Climate Variability and Change—Hydrological Impacts (Proceedings of the Fifth FRIEND World Conference held at Havana, Cuba, November 2006), IAHS Publ. 308, 2006.

Methodologies for trend detection

ZBIGNIEW W. KUNDZEWICZ[1,2] **& MACIEJ RADZIEJEWSKI**[1,3]

1 *Research Centre for Agricultural and Forest Environment, Polish Academy of Sciences, Poznań, Poland*
zkundze@man.poznan.pl; zbyszek@pik-potsdam.de

2 *Potsdam Institute for Climate Impact Research, D-14412 Potsdam, Germany*

3 *Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland*

Abstract Methodologies for detection of change in river flow data are briefly reviewed. After discussion of the issue of data for change detection, the basics of statistical testing for trend detection are presented, including such concepts as the null hypothesis, the test statistic, the significance level and the trend index. Test assumptions are thoroughly discussed. A review of practical tests (parametric, distribution-free i.e. non-parametric, and resampling: permutation and bootstrap approaches) applicable for different properties of data, is presented. The case where the records are not independent and are not normally distributed is also included. A few special problems, such as extremes and spatial data (with correlation between gauges), and detectability of change are also discussed. Assessment of a few relevant recent publications is offered.

Key words change detection; resampling; river flow data; statistical testing; trend

INTRODUCTION

Detection of changes in long time series of river flow records is an important scientific issue and of considerable practical interest since rivers are the most easily available source of water upon which, worldwide, billions of people depend. Water resources systems have been designed and operated based on the assumption of stationary conditions, i.e. that the past is a key to the future. If this assumption is abandoned, i.e. if the process of river flow is subject to changes, then existing procedures for designing water supply, flood and drought control, environmental flows, hydropower, navigation, etc., have to be revised. In order to design reservoirs, dams, dykes, etc., scheduled to serve man for decades, one has to consider changes in the river flow process. Otherwise systems may turn out to be over- or under-designed, i.e. to not serve their purpose adequately, or to be overly costly. However, for a number of reasons discussed in the following, change detection in river flow data is not a trivial task. The search for weak changes in time series of hydrological data, which are subject to strong natural variability, requires use of both data and methodology of adequate quality.

There are many possible causes for change in time series of hydrological data. Some are directly man-caused (e.g. urbanization, deforestation) and others are of natural, e.g. climatic origin. Yet, even climate change-related signatures can be seen as a manifestation of an indirect human influence via the rising concentration of greenhouse gases in the atmosphere. Often it is not easy to detect a weak, if any, climate signature hidden amidst a strong natural variability of river flow. An essential problem also is: how to distinguish a greenhouse component and a land-use change signature.

The concept of change builds upon the assumption that some kind of constancy or repeatability is possible in the process and change is a negation of such constancy. For example, one may compare the river flow process in the past and the river flow process at present as described by the statistical properties of the process in two different time periods. One can state that the river flow has changed between these two periods if the distribution of the flow rate values in the two periods is no longer the same (nor sufficiently similar).

A trend is a continued change that occurs over time. One may either view it as a manifestation of a separate time-dependent deterministic component (this requires some previous knowledge of the underlying mechanism) or simply as a continuous tendency in the statistical properties of the process.

Usually trends of simple shape (linear, low-order polynomial, piecewise linear, exponential, etc.) are considered. Different trend shapes are possible, including steeper trends similar to abrupt step-like changes, so that there is a continuum of cases and, in practice, the terms “trend” and “change” are almost interchangeable. One can also speak of trends in a non-parametric, comparative sense; e.g. an increasing trend means that the values that occur later are usually higher

than those that occur earlier. Such a trend need not have a specific shape, but may still be called strong or weak, e.g. if the probability that a later value is higher than the earlier one is close to 1 or to 0.5, respectively.

There may be short-term or long-term trends in the past observation record. Neither implies that the observed change is going to last for long. However, the above assumptions make it reasonable to expect that a long-term trend will continue at least for a short period into the future.

The task of trend detection refers to a situation with no previous knowledge of a trend component. There are two different approaches to trend detection:

– Assume a specific model with a trend component and identify the possible trend by looking at the change in statistical properties;
– Study the change in statistical properties alone.

The model in the former bullet is usually linked to some possible cause of change, while the change in statistical properties may be interpreted as the effect. Even if only the latter changes are analysed, one usually has some cause in mind or looks for possible explanations of change. In this, attempting to carry out a trend or change detection helps identify the cause–effect relationships and lets one learn about the consequences of external factors or human actions. In turn, one gains some understanding of the system, either in the form of a simplified description (e.g. finding out that it is getting wetter), or a conceptual model of attribution (e.g. increased precipitation may be a climate change effect, like warming and changes in seasonal precipitation distribution). Finally one gets better knowledge of what is going to happen and better knowledge of what to do:

– What kind of future should one prepare for?
– What can one do now to influence the future evolution of the system in a requested way?

Unfortunately, in a complex system, such as the river flow process, it is not possible to identify and separate with certainty the effects of external or new factors, especially if they are gradual (such as climate changes or land-use changes) rather than abrupt, and the normal evolution of the system that would occur without such factors as well. This distinction is the main problem of change detection and is subject to much debate.

DATA FOR CHANGE DETECTION

Availability of a long time series of reliable data is essential for any attempt to detect a change in records of river flow, which is itself an integrative process controlled by several factors, such as precipitation (liquid and solid), evaporation, infiltration, snowmelt, catchment storage, and human impacts in the drainage basin and in the river itself. Hence, selection of which stations to use in a study is important. In order to study climate change signature, flow data should be taken from pristine (baseline) rivers, where human impacts are weak, should be of high quality and extend over a long period, preferably in excess of about 50 years (Kundzewicz & Robson, 2000), to distinguish climate change-induced trends from climate variability. Detailed suggestions on how to select a network of stations for climate change detection are given in Pilon *et al.* (2000).

Data should be quality-controlled before commencing an analysis of change. Furthermore, one should be open-minded at any stage of analysis for possible data quality problems, such as: typographical errors; instruments malfunctioning (zero-drift, bias); lack of homogeneity via changes in instruments and techniques of observations, changes in gauge location, in accuracy of data, and in data conversions, e.g. altered rating curves (stage–discharge relationships), censored data, etc.

Missing values and gaps in data are frequent complicating factors. It is not easy to issue a general guidance as to whether or not to fill them, and if so, in what way. This can be particularly critical when the gaps are non-random, e.g. following equipment damage by a flood event of exceptional magnitude. Such gaps would undermine studies of change detection in floods.

The form and frequency of the data for analysis depend on the main focus of the study (Kundzewicz & Robson, 2004). The form in which the data are collected is not always the most appropriate for the study in question (e.g. high flows and low flows being of interest to flood and drought studies, respectively). It can be worth simplifying the data by reducing the frequency (aggregating) or using summary measures, such as maxima or averages.

There are many different ways in which changes in hydrological series can take place. They may have different form, may occur in the mean values and/or in variability (variance, extremes, persistence and within-year distribution, e.g. changing river flow regimes). They may have different temporal scales. In addition, there may be complicating factors such as seasonality.

Trends are being sought in river flow data collected at a range of temporal intervals: hourly, daily, monthly, annually. Data records contain either instantaneous values (e.g. of flow, stage) or totals/means in a time interval (e.g. annual mean flow). The data may also pertain to different spatial scales, from a small creek to a continental-scale river. Abrupt changes can be expected as a result of reservoir construction, installing water diversions, etc. Changes of gauging structures and of rating curves can manifest themselves as step changes in a series of flow records. Yet, step-like changes may also result from gradual changes, since nonlinear system dynamics may feature cumulative effects and thresholds. Gradual hydrological changes may be attributed to gradual causative changes such as urbanization, deforestation, climate variability and change.

Different hydrological flow-rate related indices might be of interest in change detection studies (cf. Pilon *et al*., 2000), such as: instrumental records of water level/flow, seasonal mean flow, monthly mean flow, number of ice-cover days, extreme events and their characteristics (e.g. the frequency and severity of floods and droughts, number of incidences of independent flood events or the cumulative deficit below a prescribed threshold within a time interval, spring flood volume and the duration in days of the spring flood event). Timing of seasonal events is an important index of change. This refers to the start of snowmelt season, snowmelt flood-peak time, maximum-flow time, timing of river freeze-up, timing of ice break-up, etc.

ANATOMY OF STATISTICAL TESTING FOR CHANGE DETECTION

The definition of change is necessarily vague. In most phenomena that evolve in time, no two periods are identical as far as the process behaviour is concerned. Hence one can always find some changes in statistical properties, be it minor or large. Such changes, not related to external or new factors, but resulting inherently from the internal structure of a system, are termed natural variability. In other words, a change in the statistical properties estimated from samples does not necessarily imply a true change in the process; this behaviour may be natural for the process. In change detection one is interested in identification of changes in the process beyond the natural variability. Because of the strong natural variability of the river flow process, an observation that the mean values of river flow observed in two 30-year periods (standard climatic approach) are different would be trivial, unless one can support it by further analysis.

The distinction between genuine and apparent changes is the main problem of change detection and the existence of changes in a time series is often subject to much uncertainty. This may be contrasted with the situation in communication, where one can usually separate a signal from the noise beyond doubt and successfully discover the external factor: e.g. reliably state that someone has sent a message (with the signal-to-noise ratio being very high). In the latter case, one has a fair understanding of the essential parts of the system. In the case of a complex geophysical process (such as river flow), many observed changes may be attributed to its natural variability. Moreover, a change in such a complex system might be so complicated and so subtle itself, that it would be incomprehensible and it might not be possible to give a simplified description. If one does not understand the way in which the system is changing, one is not likely to detect it. Thus one has to look for changes that one can describe, keeping in mind that there might be other changes as well, that escape the analyst's attention.

The method of significance testing offers a way to distinguish "real changes" from chance occurrences resulting from natural variability. Instead of answering the question of the presence of changes, one states how unlikely a change larger than the observed one would be under natural variability alone. The starting point for a statistical test is to define the *null hypothesis* (H_0) of no change (only natural variability) and the *alternative hypothesis* of a specific kind of change (e.g. that the mean is either increasing or decreasing over time). The null hypothesis includes a complete statistical description of our understanding of the process in question. In other words, it defines a probability space that should reflect the present knowledge of the properties of the process. The alternative hypothesis should be plausible and supported by initial analysis of the

available data. The starting point for statistical testing is to assume that the null hypothesis is true, and then to check whether the observed data is consistent with this hypothesis. The null hypothesis is rejected if the data is not consistent with it. Otherwise, there is not sufficient evidence to reject the null hypothesis (even so this is not a proof that the null hypothesis holds, only that there is not enough reason to reject it).

In order to perform a statistical test for change one selects a *test statistic*, a numerical value calculated from the data series undergoing testing, which allows a comparison of the null and alternative hypotheses, by measuring, in quantitative terms, the change asserted by the alternative hypothesis. A good test statistic should highlight the difference between the two hypotheses. Given a dataset D the value $S(D)$ reflects the strength and possibly the direction of changes in D:

– In a two-sided (two-tailed) test, both large and low values of $S(D)$ suggest that changes are present in D. Usually the large values of $S(D)$ indicate increase and low ones indicate decrease of whatever $S(D)$ measures.
– In a one-sided (one-tailed) test, only large or only low values of $S(D)$ indicate changes. For simplicity it is assumed here that large values of $S(D)$ suggest changes and low values suggest lack of changes. No information on the direction of change is supplied.

A simple example of a test statistic is the linear regression gradient, which 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 may be very different from zero. In order 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* is the main parameter of significance testing. The *observed significance level* (p-value) is defined as the probability that a value of the test statistic as extreme as, or more extreme than the observed value, would occur assuming the null hypothesis (H_0) of no change (i.e. probability that a test detects trend when none is present). Given some $\alpha > 0$ one may state that the null hypothesis is rejected, and the alternative hypothesis supported, on a *significance level* α if the observed significance level is smaller than α. Specifically, for a two-sided test, the null hypothesis is rejected if:

$$2 \min (\mathrm{P}(S(D') \geq S(D)), \mathrm{P}(S(D') \leq S(D))) < \alpha$$

where D' denotes a random realization of the process to which the null hypothesis pertains. The factor 2 comes from probable changes in the other direction that can be compared to the observed change in the probabilistic sense (then "a change of similar magnitude" means "equally improbable"). If the probability density function of $S(D')$ is symmetric, the above may be simplified to:

$$\mathrm{P}(|S(D')| \geq |S(D)|) < \alpha$$

For a one-sided test with large values indicating change, the null hypothesis is rejected if:

$$\mathrm{P}(S(D') \geq S(D)) < \alpha$$

The significance level is selected arbitrarily (usually 5% or 1%) and guarantees that the probability of obtaining a false positive in the test is at most α on average, provided that the null hypothesis is stated correctly. The significance level expresses the probability of the error of the first kind (rejecting the hypothesis of absence of trends when it is true, i.e. detecting a trend that does not exist). Selecting the significance level is a trade-off and a matter of personal judgment. It should be small, so that errors of the first kind are reduced, but making it too close to zero would prevent the detection of even very strong changes (increase of errors of the second kind—accepting the hypothesis of absence of trend when it is false, i.e. failure to detect an existing trend). A possible interpretation of the significance level is as follows: changes significant on a level above 10% may provide very little evidence against the null hypothesis. Changes significant on a level between 5 and 10%, between 1 and 5% and below 1%, can be interpreted, respectively, as possible moderate, strong, and very strong evidence against H_0. Significance testing is a useful and consistent method of eliminating false positives. Nevertheless it must be emphasized that on any significance level $0 < \alpha < 1$ there may be:

– real changes not inducing significant test results; and
– significant test results not induced by real changes.

Statistical tests can support, but not prove, hypotheses. Whenever one has deeper understanding of the system in question, one should not restrict the study to significance testing alone.

It must be noted that in an analysis of 100 time series with no changes one would detect changes on the 5% significance level in 5 time series on average. It could as well be 4 or 7. Similarly, if one applies 10 different tests to these 100 time series, the number of falsely positive test results would become 50 on average. The number of positives in an analysis may be turned into a test statistic itself. Its distribution depends on the independence of the results, i.e. the independence of time series and tests used. At the very least one must keep in mind the average number of false positives possible and look for results supported by several tests rather than by one of several tests. One should always report on the entire scope of an analysis performed, data and methods applied, even if only a part of the study revealed interesting results. Selection of only a part of the study based on the results invalidates their significance.

A statistic combined with a specific null and alternative hypotheses forms a statistical test. For several traditional tests, regions of rejection and acceptance for desired significance levels can be looked up in reference tables or calculated from simple formulae, provided that the required test assumptions apply. In general, these regions can be found if the distribution of the test statistic under the null hypothesis (i.e. assuming the null hypothesis is true) is known or can be estimated.

Typical test assumptions include:

– *A specified form of distribution* (e.g. assuming that the data are normally distributed).
– *Constancy of the distribution* (i.e. all data points have an identical distribution) This assumption is violated if there are seasonal variations or any other cycles in the data, or if there is an alteration over time in the variance or any other feature of the data that is not allowed for in the test.
– *Independence* This assumption is violated if there is autocorrelation (also referred to as serial correlation or temporal correlation) or, in the case of a multi-site study, spatial correlation.

River flow data are often strongly non-normal, and this means that tests which assume an underlying normal distribution are not adequate. The data may also show autocorrelation and/or spatial correlation and, therefore, data values are not independent. They may also display seasonality, which violates assumptions of constancy of distribution. If the assumptions made in a test are not fulfilled by the data, then the test results can be meaningless, in the sense that the estimates of significance of the results via theoretical formulae would be grossly incorrect. In such a case the test can be treated as a mere method of exploratory data analysis rather than a rigorous examination tool.

A trend index (*TI*) can be defined as a measure related to the *observed significance level* (*p*-value) α, cf. Kundzewicz *et al*., 2005, Svensson *et al*., 2005. For two-tailed tests it ranges from –100% to +100%, with negative values indicating a negative trend and positive values a positive trend. The higher the absolute value of *TI*, the higher the trend significance.

$TI = (1 - \alpha) \times 100$ [%] for positive trends

$TI = -(1 - \alpha) \times 100$ [%] for negative trends

For one-tailed trends the value $TI = (1 - \alpha)$ is always in the range from 0 to +100%. This convenient measure puts test results on a common scale and lets one readily see if the result is significant on any desired level. For example a negative trend is significant at the 10% level if $TI < -90\%$.

A summary of the main stages of a statistical analysis of change follows:

– Decide what type of series/variable to test depending on the issues of interest.
– Decide what types of change are of interest (gradual trend or step-change).
– Check out data assumptions (e.g. use exploratory data analysis, or a formal test).
– Select a statistical test, i.e. a test statistic and a method for evaluating significance levels.
– Evaluate significance of the results (e.g. trend indices).
– Investigate and interpret results.

REVIEW OF TESTS

It is recommended to start change detection studies with the exploratory data analysis (EDA), which involves visual examination using graphs to explore, understand and present data. Looking at the data can change initial preconceptions, e.g. by altering the questions to ask, and unveiling important aspects that would otherwise have not been found. Visual analysis allows one to identify such features as data problems (outliers, missing values) or seasonality, to check out test assumptions such as independence, distribution assumptions, and also aids in understanding, interpreting and presenting the results of analysis. A well-conducted EDA is such a powerful tool that it can sometimes eliminate the need for a formal statistical analysis. Statistical tests become a way of confirming whether a pattern discovered via an EDA approach is significant (Kundzewicz & Robson, 2000, 2004).

Numerous tests for trend detection have been used in studies of long time series of river flow data and more information about tests can be found in (Kundzewicz & Robson, 2000). Therefore, the issue of the satisfactory choice of a test (or a series of tests) may come about. Tests differ by their skill, power, and underlying assumptions (thus, range of applicability) and computational efficiency. The skill, i.e. assessment of how well the test detects trends, embraces evaluation of the probability of errors of the first kind or of the second kind. A test which has low probability of error of the second kind is said to be powerful.

The process of testing for change in long time series of hydrological data starts from selecting a statistical test (more than one is good practice). This selection depends on the properties of the data. A short guideline for test selection can read as follows:

– *If data are normally distributed, independent and non-seasonal* (an unlikely scenario for river flow data), any general-use parametric or non-parametric test (e.g. slope-based tests, such as linear regression) should be suitable.
– *If data are independent and non-seasonal, but are non-normal*, any of the distribution-free tests are suitable. Distribution-free (non-parametric, e. g. rank-based) methods do not require any assumptions about the form of distribution that the data derive from e.g. non-normal distribution. However, tests that are based on the normality assumption can also be applied, either by usage of transformation to normal scores or ranks, or by using a relevant test statistic and evaluating significance using resampling techniques.
– *If data are not normal and are not independent* (i.e. they exhibit serial correlation, seasonality), one could use many standard tests, but it is necessary to evaluate significance levels using block permutation or block-bootstrap methods.

Parametric tests

Among parametric tests, which are based on assumptions of normal distribution and independence are tests for step change:

– *Student's t-test* (a standard parametric test for testing whether two samples have different means and a known change-point time is assumed);
– *The Worsley likelihood ratio test* (similar to Student's t-test but suitable for use when the change-point time is unknown);

and a test for gradual trend:

– *Linear regression* (one of the most common tests for trend—it uses the regression gradient as a test statistic and assumes that data are normally distributed).

Non-parametric (distribution-free) tests

Parametric tests are based on distributional and independence assumptions. However, the majority of hydrological series are non-normally distributed and it therefore makes sense to use distribution-free testing methods. Distribution-free methods are ones in which no assumptions about the underlying distribution of the data need to be made. However, the independence assumption still remains. It is assumed that sample elements are independent and identically distributed random (i.i.d.) variables.

Among commonly used distribution-free approaches are:

– *Rank-based tests* use the ranks of the data values (not the actual data values themselves). A data point has rank *r* if it is the *r*th smallest value in a data set. There are a number of widely used and useful rank-based tests (Kundzewicz & Robson, 2000). Rank-based tests have the advantage that they are robust and usually simple to use. They are usually less powerful than tests that are directly based on the data.
– *Tests using a normal scores transformation*. Many standard tests for change rely on assumptions of normality. When data are not normally distributed, as is often the case for river flow data, these tests can still be used if the data is first transformed. The normal-scores transformation results in a data set that has a normal distribution. It is similar to using the ranks of a data series, but instead of replacing a data value by its rank, *r*, it is replaced by the typical value that the *r*th largest value from a sample of normal data would have (the *r*th normal score). The advantages of using normal scores are that the original data need not follow a normal distribution, and the test is relatively robust to extreme values. Normal scores tests are likely to give slightly improved power for detection of change relative to equivalent rank-based tests.

Among non-parametric (distribution free) tests are such tests for step change as:

– *Median change point test* / *Pettit's test for change* (a powerful rank-based test for a change in the median of a series with the exact time of change unknown, considered to be robust to changes in distributional form);
– *Wilcoxon-Mann-Whitney test* / *Mann-Whitney U test* / *Mann test* / *Rank-sum test* (rank-based test that looks for differences between two independent sample groups, based on the Mann-Kendall test statistic);
– *Distribution-free CUSUM test* (rank-based test in which successive observations are compared with the median of the series with the maximum cumulative sum (CUSUM) of the signs of the difference from the median as the test statistic);
– *The Kruskal-Wallis test* (rank-based test for equality of sub-period means);
– *Cumulative deviations and other CUSUM tests* (e.g. rescaled cumulative sums of the deviations from the mean);

and tests for trend:

– *Spearman's rho* (rank-based test for correlation between time and the ranks series);
– *Kendall's tau* / *Mann-Kendall test* (another rank-based test, similar to Spearman's rho but using a different measure of correlation which has no parametric analogue. There is a seasonal Kendall test that allows for seasonality in the data, and a modified seasonal Kendall test that additionally allows for some autocorrelation in the data).

In general, the significance level can be found if the distribution of the test statistic under the null hypothesis (i.e. assuming the null hypothesis is true) is known or can be estimated. The above tests may be used in their original (standard or basic) form, providing test assumptions are met and significance levels can be looked up in reference tables or calculated from simple formulae. Alternatively, where test assumptions cannot be met, it is recommended that only the test statistic from each basic test is used, and that the significance level is evaluated using resampling. Resampling methods are robust, require minimal assumptions to be made and are very general.

All commonly used tests assume that, under the null hypothesis, the distribution of data values does not change with time. If this is not appropriate, then more sophisticated testing approaches will be necessary.

The skill of the many existing tests is usually good just for one particular type of change. Since one does not know the pattern of variability beforehand, using a number of tests is sensible. Therefore guidance as to the applicability of tests is important and an intercomparison of available tests is a much needed activity. Useful tests should be valid for realistic assumptions.

Resampling methods are a general and flexible approach, using the data to determine observed significance levels—which means that minimal assumptions about the data need to be made. Resampling methods are based on generating random reference data from the original observed

data, e.g. by changing the order of data points, and comparing test statistics calculated on these generated series with the test statistic for the original data series. Resampling methods can be applied to almost any test statistic and provide an alternative way of obtaining TI. Resampling tests are relatively powerful, e.g. for large samples, permutation tests can be shown to be as powerful as the most powerful parametric tests. Furthermore, resampling methods can be adapted to test data which are not independent.

Permutation tests are a very useful class of tests, based on changing the order (shuffling) of data points, calculating statistics, and comparing it with the observed test statistics. This helps identify evidence for the presence of a trend in the original series.

Consider a time series of data with a possible trend and regression gradient as an example of a possible *test statistic*. Suppose first that there is no underlying trend in the data. If that is true, then it should not matter if the data is re-ordered, as the regression gradient should not change very much. Each time the data is shuffled as part of the permutation test, the selected test statistic (here: regression gradient) is recalculated. At the end of all the shuffling, there is a distribution of possible values of the test statistic under permutation, the *permutation distribution*, which depends on the data and must be recalculated for each data set. The rationale behind this approach is that under the null hypothesis of absence of change in the data, each ordering of the data set is similarly likely. Hence, the null distribution of the test statistic can be estimated from the permutation approach. If there is no trend in the observed series, then one would expect that the observed test statistic (regression gradient) for the original data is not very different to any of the generated test statistic values, i.e. it is somewhere in the middle of the permutation distribution. So to test for trend, the observed test statistic (regression gradient) is compared with the permutation distribution. If the gradient is larger (or smaller) than almost all the values in the permutation distribution (the allowed fraction of exceptions given by the significance level), one concludes that a trend is present. Conversely, if the original gradient is somewhere in the middle of the permutation distribution, one concludes that there is no evidence of trend (Kundzewicz & Robson, 2000, 2004).

Bootstrapping approaches are similar to permutation techniques. The main difference is that instead of reordering the data, the new data series are generated by sampling with replacement. For example, for a series of N values, a bootstrap sample would take N values at random from the original series: the resulting series might perhaps include two occurrences of the original first value, but none at all of the last value.

For both permutation and bootstrap methods, the generated series has the same distribution as the empirical (i.e. observed) distribution of the data. The bootstrap is generally but not always, less powerful than a permutation test. However, bootstrap methods are often to be preferred where a test is looking for change in variance. Further, bootstrap methods are also applied outside of the area of change detection, e.g., for estimating the confidence intervals of the mean and median. Permutation tests cannot be applied with these and other statistics that do not change when the data are permutated. The tests given here can be used with either method. In general, bootstrap methods are more flexible than permutation methods and can be used in a wider range of circumstances.

Summary of methods for resampling

The basic method for carrying out a permutation or bootstrap test, once the test statistic has been selected, is as follows:

- calculate the test statistic for the observed data;
- resample the data series many times (e.g. 1000) to generate new data series;
- recalculate the test statistic for each of these series;
- estimate the significance level.

To estimate the observed significance level, the data are resampled a large number of times, N. For each of these generated series, the test statistic, S, is calculated to give N artificial values of S. These are then ordered as:

$$S_1 \leq S_2 \leq \ldots \leq S_N$$

If the original test statistic is S_0 and:

$$S_k \le S_0 \le S_{k+1}$$

then the probability of the test statistic being less than or equal to S_0 under the null hypothesis is approximately:

$$p = k/N$$

Assuming that large values of S indicate departure from the null hypothesis and a two-sided test, i.e. a test in which the direction of change is assumed unknown, the significance level for this test is then:

$$2 \min(p, 1 - p) \times 100\%$$

while the trend index reads:

$$TI = (2p - 1) \times 100\%$$

Further details on the application of resampling methods can be found in Kundzewicz & Robson (2000, 2004).

Block resampling: resampling when data are not independent

As noted by Yue *et al*. (2003), the existence of spatial correlation which is not accounted for, can complicate the identification of changes—a positive serial correlation may inflate the results of change detection.

The basic resampling methods, as described above, avoid any distributional assumptions, but they still assume that data values are independent of one another. Frequently measured hydrological data (e.g. daily data) are typically not independent: they show serial dependency (autocorrelation). In block resampling, the data is permutated or bootstrapped in blocks (e.g. all values within a year are kept together). With this approach, the dependency structure within each block is built into the test and independence assumptions are thus no longer violated (see Kundzewicz & Robson, 2000). Seasonality can be taken into account by using annual, or multiples of annual, blocks of data for the bootstrapping.

SPECIAL PROBLEMS

Extremes

Testing for trend or other non-stationarity in extremes (uncommon, infrequent events) is particularly difficult (cf. Robson & Chiew, 2000). Indeed, the consequence of the tautology: extreme (rare) events are rare, is that even in a very long series of record there may only be a few really extreme values leading to catastrophic damages.

Because extremes are rare, it is necessary to construct a data series that emphasizes extremes. One option is an annual maxima series, obtained by taking the largest value in each year or season of interest. However, the information on some extremes may be insufficiently reflected in such a series (e.g. if two extreme events occur within a year). The series may contain values that are not extreme (e.g. if no extreme event occurs within a year).

A peak-over-threshold (POT) series (also called a partial duration series, PDS) consists of independent daily mean river flows that exceed a certain threshold (this threshold is the same throughout the time series). This approach has advantages over annual maxima series in that all major events are included (not just one largest in a year) and all data points in the POT series are indeed extreme events. Furthermore, POT series can be used to examine whether there is a change in either the magnitude or frequency (counts) of extreme events. To obtain a POT series, it is necessary to identify a suitable threshold and to determine whether events are independent. There should not be multiple peaks corresponding to the same event. Judging independence can be complex. The threshold should be such that the average number of events per year is within adequate bounds, e.g. on average, one or three POT events are selected per year. The former case provides information about changes in large floods, while the latter also gives knowledge about changes in moderate events. Testing more than one series gives a better picture of what is occurring.

Testing for trends in droughts is more difficult because it is often the duration of the drought that is critical. Furthermore, severe droughts may span a number of years, i.e. longer data sets are required

for change detection. Detection of effects due to climate change is likely to require much longer data sets than detection of effects with a clear anthropogenic cause (cf. Robson & Chiew, 2000).

If a data series is strongly seasonal or if changes relating to a particular season are important, then it may be appropriate to consider a seasonal extreme series, e.g. for summer.

Whenever a change is detected, it is important to seek attribution, e.g. by examining other related hydrological series. For example, if trend is seen in river flow data for a catchment that has experienced land-use change, examining a rainfall series can provide information about whether climatic conditions have been steady across the period of record.

Recent detection studies of changes in floods and droughts do not support the hypothesis of ubiquitous change in severity and frequency of extremes. Using 600 daily streamflow records in Europe, Hisdal *et al*. (2001) conclude that it is not possible to establish that drought conditions in general have become more severe or frequent. Using 195 long series of daily mean flow records to analyse annual maximum flows, Kundzewicz *et al.* (2005) did not detect ubiquitous growth of high flows. A similar result was obtained by Svensson *et al.* (2005), who examined a subset of these data using a POT approach.

Spatial analysis

Univariate analyses of long time series of flows at single sites can be extended into an approach where all the available data for the area under study are used. Regional studies of long time series of river flows may allow important corollaries to be drawn as to the observed changes in a number of neighbouring gauges, thus being a property of the area.

The assessment of regional trends in hydrological conditions can be approached from two distinct perspectives (Lins, 2000), one inherently univariate (testing for changes in series at individual sites and then performing regionalization) and one multivariate (for pre-defined, homogeneous, regions). The former approach involves applying a test for trend to the hydrological time series collected at a number of individual sites and then grouping or *regionalizing* sites having similar test results. The latter (multivariate) approach differs in that *regions* are first identified from the hydrological time series collected at multiple sites, and a new derived time series for each region is then tested for trends. The former is more applicable if the analyst wants to preserve much of the temporal information at a single site, while also identifying adjacent sites exhibiting similar behaviour. The latter is more useful in applications where the goal is to emphasize the temporal behaviour of coherent regional patterns of variability; as in a hydro-climatic analysis. Because of land-use changes, reservoir construction, and other local effects, homogeneous spatial patterns seldom emerge from regional studies of changes in flow.

The majority of studies of change detection in river flow records assume that the data at different gauges are spatially independent. However, some recent studies account spatial dependence through the application of "field" significance, which accounts the observed regional cross-correlation of river flows and allows determination of the percentage of sites that are expected to show a trend by chance. The presence of spatial correlation affects the ability of a test to assess the field significance of trends over the network. Consideration of inter-site spatial correlations (overlap in information) dramatically reduces the effective size of the sample available for trend assessments. The effect of cross-correlation in the records under study is to increase the expected number of significant trends occurring by chance. If spatial dependence (regional cross-correlation) is ignored, then significant trends are typically found in a great many more cases than with cross-correlation considered (cf. Douglas *et al.*, 2000; Burn & Hag Elnur, 2002). Using a field significance rather than significance for the individual sites is recommended for regional studies when large amounts of spatially-distributed records are available.

Naturally trendy?

Cohn & Lins (2005) demonstrated that the choice of the null hypothesis (i.e. adequate test assumptions) is critical for the results of significance testing. Significance depends on subjective assumptions about the underlying process. In fact, it can be argued that for any time series of observations with a trend one could find a null hypothesis that would render the trend insignificant. Any observed changes could thus be attributed to natural variability. It is obvious

that the null hypothesis should not include a trend component, so that it differs from the alternative hypothesis. Similarly, if the null hypothesis includes nonstationary processes with high persistence it may be barely distinguishable from the alternative hypothesis when the evidence is limited to a single time series. Therefore, it is important to support the selection of the null hypothesis, ideally by evidence from a period before the supposed trend. One must also consider and study the possibility that a test may confuse a trend with higher-than-real natural variability and persistence leading to lower detection power. It is still sensible to perform studies of changes where one expects under-detection, because conservative, lower-bound estimates of the intensity of changes may be obtained in this way and more light is shed on the degree of inherent uncertainty of the results.

A much greater danger than under-detection seems to be over-detection, due to choosing a null hypothesis that does not encompass the natural variability of the process in an appropriate way. This was discussed in detail in the section on test assumptions. Statistical tests that rely on the assumption of i.i.d. (independent, identically distributed observations) are sometimes applied without question to annual time series of river flows, because of the lack of time, the lack of a better tool, and the lack of appropriate data. While the i.i.d. assumption may be valid in some cases, the studies based on it should at least be accompanied by tests for serial dependence in the data. It would also be worthwhile to always supply a corresponding conservative result obtained using tests with more general assumptions.

In some cases it is known that the process has changed, but one may wish to know how it has changed (e.g. increased concentrations of greenhouse gases have affected the climatic system and therefore have some influence on the global temperature). The question is: what is this effect and how can it be detected? In this setting, significance testing still has a merit, namely, one can hope to say something sensible about the amount of change if the change is larger than the natural variability of the process. If the change is not significant, it means that the uncertainty about the magnitude of change is at least comparable to the observed change itself.

Unfortunately, in hydro-climatic studies the options are much more limited than, say, in public opinion sampling or medical tests, where insignificant results may usually be made more precise by a repeated test or a wider survey. Increasing the length of a time series sufficiently may not be possible if precise observations were not performed in an earlier period. The only solution then is to gather as much data as possible to confirm or disprove the existence of changes, and to give a full account of the uncertainty of the result when reporting on it.

CONCLUDING REMARKS

Change detection in hydroclimatic variables is a challenging task, because it is hard to define what change is, and one does not have a complete understanding of the statistical properties of the processes involved. Change is often bad news. The increasing number of extremes is of the worst kind and also the particularly difficult one to detect. Significance testing offers a way to distinguish “real changes” from “chance occurrences”. Although for hydroclimatic variables one can never distinguish with certainty, at least “significant changes” are well defined, based on an arbitrary selection of a significance level and probability space (null hypothesis). The results of testing depend critically both on the selection of the test object and the testing method (the test and the null hypothesis).

A few general guidelines on change detection in river flow records are in order. It is necessary to make assumptions in a change detection procedure, and one must be aware of them. Any statistical description of a process is only an idealized view of reality, and even in this idealized view there is room for surprises. One needs to make assumptions in order to apply methodological tools, keeping in mind that the results depend on the assumptions taken. It is important to remember that inappropriate test assumptions are dangerous. If the assumptions made in a statistical test are not fulfilled by the data then the test results can be meaningless. Statistical tests results express probability and not certainty, hence they provide evidence rather than proof. There is always a chance that the null hypothesis was true when a test result suggests it should be rejected, and if the null hypothesis is accepted, then this result says only that the available

evidence does not contradict the null hypothesis. Application of different methodologies to the same original data set may result in different trend estimates, hence use of several tests is recommended.

Acknowledgements This paper draws from the work accomplished within the framework of the World Climate Programme – Water Project. Financial support provided to the authors within the projects "Extreme meteorological and hydrological events in Poland" (the Ministry of Education and Science of the Republic of Poland) and ENSEMBLES (EU Sixth Framework Programme) is gratefully acknowledged. The second author was supported by the Foundation for Polish Science.

REFERENCES

Burn, D. H. & Hag Elnur, M. A. (2002) Detection of hydrologic trends and variability. *J. Hydrol.* **255**, 107–122.

Cohn, T. A. & Lins, H. F. (2005) Nature's style: naturally trendy. *Geophys. Res. Lett.* **32**, L23402, doi:10.1029/2005GL024476.

Douglas, E. M., Vogel, R. M. & Kroll, C. N. (2000) Trends in floods and low flows in the United States: impact of spatial correlation. *J. Hydrol.* **240**, 90–105.

Hisdal, H., Stahl, K., Tallaksen, L. M. & Demuth, S. (2001) Have streamflow droughts in Europe become more severe or frequent? *Int. J. Climatol*. **21**, 317–333.

Kundzewicz, Z. W. & Robson, A. (ed.) (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 river flow records – review of methodology. *Hydrol. Sci. J.* **49**(1), 7–19.

Kundzewicz, Z. W., Graczyk, D., Maurer, T., Pińskwar, I., Radziejewski, M., Svensson, C. & Szwed, M. (2005) Trend detection in river flow series: 1. Annual maximum flow. *Hydrol. Sci. J.* **50**(5), 797–810.

Lins, H. F. (2000) Spatial/regional trends. In: *Detecting Trend and Other Changes in Hydrological Data* (ed. by Z. W. Kundzewicz & A. Robson), chapter 9. World Climate Programme – Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD – no. 1013. World Meteorological Organization, Geneva, Switzerland.

Pilon, P. Kundzewicz, Z. W. & Parker, D. (2000) Hydrological data for change detection. In: *Detecting Trend and Other Changes in Hydrological Data* (ed. by Z. W. Kundzewicz & A. Robson), chapter 3. World Climate Programme – Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD – no. 1013. World Meteorological Organization, Geneva, Switzerland.

Robson, A. & Chiew, F (2000) Detecting changes in extremes. In: *Detecting Trend and Other Changes in Hydrological Data* (ed. by Z. W. Kundzewicz & A. Robson), chapter 7. World Climate Programme – Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD – no. 1013. World Meteorological Organization, Geneva, Switzerland.

Svensson, C., Kundzewicz, Z. W. & Maurer, T. (2005) Trend detection in river flow series: 2. Flood and low-flow index series. *Hydrol. Sci. J.* **50**(5), 811–824.

von Storch, H. (1995) Misuses of statistical analysis in climate research. In: *Analysis of Climate Variability Applications of Statistical Techniques* (ed. by H. von Storch & A. Navarra), 11–26. Springer Verlag, Berlin, Germany. http://w3g.gkss.de/staff/storch/pdf/misuses.pdf.

Yue, S., Pilon, P. & Phinney, B. (2003) Canadian streamflow trend detection: impacts of serial and cross-correlation. *Hydrol. Sci. J.* **48**(1), 51–64.