Identifying discontinuities of flood frequency curves

Arianna Miniussi[1,1], Stefano Basso[1,1], Ralf Merz[2,2], and Lisa Kaule[3,3]

[1]Helmholtz Centre for Environmental Research
[2]Helmholtz Centre for Environmental Research (UFZ)
[3]University of Bayreuth

November 30, 2022

Abstract

Discontinuities in flood frequency curves, here referred to as flood divides, hinder the estimation of rare floods. In this paper we develop an automated methodology for the detection of flood divides from observations and models, and apply it to a large set of case studies in the USA and Germany. We then assess the reliability of the PHysically-based Extreme Value (PHEV) distribution of river flows to identify catchments that might experience a flood divide, validating its results against observations. This tool is suitable for the identification of flood divides, with a high correct detection rate especially in the autumn and summer seasons. It instead tends to indicate the emergence of flood divides not visible in the observations in spring and winter. We examine possible reasons of this behavior, finding them in the typical streamflow dynamics of the concerned case studies. By means of a controlled experiment we also re-evaluate detection capabilities of observations and PHEV after discarding the highest maxima for all cases where both empirical and theoretical estimates display flood divides. PHEV mostly confirms its capability to detect a flood divide as observed in the original flood frequency curve, even if the shortened one does not show it. These findings prove its reliability for the identification of flood divides and set the premises for a deeper investigation of physiographic and hydroclimatic attributes controlling the emergence of discontinuities in flood frequency curves.

Identifying discontinuities of flood frequency curves

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

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

[2]Department of Hydrology, University of Bayreuth, Bayreuth, Germany

Highlights:

- We develop an automated method to detect discontinuities of flood frequency curves
- We test it on observed and physically-based theoretical flood frequency curves
- We discuss the reliability of the physically-based approach to detect discontinuities

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

Abstract

Discontinuities in flood frequency curves, here referred to as flood divides, hinder the estimation of rare floods. In this paper we develop an automated methodology for the detection of flood divides from observations and models, and apply it to a large set of case studies in the USA and Germany. We then assess the reliability of the PHysically-based Extreme Value (PHEV) distribution of river flows to identify catchments that might experience a flood divide, validating its results against observations. This tool is suitable for the identification of flood divides, with a high correct detection rate especially in the autumn and summer seasons. It instead tends to indicate the emergence of flood divides not visible in the observations in spring and winter. We examine possible reasons of this behavior, finding them in the typical streamflow dynamics of the concerned case studies. By means of a controlled experiment we also re-evaluate detection capabilities of observations and PHEV after discarding the highest maxima for all cases where both empirical and theoretical estimates display flood divides. PHEV mostly confirms its capability to detect a flood divide as observed in the original flood frequency curve, even if the shortened one does not show it. These findings prove its reliability for the identification of flood divides and set the premises for a deeper investigation of physiographic and hydroclimatic attributes controlling the emergence of discontinuities in flood frequency curves.

1 Introduction

Despite considerable efforts to achieve reliable estimation of rare floods, these events are still among the most common natural disasters (Wallemacq & House, 2018). The evaluation of their hazard is however crucial for several applications, including the design of hydraulic structures, risk planning and mitigation, and computation of premiums in the insurance industry. Appraisal of the flood hazard is especially difficult when the magnitude of the rarer floods can take values which are several times to orders of magnitude larger than commonly observed floods, resulting in a marked uprise of the flood frequency curve beyond certain return periods (Rogger et al., 2012; Smith et al., 2018).

Cognitive biases often lead to downplay the occurrence of such extreme events (B. Merz et al., 2015, 2021), although the scientific literature repeatedly signalled the pervasiveness of these behaviors terming them in various ways. In fact, heavy-tailed distributions of floods (Farquharson et al., 1992; Bernardara et al., 2008; Villarini & Smith, 2010), inversions of concavity and step changes in flood magnitude-frequency curves (Rogger et al., 2012; Guo et al., 2014; Basso et al., 2016) and large values of the ratios between the maximum flood of record and the sample flood with a specified recurrence time (Smith et al., 2018) and between empirical high flow percentiles (Mushtaq et al., 2022) are all manifestations of a marked increase of the magnitude of the rarer floods highlighted by means of different approaches. To further stress the common nature of all these phenomena, in this study we favor none of the previous locutions and instead label them as *flood divides*. The term was chosen to highlight the existence of a discharge threshold which marks the rise of progressively larger floods (red square in Figure 1d) and thus distinguishes between common and increasingly extreme floods that may occur in river basins.

Rogger et al. (2012) investigated marked uprises (i.e., discontinuities in the slope) of flood frequency curves, which they called step changes, by leveraging information collected from field surveys in two small alpine catchments to calibrate a distributed deterministic rainfall-runoff model. They suggested that step changes occur when a threshold of the catchment storage capacity is exceeded, and performed a synthetic experiment (Rogger et al., 2013) to examine the effect of catchment storage thresholds and combined multiple controls (e.g., the temporal variability of antecedent soil storage and the size of the saturated regions) on the return period of the step change.

Guo et al. (2014) and Basso et al. (2016) instead linked different shapes of flood frequency curves and a marked growth of the magnitude of the rarer floods to the catchment water balance. The former justified these features through 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. The latter explained them by means 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, thus causing the emergence of flood divides.

Smith et al. (2018) computed the ratio between the maximum flood of record and the sample 10-year flood for thousands of gauges across the USA, finding large values for a substantial amount of them. Different flood-generating processes (R. Merz & Blöschl, 2003; Berghuijs et al., 2014; Tarasova et al., 2020) or mixtures of flood event types (Hirschboeck, 1987; Villarini & Smith, 2010; Smith et al., 2018) were indicated by other studies as possible causes of these marked increases of the magnitude of the rarer floods.

Finally, a rather common approach to study this phenomenon consists in evaluating the shape parameter of Generalized Extreme Value distributions fitted to observed annual maximum series (Farquharson et al., 1992; Bernardara et al., 2008; Villarini & Smith, 2010; Smith et al., 2018). Notwithstanding the drawbacks of such a parametric approach applied in association with limited records of annual maxima, these studies highlighted the ubiquitous occurrence of flood divides and flood distributions characterized by thick upper tails, as indicated by widespread positive values of the shape parameter. Moreover, Smith et al. (2018) showed that the values of the shape parameter significantly increase with longer data records. Their findings thus suggest that uprises of flood frequency curves may be the norm rather than rare conditions, pointing to the limited data record as the reason for the latter belief.

Although former research hints at the ubiquitousness of flood divides in flood frequency curves and provide indications of their possible drivers, a quantitative methodology to identify flood divides, which is robust to sampling uncertainty and tested in a large set of case studies, is still lacking. The relevance of our study is thus twofold: (i) we develop such a methodology for the detection of flood divides and evaluate their emergence across the US and Germany, in a large set of catchments with contrasting physio-climatic features; (ii) we examine the reliability of a process-based stochastic framework for the estimation of flood frequency curves to detect flood divides and infer their occurrence, benchmarking its results against observations.

2 Methodology and Data

2.1 The Physically-based Extreme Value distribution of river flows

2.1.1 Theoretical framework

The PHysically-based Extreme Value (PHEV) distribution of river flows is a parsimonious mechanistic-stochastic formulation of flood frequency curves (Basso et al., 2016, 2021) 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, daily 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 $\lambda < \lambda_P$ $[T^{-1}]$ 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, Schirmer, & Botter, 2015).

The above-summarized mechanistic-stochastic description of runoff generation processes allows for expressing the probability distributions of daily flows (Botter et al., 2009) and peak flows (i.e., local flow peaks occurring as a result of streamflow-producing rainfall events) as a function of a few physically meaningful parameters (Basso et al., 2016). It also enables classifying hydrologic regimes according to their typical streamflow dynamics, which are summarized by the persistency index (Botter et al., 2013). This is defined as the ratio between runoff frequency and mean catchment response time. An erratic regime (low persistency index), 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 (high persistency index), typically occurring in cold-humid seasons and lowland catchments, is characterized by frequent rainfall events and a rather constant water supply to the catchment.

Considering that peak flows in a given reference period (e.g., a season) are Poisson distributed and postulating their independence yield the probability distribution of flow maxima (i.e., maximum values in a specified timespan). The return period is finally obtained as the inverse of the exceedance cumulative probability of flow maxima, thus providing an expression of the flood frequency curve which reads (Basso et al., 2016):

$$T_r(q) = \frac{1}{1 - exp\left[-\lambda \tau D_j(q)\right]} \tag{1}$$

where τ [T] is the duration in days of the reference period used in the analyses; $D_j(q) = \int_q^\infty p_j(q)\, dq$ is the exceedance cumulative probability of peak flows; p_j is the probability density function of peak flows, $p_j(q) = Cq^{1-a} \exp(\frac{\lambda q^{1-a}}{K(1-a)} - \frac{q^{2-a}}{\alpha K(2-a)})$; α and λ are the aforementioned parameters describing Poisson-distributed runoff events, a and K are the parameters of the power-law storage-discharge relation, and C is a normalization constant.

2.1.2 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 mean daily rainfall depth in rainy days, while λ (frequency of streamflow-producing rainfall) as the ratio between the long term mean daily flow $< q >$ and α (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). 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 et al., 2016).

2.2 Identification of Flood Divides

To identify flood divides, we adopt the method proposed by Rogger et al. (2013): a flood divide is defined as the sharpest bend of the flood frequency curve, here considered in terms of rescaled streamflow maxima (i.e., seasonal maxima divided by the long term mean daily flow, $< q >$) as a function of the return period, the latter represented in logarithmic scale. We develop a methodology dedicated to its identification from both empirical estimates of the flood frequency curve obtained by means of Weibull plotting position and PHEV.

The resulting approach, which can be employed without depending on subjective evaluation, is detailed in the following.

1. The curvature of the flood frequency curve, of which we show an example in Figure 1, is computed as $logTr''/(1 + logTr'^2)^{(3/2)}$ (where the apex indicates the derivation operation with respect to the rescaled streamflow) 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.

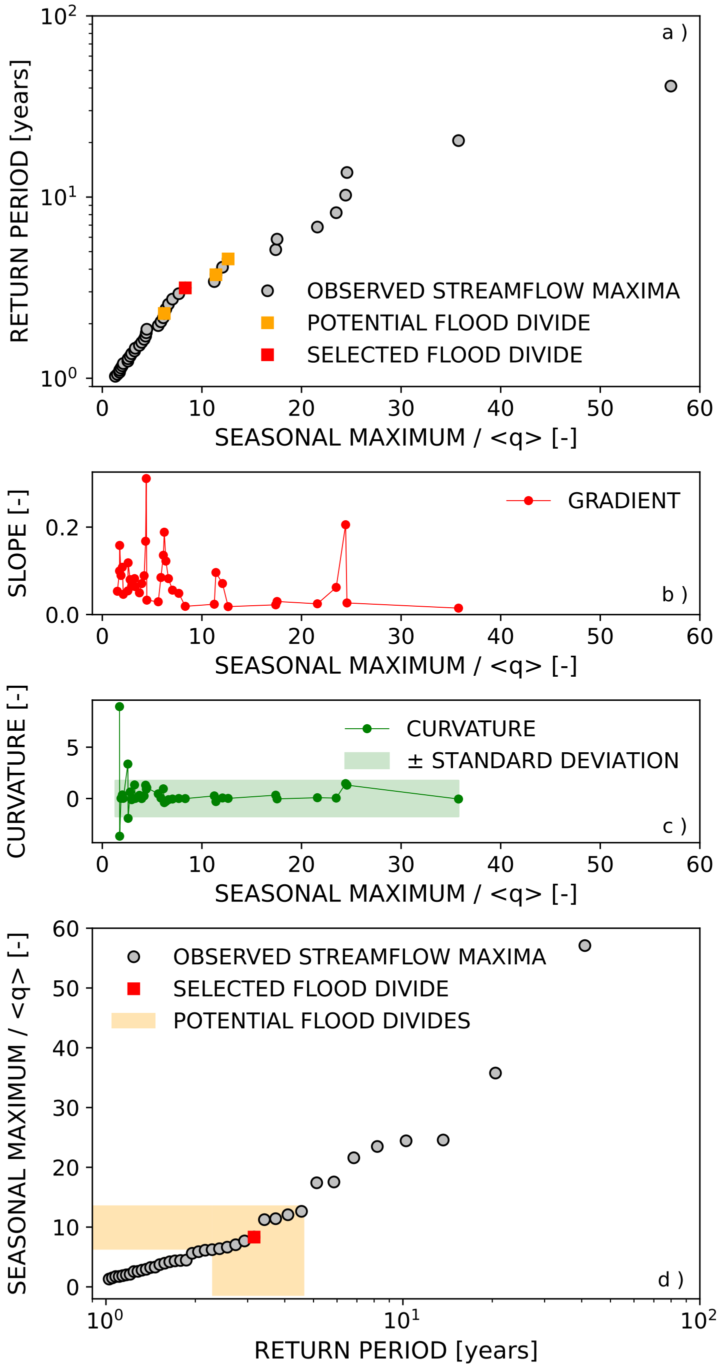

Figure 1: Exemplary application of the proposed methodology to detect flood divides to the Rott river at Kinning, Bavaria (ID: 18801005), in the summer season. a) Visualization of how the approach is actually applied, i.e., expressing the logarithm of the return period as a function of the rescaled seasonal maxima (gray filled circles). Potential flood divides (i.e., all the points with a p-value of the Mann-Whitney U-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 U-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 flood divide, while the orange shaded area represents the range of potential flood divides.

2. As the noise associated to computing the curvature on a discrete and rather sparse set of points (seasonal maxima) might lead to identification errors, a heuristic 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 1c);
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 flood divide identified at point 2 to check if their distributions are statistically different at a significant level equals to 0.05 (in other words, if the slope of the curve significantly differs between the left and right-hand side of the flood divide); 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 $\Delta\mathrm{q} = 0.05$ up to a value of rescaled streamflow equal to 200, i.e., 200 times the long-term average streamflow. The relative increment of the slope between the left and right-hand side of a potential PHEV flood divide is also evaluated within the observational range.
4. We finally identify as flood divide 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 (moderate effect size; Gignac and Szodorai (2016); Lovakov and Agadullina (2021)) and the slope increment exceeds a value of 1%.

Figure 1 visually exemplifies the application of the developed approach for flood divides detection to the flood frequency curve of the Rott river at Kinning, Bavaria (ID: 18801005), in the summer season. In Figure 1a 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 rescaled seasonal maxima on the x-axis), as streamflow is the independent variable in Eq. (1). The red square in Figure 1a-d represents the selected flood divide, 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 1b) and fulfilling the additional criterion on the Cohen's d. We also show points that are initially analyzed as potential flood divides (i.e., all the points with a Mann-Whitney p-value lower than 0.05, orange squares in Figure 1a).

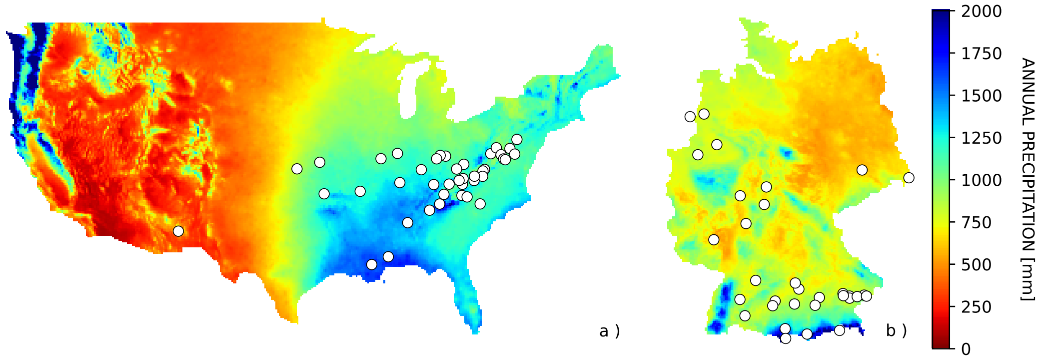

Figure 2: Select river basins (white filled circles) from the (A) MOPEX and (B) German datasets. The background of the maps represents 30-years annual precipitation normals (1981-2010 for the US and 1991-2020 for Germany).

2.3 Datasets

We use daily rainfall and 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), performing all analyses in a seasonal time frame to account for the seasonality of rainfall and runoff (Allamano et al., 2011; Baratti et al., 2012). To assure that PHEV suitably represents the key processes of streamflow generation in the set of case studies, we only consider catchments with low human impact, weak or absent inter-seasonal snow dynamics (Botter et al., 2013; Wang & Hejazi, 2011) and hydrograph recession properties which are independent of the peak flow (Basso et al., 2021). Similarly to previous studies (R. Merz et al., 2020), we as well restrict our analysis to cases for which the root mean square error (RMSE) between the predicted and observed flood frequency curve is limited (i.e., lower than 0.3), as a fairly accurate estimation of the flood frequency curve is a precondition to investigate if PHEV is able to correctly identify flood divides and whether their occurrence is affected by physio-climatic catchment attributes. RMSE is here calculated as $\sqrt{\left[\sum_{i=1}^{N}(logTr_{ds} - logTr_{PHEV})^2\right]/N}$, where Tr_{ds} and Tr_{PHEV} are empirical and PHEV estimates of the return periods of seasonal maxima, and N is the number of values in their observed sample. This selection yields a set of 101 case studies (i.e., catchment-season combinations), divided into 23, 29, 23 and 26 cases respectively in the spring, summer, autumn and winter seasons. Their catchment areas vary between 43 and 9052 km^2 (median: 865 km^2). The locations of their outlets are displayed in Figure 2.

3 Results and Discussion

We apply the methodology for the identification of flood divides introduced in the previous section to each observed and analytic seasonal flood frequency curve, thus allowing for evaluating the flood divide detection of PHEV against observations, which we consider as benchmark (Figure 3). The bar plots in Figure 3 show the percentages of case studies for which a flood divide is identified from both PHEV and the observational records (true positives, dark green color), those which display a flood divide neither in the empirical nor in the analytic flood frequency curves (true negatives, light green), the percentages of cases where a flood divide 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 flood divide which is not confirmed by the available observations (false positives, orange). The existence of both true positives and true negatives emphasizes the capability of PHEV to mimic varied observed shapes of flood frequency curves (Basso et al., 2016) and to identify both the presence and the absence of a flood divide.

The bar plots in Figure 3a and 3b differ for the criteria applied in the flood divide identification methodology. In Figure 3a only the controls on the p-value of the Mann-Whitney U-test mentioned in Section 2.2 are considered, whereas the additional requirements on the effect size and slope increment are as well used in Figure 3b. True positives (dark green) prevail in the summer and autumn seasons of Figure 3a, amounting to about 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 flood divides are used, by accounting for the mentioned additional criteria, 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 flood divide, thus not representing a noteworthy hazard. Most of them however become false positives (orange color in Figure 3b) as the identified changes of the slope of the flood frequency curve are not substantial according to the limited amount of available observations, whereas PHEV confirms the existence of a flood divide thanks to its evaluation in an unlimited number of points. 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 flood divides with confidence from plain observations.

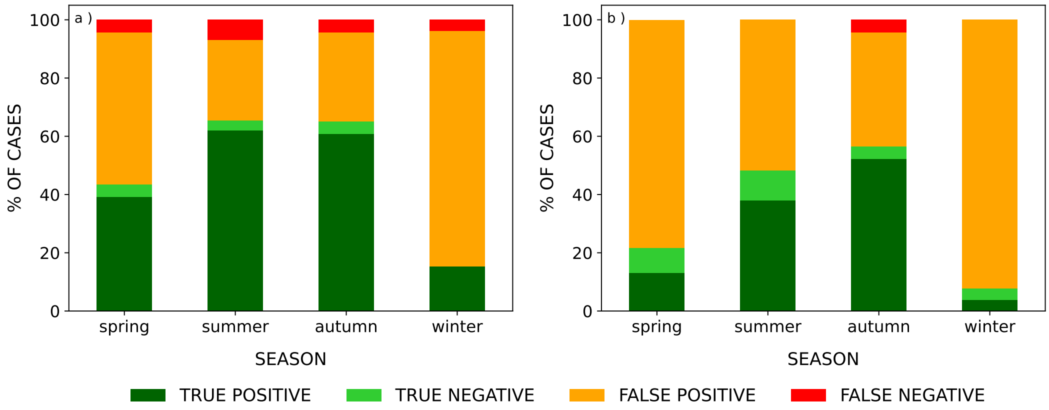

Figure 3: Performance of the PHysically-based Extreme Value (PHEV) distribution of river flows in the detection of flood divides when only the controls on the Mann-Whitney U-test are considered (see Section 2.2, panel a) and when the whole methodology for detecting flood divides is applied (see Section 2.2, panel b). Percentages are calculated on the overall number of case studies, which amount to 23, 29, 23 and 26 cases respectively in the spring, summer, autumn and winter seasons. True positives (dark green color) and true negatives (light green) indicate coherence between PHEV and observations, i.e., flood divides 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 flood divide that was not identified by the observations or the observations display a flood divide which is not detected by PHEV. The reasons for the presence of false positives are further investigated in the study and clarified in the text and figures.

The predominance of false positives in spring and winter (orange color in Figure 3) calls for further investigation of their causes. We therefore hypothesize 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 flood divides 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 flood divide, see Figure 1a) 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. We then perform the following experiment to test this hypothesis. 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 flood divide) 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 flood divides on these fictitious flood frequency curves and the corresponding shortened data records.

PHEV detects a true flood divide (i.e., true positives) in 81% of the cases even when the largest points are removed, whereas the observations only in 40%. The maps in Figure 4b and 4c summarize this result: half circles are colored either in green, if a flood divide 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 most left halves of the circles are colored in green and most of the right ones are instead red, thus indicating a high success rate of PHEV and a significantly lower one of observations in inferring the emergence of flood divides from shortened records. A similar result is obtained by discarding maxima with return period greater than 10 years (i.e., discarding about five-six points instead of the highest ten), when PHEV correctly detects 85% of true flood divides in comparison to a correct detection rate from observations of 60%. 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 flood divides from short series 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 flood divide which shall be expected is provided by considering typical streamflow dynamics occurring for distinct river flow regimes, here characterized by means of the persistency index (Botter et al., 2013). 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., 2015). In the context of this study, false positives shall therefore mostly occur for persistent regimes, as such large events enabling detection of flood divides from empirical flood frequency curves are less likely to have been observed during the available data record.

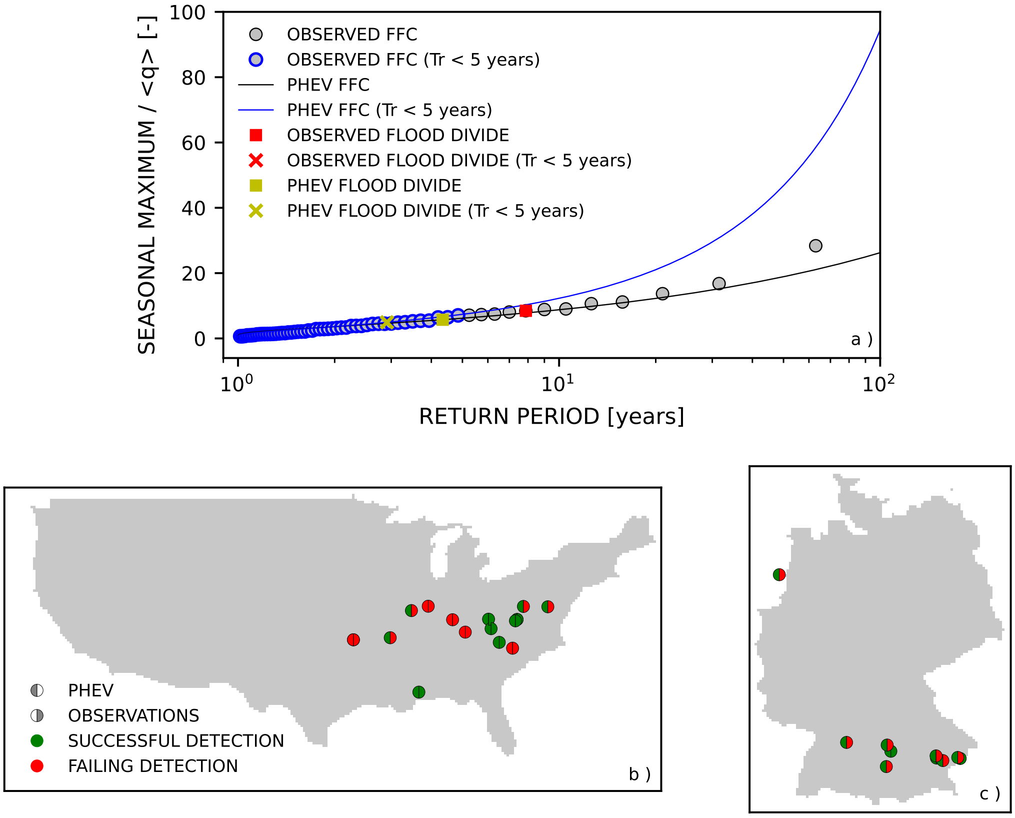

Figure 4: Visual explanation and results of an experiment aimed at testing hypotheses on the emergence of false positives. a) Gray dots with black (blue) contour represent the complete (shortened, until a return period of 5 years) observed seasonal maxima series of the Wörnitz river at Harburg, Bayern (ID:11809009), in the summer 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 flood divide detected from the observations (by PHEV) using the complete series, while the corresponding crosses (the red one is not visible in the plot as no flood divide was detected after shortening the observations) represent the observed and analytic flood divides 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 flood divide 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, most of the left halves are green (PHEV detects true flood divides even from the shortened series in the majority of the cases), whereas most of the right ones are red (flood divides are not always identified from observations when the shortened records are used).

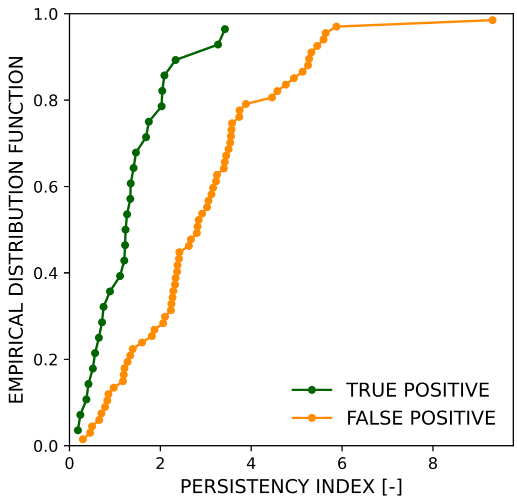

Figure 5: Empirical cumulative distribution functions of the persistency index for true positive (dark green) and false positive (orange) cases. The distributions are significantly different in a statistical sense (the p-value of the 2-samples Kolmogorov-Smirnoff test is lower than 0.01.)

In Figure 5 we compare the cumulative distributions of the persistency index for the sets of true positives (dark green) and false positives (orange) to verify whether this is the case. The distributions clearly differ and false positives mostly occur for persistent regimes. This qualitative evaluation is validated 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. Remarkably, the seasons characterized by the larger portion of false positives are spring and winter, during which regimes tend to be more persistent.

The physical explanation provided here of the different telling power of streamflow data for rivers characterized by distinctively different streamflow dynamics agrees with the results of previous research. For example, Botter et al. (2013) showed less variable streamflow distributions across years in erratic regimes compared to persistent ones, which determines higher representativeness of their estimates in the former case for a given length of the data record. Smith et al. (2018) also demonstrated that upper tail ratios grow with the length of data and, for a given data length, are larger (i.e., flood divides are more often identified) in arid and semiarid regions than in humid ones. Their results jointly suggest that, given similarly long data records, the typical (erratic) flow dynamics of drier areas enable more reliable characterization of the whole range of values possibly spanned by streamflow and of the presence or absence of flood divides according to the physical explanation provided above.

4 Concluding Remarks

In this work we examine the occurrence of marked uprises of flood frequency curves (termed flood divides), which are pivotal for a correct estimation of river flood hazard. We develop a robust methodology to identify them from observational records and models, and evaluate the capability of the PHysically-based Extreme Value distribution of river flows (PHEV) to reliably detect flood divides.

Results show that PHEV is consistently able to recognize the presence/absence of flood divides 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 flood divide detection analysis on the shorter series, showing that PHEV can foresee the emergence of flood divides even if the shortened observations do not display it. The result supports claims of the dependability of flood divides 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 flood divides, analogously to what occurs when we artificially reduce the size of the observational sample.

The present analysis, 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 flood divides and consequently unveil the propensity of rivers to large floods. The method is especially relevant in data scarce conditions, although limitations linked to the domain of applicability of this tools exist and have been recalled in this work. The study lays the foundations for a better comprehension of climate and landscape controls of observed marked rises of the magnitude of the rarer floods, which is the subject of ongoing research.

Acknowledgments

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project Number 421396820 'Propensity of rivers to extreme floods: climate-landscape controls and early detection (PREDICTED)' and Research Group FOR 2416 'Space-Time Dynamics of Extreme Floods (SPATE)'. The financial support of the Helmholtz Centre for Environmental Research - UFZ is as well 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*, 3207–3215. 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., Botter, G., Merz, R., & Miniussi, A. (2021). Phev! the physically-based extreme value distribution of river flows. *Environ. Res. Lett.*, *16*, 124065. doi: 10.1088/1748-9326/ac3d59

Basso, S., Frascati, A., Marani, M., Schirmer, M., & Botter, G. (2015). Climatic and landscape controls on effective discharge. *Geophysical Research Letters*, *42*, 8441–8447. 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

Bernardara, P., Schertzer, D., Eric, S., Tchiguirinskaia, I., & Lang, M. (2008). The flood probability distribution tail: How heavy is it? *Stochastic Environmental Research and Risk Assessment*, *22*, 5638 - 5661. doi: 10.1002/2014WR015692

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.

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

Farquharson, F. A. K., Meigh, J. R., & Sutcliffe, J. (1992). Regional flood frequency analysis in arid and semi-arid areas. *J. Hydrol.*, *138(3)*, 487-501. doi: 10.1016/0022-1694(92)90132-F

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78. doi: 10.1016/j.paid.2016.06.069

Guo, J., Li, H.-Y., Leung, L. R., Guo, S., Liu, P., & 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

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.

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

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485-504. doi: 10.1002/ejsp.2752

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

Merz, B., Blöschl, G., Vorogushyn, S., & et al. (2021). Causes, impacts and patterns of disastrous river floods. *Nat. Rev. Earth Environ.*, *2*, 592 – 609. doi: 10.1038/s43017-021-00195-3

Merz, B., Vorogushyn, S., Lall, U., Viglione, A., & Blöschl, 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., & Blöschl, G. (2003). A process typology of regional floods. *Water Resources Research*, *39*, 9578 – 9591. doi: 10.1029/2002WR001952

Merz, R., Tarasova, L., & Basso, S. (2020). Parameter's controls of distributed catchment models—how much information is in conventional catchment descriptors? *Water Resources Research*, *56*(2), e2019WR026008. doi: 10.1029/2019WR026008

Mushtaq, S., Miniussi, A., Merz, R., & Basso, S. (2022). Reliable estimation of high floods: A method to select the most suitable ordinary distribution in the metastatistical extreme value framework. *Advances in Water Resources*, *161*, 104127. doi: 10.1016/j.advwatres.2022.104127

Porporato, A., Daly, E., & Rodriguez-Iturbe, I. (2004). Soil water balance and ecosystem response to climate change. *The American Naturalist*, *164*, 625–632. doi: 10.1086/424970

Rogger, M., Pirkl, H., Viglione, A., Komma, J., Kohl, B., Kirnbauer, R., ... Blöschl, 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., & Blöschl, G. (2013). Quantifying effects of catchments storage thresholds on step changes in the flood frequency curve. *Water Resources Research*, *49*, 6946–6958. doi: 10.1002/wrcr.20553

Schaake, J., Duan, Q., Andréassian, 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

Smith, J. A., Cox, A. A., Baeck, M. L., Yang, L., & Bates, P. D. (2018). Strange floods: The upper tail of flood 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

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