Figure1.

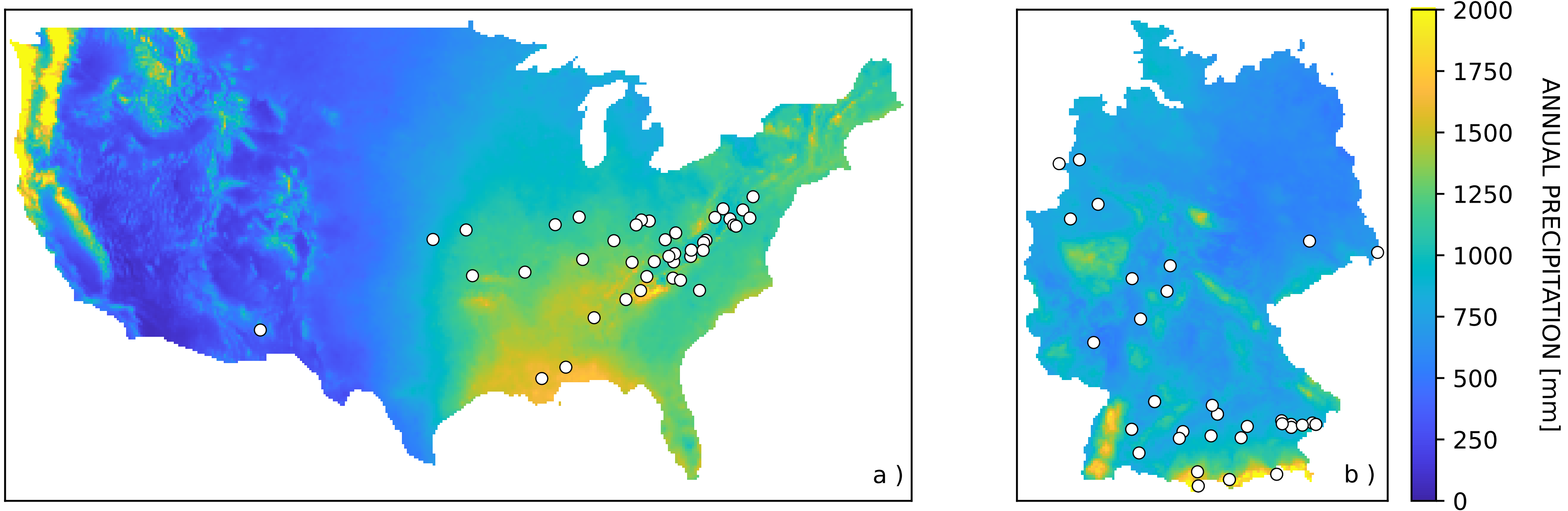

a)
b)
2000
1750
1500
1250
1000
750
500
250
0
ANNUAL PRECIPITATION [mm]

Figure2.

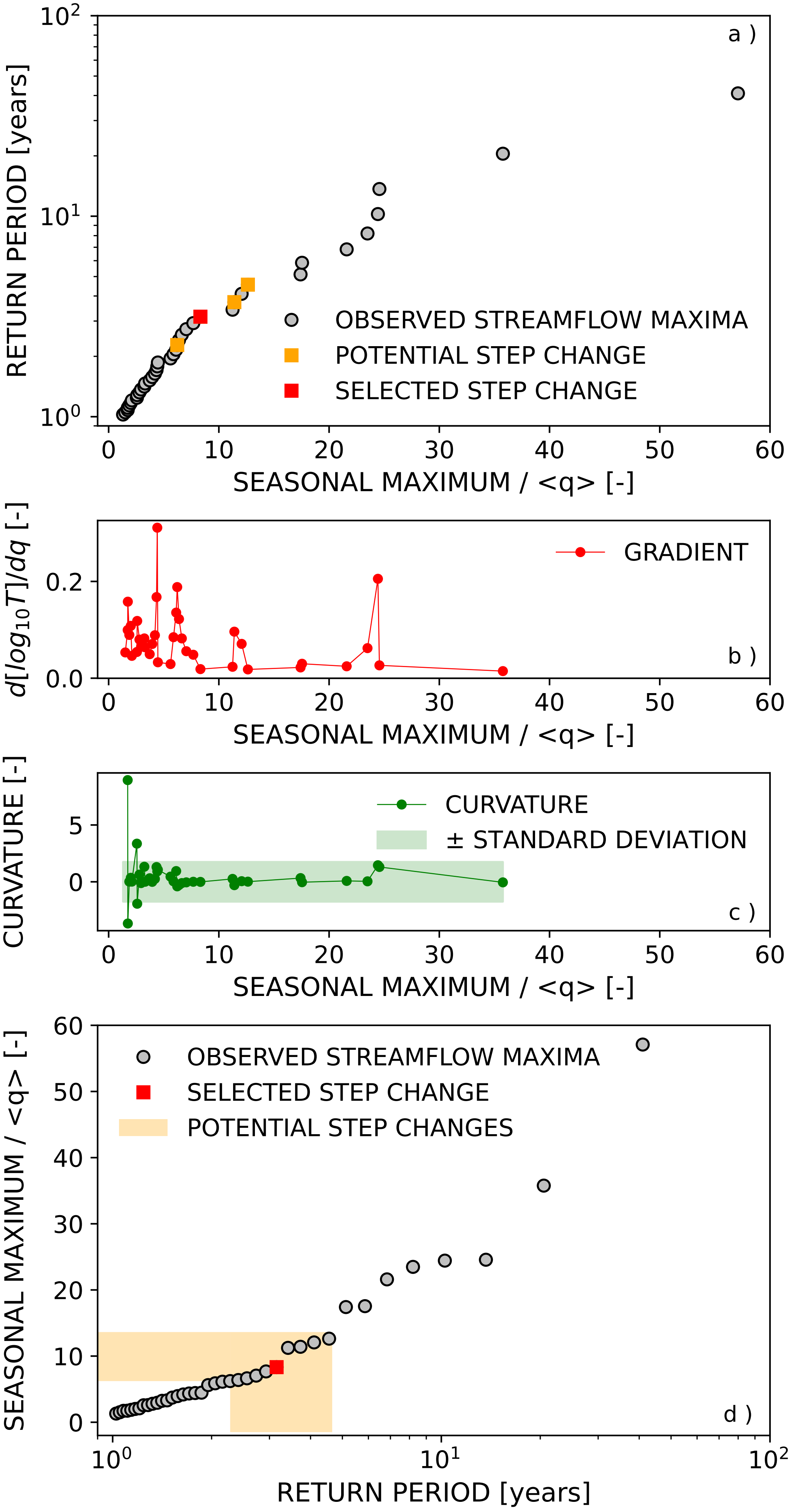
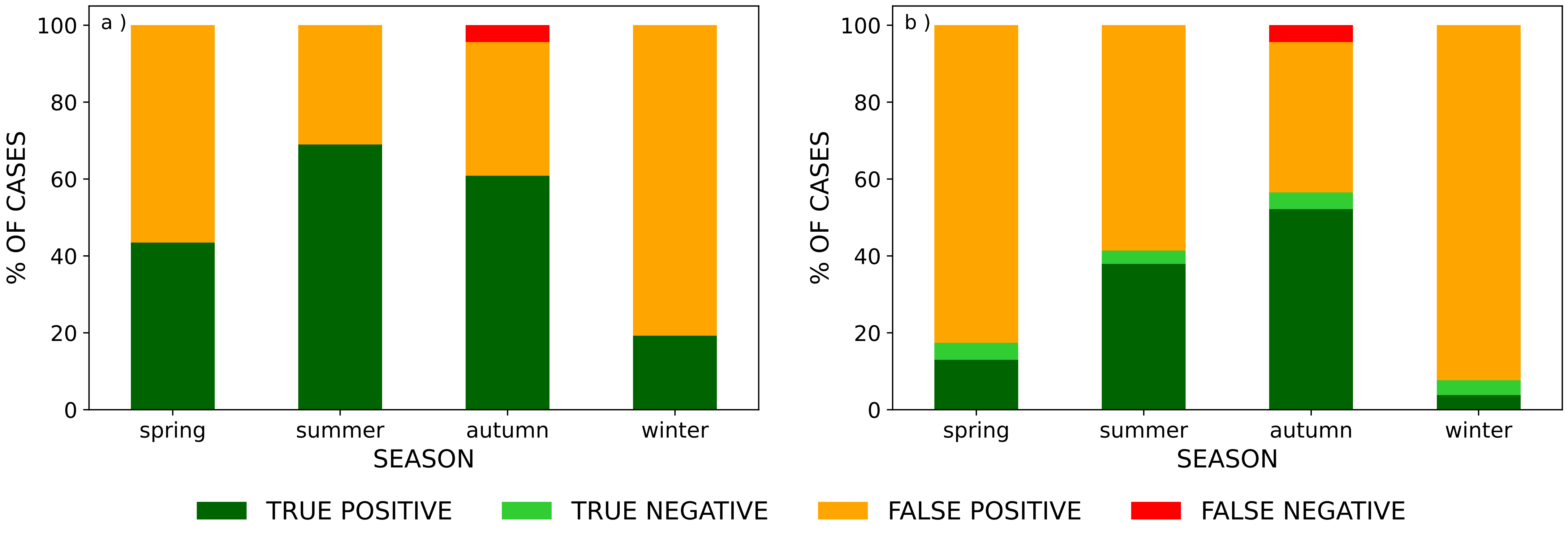
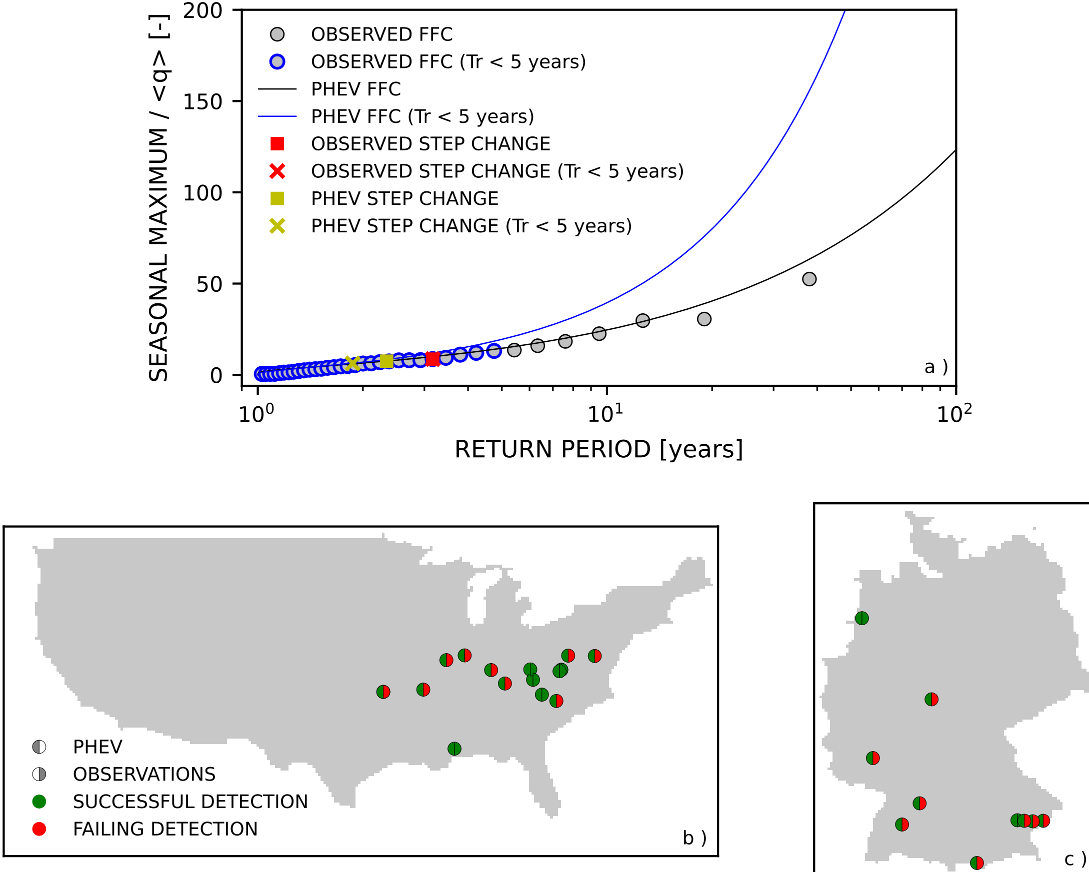

a)
RETURN PERIOD [years]
SEASONAL MAXIMUM / <q> [-]
OBSERVED STREAMFLOW MAXIMA
POTENTIAL STEP CHANGE
SELECTED STEP CHANGE
b)
$d[log_{10}T]/dq$ [-]
GRADIENT
c)
CURVATURE [-]
CURVATURE
± STANDARD DEVIATION
d)
SEASONAL MAXIMUM / <q> [-]
RETURN PERIOD [years]
OBSERVED STREAMFLOW MAXIMA
SELECTED STEP CHANGE
POTENTIAL STEP CHANGES

Figure3.



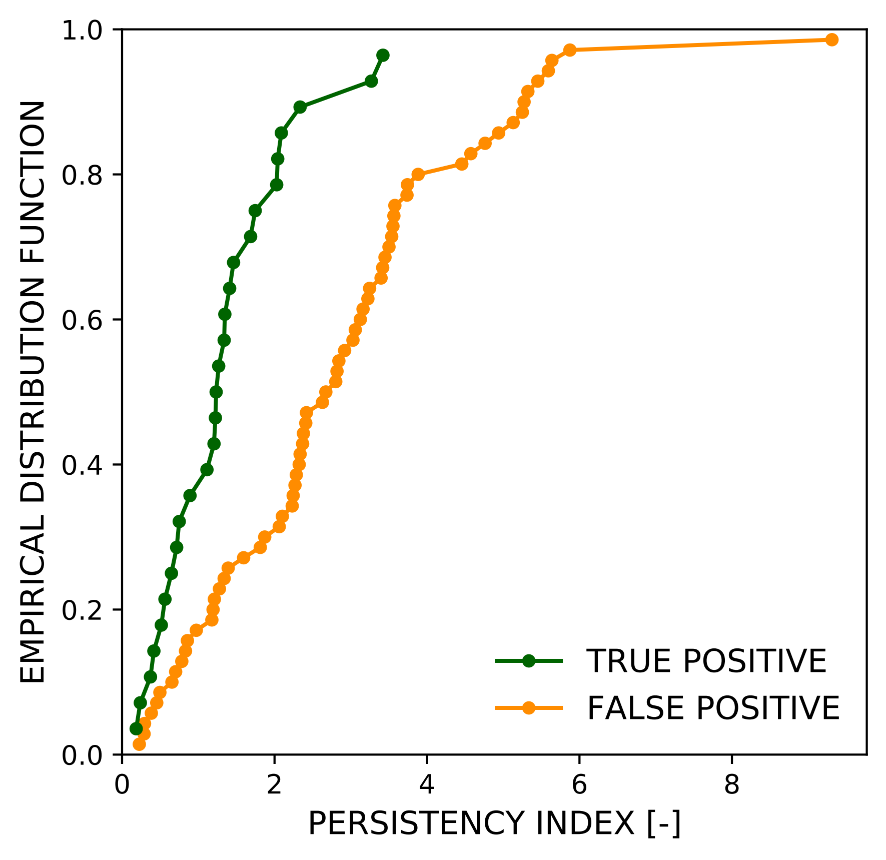
a)
b)
% OF CASES
SEASON
spring
summer
autumn
winter
0
20
40
60
80
100
TRUE POSITIVE
TRUE NEGATIVE
FALSE POSITIVE
FALSE NEGATIVE

Figure4.



SEASONAL MAXIMUM / <q> [-]
RETURN PERIOD [years]
200
150
100
50
0
10^0
10^1
10^2
OBSERVED FFC
OBSERVED FFC (Tr < 5 years)
PHEV FFC
PHEV FFC (Tr < 5 years)
OBSERVED STEP CHANGE
OBSERVED STEP CHANGE (Tr < 5 years)
PHEV STEP CHANGE
PHEV STEP CHANGE (Tr < 5 years)
a)
PHEV
OBSERVATIONS
SUCCESSFUL DETECTION
FAILING DETECTION
b)
c)

Figure5.



EMPIRICAL DISTRIBUTION FUNCTION
PERSISTENCY INDEX [-]
TRUE POSITIVE
FALSE POSITIVE
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8

Identifying discontinuities of flood frequency curves

Arianna Miniussi[1], Ralf Merz[1], Lisa Kaule[1,2], Stefano Basso[1]

[1]Department Catchment Hydrology, Helmholtz Centre for Environmental Research - UFZ, Halle (Saale), Germany
[2]Department of Hydrology, Unversity of Bayreuth, Bayreuth, Germany

Key Points:

- Discontinuities (aka step changes) in flood frequency curves hinder the estimation of rare and extreme floods
- We provide an objective and robust methodology for detecting discontinuities of observed flood frequency curves
- We prove the reliability of a physically-based distribution of river flows in detecting the emergence of step changes

Corresponding author: Stefano Basso, stefano.basso@ufz.de

Abstract

Discontinuities in flood frequency curves, here referred to as step changes, hinder the estimation of high return levels of streamflow. In this paper, we develop a robust and objective methodology for the detection of step changes, and apply it to a large dataset of catchments in the USA and Germany. Given the statistical uncertainty of observed time series due to their limited sample size, we then assess the reliability of a PHysically-based Extreme Value (PHEV) distribution of river flows to identify catchments that might experience a step change. Results show that PHEV is suitable for step changes detection, with a high correct detection rate especially in the autumn and summer seasons, whereas it tends to often show a step change not visible in the observations in spring and winter (seasons typically characterized by persistent flow regimes with reduced likelihood of exhibiting relatively large floods), for which we examine the possible reasons. By means of a controlled experiment we re-evaluate the step change detection method on true positive cases (i.e., when both observations and PHEV display a step change) discarding the highest maxima. PHEV confirms its capability to detect a step change, as observed in the original flood frequency curve, even if the shortened one does not show it. These findings prove the reliability of PHEV for the identification of step changes, especially relevant in scarce data regions, and set the premises for a deeper investigation of physiographic and hydroclimatic attributes controlling the emergence of discontinuities in flood frequency curves.

Plain Language Summary

Estimation of rare and extreme floods is an on-going challenge and is crucial for many applications, for example the design of hydraulic structures, planning of mitigation measures and the underwriting process in the insurance industry. In several cases flood frequency curves, a widely used tool representing the magnitude of flow maxima versus their average observed recurrence interval (named return period) display a discontinuity, meaning that some high floods occur more frequently than expected. This feature is hard to estimate by analyzing observations only, since the uncertainty due to the limited length of their time series prevails. Here, we first identify discontinuities (step changes) in the observed flood frequency curves, and then leverage a parsimonious physically-based extreme value distribution (PHEV) for their detection. Being PHEV suitable for detecting the emergence of step changes, this analysis prepares the ground for a more in-depth investigation of the physical/climatic features that make some catchments more unpredictable than others.

1 Introduction

Despite the efforts of researchers for achieving more reliable estimation of rare floods, these events are still the primary natural disasters (Wallemacq & House, 2018). The evaluation of their hazard is yet crucial for several applications, among which the design of hydraulic structures, risk planning and mitigation, or computation of premiums in the insurance industry. Floods come unexpectedly and the surprise element is often neglected when evaluating their risk and planning their management (B. Merz et al., 2015). Understanding the reasons that lead to their occurrence and timely identifying hazardous catchments is therefore a fundamental goal of flood research (e.g., B. Merz et al., 2010; Wing et al., 2018).

In recent years, alternatives have been proposed to the standard flood frequency analysis, i.e., the fitting of a distribution on streamflow maxima and its performance evaluation by means of goodness-of-fit tests, whose criticism dates back to Kleme (1993). Some studies proposed to consider flood peaks as the product of two random variables, namely the runoff coefficient and a characteristic rainfall, yet focusing on synthetic experiments (Gaume., 2006; Viglione et al., 2009). R. Merz and Blschl (2008a, 2008b) and Viglione et al. (2013) proposed to extend the traditional flood frequency analysis by a systematic expansion of information

beyond the flood sample, to fully capture the subtleties of the flood characteristics. Others attempt to provide a more accurate description of the processes underlying the emergence of extremes from either a statistical (Marani & Ignaccolo, 2015; Zorzetto et al., 2016; Miniussi et al., 2020) or physically-based with stochastic components (Botter et al., 2009; Basso, Schirmer, & Botter, 2016) viewpoint.
Rogger et al. (2012) focused their analyses on sudden increments of the flood frequency curve, which they called step change. In fact, if step changes are not mere statistical artefacts, caused for instance by the limited length of the available observational sample, they might be highly critical for the appraisal of flood hazard (Rogger et al., 2013). Rogger et al. (2012) first researched step changes in flood frequency curves conducting tailored field studies in two small (i.e., with area lower than 100 km^2) alpine catchments with the aim of understanding the reasons behind their emergence. Leveraging detailed information collected from field surveys, they calibrated a distributed deterministic rainfall-runoff model and suggested that the step change occurs when a threshold of the storage capacity of the catchment is exceeded. Based on this preliminary work, Rogger et al. (2013) performed a synthetic experiment to quantitatively examine the effects of catchment storage thresholds on step changes in the flood frequency curve and analyzed the combined effect of multiple controls (among others, the temporal variability of antecedent soil storage and the size of the saturated regions) to investigate how they influence the return period of occurrence of the step change.
From another perspective, Guo et al. (2014) linked the shape of the flood frequency curve with the aridity index (i.e., the ratio between mean annual potential evaporation and precipitation, Budyko (1974)), showing that flood frequency curves characterized by increasing aridity index are steeper. Along this line, Metzger et al. (2020) showed that the tail of extreme value distributions in arid/semi-arid watersheds is heavier compared to the one describing Mediterranean watersheds. Basso, Schirmer, and Botter (2016) instead explained different shapes of the flood frequency curve in terms of the persistency index (i.e., the ratio between mean catchment response time and runoff frequency, Botter et al. (2013)), and highlighted that the concavity of the flood frequency curve changes from downward to upward shifting from persistent to erratic regimes. Diverse shapes of the flood frequency curve were also linked to different flood-generating processes (R. Merz & Blschl, 2003; Berghuijs et al., 2014; Tarasova et al., 2020) or mixtures of flood event types (Hirschboeck, 1987; Villarini & Smith, 2010; Smith et al., 2018).

Former research found some indications of the possible role played by varied drivers in determining the shape of flood frequency curves. However, a quantitative and robust (i.e., tested in a large set of case studies) methodology to identify step changes, which encompasses the critical interactions among hydrological processes in river basins, is still lacking. In search for a suitable approach to characterize these behaviors, we observe that purely statistical models, whose parameters are fitted on observed flood values, either tend to significantly underestimate rare events or provide estimates which remarkably vary with the available observational sample (e.g., Laio et al., 2010; Viglione et al., 2013). High dimensional hydrological models as well suffer from large uncertainty in the identification of their parameters due to the limited information available (Beven, 2006; Her et al., 2019; Seibert et al., 2019), questioning their reliability in understanding the processes underlying the occurrence of extremes. Conversely, process-based stochastic models proved themselves powerful tools in water science for inferring behaviors of complex systems (e.g. Porporato (2021); Montanari and Koutsoyiannis (2014); McGrath et al. (2019); Bertassello et al. (2020)). By accounting for both the stochastic character of climate conditions and the internal catchment dynamics, such approaches might provide valuable insights for the identification of step changes of flood frequency curves.

The relevance of our study is twofold: (i) we develop an objective robust methodology for the detection of step changes and evaluate their emergence across the US and Germany, in a large set of catchments with contrasting climatic and physiographic characteristics; (ii) we examine the reliability of a process-based stochastic framework for the estimation of

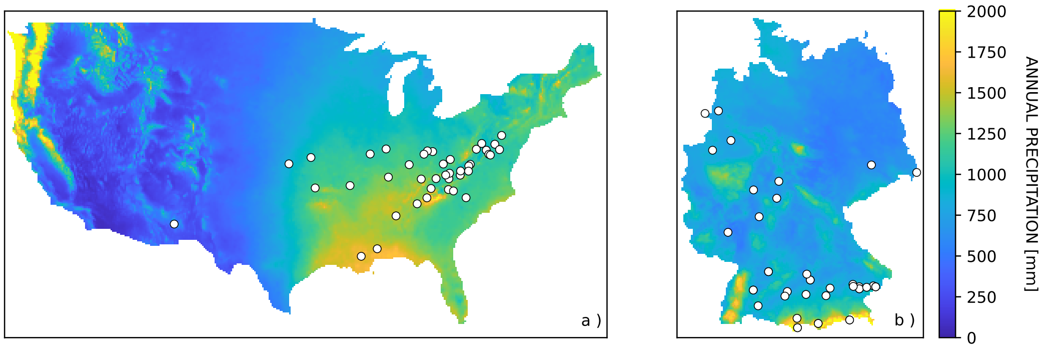


Figure 1. Select river basins (white filled circles) from the (A) MOPEX and (B) German datasets. The background of the maps represents 30-years annual precipitation normals (1981-2010 for the US and 1991-2020 for Germany).

flood frequency curves to detect step changes and infer their occurrence.
After describing the data analyzed, we briefly summarize the theoretical foundations of such a mechanistic-stochastic framework (i.e., the PHysically-based Extreme Value (PHEV) distribution of river flows) and explain the methodology for step changes identification. We then present and discuss our findings, and recapitulate main results and future developments in the conclusions.

2 Data and Methodology

2.1 Datasets

We analyze daily streamflow time series from the Model Parameter Estimation Experiment (MOPEX) dataset (Duan et al., 2005; Schaake et al., 2006) and from Germany (Tarasova et al., 2018) at a seasonal scale, to account for the seasonality of rainfall and runoff (Allamano et al., 2011; Baratti et al., 2012). In order to comply with the physical assumptions underlying PHEV, we consider only catchments with low human impact and weak or absent inter-seasonal snow dynamics (Botter et al., 2013; Wang & Hejazi, 2011). Hydrograph recession properties should as well not consistently vary with the peak flow (Basso et al., 2015). A fairly accurate estimation of the flood frequency curve by PHEV is a precondition to investigate if it is able to correctly identify the occurrence of step changes, and for understanding their possible relation with catchment properties embedded in PHEV. We will thus limit our analysis to cases in which the root mean square error between the whole predicted and observed flood frequency curve is limited (lower than 0.3), aiming to minimize step changes missclassifications due to error noise.
The gauges remaining after this cascade selection process are shown in Figure 1.

2.2 The PHysically-based Extreme Value (PHEV) distribution of river flows

PHEV is a parsimonious mechanistic-stochastic formulation of flood frequency curves that stems from a rigorous mathematical description of catchment-scale daily soil moisture and streamflow dynamics in river basins (Laio et al., 2001; Porporato et al., 2004; Botter et al., 2007). In this framework, precipitation is represented as a marked-Poisson process with frequency λ_P $[T^{-1}]$ and exponentially-distributed depths with average value α $[L]$. Soil moisture decreases due to evapotranspiration and is replenished by precipitation events that eventually trigger runoff pulses when an upper wetness threshold is crossed. These pulses, which feed water to a hydrologic storage, are also a Poisson process with frequency λ_P $[T^{-1}]$ and an exponential distribution of magnitudes with mean α $[L]$. A non-linear (i.e., power-law) storage-discharge relation epitomizes the hydrological response of the catchment and encompasses the joint effect of different flow components (Brutsaert & Nieber, 1977; Basso et al., 2015).

The above-summarized mechanistic-stochastic description of runoff generation processes allows for expressing the probability distributions of daily flows (Botter et al., 2009), peak flows and flow maxima (Basso, Schirmer, & Botter, 2016) as a function of a few physically meaningful parameters. In particular, the PHysically-based Extreme Value distribution of river flows reads:

$$p_M(q) = \lambda\tau \exp(-\lambda\tau D_j(q))\, p_j(q) \tag{1}$$

where τ [T] is the duration in days of the time frame chosen for the analyses (e.g., a season); $p_j(q)$ is the probability distribution of peak flows, $p_j(q) = Cq^{1-a}\exp(\frac{\lambda q^{1-a}}{K(1-a)} - \frac{q^{2-a}}{\alpha K(2-a)})$; $D_j(q) = \int_q^{\infty} p_j(q)\, dq$ is the exceedance cumulative probability of peak flows; α and λ are the aforementioned parameters describing Poisson-distributed effective rainfall events, a and K are the parameters of the power-law storage-discharge relation, and C is a normalization constant.

2.2.1 Parameter Estimation

The four parameters of PHEV (α, λ, a, K) are rather straightforward to estimate at the catchment scale. They are indeed directly derived from the observed time series of precipitation and streamflow: α is computed as the average rainfall depth during events, while λ (frequency of streamflow-producing rainfall) as the ratio between α and the long term average streamflow $< q >$ (Botter et al., 2007). The parameters of the power-law storage-discharge relation (i.e., the recession exponent a and coefficient K) are estimated through hydrograph recession analysis (Brutsaert & Nieber, 1977) following the approach proposed by Biswal and Marani (2010), who noted that dQ/dt vs Q curves in a log-log plot can show signicant deviations from one another within the same catchment. Finally, the recession coefficient is not directly used as input in Eq. (1), but it is replaced by its maximum likelihood estimation on the observed seasonal flood frequency curve (Basso, Schirmer, & Botter, 2016). Therefore, although being a 4-parameter distribution, PHEV requires the actual calibration of one parameter only. Further details concerning parameter estimation methods are available in Basso, Schirmer, and Botter (2016) and Dralle et al. (2017).

2.3 Step Change Identification

Following Rogger et al. (2013), a step change is here defined as the sharpest bend of the flood frequency curve. We thus propose a methodology dedicated to its identification from both empirical estimates of the flood frequency curve obtained by means of Weibull plotting position and the analytical model. Some modifications are yet needed when applying it to the observations, in order to deal with the noise affecting the computation of derivatives when only a discrete and rather sparse set of observations is available. The resulting approach can be therefore further employed without depending on subjective evaluation.

1. The curvature of the flood frequency curve, of which we show an example in Figure 2, is computed as $logTr''/(1 + logTr'^2)^{(3/2)}$ (where the apex indicates the derivation operation with respect to q) for both the observations and PHEV. In the former case, we use the method developed by Jianchun et al. (1994) for computing derivatives in non-equally spaced points, while for PHEV we employ the Python routine from the Scipy library (*misc.derivative*), which uses a central difference formula with spacing dx to compute the n^{th} derivative at a specified point.
2. As the noise associated to computing the curvature on a non smooth set of points (seasonal maxima) might lead to identification errors, a filter is applied on the curvature calculated from observations: only points on the right-hand side of the last value

of the curvature exceeding the range $\pm\sigma$ (where σ indicates the standard deviation of the curvature itself), are considered (Figure 2c);

3. The Mann-Whitney U-test (Mann & Whitney, 1947) is applied on the values of the first derivatives on the left and right-hand sides of each potential step change identified at point 2 to check if their distributions are statistically different at a significant level α equals to 0.05 (in other words, if the slope of the curve significantly differs between the left and right-hand side of the step change); the effect size is then computed by means of the Cohen's d (Cohen, 1974) to evaluate if the magnitude of the difference is relevant (Sullivan & Feinn, 2012). For PHEV, this step is performed on a dense set of values, equally spaced with an interval Δq = 0.05 up to a value of normalized streamflow equal to 200, i.e., 200 times the long-term average streamflow.
4. We finally select the point for which the p-value of the Mann-Whitney test is the lowest, provided that the Cohen's d is greater than 0.4 (small to medium effect size, according to Cohen (1974)).

Figure 2 visually exemplifies the application of the developed approach for step changes detection to the flood frequency curve of the Rott river at Kinning, Bavaria (ID: 18801005), in the summer season. In Figure 2a the flood frequency curve is represented with switched axes (i.e., the logarithm of the return period is represented on the y-axis whereas the normalized seasonal maxima on the x-axis), as streamflow is the independent variable in Eq. (1). The red square in Figure 2a-d represents the selected step change, i.e., the one associated to the lowest p-value of the Mann-Whitney U-test applied to the distributions of the first derivatives (Figure 2b) and fulfilling the additional criterion on the Cohen's d value, which must be greater than 0.4. We also show the points that could be considered as potential step changes (i.e., all the points with a Mann-Whitney p-value lower than 0.05, orange squares in Figure 2a). It is though important to stress that here we are most interested in the presence/absence of a step change, rather than in its exact position.

3 Results and Discussion

We apply the methodology for the identification of step changes introduced in the previous section to each observed and analytic seasonal flood frequency curve, thus allowing for the evaluation of the step change detection rate of PHEV with respect to the observations (Figure 3). The bar plots in Figure 3 show the percentage of case studies for which a step change is identified from both PHEV and the observational records (true positives, dark green color), the fraction of cases which display a step change neither in the empirical nor in the analytic flood frequency curves (true negatives, light green), those where a step change is detected from the observations but not from the analytical model (false negatives, red), and those where the analytical model has foreseen the occurrence of a step change which is not confirmed by the available observations (false positives, orange).
The bar plots in Figure 3a and 3b differ for the criteria applied in the step changes identification methodology. In Figure 3a only the controls on the p-value of the Mann-Whitney U-test mentioned in Section 2.3 are considered, whereas the additional requirement on the effect size is as well used in Figure 3b. True positives (dark green) prevail in the summer and autumn seasons of Figure 3a, amounting to about 70% and 60% of the cases. False positives constitute instead a sizable share of the cases in spring and winter. When more stringent requirements for the identification of step changes are used, by accounting for the additional criterion on the effect size, the percentage of true positives decreases (Figure 3b, dark green). A few cases of those shifting category become true negatives, indicating that the slope of the flood frequency curve does not substantially increases on the right-hand side of the potential step change (although its change is statistically significant), thus not representing a noteworthy hazard. Most of them however become false positives (orange color in Figure 3b), as PHEV confirms the existence of a step change thanks to its evaluation in an arbitrary and not limited number of points. Conversely, the limited lengths of the data records prevent us from identifying statistically robust substantial changes of the slope of

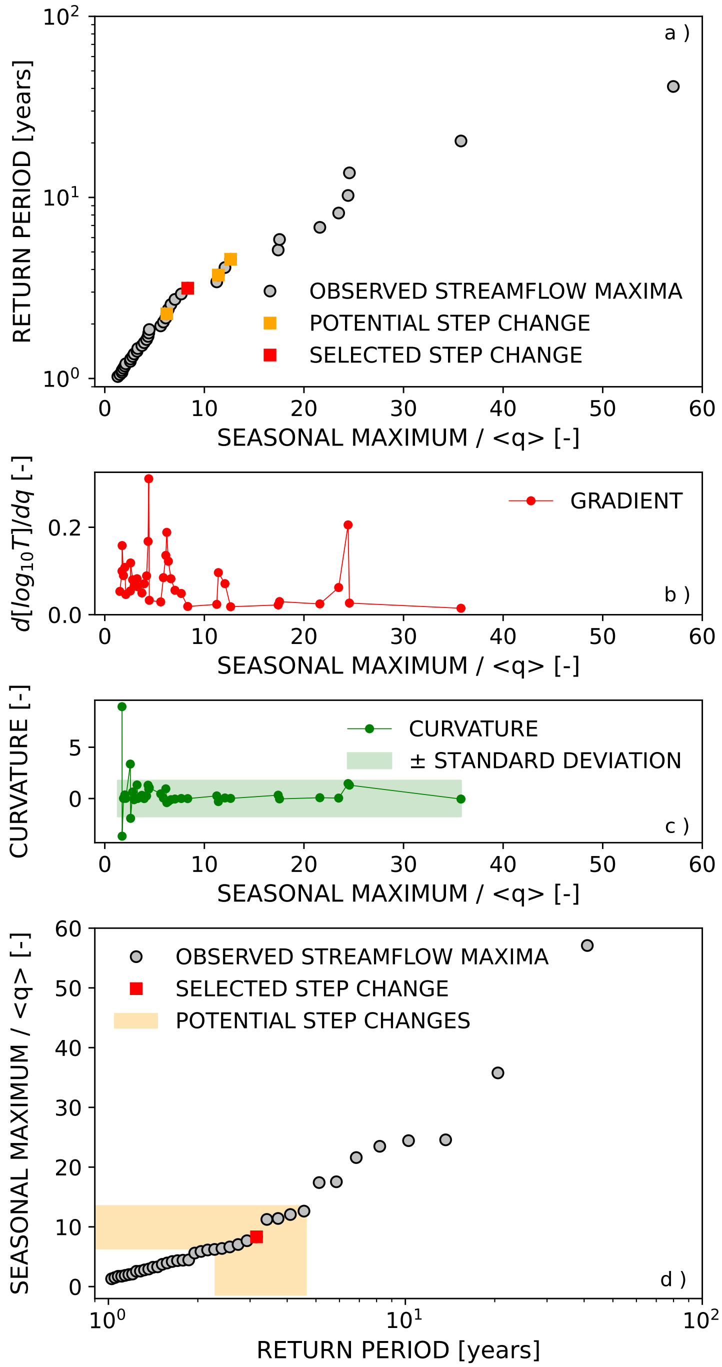


Figure 2. Exemplary application of the proposed methodology to detect step changes to the Rott river at Kinning, Bavaria (ID: 18801005), in the summer season. a) Visualization of how the approach is actually applied, i.e., expressing the logarithm of the return period as a function of the normalized seasonal maxima (gray filled circles). Potential step changes (i.e., all the points with a p-value of the Mann-Whitney test lower than 0.05) are represented by orange squares, while the selected one (i.e., the one exhibiting the minimum p-value of the Mann-Whitney test and Cohen's d greater than 0.4) is depicted with a red square. b) First derivative computed on observations. c) Curvature computed on observations, with the shaded area representing twice its standard deviation. d) Standard representation of the flood frequency curve, namely observed maxima as a function of the logarithmic value of the return period (gray filled circles). The red square indicates the selected step change, while the orange shaded area represents the range variability of the potential step changes.

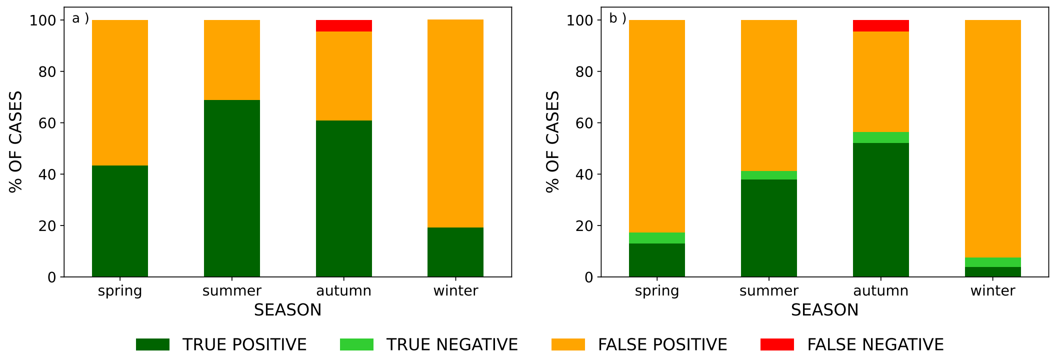


Figure 3. Performance of the PHysically-based Extreme Value (PHEV) distribution of river flows in the detection of step changes when only the controls on the Mann-Whitney U-test are applied (see Section 2.3, panel a) and when both the Mann-Whitney U-test and the Cohen's d are considered (panel b). True positives (dark green color) and true negatives (light green) indicate coherence between PHEV and observations, i.e., step changes are either detected or not from both PHEV and the observed records. These constitute a large number of cases in summer and autumn. False positives (orange) and false negatives (red) represent the cases in which either PHEV detects a step change that was not identified by the observations (in the majority of the remaining cases) or the observations display a step change which is not detected by PHEV (only one case). The reasons for the presence of false positives are further investigated in the study and clarified in the text and figures.

the flood frequency curves from plain observations. Consistent results are also found when considering different significant levels for the Mann-Whitney test: the strictest the level the highest the share of cases shifting between true and false positives, which once again points to the unfeasibility of detecting step changes with confidence from plain observations. The existence of both true positives and true negatives emphasizes the capability of PHEV to mimic varied observed shapes of flood frequency curves (Basso, Schirmer, & Botter, 2016) and to identify both the presence and the absence of a step change. One single false negative (i.e., a case where a step change is detected from observations but not foreseen by the model, red color in Figure 3b) occurs in the entire dataset. Whereas either PHEV limitations or statistical uncertainty intrinsic to relatively short data samples might cause such an occurrence, the existence of only one such case further confirms the remarkable capabilities of PHEV to spot the occurrence of step changes, especially when these are also detected in the available observations.

The predominance of false positives in spring and winter (orange color in Figure 3) calls for further investigation of their causes. Our hypothesis is that PHEV, by leveraging the embedded mechanistic description of hydro-climatic dynamics taking place in watersheds and the information gained from analyzing daily rainfall and streamflow series, might indicate the possible emergence of step changes that are not yet displayed by the observed flood frequency curves. In fact, these empirical estimates are likely affected by small sizes of the samples of large events (i.e., those on the right-hand side of each potential step change, see Figure 2a) and by the specific character of catchments, which may have a more or less enhanced propensity to exhibit extreme floods and thus display them in a limited data record. To test this hypothesis, we perform the following experiment. We consider the set of true positives (i.e., the 27 cases for which both PHEV as well as the observed flood frequency curve show a step change) and retain only maxima with return periods below 5 years (see an explanatory example in Figure 4a, where the maxima retained are represented by gray filled

circles with blue contours). In so doing, we approximately discard in each case the largest ten points and their corresponding years of occurrence. Thereby, fictitious flood frequency curves only comprising maxima with smaller magnitudes (and return periods) are created, thus reproducing the conditions we hypothesized as possible reasons of the emergence of false positives. We then apply the usual methodology for identifying step changes on these fictitious flood frequency curves and the corresponding shortened data records.
PHEV detects a true step change (i.e., true positives) in 100% of the cases even when the largest points are removed, whereas the observations only in 37%. The maps in Figure 4b and 4c summarize this result: half circles are colored either in green, if a step change is successfully detected from the shortened flood frequency curve, or in red in the opposite case. The left half of the circle depicts the detection capability of PHEV, while the right side the results obtained from the observations. It can be easily seen that all left halves of the circles are colored in green and most of the right ones are instead red, thus indicating a 100% success rate of PHEV and a significantly lower success rate of observations in inferring the emergence of step changes from shortened records. A similar result is obtained when discarding maxima with return period greater than 10 years (i.e., discarding about five-six points instead of the highest ten). Only when retaining but the very left part of the flood frequency curve (return periods lower than 2 years) PHEV detection rate decreases. Also in this case, however, PHEV notably detects a discontinuity in 60% of the cases.
The outcome of this experiment strongly suggests that the detected false positives (orange color in Figure 3) indeed arise because of the statistical uncertainty of limited data records and the capability of PHEV to infer the occurrence of step changes rather than by its inability to correctly identify inflection points which were detected (or not) in the observed flood frequency curves.

A physical explanation of the reason why some observational series might not exhibit a step change, although this is expected according to PHEV estimates, is provided by considering typical streamflow dynamics occurring for distinct river flow regimes, here characterized by means of the persistency index φ (Botter et al., 2013). This index classifies hydrologic regimes into erratic ($\varphi < 1$) and persistent ($\varphi > 1$). An erratic regime, which is commonly found during dry seasons, very hot humid seasons with intense evapotranspiration or in fast responding catchments, is characterized by periods between the arrival of runoff-producing rainfall events which are longer than the typical duration of flow pulses. Conversely a persistent regime, typically occurring in cold-humid seasons and lowland catchments, is characterized by frequent rainfall events and a rather constant water supply to the catchment. In terms of flow dynamics, the behaviors of these different flow regimes can be explained as follows: when streamflow values weakly oscillate around their mean (persistent regimes), the probability of occurrence of relatively large flows is very low, and extreme events are unlikely to be captured by short time series. On the contrary, erratic regimes are composed of a sequence of high flows interspersed in between prolonged periods of low flows. Events which are several times (i.e., order of magnitudes) higher than the average flow are thus more likely to occur in these regimes (Basso, Frascati, et al., 2016). In the context of this study, false positives shall does mostly occur for persistent regimes, as such large events allowing for the detection of step changes from empirical flood frequency curves are less likely to have been observed during the available data record.
In Figure 5 we compare the cumulative distributions of the persistency index for the sets of true positives (dark green) and false positives (orange). Clearly, the distributions differ and false positives mostly occur for persistent regimes. This qualitative evaluation is verified by applying the 2-sample Kolmogorov-Smirnoff test, which evaluates if two samples come from the same distribution (null-hypothesis): in this case we can reject the null-hypothesis at the 0.01 significance level, meaning that the two samples are drawn from different distributions and false positives are significantly more likely to occur for persistent regimes, according to the physical explanation provided above. Remarkably, the seasons characterized by the larger portion of false positives are spring and winter, during which regimes tend to be more persistent thanks to a fairly constant water supply.

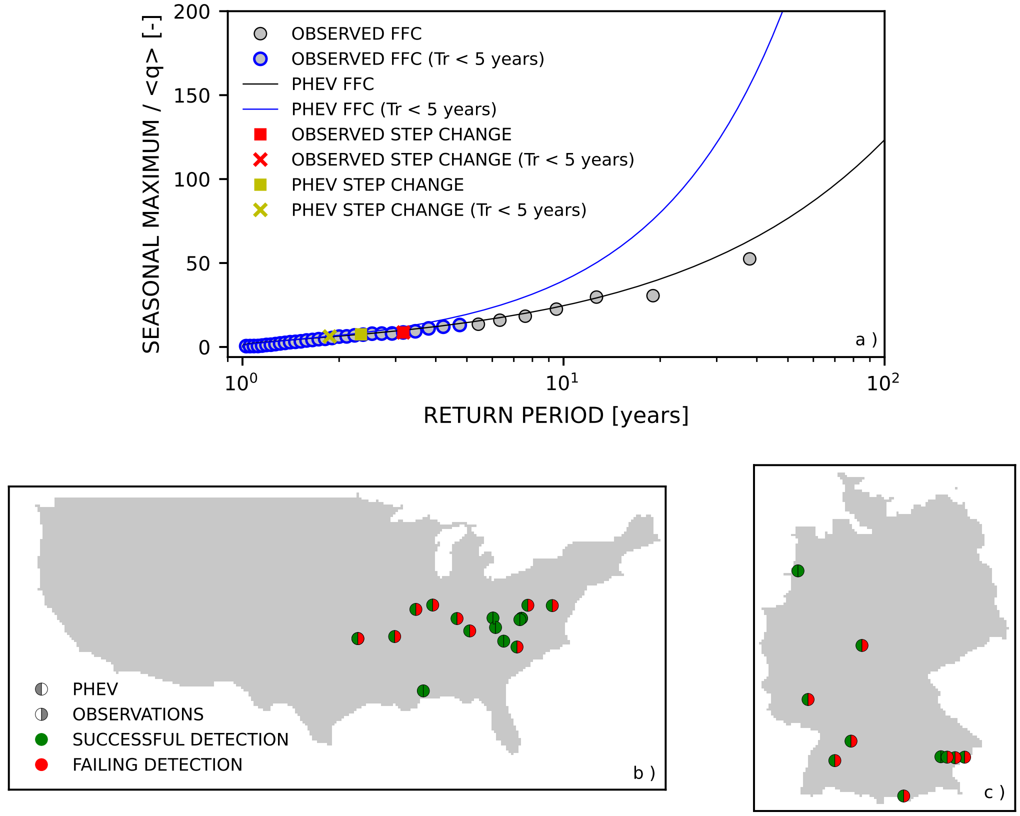


Figure 4. Visual explanation and results of the experiment we perform to test the hypothesis on the emergence of false positives. a) Gray dots with black (blue) contour represent the complete (shortened, until a return period of 5 years) observed seasonal maxima series of the Tug Fork river near Kermit, WV (USGS ID: 03214000), in the autumn season. The solid black (blue) line displays the analytic flood frequency curve (i.e., PHEV) whose parameters are estimated from the complete (shortened) time series. The red (yellow) square indicates the step change detected from the observations (by PHEV) using the complete series, while the corresponding crosses (the red one is not visible in the plot as it overlaps the red square) represent the observed and analytic step changes detected on the shortened flood frequency curve. b-c) Locations of the true positives in the US (panel b) and Germany (panel c). The left (right) half of the circles represent PHEV (observations) ability to detect a step change when the shortened flood frequency curves (i.e., maxima characterized by return period below 5 years) are used. The green (red) colored halves indicate successful (failing) detection. Remarkably, all the left halves are green (PHEV always detects true step changes, i.e., true positives, even from the shortened series), whereas most of the right ones are red (step changes are not always identified from observations when the shortened records are used).

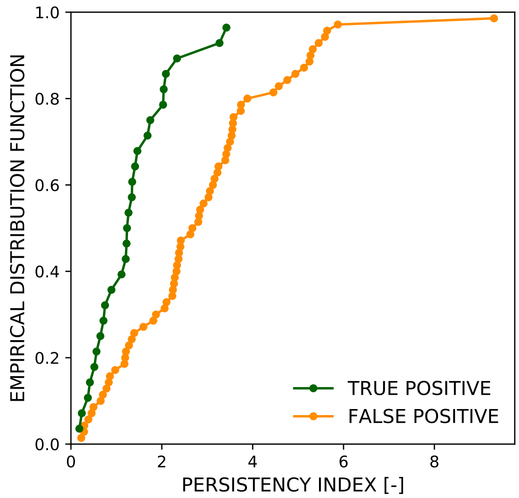


Figure 5. Empirical cumulative distribution functions of the persistency index values dividing the case studies into true positives (dark green) and false positives (orange). The distributions are significantly different in a statistical sense (the p-value of the 2-samples Kolmogorov-Smirnoff test is lower than 0.01.)

4 Concluding Remarks

In this work we examine the occurrence of sharp uprises in flood frequency curves (termed step changes), which are pivotal for a correct estimation of river flood hazard. We develop a robust and objective methodology to identify them from observational records, and employ a parsimonious PHysically-based Extreme Value distribution of river flows (PHEV) to evaluate its capability to reliably detect step changes, thus providing an agile methodology for their identification.

Results show that PHEV is able to recognize the presence/absence of step changes with good coherence in a large set of case studies from the US and Germany. Possible reasons for the occurrence of a sizeable number of false positives are investigated by accounting for both the statistical uncertainty of relatively short observational records and the typical hydro-climatic variability of different river basins, which affects the information content of these limited data series.

To this end, we perform a controlled experiment in which we remove the highest flow maxima in the flood frequency curves of the true positive cases and repeat the step change detection analysis on the shorter series, showing that PHEV can foresee the emergence of step changes even if the shortened observations do not display it. The result supports claims of the dependability of step changes initially classified as false positives.

An investigation of the intrinsic dynamics of streamflows in the set of true and false positives further elucidates the issue. The majority of cases for which false positives are detected feature markedly persistent regimes that, by their nature, rarely exhibit large extreme flow values. The limited length of the available observed time series might be thus constraining the possibility to observe expected step changes, analogously to what occurs when we artificially reduce the size of the observational sample.

The present analysis, thoroughly performed on a wide set of catchments characterized by different hydroclimatic features, reveals PHEV as a reliable tool to identify and foresee the occurrence of step changes and consequently to unveil the propensity of rivers to large floods, with great relevance especially in data scarce conditions. The study lays the foundations for a better comprehension of climate and landscape controls on the emergence of step changes in flood frequency curves, which is the subject of current work.

Acknowledgments

The financial support of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through project number 421396820 "Propensity of rivers to extreme floods: climate-landscape controls and early detection (PREDICTED)", as well as the Helmholtz Centre for Environmental Research - UFZ, are gratefully acknowledged. We thank the Bavarian State Office of Environment (LfU, `https://www.gkd.bayern.de/de/fluesse/abfluss`) and the Global Runoff Data Centre (GRDC) prepared by the Federal Institute for Hydrology (BfG, `http://www.bafg.de/GRDC`) for providing the discharge data for Germany. The MOPEX dataset is available at `https://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/`. 30-year normal precipitation gridded data for the US are provided by the PRISM Climate Group, Oregon State University, `http://prism.oregonstate.edu` (downloaded on June, 1st 2021); 30-year normal precipitation gridded data for Germany are provided by the Deutsche Wetter Dienst (DWD) at `https://opendata.dwd.de/climate_environment/CDC/grids_germany/multi_annual/precipitation/`

References

Allamano, P., Laio, F., & Claps, P. (2011). Effects of disregarding seasonality on the distribution of hydrological extremes. *Hydrol. Earth Syst. Sci.*, *15*, 32073215. doi: doi:10.5194/hess-15-3207-2011

Baratti, E., Montanari, A., Castellarin, A., Salinas, J. L., Viglione, A., & Bezzi, A. (2012). Estimating the flood frequency distribution at seasonal and annual time scales. *Hydrol. Earth Syst. Sci.*, *16*, 4651-4660. doi: 10.5194/hess-16-4651-2012

Basso, S., Frascati, A., Marani, M., Schirmer, M., & Botter, G. (2016). Climatic and landscape controls on effective discharge. *Geophysical Research Letters*, *42*, 84418447. doi: 10.1002/2015GL066014

Basso, S., Schirmer, M., & Botter, G. (2015). On the emergence of heavy-tailed streamflow distributions. *Advances in Water Resources*, *82*, 98 - 105. doi: 10.1016/j.advwatres.2015.04.013

Basso, S., Schirmer, M., & Botter, G. (2016). A physically based analytical model of flood frequency curves. *Geophysical Research Letters*, *43*(17), 9070-9076. doi: 10.1002/2016GL069915

Berghuijs, W. R., Sivapalan, M., Woods, R. A., & Savenije, H. (2014). Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resources Research*, *50*, 5638 - 5661. doi: 10.1002/2014WR015692

Bertassello, L. E., Rao, P. S. C., Jawitz, J. W., Aubeneau, A. F., & Botter, G. (2020). Wetlandscape hydrologic dynamics driven by shallow groundwater and landscape topography. *Hydrological Processes*, *34*, 1460-1474. doi: 10.1002/hyp.13661

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*. doi: 10.1016/j.jhydrol.2005.07.007

Biswal, B., & Marani, M. (2010). Geomorphological origin of recession curves. *Geophysical Research Letters*, *37*(24). (L24403) doi: 10.1029/2010GL045415

Botter, G., Basso, S., Rodriguez-Iturbe, I., & Rinaldo, A. (2013). Resilience of river flow regimes. *Proceedings of the National Academy of Sciences*, *110*(32), 12925-12930. doi: 10.1073/pnas.1311920110

Botter, G., Porporato, A., Rodriguez-Iturbe, I., & Rinaldo, A. (2007). Basin-scale soil moisture dynamics and the probabilistic characterization of carrier hydrologic flows: Slow, leaching-prone components of the hydrologic response. *Water Resources Research*, *43*(2). doi: 10.1029/2006WR005043

Botter, G., Porporato, A., Rodriguez-Iturbe, I., & Rinaldo, A. (2009). Nonlinear storage-discharge relations and catchment streamflow regimes. *Water Resources Research*, *45*(10). doi: 10.1029/2008WR007658

Brutsaert, W., & Nieber, J. L. (1977, 6). Regionalized drought flow hydrographs from a mature glaciated plateau. *Water Resources Research*, *13*(3), 637-643. doi: 10.1029/WR013i003p00637

Budyko, M. (1974). *Climate and life.* Academic Press.

Cohen, J. (1974). *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates.

Dralle, D. N., Karst, N. J., Charalampous, K., Veenstra, A., & Thompson, S. E. (2017). Event-scale power law recession analysis: quantifying methodological uncertainty. *Hydrol. Earth Syst. Sci.*, *21*(1), 65–81. doi: 10.5194/hess-21-65-2017

Duan, Q., Shaake, J., Andreassian, V., S., F., Goteti, G., Gupta, H., ... Wood., E. (2005). Model parameter estimation experiment (mopex): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, *320*, 3 - 17. doi: 10.1016/j.jhydrol.2005.07.031

Gaume., E. (2006). On the asymptotic behavior of flood peak distributions. *Hydrol. Earth Syst. Sci.*, *10*, 233-243.

Guo, J., Li, H.-Y., Leung, L. R., S. Guo, P. L., & Sivapalan, M. (2014). Links between flood frequency and annual water balance behaviors: A basis for similarity and regionalization. *Water Resources Research*, *50*. doi: 10.1002/2013WR014374

Her, Y., Yoo, S.-H., Cho, J., S., H., Jeong, J., & Seong, C. (2019). Uncertainty in hydrological analysis of climate change: multiparameter vs. multi-gcm ensemble predictions. *Nature*, *9*. doi: /10.1038/s41598-019-41334-7

Hirschboeck, K. (1987). Hydroclimatically-defined mixed distributions in partial duration flood series. In *In: Singh v.p. (eds) hydrologic frequency modeling* (p. 199-212). Louisiana State University, Baton Rouge, U.S.A: Springer, Dordrecht. doi: 10.1007/978-94-009-3953-0_13

Jianchun, L., Pope, G. A., & Sepehrnoori, K. (1994). A high-resolution finite-difference scheme for nonuniform grids. *Appl. Math. Modelling*, *19*, 162 - 172.

Kleme, V. (1993). Probability of extreme hydrometeorological events - a different approach. In *Extreme hydrological events: Precipitation, floods and droughts (proceedings of the yokohama symposium)* (p. 167-176). Oxfordshire.

Laio, F., Allamano, P., & Claps, P. (2010). Exploiting the information content of hydrological outliers for goodness-of-fit testing. *Hydrol. Earth Syst. Sci.*, *14*, 1909-1917. doi: 10.5194/hess-14-1909-2010

Laio, F., Porporato, A., Ridolfi, L., & Rodriguez-Iturbe, I. (2001). Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress: Ii. probabilistic soil moisture dynamics. *Advances in Water Resources*, *24*(7), 707 - 723. doi: 10.1016/S0309-1708(01)00005-7

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50 – 60. doi: 10.1214/aoms/1177730491

Marani, M., & Ignaccolo, M. (2015). A metastatistical approach to rainfall extremes. *Advances in Water Resources*, *79*, 121–126. doi: 10.1016/j.advwatres.2015.03.001

McGrath, G., Kaeseberg, T., Silva, R., J. D., J. W., Jawitz, Blumensaat, F., Borchardt, D., ... Rao, P. S. C. (2019). Network topology and rainfall controls on the variability of combined sewer overflows and loads. *Water Resources Research*, *55*, 9578 9591. doi: 10.1029/2019WR025336

Merz, B., Hall, J., Disse, M., & Schumann, A. (2010). Fluvial flood risk management in a changing world. *Nat. Hazards Earth Syst. Sci.*, *10*, 509 - 527. doi: 10.5194/nhess-10-509-2010

Merz, B., Vorogushyn, S., Lall, U., Viglione, A., & Blschl, G. (2015). Charting unknown waters on the role of surprise in flood risk assessment and management. *Water Resources Research*, *51*, 6399 6416. doi: 10.1002/2015WR017464

Merz, R., & Blschl, G. (2003). A process typology of regional floods. *Water Resources Research*, *39*, 9578 9591. doi: 10.1029/2002WR001952

Merz, R., & Blschl, G. (2008a). Flood frequency hydrology 1: Temporal, spatial and causal expansion of information. *Water Resources Research*, *44*. doi: doi:10.1029/2007WR006744

Merz, R., & Blschl, G. (2008b). Flood frequency hydrology 2: Combining data evidence. *Water Resources Research*, *44*. doi: doi:10.1029/2007WR006745

Metzger, A., Marra, F., Smith, J., & Morin, E. (2020). Flood frequency estimation and uncertainty in arid/semi-arid regions. *Journal of Hydrology*, *590*. doi: 10.1016/j.jhydrol.2020.125254

Miniussi, A., Marani, M., & Villarini, G. (2020). Metastatistical Extreme Value Distribution applied to floods across the continental United States. *Advances in Water Resources*, *136*. doi: 10.1016/j.advwatres.2019.103498

Montanari, A., & Koutsoyiannis, D. (2014). Modeling and mitigating natural hazards: Stationarity is immortal! *Water Resources Research*, *50*, 97489756. doi: 10.1002/2014WR016092

Porporato, A. (2021). `https://blogs.egu.eu/divisions/hs/2020/07/15/amilcare-porporato-2020-john-dalton-medallist-on-agile-models-for-complex-systems-in-the-environmental-sciences/`.

Porporato, A., Daly, E., & Rodriguez-Iturbe, I. (2004). Soil water balance and ecosystem response to climate change. *The American Naturalist*, *164*, 625632. doi: 10.1086/424970

Rogger, M., Pirkl, H., Viglione, A., Komma, J., Kohl, B., Kirnbauer, R., ... Blschl, G. (2012). Step changes in the flood frequency curve: process controls. *Water Resources Research*, *48*, W05544. doi: 10.1029/2011WR011187

Rogger, M., Viglione, A., Derx, J., & Blschl, G. (2013). Quantifying effects of catchments storage thresholds on step changes in the flood frequency curve. *Water Resources Research*, *49*, 69466958. doi: 10.1002/wrcr.20553

Schaake, J., Duan, Q., Andrassian, V., Franks, S., Hall, A., & Leavesley, G. (2006). The model parameter estimation experiment (mopex). *Journal of Hydrology*, *320*(1), 1 - 2. (The model parameter estimation experiment) doi: 10.1016/j.jhydrol.2005.07.054

Seibert, J., Staudinger, M., & van Meerveld, H. (2019). Computer simulation validation. simulation foundations, methods and applications. In B. C. & S. N. (Eds.), (chap. Validation and Over-ParameterizationExperiences from Hydrological Modeling). Springer,Cham.

Smith, J. A., Cox, A. A., Baeck, M. L., Yang, L., & Bates, P. D. (2018). Strange oods: The upper tail of ood peaks in the united states. *Water Resources Research*, *54*, 6510-6542. doi: 10.1029/2018WR022539

Sullivan, G., & Feinn, R. (2012). Using effect size or why the p value is not enough. *Journal of Graduate Medical Education*, 279-282. doi: 10.4300/JGME-D-12-00156.1

Tarasova, L., Basso, S., & Merz, R. (2020). Transformation of generation processes from small runoff events to large floods. *Geophysical Research Letters*, *47*(22), e2020GL090547. doi: 10.1029/2020GL090547

Tarasova, L., Basso, S., Zink, M., & Merz, R. (2018). Exploring controls on rainfall-runoff events: 1. Time series-based event separation and temporal dynamics of event runoff response in Germany. *Water Resources Research*, *54*, 7711 - 7732. doi: 10.1029/2018WR022587

Viglione, A., Merz, R., & Blschl, G. (2009). On the role of the runoff coefficient in the mapping of rainfall to flood return periods. *Hydrol. Earth Syst. Sci.*, *13*, 577593.

Viglione, A., Merz, R., Salinas, J., & Blschl, G. (2013). Flood frequency hydrology: 3. a bayesian analysis. *Water Resources Research*, *49*, 675-692. doi: 10.1029/2011WR010782

Villarini, G., & Smith, J. (2010). Flood peak distributions for the eastern united states. *Water Resources Research*, *46*. doi: 10.1029/2009WR008395

Wallemacq, P., & House, R. (2018). *Economic Losses, Poverty & Disasters 1998-2017* (Tech. Rep.). United Nations Office for Disaster Risk Reduction (UNDRR) and Centre for Research on the Epidemiology of Disasters (CRED).

Wang, D., & Hejazi, M. (2011). Quantifying the relative contribution of the climate and direct human impacts on mean annual streamflow in the contiguous United States. *Water Resources Research*, *47*. doi: 10.1029/2010WR010283

Wing, O., Bates, P., Smith, A., Sampson, C., Johnson, K., Fargione, J., & Morefield, P. (2018). Estimates of present and future flood risk in the conterminous united states. *Environmental Research Letters*, *13*. doi: 10.1088/1748-9326/aaac65

Zorzetto, E., Botter, G., & Marani, M. (2016). On the emergence of rainfall extremes from ordinary events. *Geophysical Research Letters*, *43*, 8076–8082. doi: 10.1002/2016GL069445