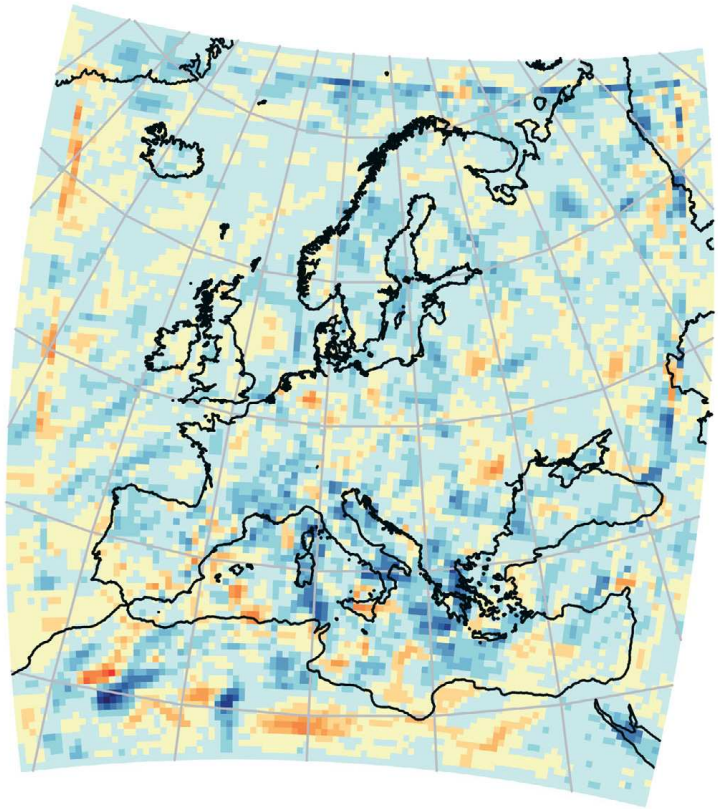
Technical Report No. 13

INTERCOMPARISON OF METHODS FOR TREND DETECTION IN HYDROLOGICAL EXTREMES DERIVED FROM GRIDDED DATA



Author names: Maciej Radziejewski and Zbigniew W. Kundzewicz

Date: October 2008

WATCH is an Integrated Project Funded by the European Commission under the Sixth Framework Programme, Global Change and Ecosystems Thematic Priority Area (contract number: 036946).
The WACH project started 01/02/2007 and will continue for 4 years.

Title:	Intercomparison of methods for trend detection in hydrological extremes derived from gridded data
Authors:	Maciej Radziejewski and Zbigniew W. Kundzewicz
Organisations:	Research Centre for Agricultural and Forest Environment, Polish Academy of Sciences, Poznań, Poland
Submission date:	October 2008
Function:	This report is an output from Work Block 4; Task 4.2.2
Deliverable	WATCH Deliverable D 4.2.2

1 Introduction – univariate case

This document describes several approaches to detection of changes in hydrological extremes derived from gridded data.

Detection of changes in hydrological records is a challenging scientific issue of considerable practical interest and importance. However, even for the univariate case (e.g. data observed at one station) it is not a trivial task, because changes in the time series of record can be weak and non-uniform (dependent on time window), and natural variability can be strong. Hence requirements for both adequate data and methodology are severe.

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 process (as described by its statistical properties) in two different time periods. A trend is a continued change that occurs over time. One may either view it as a manifestation of a time-dependent deterministic component (this requires some understanding of the underlying mechanism) or simply as a continuous tendency in the statistical properties of the process (Kundzewicz and Radziejewski, 2006).

Usually trends of simple shape (linear, low-order polynomial, piecewise linear, exponential, etc.) are considered. Different trend shapes are possible, 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.

For discussion of univariate change detection, including the issue of appropriate data, anatomy of testing (null *vs* alternative hypothesis, test statistic, significance level), test assumptions, errors, see Kundzewicz and Radziejewski (2006). They also proposed a review of tests (parametric and non-parametric, i.e. distribution-free, and based on resampling).

Testing for change 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 flood-related records 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. In the case of floods, one option is to use 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). Alternatively, a peak-over-threshold (POT) series (also called a partial duration series, PDS) can be used, consisting of independent data 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. 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 on extremes 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).

A few general guidelines on change detection 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 (that there is, or that there is no, ground to reject the null hypothesis) 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.

2 Multivariate case

Univariate analyses of hydrological data at single sites can be extended into an approach where all the available data for the area under study are used. The assessment of regional trends can be approached from two distinct perspectives (Lins, 2000), one being inherently univariate (testing for changes in series at individual sites and then performing regionalization) and one being multivariate (for pre-defined, homogeneous, regions). The former approach involves applying a test for trend to the records at 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 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 hydroclimatic analysis (Kundzewicz and Radziejewski, 2006).

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.

In addition to multivatiate point records, gridded hydrological data are being increasingly used. If series of hydrological data are given in a number of discrete points (stations), for a common time interval, gridded maps can be produced via interpolation. Precipitation data, the main input to global hydrological cycles and climate models, are based on rain gauge records. Global Precipitation Climatology Centre (GPCC) operated by the Deutscher Wetterdienst (National Meteorological Service of Germany) collects monthly gridded area-mean rainfall totals on a 1° x 1° global grid. Remotely-sensed (e.g., satellite) data offering spatial coverage are naturally gridded. The reanalysis projects (e.g. NCEP/NCAR Reanalysis or ECMWF 40 Year Re-analysis (ERA-40) Data Archive) produce gridded information with the help of state-of-the-art data assimilation, blending several sources of observations (land surface, ship, sonde, aircraft, satellite and other data) and with advanced quality control, using data from 1948 (from 1957, respectively) to the present. ERA-40 has resolution of 2.5° for “basic 2.5° atmospheric” and a higher space resolution for “full resolution atmospheric”. Climate model results are given for broadly-spaced grids. Gridded hydrological data include such variables as for example: precipitation, snow cover, snow water equivalent, potential evapotranspiration, soil moisture, river runoff, groundwater recharge.

Gridding of river runoff, based on river discharge records in cross-sections, is also practicised.

Detection of change in spatial-temporal gridded data fields is considerably more difficult to interpret than for univariate case. A simplistic approach is to perform simple statistical tests separately for each grid cell. As a result, one can have a gridded map illustrating for which grid cells the tests are statistically significant. But many geophysical random fields, including hydrometeorological variables have substantial spatial correlations (von Storch and Zwiers, 1999).

A single aggregated inference is needed about the whole spatial test. Decision made at one grid cell (accept/reject the null hypothesis) may not be statistically independent from decisions made at other locations. It is not easy to determine regions of significant change and the concept of field significance comes about.

In the present paper, typically the input data used is surface or near-surface data from reanalysis or climate model output, i. e.:

- precipitation,
- temperature (e. g., T, Tmin, Tmax), and, possibly,
- other climate variables.

The spatial grid is two-dimensional with time as the third dimension. Typically the temporal resolution is daily or higher. One can also use spatial, two-dimensional data describing the surface properties, such as elevation data and geographical data.

The climate data may be available as:

- One sample,
  i.e. one three-dimensional array, e. g., reanalysis data or climate model simulation for a continuous time period.
  In this case a statistical test for change in a time series will be used at some point as a part of a testing method. Even when the data is not available for a continuous time period, for example simulations of periods of 10 years every 30 years throughout a longer period, essentially the same methods may be used.
- Two samples,
  i.e. two three-dimensional arrays, e. g., two climate model simulations for two time periods (past and future), or for two scenarios (possibly with the same initial/boundary conditions).
  In this case a corresponding two-sample test would be used instead of a time series test.

In general the majority of methods described herein may be adapted to either of the above cases, as will be seen throughout. Figure 1 illustrates comparison of indices for two time windows. This two-sample view can be selected even if model simulation is available for the complete time period 1961-2099.

Climate model output is typically available for a number of runs, sometimes using a number of models for a given scenario, and a number of scenarios. These may be used for uncertainty assessment and comparison between scenarios. Nevertheless, testing for trends will come down to either one-sample or two-sample testing and comparison of results.

Several methods for detection of trends in hydrological extremes are described in the first four sections. Therein we focus on the problem of trend detection in a single dataset/pair of datasets. We start with the methods where testing is performed on raw precipitation data, and finish with those where hydrological modeling is involved. The subsequent section shows how to study results from a number of

simulations/models to compare scenarios and perform uncertainty assessment. The last section treats the technical side of computations with gridded data.

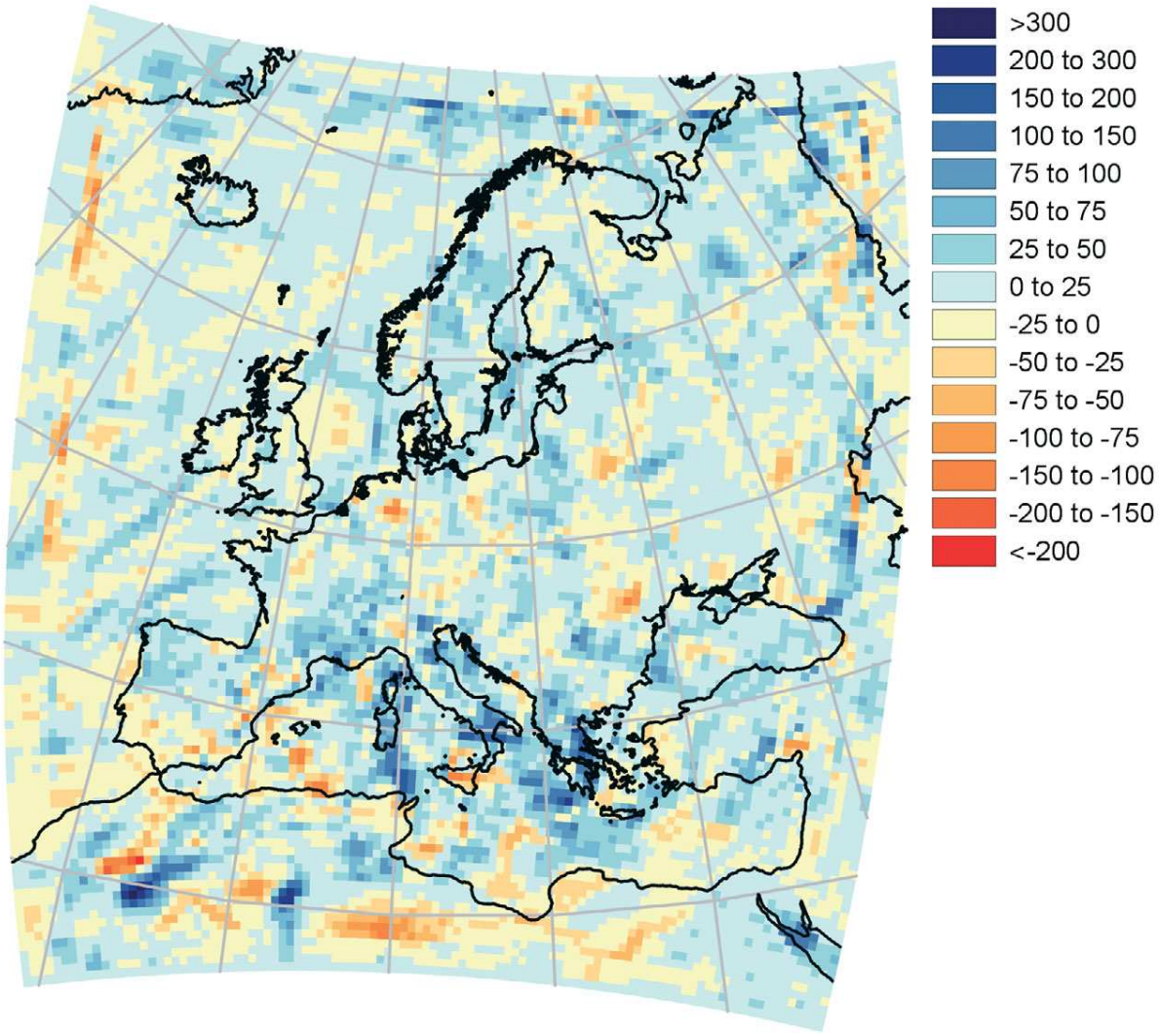


Fig. 1 Difference in annual maximum daily precipitation (mm) over Europe between the control period (1961–1990) and future projection (2070–2099) (HadRM3-P, SRES A2 scenario). Source: Kundzewicz et al. (2006).

This document does not describe the general theory of change detection, or specific statistical tests, or the details of how to study changes in extremes. Please refer to (Radziejewski and Kundzewicz, 2004) for this information. The focus here is on how to deal with gridded data and with a number of simulations/models.

3 Trend detection in gridded precipitation data

The most straightforward approach described here is to test for changes in precipitation data itself. Daily precipitation data for a given grid box may be treated as a time series (or as two time series, in case of two samples). An appropriate one-sample or two-sample statistical test for change detection may then be used for each grid box. In order to allow testing for different kinds of changes the test may be combined with other operations on data, such as:

- computation of annual indices of extremes from daily data (e.g., the total number of days above/below a given threshold, the length of the longest time period above/below a given threshold, or more sophisticated indices),
- restriction to a specific season, etc.

The statistical test of choice may then be applied to such indices for each grid box. In any case a point-wise (local) test result (strength, direction, and significance of changes) is obtained for each grid box. The number of local results is simply too large to quote each individually. It is always advisable to present the local results (including local significance) on a map for exploratory analysis/expert assessment. It must be noted that local significance levels cannot be interpreted as a global significance level would be. However, the global result may be complex, with different directions of changes in different areas, so a map provides a quick answer on what changes have been detected.
The remaining question is thus how to compute the global significance of changes , so called *field significance*, from the array of local results. The methods of computing the field significance (von Storch and Zwiers, 1999, Section 6; partly after Livezey and Chen, 1983) allows us to assign one global significance value to the set of all local test results on a grid. Thus we can turn an array of local test to one global, aggregate test result that should be easier to describe and interpret. This problem is addressed in the remainder of this section, with some of the methods quoted from (von Storch and Zwiers, 1999) and some developed anew.

3.1 Univariate testing

We start with methods that are essentially univariate, as they are based on estimating local (time-series-based) results, and then operate on those results alone.

3.1.1 Conservative significance estimation for a small region (cluster)

We use the notion of *Trend index* (*TI*), a convenient measure that encompasses both observed significance (the *p*-value, denoted *p*) and the direction of changes, cf. (Radziejewski and Kundzewicz, 2004). The sign of *TI* corresponds to the direction of change (positive indicates an increase and negative a decrease) and the absolute value is:

$|TI| = 1 - p$.

When performing detection on gridded data in a small region, one may expect the results to be highly correlated, because of geographical proximity and similarity in conditions. Such a small region (cluster) should be defined by an expert. Results within one cluster can be replaced by one (cluster-wide) result in one of two ways:

1. The cluster result is considered significant (on a given level) only if all the local results are significant and agree in direction, or
2. The cluster-wide *TI* is computed as an average of all local *TI*.

Both methods may be called conservative, or cautious, in that lack of strong dependence between the local results would lead to underdetection of changes (decrease in test power). Method 1. is more conservative of the two, and 2. seems preferable.

It is possible to focus the detection on key areas, such as river sources (of selected large rivers). If the results agree in direction and significance on a key area selected previously (before the results are known) one may consider such results valid for the key area. However, if several such areas are tested for changes, all of the results (significant and insignificant) should be quoted.

3.1.2 Cluster-based trend detection for larger regions

One possible way to compute field significance for larger grids would be divide the grid to a number of clusters. Data within the cluster may be assumed dependent and treated as in the previous subsection. Data in different clusters may be assumed independent and thus the cluster-wide trend indices may be treated as independent, uniformly-distributed variables. The global significance of the results may be based on the number of significant cluster-wide results as compared to the appropriate binomial distribution. An aggregate trend index and resampling of cluster-wide trend indices may also be used (please see the next section for an explanation). In the latter case the cluster-wide trend indices may be weighted according to each cluster's area or population, depending on the problem being considered. Methods of relaxing the null hypothesis (and obtaining stronger conclusions) described in the next section also apply in this case.

Cluster-based approach has some advantages: it is simple and just having the clusters and results stated for them is a valuable asset. The resulting message is easily understood. However, clusters require an expert to define, meaning that we have to take guesses to define them. Moreover, too small clusters might result in overdetection of trends (only "might", because the methd used within them is conservative), while too large clusters would most probably result in underdetection. It may be desirable to blend the conservative and the optimistic approach (optimistic = assuming independence) and define two or more levels of clusters. This may be a subject of a further study.

Cluster-based approach described here is similar to the spatial patterns approach (von Storch and Zwiers, 1999, Section 6), except there the authors suggest aggregation of (or applying the patterns to) the data itself, while in the approach described here only the trend indices are aggregated.

3.1.3 Decreasing the number of degrees of freedom

Similarly to the situation in previous subsection, if a number of (local or regional) trend indices are assumed independent, the global significance may be obtained from them by comparing the number of significant results, or, more precisely, the fraction of significant results among them, with the appropriate binomial distribution. The total number of results is one of the paramaters of the binomial distribution, also called the number of degrees of freedom.

Local results are usually not independent. One method to deal with spatial dependence (von Storch and Zwiers, 1999, Section 6.8; after Livezey and Chen, 1983) is to reduce the number of degrees of freedom in the binomial distribution. In fact, the smallest number of degrees of freedom for which the result is still significant may be computed, and then an expert is to judge whether or not the grid may be divided to this many, more-or-less equal, parts that may be assumed independent. This approach is similar to cluster-based approach. However, on one hand, it is much easier to implement, on the other hand it does not provide such rich information on what changes occur where, and involves expert guesses to a greater degree.

3.2 Multivariate testing

3.2.1 Multivariate Hotelling test

Another method suggested by (von Storch and Zwiers, 1999, Sections 6.6.10 and 6.5) is the multivariate Hotelling test (T^2). This test may be applied in a two-sample case, provided the distributions of the variables are normal with identical (unknown) covariance matrices. Moreover, the total number of

time units in the samples must be greater than the number of grid points in order to estimate the inverse of the covariance matrix. These assumptions are usually not satisfied by climate model precipitation fields. The authors suggest use of spatial patterns, i.e., performing a weighted spatial pre-aggregation of data, in order to reduce the number of spatial degrees of freedom. Such pre-aggregation may also make the distribution closer to normal.

The main problem with the Hotelling test seems to be that it tests the null hypothesis that the mean has not changed in *any* of the points/patterns considered. Thus, if the null hypothesis is rejected, the results support the alternative hypothesis that the mean has changed in *at least one grid box or pattern*. This is certainly not a finding of great interest.

3.2.2 Aggregate Trend Index

Given a field of local test results (in a one-sample or two-sample case) an aggregate measure of the strength of changes may be computed for the entire field as the mean of absolute values of local T_l-s, or as the mean of their squares. Higher powers could also be used theoretically, if the detection focuses on strong changes, while in the sum of the absolute values even statistically insignificant local changes may contribute to a significant global result, simply because there is more of them than would be expected. The classical measure used by Livezey and Chen (1983) is the number of significant local results on the 5% significance level.

The aggregate trend index thus obtained does not, yet, provide information on the significance of changes. It is only a comparative measure. It is meant to be compared with values of aggregate trend indices obtained in the course of resampling. In fact, both the local trend indices (and the aggregate trend index) and the final global trend index should be computed from the same resampled data sets using Monte-Carlo methods. The choice of a resampling method determines the kind of null and alternative hypothesis that we test for.

3.2.3 Methods of multivariate resampling

It must be stressed that the resampling method to be used should be appropriate for the problem studied. Therefore specifics are not and cannot be given in this section. For example, in a two-sample case the samples may be related in various ways: they can be realizations of the same model with the same initial conditions, only slightly different parameters, or they can come from the same simulation in different time periods, or they can be completely independent. The resampling method (e.g., the way they are permuted) would be different in every case and, in general, should reflect the way they are related with the exception of the features being tested. It is dangerous to apply a ready-made method blindly to any data – the method should always be re-considered, so it leads to viable results and not artifacts.

3.2.3.1 Climate-model-based resampling

In general, resampling methods may be divided to model-based and data-based. Whether we test for trends in reanalysis or climate model output, the best model we have is the climate model itself, so resampling would involve running a large number of climate model simulations. At present this approach is not yet computationally feasible, but it may be possible in the future.

When testing for trends in a one-sample case the model may be used to generate controlled, trend-free simulations (with the same resolution as original data) whose aggregate trend indices may be

compared with the original one. In the two-sample case it is possible to generate a large number of simulations corresponding to both samples.

3.2.3.2 Data-based resampling

Methods of resampling based on data include various permutation/block permutation techniques as well as sampling with replacement (and its block versions). One may argue that blocks on length of one year (in the time dimension) are sufficient to account for autocorrelations of climate variables in time. Spatial correlations may similarly be taken into account by using appropriate block sizes in spatial dimensions. The block size is, again, a matter of expert judgment, or, better, an effect of analysis of spatial correlation structure of the variable in question.

An "extremely cautious" approach in this case would be to assume complete spatial auto-correlation and use the entire field as one block, i.e. permute or resample blocks spanning the entire spatial extent of the grid and appropriate periods of time. In fact, this is, essentially, what Livezey and Chen (1983) did in their Monte-Carlo approach. They performed a two-sample analysis (of correlation) where one of the samples was a time series. Their resampling method was to permute the time series, which is equivalent to permuting both independently with complete spatial fields of the grid variable preserved. [Except they used a slightly different method for estimating local results.] This approach raises interesting methodological questions, as to what the resulting global result really measures: the test results with autocorrelation accounted for (as intended), the autocorrelation of the variable itself, or the interplay between the different ways to count test results within the method. In fact, the author of the present text does believe that this method is perfectly sound, but in order to use it is best to analyze its properties and power in a controlled situation.

Another possibility to account for autocorrelation in space and time is to use spectral methods. In this approach an appropriate three-dimensional Fourier transform of the data should be performed, the moduli (power spectrum) of the coefficients should be kept fixed, and the arguments (phases) appropriately randomized, depending, again, on the problem in hand. Spectral resampling of two samples would usually involve the estimation of the same moduli based on both samples and independently randomized arguments, however this is, again problem-dependent.

3.2.4 Drawing stronger conclusions

As already noted above, classical field significance methods really test for a very specific null hypothesis of no changes anywhere. Rejecting it leads only to very uninteresting results like: *There are changes in at least one place, somewhere in the World (but we thing they are global, in fact)*. It may be much more interesting to obtain answer to questions like: *Does GHG induce changes in precipitation over more than a half of the World?*, or: *Does the addition of GHG into a simulation induce changes in precipitation over more than a half of the World?*

The null hypothesis to address such question would have to involve a possibility of a given percentage of points having significant changes. Suppose we allow for changes in one half of the grid. When testing the original data, the 50% of points with most significant local changes are identified. Then, when resampling, local trend indices at those points are always replaced with the original ones. This definitely makes it harder to detect changes and the greater the percentage of points where changes are allowed (instead of 50% used here), the lower the resulting global trend index or *p*-value. Thus we obtain a relationship between the area percentage and the resulting global *p*-value.

Given the area percentage - *p*-value relationship we may integrate it to obtain a dimensionless measure between 0 and 1. It may be interpreted as "the area percentage" affected by changes. However, the interpretation of such a measure would have to be studied first. The entire area percentage - p-value relationship would be interesting in any way. As it depends on the function used in constructing the aggregate trend index, by the use of different functions we may draw conclusions like: *Increased GHG concentration affects precipitation over most of the World, however, the results are often insignificant locally, and affect the entire system rather than single points*.

Usually we are interested in the direction of changes, not just if they occur. Although the complete information on the direction of changes would be presented on a map, the map may be too rich in information if a large number of such maps should be analyzed. For example we are more likely to draw conclusions like *Models agree that it gets wetter over most of the World in [season]* than *Models agree that this is a map of projected changes*. Such conclusions may be obtained by, again, modifying the detection procedure. In this case one should disregard all the positive or all the negative trend indices when computing the aggregate trend index, to focus only on the changes in one direction.

4 Aggregation of local distributions based on clusters

The approach described in this section was developed and is being used[1] for rough estimation of regional flood damages, but it may be used with other data as well. The general idea is to perform structured aggregation of point-wise distributions in order to obtain an adequate regional estimate. This approach is suited to two-sample analysis, so if a single sample is to be studied, then it should be divided into two sub-periods and the distributions should be derived for each sub-period separately. The main constraint here is that it is the distributions that are aggregated. Please note that it makes no sense to use this method to compare, e. g., the distribution of daily precipitation or any climate variable where we have the information on each value and time of its occurrence. *This approach only becomes useful when aggregating distributions alone*. We faced the problem of aggregating damage distributions that were only possible to compute locally, based on flood zones. This is the specific problem described here. For each box in a rectangular grid we get four numbers, representing a 20-year, 50-year, 100-year, and 500-year floods. They describe upper tails of point-wise distributions. We needed to aggregate these values to derive flood losses distribution for a region of interest. It is equally possible (although computationally more demanding) to aggregate entire distributions, not just the tails.

4.1 Proposed method

The input data is the distribution (or the tail of the distribution) of a variable of interest for each grid box. The simplest methods of aggregating point-wise distributions are convolution (if independence between grid boxes is assumed) and simple summation of quantiles (corresponding to the assumption of complete dependence between grid boxes). If the variable has positive spatial autocorrelation we know that treating its values in different boxes as independent would lead to underestimation of extremes, while simply summing the extreme values across a region would overestimate the regional extremes. This method proposes a tradeoff between independence and complete dependence.

Solution: Define clusters of grid boxes and treat events in one cluster as completely dependent, events in different clusters as independent. Clusters should (roughly) correspond to a typical spatial extent of a flood event, which does depend on the flood size, therefore larger clusters are needed for bigger floods. Flood distributions for smaller clusters are aggregated by a mix of convolution and addition to get the distribution in a larger cluster.

[1] in the ADAM (Adaptation and Mitigation Strategies) EU 6FP Integrated Project

4.2 Defining the clusters

We need to define clusters of levels 1 through 4, L1 corresponding to the likely extent of a 20-year flood event, L2 to that of a 50-year flood, L3 – 100-year, and L4 – 500-year. We also define L0 clusters to be individual grid boxes and one L5 cluster: the region of interest, for which we derive the distribution. So we have 6 levels altogether. The layout of clusters must be hierarchical, i.e. each level-k cluster (k = 1, …, 5) must consist entirely of level-(k – 1) clusters. This should not be a problem, because lower-level clusters can always be split or cut along the higher-level cluster boundaries.

An expert should be able to give a rough estimate of the likely shape and extent of flood events given the flood zones. Then, the actual clusters probably need to be defined using the information on terrain features.

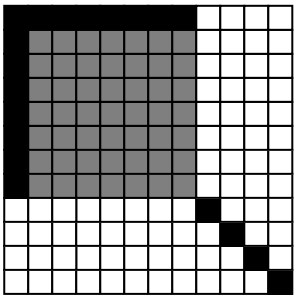
4.3 Representing the upper tail of the distribution

Addition of independent variables usually involves Abel's convolution of probability density functions. However, because of the special situation here, where more extreme events are assumed dependent, it is better to represent a distribution as a sequence of quantiles rather than a PDF. While PDFs might (or might not) be computationally more efficient, the quantiles approach is simpler to describe and easier to work out.

Let N equal 500. This happens to be equal to the "rareness" of the most extreme event considered here, however, it is really the least common multiple of 20, 50, 100, and 500. We can represent the distribution of a random (flood) variable X using a vector of 50 numbers (q_1, …, q_{50}) such that q_{50} is the 500-year flood, q_{46} the 100-year flood, q_{41} the 50-year flood, and q_{26} the 20-year flood. In general, a d-year flood is in $q_{51\text{-}500/d}$ for d > 10, but we put q_1 = 0 for all "10-year and less" events. If we start with just the four d-year floods for a gridbox, I would set q_{50}, q_{46}, q_{41}, q_{26}, and q_1 accordingly and fill the blanks by linear interpolation. Then we can think of X as sampled randomly from (q_1, …, q_{50}) where each q_i is selected with probability p_i with $p_1 = 1 - 49/N$, and $p_2 = p_3 = \ldots = p_{50} = 1/N$. This is a crude approximation, but it should be ok when the focus is on high tails.

4.4 Clumping two distributions

We are given two random variables X and Y whose distributions are represented using 50 quantiles: (a_1, …, a_{50}) for X and (b_1, …, b_{50}) for Y. The probabilities are as above. We are interested in the distribution of $X + Y$. For this we need to construct a matrix of joint probabilities $p_{ij} = P(X = a_i, Y = b_j)$. This matrix should be symmetric and satisfy $p_i = \sum_{j = 1, \ldots, 50} p_{ij}$. It should look like this:



where the black rectangles represent higher probabilities and white is zero. The picture shows a situation when events up to some size (corresponding to quantiles up to index m) are independent and larger events (indices $m + 1$ and above) are dependent. Then the matrix would be defined like this:

$$\begin{array}{ll} p_{ij} = p_j, & \text{for } i = j > m, \\ 0, & \text{for } i \neq j, \max(i, j) > m, \\ p_i p_j / (p_1 + \ldots + p_m), & \text{for } i, j \leq m. \end{array}$$

Other choices for the matrix p_{ij} are possible. I thought about gentler blending between dependence and independence, but this gets much harder to implement and there is no obvious best choice for such blended matrix, so I have decided to stick with the above.

Note that the choice of m depends on the cluster level, so you only need to construct such a matrix once per level. Once you have the matrix you simply take all the pairs $(a_i + b_j, p_{ij})$ with non-zero p_{ij} and order them by the first coordinate to get the values and corresponding probabilities of $X + Y$, analogous to the 50-quantiles representation, except it is somewhat denser. This is then converted to the 50-quantile format by averaging and interpolation.

4.5 Computational cost

Calculation speed may or may not be a concern, so you may well ignore this paragraph. The main computational cost in clumping comes from sorting the (sum, probability) pairs. At first it may seem like you have to sort 2500 items each time, but it is usually much less. Most of the clumping will be done on level 0 (other clusters are smaller in number), and there will be very few non-zero probabilities in the joint distribution for this level.

4.6 Aggregation over a cluster

Let $X_1, \ldots, X_n$ be the random variables associated to the n level-k clusters ($k = 0, \ldots, 4$) that comprise one level-$(k + 1)$ cluster. We want to compute the upper tail of the distribution of

$$Y = X_1 + \ldots + X_n$$

from the corresponding tails of $X_1, \ldots, X_n$. All the distributions are represented in the 50-quantile format.

First we choose an independence threshold m. The threshold depends on k as indicated in the table:

Level (k)	Independence threshold (m)
0	1
1	34
2	44
3	48
4	50

The thresholds for $k = 1, 2, 3$ are rather arbitrary, chosen in the middle between the corresponding quantile indices. The threshold of 1 for $k = 0$ means that everything is dependent within level-1 clusters. The thresholds may be modified as necessary.

The joint probability matrix is constructed for the chosen m. Then the distributions of X_1 and X_2 can be clumped and replaced by the resulting distribution of $X_1 + X_2$. Similarly X_3 and X_4 get replaced by $X_3 + X_4$ etc. Then the resulting sums are clumped until the approximate distribution of $Y = X_1 + \ldots + X_n$ is

obtained. Successive aggregations over clusters lead to the final regional distribution represented by the 50 quantiles.

4.7 Trend detection

Change detection may be performed by running a statistical test to compare the resulting (regional) distributions. An appropriate variation of the Kolmogorov-Smirnoff test may be used to deal with just the tails of the distributions – in fact the values of unknown quantiles may all be assumed zero in case of damages.

5 Simple modeling of river flow

A simple scheme of hydrological modeling may be implemented:

1. Get Hydro1K elevation data. It is hydrologically correct.
2. Get precipitation and temperature fields from reanalysis and climate model output.
3. Compute evapotranspiration from precipitation and temperature, based on the formula:
4. Compute total runoff as the difference: precipitation minus evapotranspiration.

This leads to a dynamic hydrological model that may be called over-simplified. However, it need not be appropriate for operational use. The important feature is that it does capture the direction of projected changes in the properties of river flow.

Given a point of interest the time series (or two time series in case of two-sample analysis) may be studied for trends using time series methods. The analysis may also be applied to a number of points. In that case, again, a global result may be obtained using multivariate resampling, as described in the first chapter. In this case it would be precipitation and temperature fields that one should resample, and then, for each randomization, the hydrological model would be run again and reference hydrological time series would be derived. For this reason alone, the hydrological model employed must be very simple.

6 Advanced hydrological modeling

Much more advanced hydrological models exist and are well described. Such models may reflect river flow very well, but are generally possible to construct on a local scale only, as they demand a lot of effort (specific to the river) and data (available locally or at least not globally). In the future it may be possible to construct a universal hydrological model that performs well for all known rivers. However, for the present project such models are not considered.

7 Uncertainty assessment and comparison between scenarios

If a large number of climate-model simulations are available, they may be utilized depending on their number. A large number of different runs of a single model under the same scenario may serve as a basis for Monte-Carlo methods, as described in the first chapter. At present such a large number is usually not yet available. More probably several (say, three) simulations would be available for a given model and scenario. They may be used as an aid in Monte-Carlo approach, i.e. each may serve as a basis for resampling. Another possible approach (to be used in particular when it is important to preserve the relations between the data coming from the same model run) is to compute the mean of the global trend indices of the three runs if the results are similar (i.e. the maps of local trends are similar). Similarly, one may compute the mean of *p*-values in the area – *p*-value relationships of the runs. Comparison of results obtained using different models is only possible on the level of conclusions. Computing the mean global *p*-value may be too restrictive in this case. It would make more sense to

present the distribution of *p*-values across models, provided the results agree. Results in the form of a statement like *It gets drier over most of the World in [season]* can be summarized between the models and thus a frequency table could summarize results from a large number of different models. Such an approach could be used with model data from the Ensembles project.

8 Computations with gridded datasets

Computations with gridded data sets often require specific technical problems to be solved. The main problem is that the size of the data is much larger than the available memory. The other is that data are available in different formats (text, NetCDF, other binary formats) and conversion is not always an option. Third, libraries to manipulate such data are not always easy to use, i.e. a large part of the programmer's effort goes into reading/writing the data even using a designated library. Then, using a different dataset in a different format requires major changes in the program.

To address these problems a dedicated library is being developed. It will allow reading/writing of gridded data in ASCII format and in NetCDF format (through the official NetCDF library) with a possibility of adding other formats as necessary. The library will be easy to use: simply passing the name of a data file will make its data available as a collection of virtual multi-dimensional arrays from which appropriate slices may be read into memory. The library takes care of appropriate caching of data, so that the resulting performance is very good. It must be noted that, while the NetCDF file itself is self-describing, ASCII data must be accompanied by a description file with references to the actual data files (using relative paths, so data can be read over network). As a special function, the library may also be able to analyze and auto-detect the format of and ASCII grid file.

Bibliography

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

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

Kundzewicz Z.W. and Radziejewski M. Methodologies for trend detection. In: *Climate Variability and Change-Hydrological Impacts (FRIEND*). S. Demuth, A. Gustard, E. Planos, F. Scatena & E. Servat (Eds), IAHS Publ. 308, 538-550, 2006.

Kundzewicz, Z. W., Radziejewski, M., Pińskwar, I., Precipitation extremes in the changing climate of Europe. *Clim. Res*. 31: 51–58.

Kundzewicz, Z. W. and Robson, A. (eds.) *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, 2000.

Lins, H. F. 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, 2000.

Livezey R.E. and Chen W.Y. Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, 111: 46–59, 1983.

Radziejewski M., About trend detection in river floods, In: *IN EXTREMIS, Extremes, Trends, and Correlations in Hydrology and Climate,* Juergen P. Kropp and Hans Joachim Schellnhuber (Eds), Springer, accepted for publication.

Radziejewski M. and Kundzewicz Z. W. Detectability of changes in hydrological records. *Hydrol. Sci. J.* **49**(1), 39-51, 2004.

Robson, A. and Chiew, F. 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, 2000.

von Storch H. and Zwiers F. W. *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge, 1999.