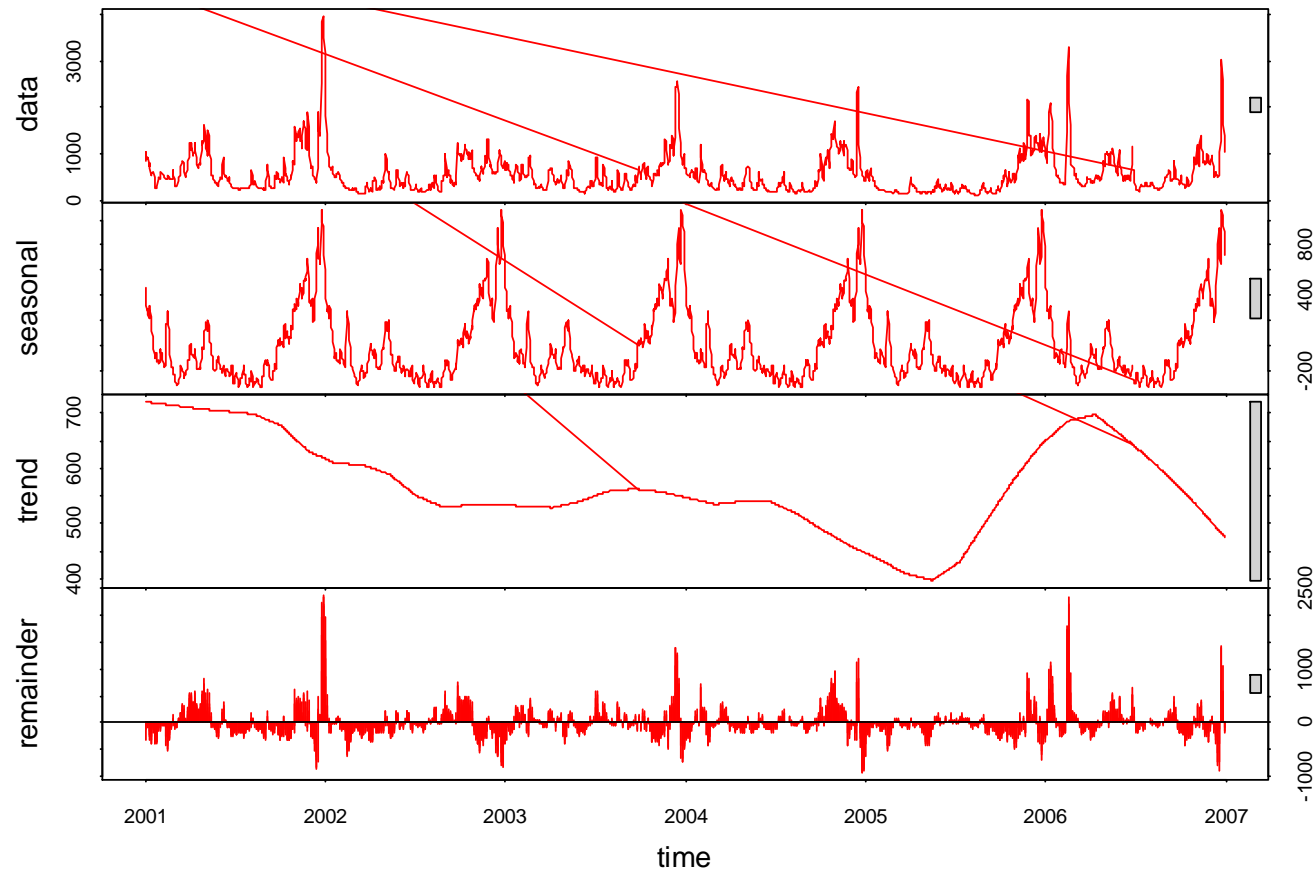
A time series represents the values of a variable at successive points in time:

- At point locations, from station observations
- As area-averages, e.g. average rainfall in a river basin

In hydrology and hydrometeorology, time series are discrete, i.e. values at fixed calendar time intervals:

- Fractions of a day, e.g. 30 minutes, 1 hour, 3 hours, ...
- Fractions of a year, e.g. 1 day, 1 month, 3 months, ...
- Annual series, usually mean value or sum for each year

Daily discharge at Lubok Paku, Malaysia, 2001-2006



STL decomposition (Cleveland et al., 1990)

Components of time series

1. Periodic variation:
 - daily cycles
 - Seasonal cycles
2. Trend:
 - Secular trend, systematic increase/decrease over long time periods
 - Cyclic trends, irregular variations, e.g. sequences of wet and dry years
3. Episodic variations due to extreme weather, usually small in number
4. Random fluctuations, often dominant source of variation in hydrology

Statistical description

- Mean – measure of location
- Standard deviation – measure of variation about the mean
- Skewness – measure of symmetry about the mean
- Standard error (i.e. the standard deviation of the mean):

$$SE = \frac{SD}{\sqrt{N}}$$

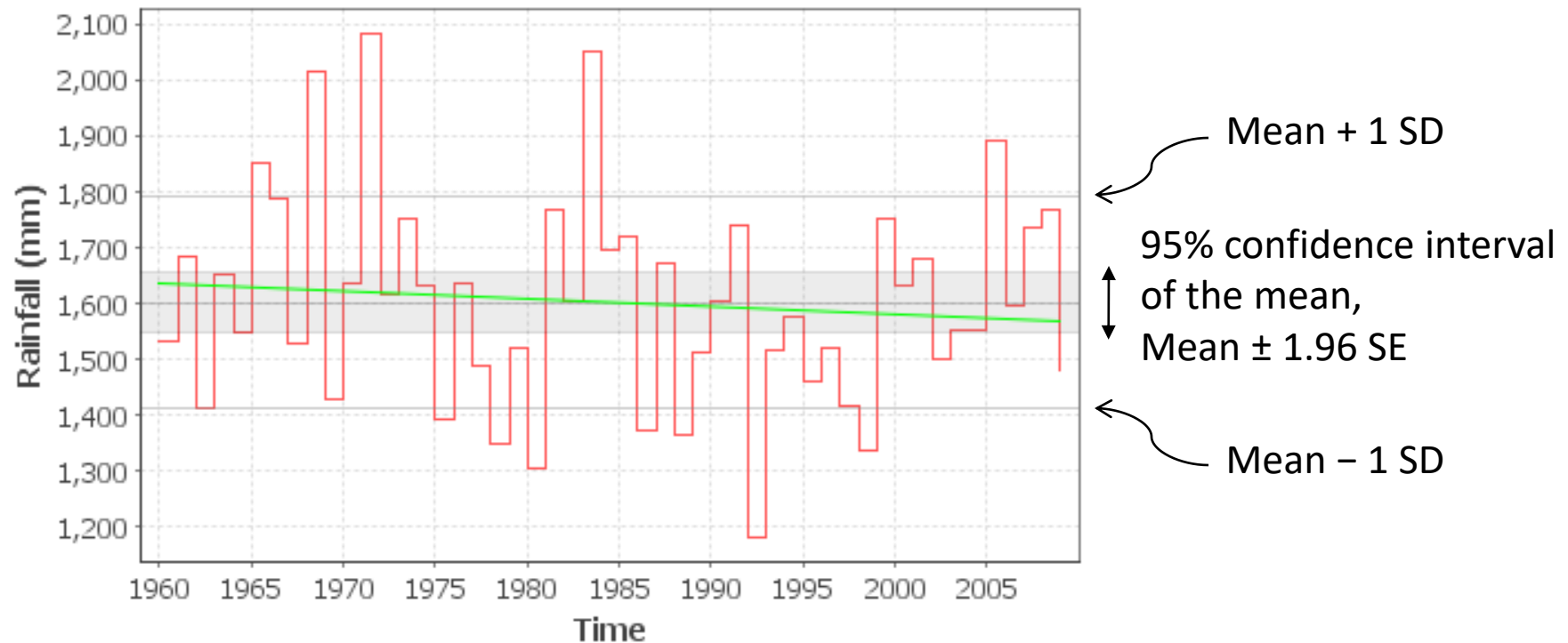
- Coefficient of variation:

$$CV = \frac{SD}{\text{Mean}}$$

See also: DID Manual Vol 4, Ch 5 (2009); Helsel and Hirsch (2002)

Example: 50 y annual rainfall, Lancang station, China

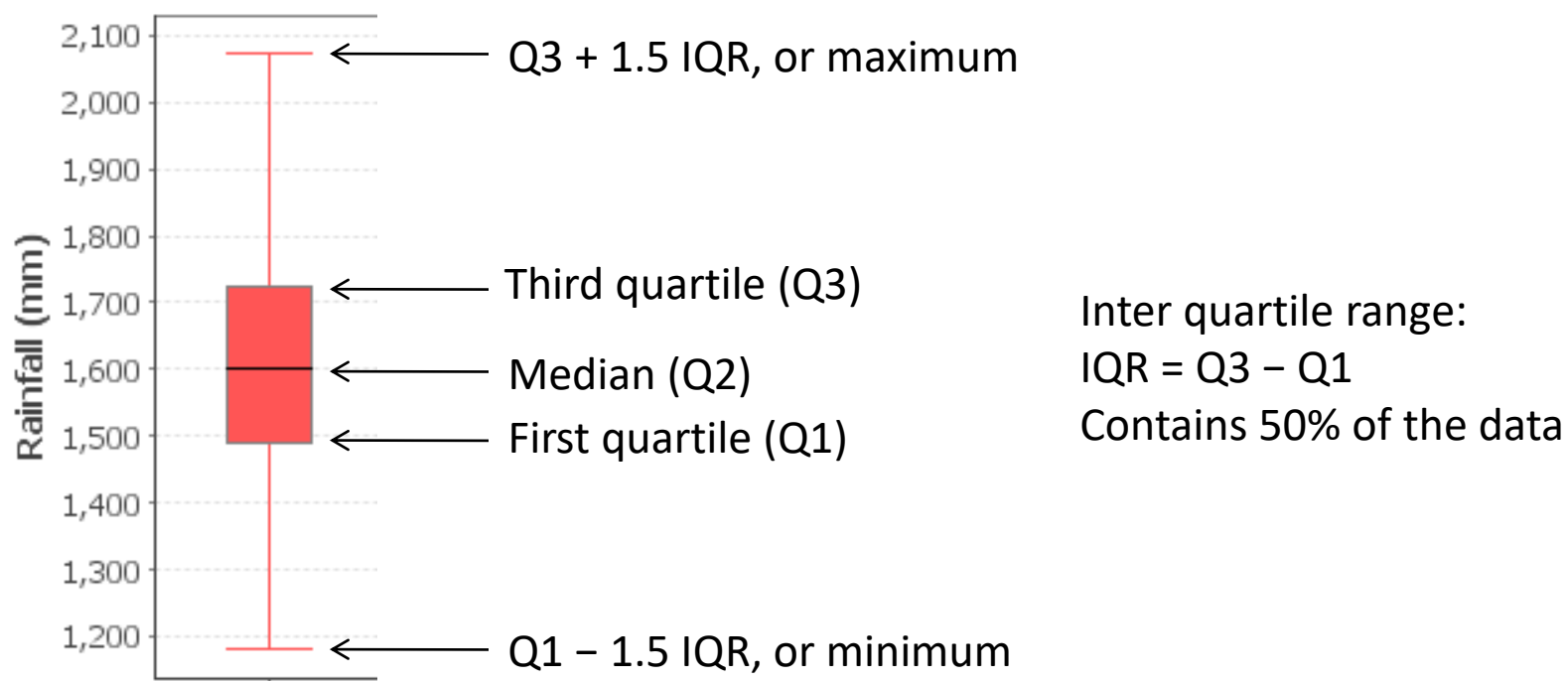
Mean	SD	Skew	SE	CV
1603	187.8	0.4344	26.56	0.1171



The green line is the apparent trend – needs to be tested if statistically significant

Tukey's five-number summary and box plot (Lancang rainfall)

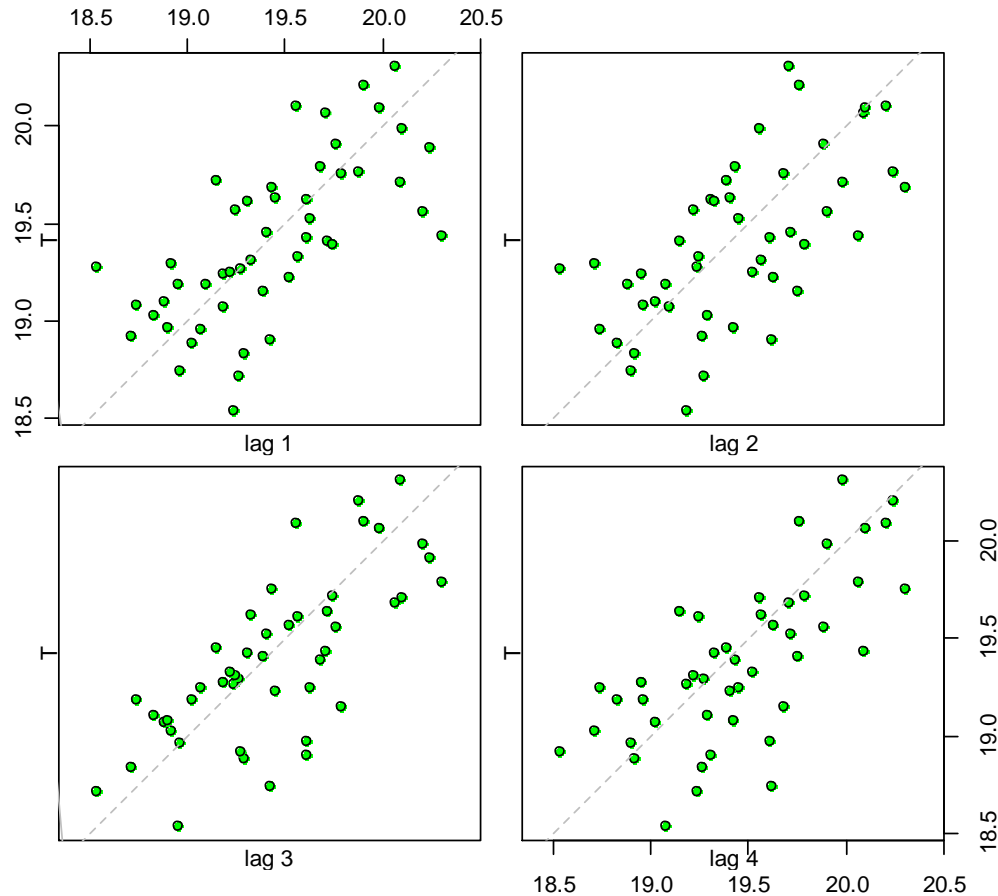
Min	Q1	Median	Q3	Max
1183	1490	1600	1723	2085



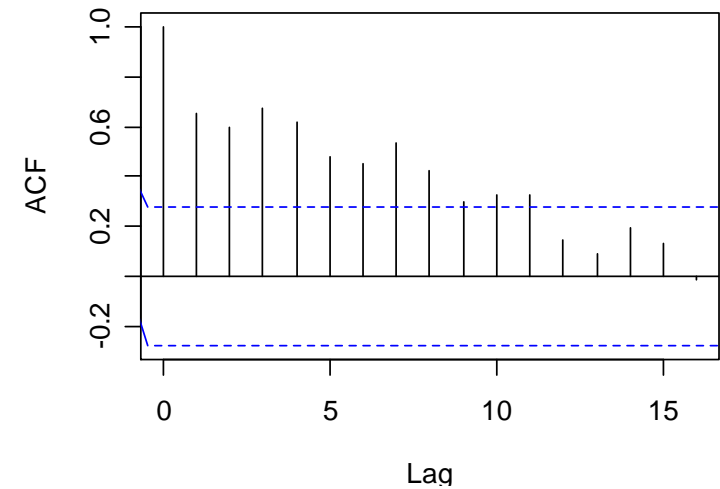
Box plots are useful to visualize and compare distributions. See Helsel and Hirsch (2002) for further details.

Autocorrelation

Lancang mean temperature



Correlogram



- Autocorrelation function (ACF) is the Pearson R for each lag
- The lag 1 R is a measure of persistence

Statistical testing

Statistical tests aim to accept or reject a null-hypothesis, with a specified level of significance, α

A null hypothesis for a specific test can be:

H_0 : There is no trend at significance level α

which leads to the alternative hypothesis:

H_1 : There is a trend at significance level α

A common value for α is 0.05, sometimes 0.1 or 0.01 is also used.

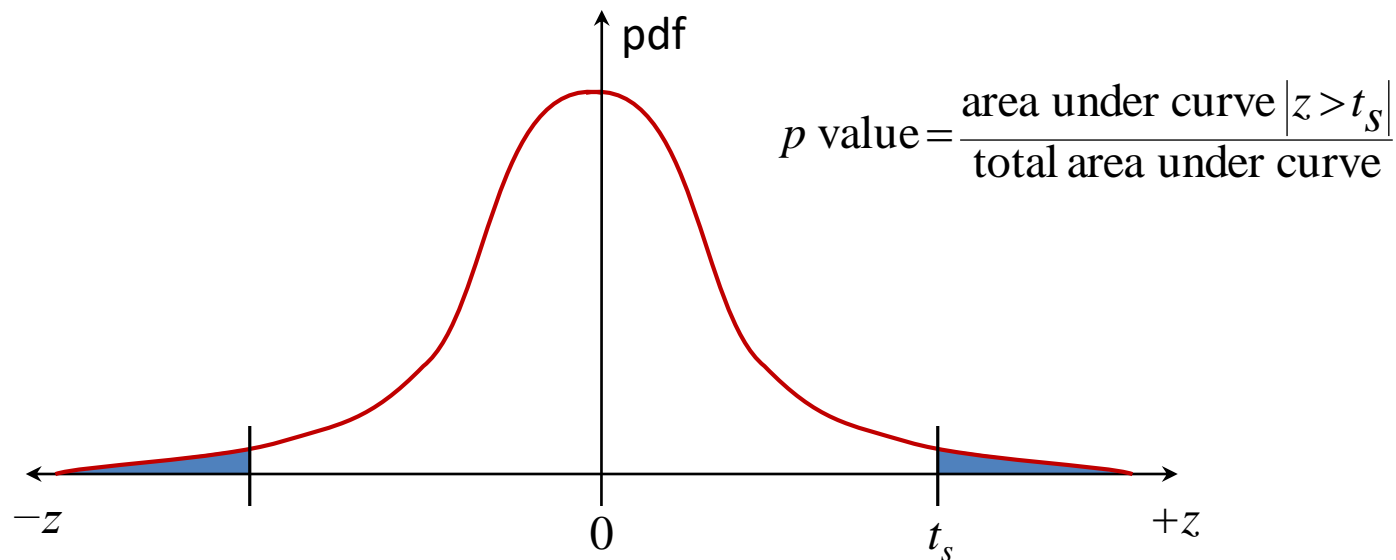
More specific, α is the probability of rejecting H_0 while it was actually true. This is called a Type 1 error.

Test computations (by software)

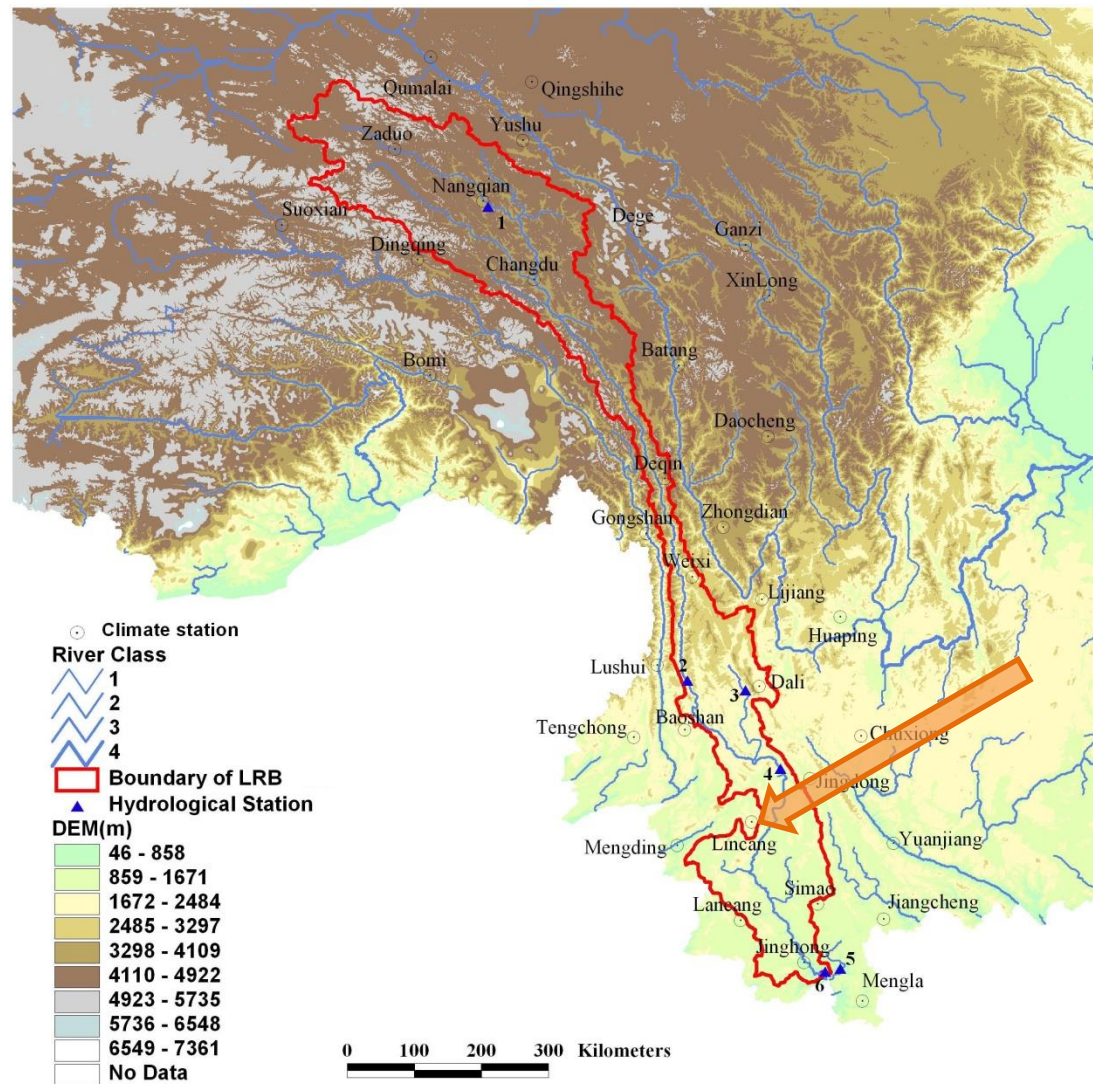
For a given test, we compute the test statistic from the data, say t_s , which follows a certain statistical distribution (e.g. Normal, Student-t)

Then we compute the p value from the probability density function (pdf) of the distribution for test statistic

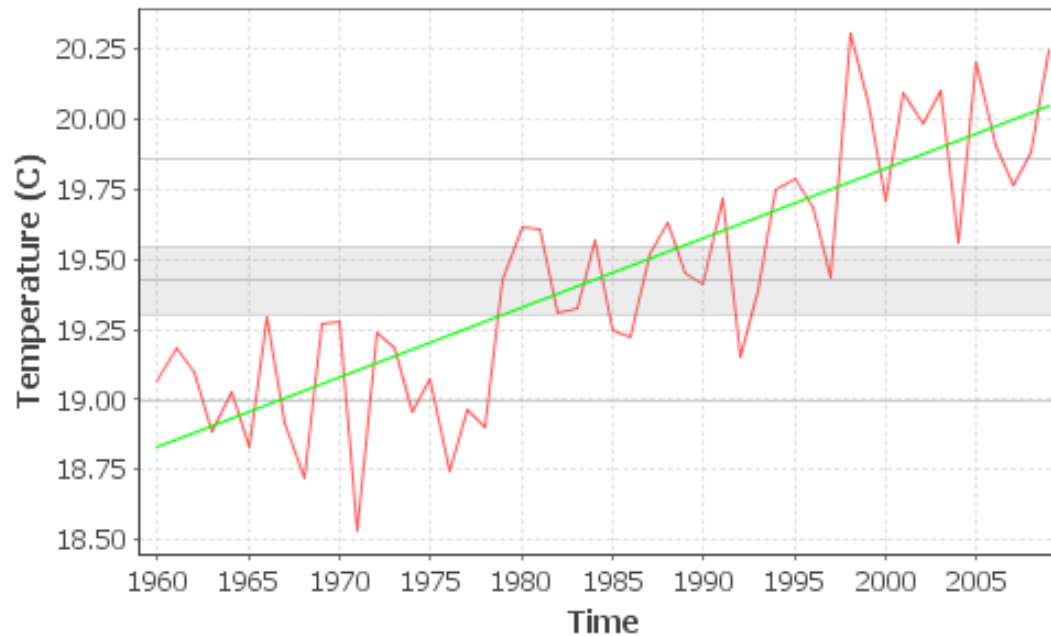
The null hypothesis H_0 is rejected if $p \leq \alpha$



Example: temperature at Lancang station, China

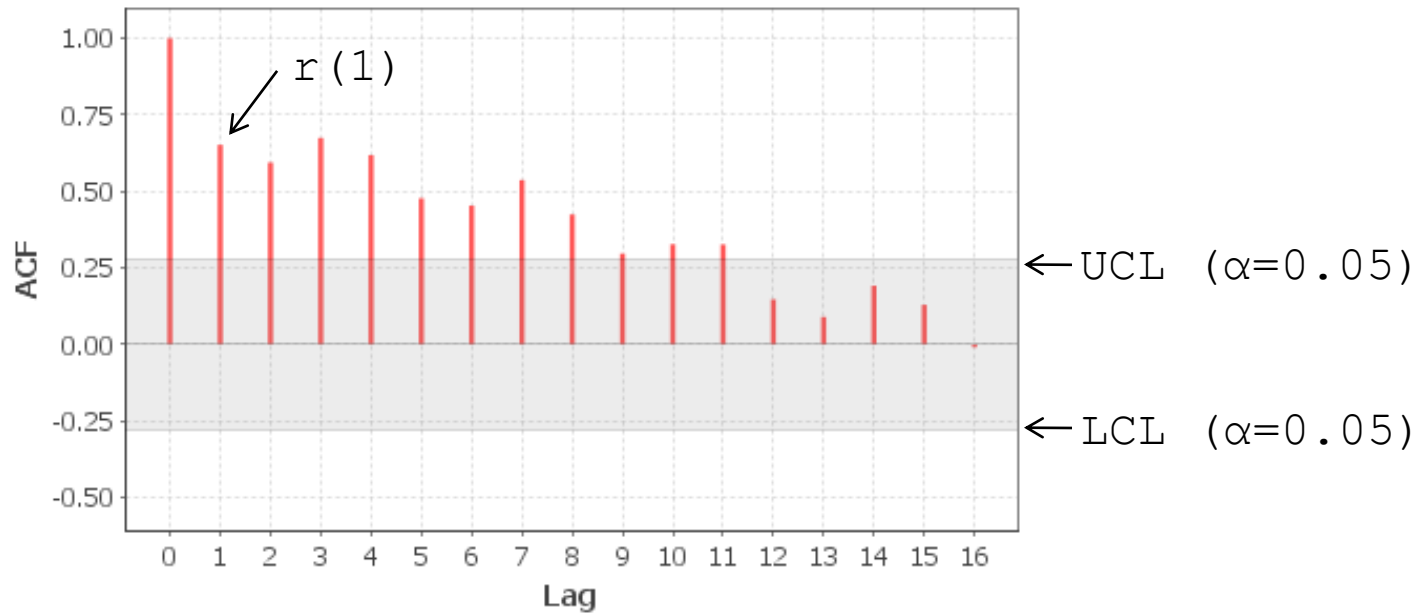


Step 1: Plot the data and obtain some statistics:



Mean	SD	Skew	SE	CV
19.42	0.4322	0.1724	0.06112	0.02225
Min	Q1	Median	Q3	Max
18.53	19.10	19.40	19.72	20.31

Step 2: test for absence of persistence (no significant lag 1 autocorrelation)



$$r(1) = 0.652$$

Alpha	0.10	0.05	0.02	0.01
UCL	0.233	0.277	0.329	0.364
LCL	-0.233	-0.277	-0.329	-0.364

Since $r(1)$ is outside the range $[LCL; UCL]$, we conclude that there is persistence

Intermezzo: Pre-whitening

If significant persistence is present, we have to pre-whiten the data when testing for trend

Pre-whitening removes lag 1 autocorrelation from the series by:

$$y'_t = \begin{cases} (1-r_1)y_t & \text{for } t=1 \\ y_t - r_1 y_{t-1} & \text{for } t > 1 \end{cases}$$

Then, testing persistence for the pre-whitened series:

$$r(1) = -0.153$$

Alpha	0.10	0.05	0.02	0.01
UCL	0.233	0.277	0.329	0.364
LCL	-0.233	-0.277	-0.329	-0.364

The $r(1)$ of the pre-whitened series is inside the range [LCL;UCL]

Step 3: tests for absence of trend should be *non-parametric*. That is, no underlying statistical distribution of the data is assumed. There are two tests:

1. Test based on Spearman's rank-correlation coefficient, ρ (rho)
2. Test based on Kendall's correlation coefficient, τ (tau)

The latter, the Mann-Kendall trend test is common in climatology

The test result for our (pre-whitened) temperature data is:

Mann-Kendall trend test (pre-white) :

Series Start = 1960 End = 2009 Length = 50

Null hypothesis: There is no trend in the series

S = 353

tau = 0.2882

p = 0.002875 ***

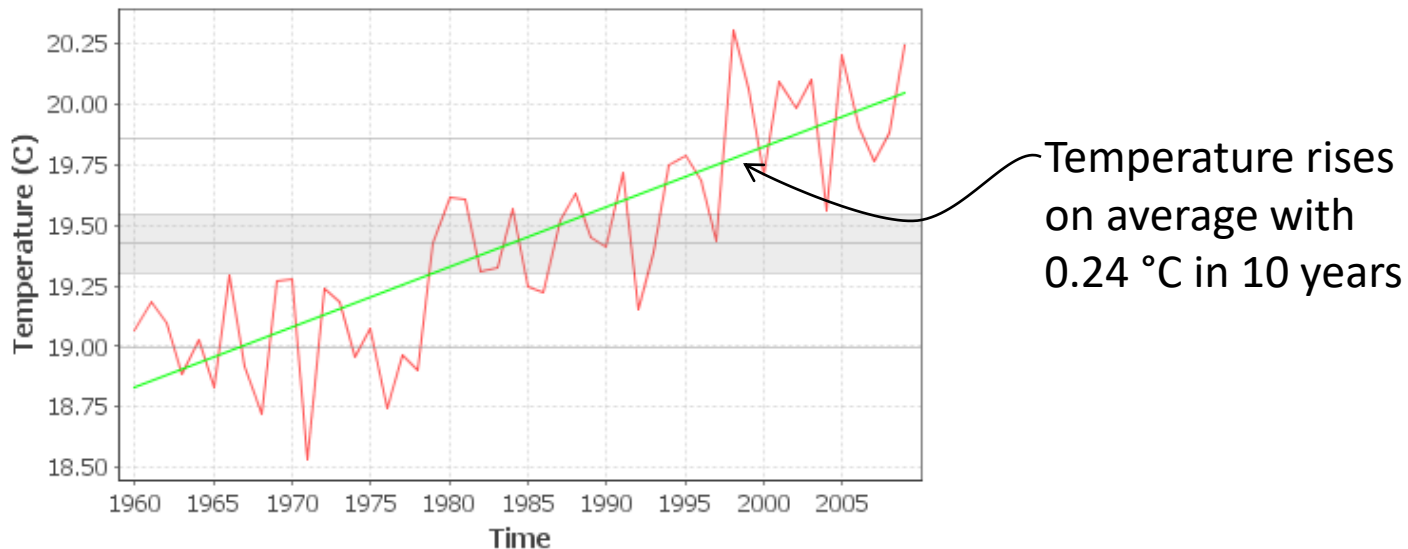
Since $p < 0.05$ (< 0.01), H_0 is rejected \rightarrow Temperature increased significantly

Step 4: determine the warming rate

If a significant trend is present, the average rate of increase of decrease can be obtained from the slope of a simple linear regression:

	Estimate	SE	t-stat	Pr(> t)	
Const	18.83	0.06900	272.9	<0.0001	***
Slope	0.02442	0.002427	10.06	<0.0001	***

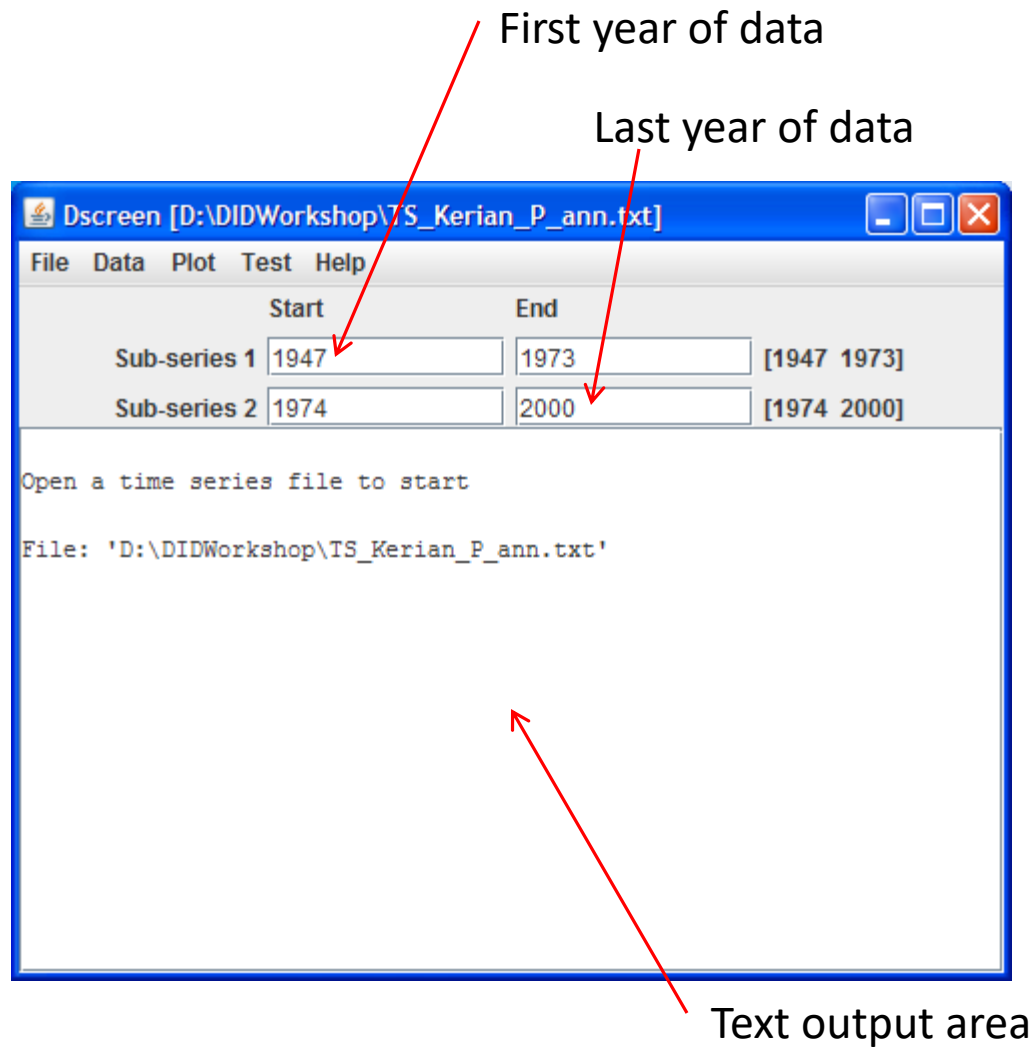
Residual SE: 0.2476 Regression DF: 48
R-squared: 0.6784 Adj R-squared: 0.6717



Using DScreen software: Sungai Kerian, Malaysia

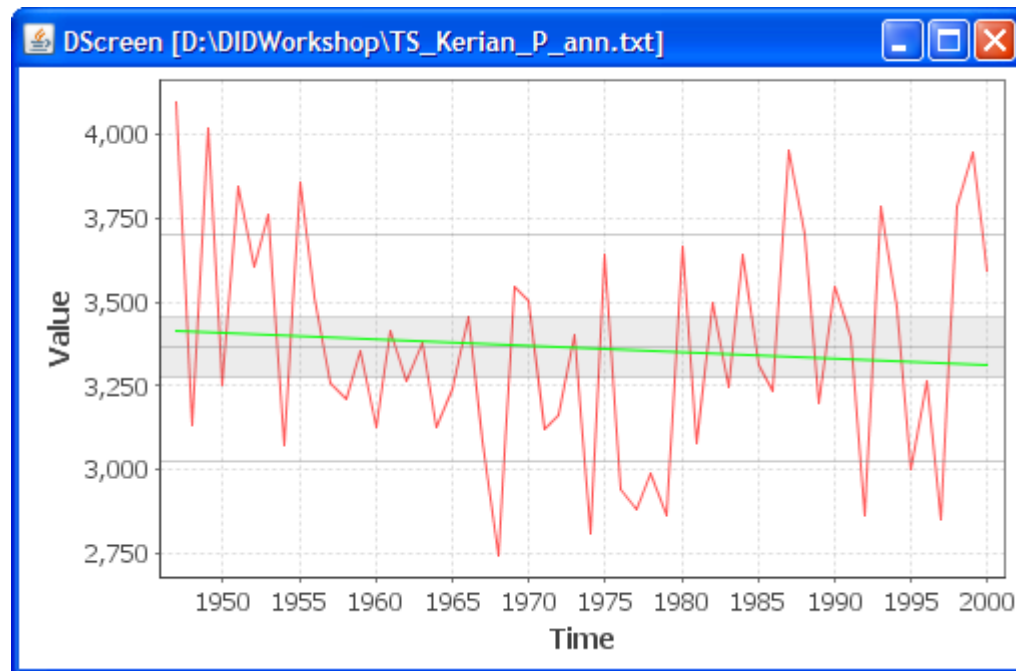
Data: annual rainfall (mm) in the Sg. Kerian catchment for 1947-2000
(DID Manual, Vol. 4, 2009)

1. In Windows explorer, go to the **DIDWorkshop** folder
2. Double click on the file **DScreen.jar**
3. Select **File – Open** from the menu bar
4. Choose the data file **TS_Kerian_P_ann.txt**



Step 1: plot the time series

Select **Plot – Series – Continuous** from the menu

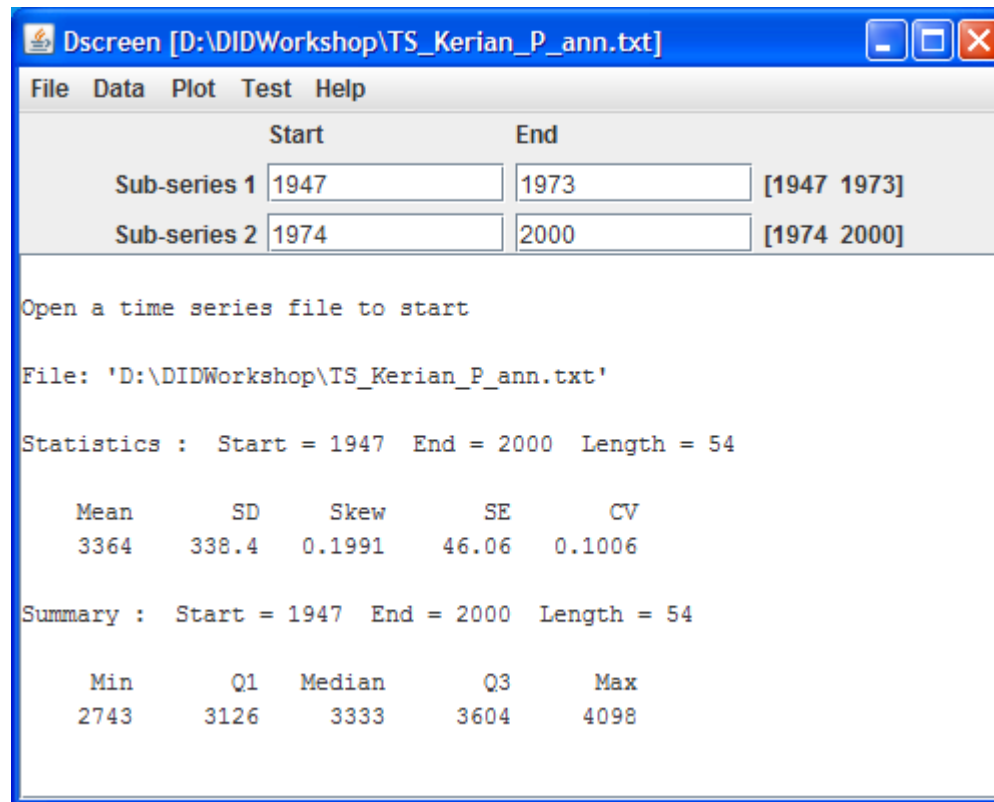


(Plot windows stick on the screen until you close them
You can move them to a convenient place
Right-click in plot window opens a menu with options)

Step 1a: Describe the data

Select **Data – Statistics**

Select **Data – Summary**



The screenshot shows the Dscreen application window with the title bar 'Dscreen [D:\DIDWorkshop\TS_Kerian_P_ann.txt]'. The menu bar includes 'File', 'Data', 'Plot', 'Test', and 'Help'. Below the menu bar is a table with two rows for 'Sub-series 1' and 'Sub-series 2', each with 'Start' and 'End' date fields and a resulting date range. The main text area displays the following information:

Open a time series file to start

File: 'D:\DIDWorkshop\TS_Kerian_P_ann.txt'

Statistics : Start = 1947 End = 2000 Length = 54

Mean	SD	Skew	SE	CV
3364	338.4	0.1991	46.06	0.1006

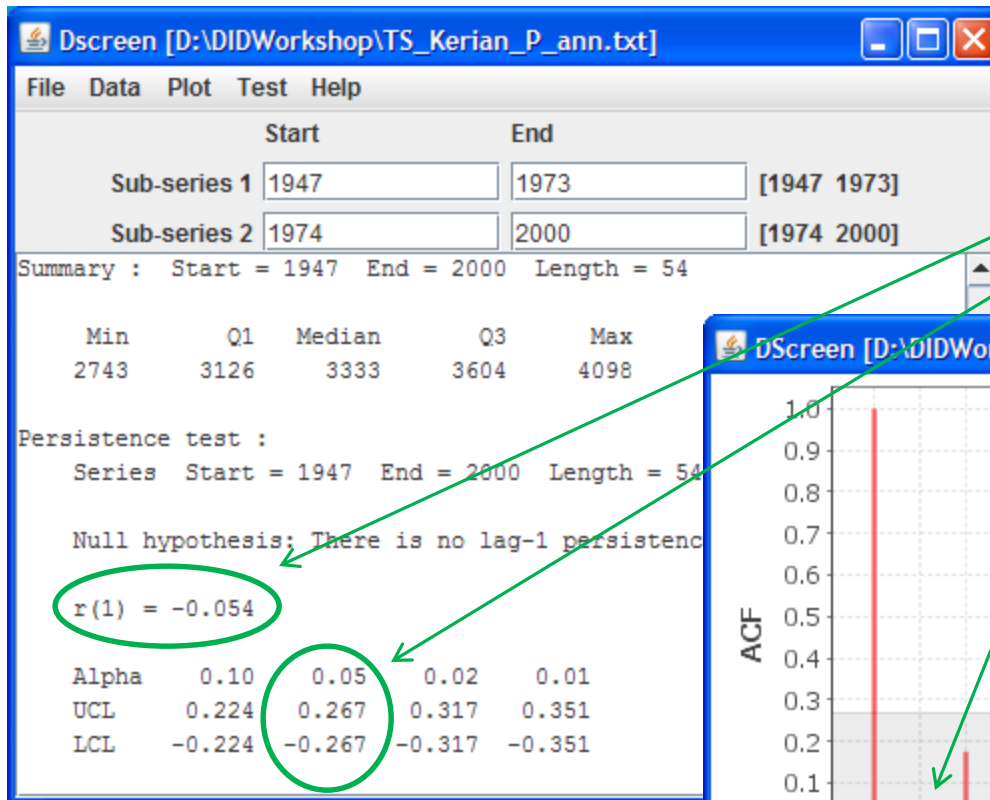
Summary : Start = 1947 End = 2000 Length = 54

Min	Q1	Median	Q3	Max
2743	3126	3333	3604	4098

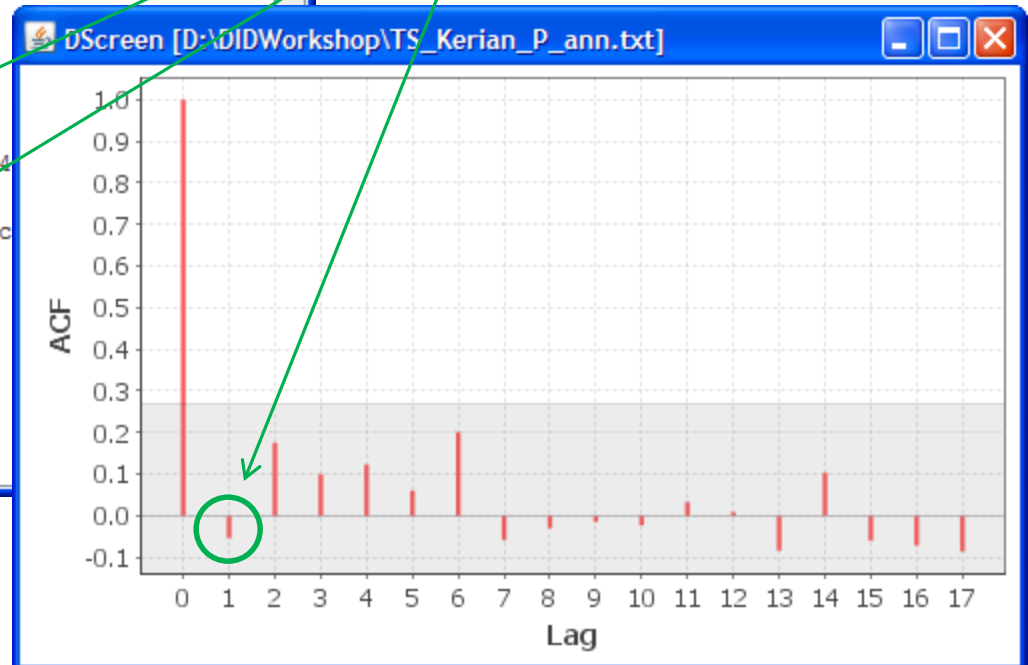
Step 2: Check autocorrelation / test persistence

Select **Plot – Correlogram**

Select **Test – Autocorrelation – Persistence** (no pre-white)



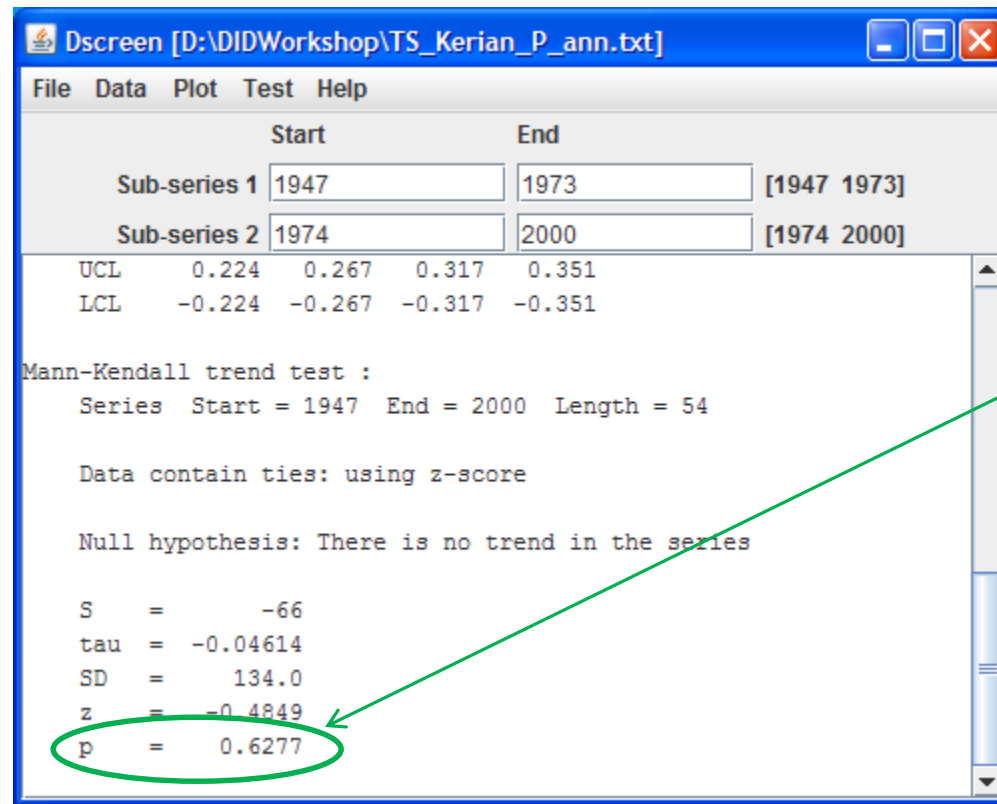
Verify absence of persistence



Step 3: Test for trend

Since there is no persistence, pre-whitening is not required

Select **Test – Secular trend – Mann-Kendall**



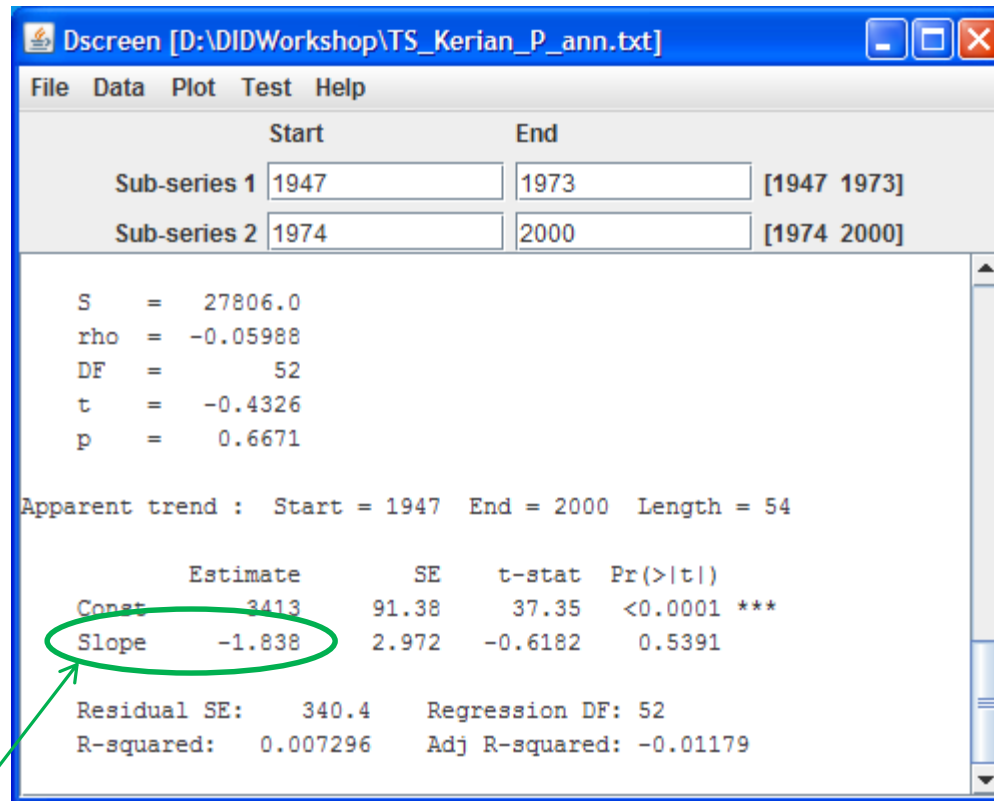
There is a significant trend if we can reject the null hypothesis

Can we?

Note: we can cross-verify using the Spearman test

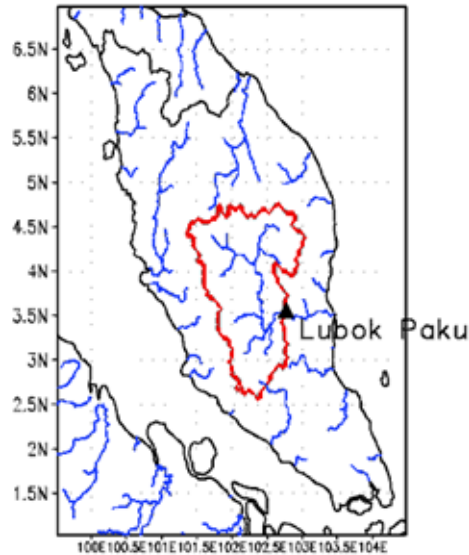
How much is the rainfall amount changing?

Select Data – Apparent trend



Decrease in rainfall of 18 mm per 10 years, too small to be significant

Pahang Basin



Application to Pahang, Malaysia

We carry out a trend analysis for each of the four data sets below

Summarize the outcome

Whole basin at Lubok Paku (25,600 km²):

- TS_Pahang_T_ann.txt: annual temperature (°C, 31 y)
- TS_Pahang_P_ann.txt: annual rainfall (mm, 31 y)
- TS_Pahang_Q_ann.txt: annual discharge (m³/s, 23 y)

Pejabat JPS station (Sg. Kuantan):

- TS_Pejabat_P_30mi-max.txt: annual maximum 30 min rainfall (mm, 39 y)

Source: DID data base; Wong et al., Hydrol. Proc. (2010); DID Manual Vol 4 (2009)