Training Course

Climate Change and its Impact on Water Resources

[May 17 - 21, 2010]

LECTURE - 2

TREND ANALYSIS OF HYDRO-METEROLOGICAL DATA

By

VIJAY KUMAR

Organised by

National Institute of Hydrology
Roorkee- 247 667
&
Indian Institute of Technology
Roorkee -247 667

TREND ANALYSIS OF HYDRO-METEROLOGICAL DATA

1. Introduction

Water is indispensable to sustain life on Earth. It is needed in large volumes virtually in any human activity. Therefore, adverse changes in availability of water, in both quantity and quality contexts (too little, too much, too dirty) are of considerable concern. Changes in climatic variables, such as temperature and precipitation, have significant impacts on water resources and hence on societies and ecosystems.

The IPCC observed that global average air temperature near earth's surface rose to 0.74±0.18°C in the last century. Eleven of the last twelve years rank among the 12 warmest years in the instrumental record of global surface temperatures since 1850. Globally, average sea level rose at an average rate of 1.8 mm per year over 1961 to 2003. The rate was higher over 1993 to 2003, about 3.1 mm per year (*IPCC, 2007*). The projected sea level rise by the end of this century is likely to be 0.18 to 0.59 metres. IPCC predicts global temperatures will rise by 2-4.5°C by the end of this century and for the next two decades a warming of about 0.2°C per decade is projected. Even if the concentrations of all greenhouse gases and aerosols had been kept constant at year 2000 levels, a further warming of about 0.1°C per decade would be expected. This unprecedented increase is expected to have severe impact on global hydrological systems, ecosystems, sea level, crop production and related processes. The impact would be particularly severe in the tropical areas, which mainly consist of developing countries, including India.

Over recent decades the scientific community has amassed a wealth of observational data. The last one hundred years have been a period of rapid climate change, partly in response to human influences. Social, economic, industrial, and land use developments all contribute to human impact on our climate, locally, nationally and globally. The changes already observed have had, and continue to have, impacts on many aspects of society, including health, agriculture, water resources and energy demand. In order to make appropriate plans for the future it is vital to investigate observed changes in climate. In doing so, models of past and present climate can be validated and scenarios of future climate put into the context of any change already recorded. In order to plan for adaptation to climate change there is a need to know the degree of change already experienced in specific locations throughout the seasons. Having sufficient information about climatic change in the recent past is necessary to improve the certainty and accuracy of estimates about the future, and the role of this information is particularly important in assessments of regional climate change. Identification of long-term trends in climate change is providing information for decision-makers and resource managers that allows them to better anticipate and plan for the potential impacts of climate variability and change.

A temporal trend is the general increase or decrease in observed values over time. The trend describes the long smooth movement of the variable lasting over the span of observations, ignoring the short term fluctuations. The purpose of a trend test is to determine whether the values of a series generally increase or decrease. Trend analysis is used to determine the significance of a trend and to estimate the magnitude of that trend. Trend analysis determines whether the measured values of a variable increase or decrease during a time period.

Detection of trends in long time series of hydrological data is of paramount scientific and practical significance. Water resources systems have been designed and operated based on the assumption of stationary hydrology. If this assumption is incorrect then existing procedures for designing levees, dams, reservoirs, etc. will have to be revised. Without revision there is a danger that systems are over or under designed and either do not serve their purpose adequately or are overly costly. Studies of change are also of importance because of our need to understand the impact that man is having on the "natural" world. Urbanisation, deforestation, emissions of greenhouse gases, changes in agricultural practice and dam construction are just a few examples of anthropogenic activities that may be altering important aspects of the hydrological cycle. The principal water-related problems have been always related to having too much water (floods) or too little water (low flows or droughts). This means that studying changes in characteristics of hydrological extremes is of major importance.

There are many parametric and non-parametric methods that have been applied for detection of trends. Parametric testing procedures are widely used in classical statistics. In parametric testing, it is necessary to assume an underlying distribution for the data (often the normal distribution), and to make assumptions that data observations are independent of one another. For many hydrological series, these assumptions are not appropriate. Firstly hydrological series rarely have a normal distribution. Secondly, there is often temporal dependence in hydrological series, particularly if the time series interval is short (e.g. today's flow tells us quite a bit about what tomorrow's flow is likely to be). If parametric techniques are to be used, it may be necessary to (a) transform data so that its distribution is nearly normal and (b) restrict analyses to annual series, for which independence assumptions are acceptable, rather than using the more detailed monthly, daily or hourly flow series. In non-parametric and distribution-free methods, fewer assumptions about the data need to be made. With such methods it is not necessary to assume a distribution. However, many of these methods still rely on assumptions of independence.

2. The Process of Analysis

The main stages of an analysis are

- Obtaining and preparing a suitable dataset
- Exploratory analysis of the data
- Application of statistical tests
- Interpretation of the results

In many studies, a specific dataset is the focus of study and the question is whether the data shows any evidence of trend or other change. In other cases, one has a particular question in mind and is seeking the right data to best answer the question. Even when there is a specific dataset of interest, it is still important to consider other available sources of data. For example, when investigating change in flow series it is often helpful to obtain rainfall data too. Obtaining a suitable dataset sounds straightforward but, in practice, it can require care and skill. There are many important aspects that may need to be considered when obtaining and preparing data. These include

- Data should be quality controlled before commencing an analysis of change.
- Data series should be as long as possible. For investigation of climate change, a minimum of 50 years of record is suggested - even this may be not be sufficient.
- Missing values and gaps in a data series make analysis harder and raise questions of data quality. It is important to consider whether gaps are truly random, or whether they are perhaps associated with major flooding making the remaining data unrepresentative.
- Very frequent data contains more information but can also be harder to analyse both computationally and because more restrictive assumptions must be made.

Exploratory data analysis (EDA) is a very powerful graphical technique that is a key component of any data analysis. Exploratory data analysis is itself an iterative process and it should be used at more than one stage of an analysis. Its first use is to examine the raw data. This may identify further interesting aspects of the data, such as seasonality, which in turn invite further investigation. Exploratory data analysis also has an important role in helping to check out test assumptions. For example, having fitted a trend to the data, exploratory data analysis can be used to examine the residuals to check for independence. In some cases, this may mean that the model needs to be altered then revisited using EDA. Finally EDA can provide a very valuable means of presenting both the data and the results in a way that maximises understanding and impact.

Exploratory data analysis allows a much greater appreciation of the features in data than tables of summary statistics and statistical significance levels. This is because the human brain and visual system is very powerful at identifying and interpreting patterns. It is often able to see important features, structures or anomalies in a data series that would be very difficult to detect in any other way. Just looking at the data can change initial preconceptions, can alter the questions that it is sensible to ask, and can uncover important aspects that would never otherwise have been found. Exploratory data analysis involves:

- plotting graphs
- studying the graphs
- re-plotting graphs to improve the display of important features
- identifying further graphs that are needed
- iterating through the above.

3. Some Basics of Statistical Testing

Hypotheses

The starting point for a statistical test is to define the null and alternative hypotheses; these are statements that describe what the test is investigating. For example, to test for trend in the mean of a series the null hypothesis (H_0) would be that there is no change in the mean of a series, and the alternative hypothesis (H_1) would be that the mean is either increasing or decreasing over time. To test for step-change in the mean of a series, the null hypothesis would again be that there is no change in the mean of the series, but the alternative hypothesis would be that the mean of the series has suddenly changed. The starting point for statistical testing is to assume that the null hypothesis is true, and then to check whether the observed data are consistent with this hypothesis. The null hypothesis is rejected if the data are not consistent.

Test statistic

To compare between the null and alternative hypotheses, a test statistic is selected and then its significance is evaluated, based on the available evidence. The test statistic is a means of comparing the null and alternative hypotheses. It is just a numerical value that is calculated from the data series that is being tested. A good test statistic is designed so that it highlights the difference between the two hypotheses. A simple example of a test statistic is the linear regression gradient: this can be used to test for a trend in the mean. If there is no trend (the null hypothesis) then the regression gradient should have a value near to zero. If there is a large trend in the mean (the alternative hypothesis) then the value of the regression gradient would be very different from zero. More formally, to carry out a statistical test it is necessary to compare the observed test statistic with the expected distribution of the test statistic under the null hypothesis. The significance level of a test statistic expresses this concept more formally.

Significance level

The significance level is a means of measuring whether a test statistic is very different from values that would typically occur under the null hypothesis. Specifically, the significance level is the probability of a value as extreme as, or more extreme than the observed value, assuming “no change” (the null hypothesis). In other words, significance is the probability that a test detects trend when none is present. Thus a 5% significance level would be interpreted as strong evidence against the null hypothesis— with a 1 in 20 chance of that conclusion being wrong.

A possible interpretation of the significance level might be:

- Significance level >10% - very little evidence against the null hypothesis (*Ho*)
- 5 % to 10 % - possible evidence against H_0
- 1 % to 5 % - strong evidence against H_0
- below 1 % - very strong evidence against H_0.

Note that when reporting results the actual significance levels should normally be quoted (e.g. a significance level of 5 %). For many traditional statistical methods, significance levels can be looked up in reference tables or calculated from simple formulae, providing the required test assumptions apply. No statistical test is perfect, even if all test assumptions are met. A 5% significance level means that we will make an error 5% of the time: i.e. if the null hypothesis was true then 1 in 20 test results will have a significant (and incorrect) result. It is important to remember this when interpreting results.

Power and errors

There are two possible types of error that can occur in a test result. The first is that the null hypothesis is incorrectly rejected (type I error) - the significance level expresses the probability of this error. The second is that the null hypothesis is accepted when the alternative hypothesis is true — type II error. A test which has low type II error probability is said to be powerful. In general more powerful tests are to be preferred. The power of the test is the probability of correctly detecting a trend when one is present.

4. Methods of Trend Detection

Trend testing for hydro-meteorological variables such as precipitation, temperature and streamflow has been of particular interest to hydrologists and researchers for several decades. A comprehensive review of statistical approaches used for trend analysis of water resources time series is provided by Helsel and Hirsch (1992). Recent studies indicate that the most widely used method is the non-parametric Mann-Kendall trend test.

Regression Analysis

This is a parametric test that assumes normally distributed data. It is used to test for linear trend by the linear relationship between time and the variable of interest. The correct application of this method requires the variables to be normally distributed and temporally and spatially independent. However, this method has been applied to assessing significance of linear trends for a wide range of variables, frequently without discussing the normality of their distributions and/or their temporal and spatial independence. The other main disadvantage of the method is that it can not reject outliers properly. Also, the impact of time-dependent missing data may bias the parametric rainfall trends if assumed to be zero or at the daily average for the month.

The regression analysis can be carried out directly on the time series or on the anomalies (i.e. deviation from mean). Trend in any time series (say rainfall) at a particular station can be examined by applying the regression analysis with time as the independent variable and annual rainfall as the dependent variable. A linear equation, $y = mt + c$, defined by c (the intercept) and trend m (the slope), which represents the rate of increase or decrease of the variable, can be fitted by regression. Intercept and slope of the estimated regression line can be obtained from the sample data with the least squares criterion. Slope of the regression line indicates per year increase or decrease in rainfall. The significance of the estimated trend is tested using the t-test with the null hypothesis (H_0) and alternative hypothesis (H_1) as followings:

H_0: m =0

H_1: m $\neq$ 0

t-calculated is obtained as

$$t = \left| \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \right| \tag{1}$$

H_0 is rejected, if t-calculated is less than -tn, $\alpha/2$ or more than tn, $\alpha/2$. Here, tn,$\alpha/2$ is the tabulated value of t-statistics with n-degrees of freedom and α-level of significance of a two-tailed distribution. Alternatively, if the calculated t-statistics lies within the range of the tabulated value, then H_0 is accepted and no trend is considered.

Mann-Kendall Test

One of the widely used non-parametric tests for detecting a trend in hydro-climatic time series is the Mann–Kendall (MK) test. Kendall (1938) proposed a measure *tau* to measure the strength of the monotonic relationship between *x* and *y*. Mann (1945) suggested using the test for significance of Kendall's *tau*, where one of the variables is time as a test for trend. The test is well known as Mann-Kendall's test, which is powerful for uncovering deterministic trends. It accepts or rejects the null hypothesis of randomness against the alternative of a monotonic trend. The advantage of the Mann-Kendall test is that it relies only on a few assumptions: the potential trend may be either linear or nonlinear, and no assumptions are made regarding the underlying statistical distribution. Nevertheless, it is well recognised that the Mann-Kendall test is not robust against autocorrelation in the sense that false positive trend identifications get more likely. With a positively auto-correlated series, there are more chances of a series being detected as having trend while there may be actually none. The case is reverse for negatively auto-correlated series, where trend fails to get detected. This effect depends on the sample size as well as on the magnitude of the trend to be identified. Pre-whitening has been used to detect trend in time series in presence of autocorrelation. Pre-whitening techniques introduced to remove effects induced by autocorrelation may also bias the Mann-Kendall test result.

For pre-whitening, the data series is tested for serial correlation. If the lag-1 auto-correlation (r_1) is found to be non-significant at 95% confidence level, then Mann-Kendall test is applied to the original data series (x_1, x_2,.....,x_n), otherwise, Mann-Kendall test was applied on 'pre-whitened' series obtained as (x_2-r_1x_1, x_3-r_1x_2,.....,x_n-r_1x_{n-1}).

The MK statistics (S) is defined as (Salas, 1993)

$$S = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \operatorname{sgn}(x_j - x_i) \tag{2}$$

where N is number of data points. Assuming (x_j-x_i) = θ, the value of *sgn(θ)* is computed as follows:

$$\operatorname{sgn}(\theta) = \begin{cases} 1 & if \quad \theta > 1 \\ 0 & if \quad \theta = 1 \\ -1 & if \quad \theta < 1 \end{cases} \tag{3}$$

This statistics represents the number of positive differences minus the number of negative differences for all the differences considered. For large samples (N>10), the test is conducted using a normal distribution (Helsel and Hirsch, 1992) with the mean and the variance as follows:

$$E[S] = 0 \tag{4}$$

$$Var(S) = \frac{N(N-1)(2N+5) - \sum_{k=1}^{n} t_k(t_k - 1)(2t_k + 5)}{18} \tag{5}$$

where n is the number of tied (zero difference between compared values) groups, and t_k is the number of data points in the *k*th tied group. The standard normal deviate (Z-statistics) is then computed as (Hirsch *et al.*, 1993):

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{Var(S)}} & if \quad S > 0 \\ 0 & if \quad S = 0 \\ \dfrac{S+1}{\sqrt{Var(S)}} & if \quad S < 0 \end{cases} \qquad (6)$$

The value of Z is computed and if the value lies within the limits ±1.96, the null hypothesis of having no trend in the series cannot be rejected at 95% level of confidence.

Sen's Estimator

The magnitude of trend in a time series can be determined using non-parametric method known as Sen's estimator (Sen, 1968). This method assumes a linear trend in the time series. In this method, the slopes (T_i) of all data pairs are calculated first by

$$T_i = \frac{x_j - x_k}{j-k} \qquad \text{for } i = 1,2,\ldots\ldots..,N \qquad (7)$$

where x_j and x_k are data values at time j and k (j>k) respectively. The median of these N values of T_i gives the Sen's estimator of slope (β). A positive value of β indicates an upward trend and a negative value indicates a downward trend in the time series.

5. Case Study

A case study on trend analysis of rainfall data of Kashmir region is enclosed herewith.

References

Helsel, D.R., and Hirsch, R.M., (1992). Statistical Methods in Water Resources, Elsevier Science Publishing, New York, pp. 522.

Kendall M.G. (1938). A new measure of rank correlation. *Biometrika*, 30, 81-93.

Kendall, M.G., (1975). Rank Correlation Measures, Charles Griffin, London.

Kundzewicz, Z. W. & Robson, A. (eds) (2000) *Detecting Trend and Other Changes in Hydrological Data.* World Climate Programme—Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD no. 1013. World Meteorological Organization, Geneva, Switzerland.

Kundzewicz, Z. W. & Robson, A. (2004). Change detection in hydrological records - a review of the methodology. *Hydrological Sciences Journal,* 49(1), pp. 7-14.

Mann, H.B., (1945) Non-parametric tests against trend, Econometrica 13, MathSciNet, pp. 245-259.

Salas J.D., (1992). Analysis and modeling of hydrologic time series. In *Handbook of Hydrology,* Maidment DR (ed). McGraw-Hill, New York; 19.1-19.72.

Sen, P.K. (1968). "Estimates of the regression coefficient based on Kendall's tau". *Journal of the American Statistical Association,* 63:1379-1389.