PHEV! The PHysically-based Extreme Value distribution of river flows

S Basso[1], G Botter[2], R Merz[1] and A Miniussi[1]

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

[2] Department of Civil, Environmental and Architectural Engineering, University of Padova, Padua, Italy.

E-mail: `stefano.basso@ufz.de`

Abstract. Magnitude and frequency are prominent features of river floods informing design of engineering structures, insurance premiums and adaptation strategies. Recent advances yielding a formal characterization of these variables from a joint description of soil moisture and daily runoff dynamics in river basins are here systematized to highlight their chief outcome: the PHysically-based Extreme Value (PHEV) distribution of river flows. This is a physically-based alternative to empirical estimates and purely statistical methods hitherto used to characterize extremes of hydro-meteorological variables. Capabilities of PHEV for predicting flood magnitude and frequency are benchmarked against a standard distribution and the latest statistical approach for extreme estimation in two ways. The methods are first applied to an extensive dataset to compare their skills for predicting observed flood quantiles in a wide range of case studies. Synthetic time series of streamflow, generated for select river basins from contrasting hydro-climatic regions, are later used to assess performances for rare events. Both analyses reveal fairly unbiased capabilities of PHEV to estimate flood magnitudes corresponding to return periods much longer than the sample size used for calibration. The results also emphasize reduced prediction uncertainty of PHEV for rare floods when the mechanistic hypotheses postulated by the method are fulfilled, notably if the flood magnitude-frequency curve displays an inflection point. These features, arising from the mechanistic understanding embedded in the novel distribution of the largest river flows, are key for a reliable assessment of the actual flooding hazard associated to poorly sampled rare events, especially when lacking long observational records.

Keywords: flood hazard, runoff dynamics, extreme events, flood frequency, large scale benchmarking.

Submitted to: *Environ. Res. Lett.*

1. Introduction

Reliable estimates of magnitude and frequency of river floods are crucial for a wide range of social and economic activities. For instance, they inform design of engineering structures, insurance premiums, urban planning and adaptation strategies. But despite hundreds of years of efforts, flooding is still the most common natural disaster (Wallemacq and House, 2018). Hazard assessment is indeed hampered by processes that might be more variable than suggested by observed records, let alone ongoing global changes (The Economist, 2017; Bevere et al., 2020). This is a problem, as both purely statistical methods hitherto used for characterizing extremes of hydro-meteorological variables (Katz et al., 2002; Morrison and Smith, 2002; England et al., 2019; Metzger et al., 2020) and complex hydrological models encoding state-of-the-art knowledge of physical processes (Maxwell and Miller, 2005; Knijff et al., 2010; Hirpa et al., 2018; Kuffour et al., 2020) heavily rely on observations. Tools pairing physical understanding of the mechanisms producing extreme events with easily tractable mathematical descriptions of them and less demanding data requirements are therefore vital to a more accurate flood risk estimation (Klemeš, 1989; Barth et al., 2019).

This letter systematizes and tests advances in the mechanistic-stochastic description of soil moisture and runoff dynamics in river basins to emphasize what might be such a tool: the PHysically-based Extreme Value (PHEV) distribution of river flows. Perks of this physically-grounded alternative to statistical and hydrological methods hitherto used for characterizing extremes of hydro-meteorological variables are its being simple as a statistical distribution, mechanistic as a fully-fledged model, and stochastic as nature is. Capabilities of PHEV for predicting flood magnitude and frequency are here benchmarked against observation-based estimates and leading flood hazard assessment methods, a pivotal milestone to further adoption of this tool in the community of researchers, professionals and policy markers.

2. Methods

2.1. The physically-based extreme value distribution of river flows

The physically-based extreme value distribution of river flows arises from advances in the mathematical description of catchment-scale daily soil moisture and runoff dynamics in river basins (Laio et al., 2001; Porporato et al., 2004; Botter et al., 2007). These well established scientific theories represent rainfall as a marked Poisson process with frequency $\lambda_P[T^{-1}]$ and exponentially distributed depths with average $\alpha[L]$. Infiltration of rainfall into the soil determines stochastic increments of the moisture content, whereas evapotranspiration causes loss of soil moisture from the root zone. Losses linearly vary with the soil moisture between the wilting point and a critical upper threshold akin to the water holding capacity of the soil. The exceedance of this threshold triggers runoff pulses occurring with frequency $\lambda < \lambda_P[T^{-1}]$ and whose magnitudes follow an exponential distribution with average α [L] (Botter et al., 2007). These pulses feed

a single catchment hydrologic storage, which is finally drained by the river network as streamflow. A non-linear storage-discharge relation echoes a hydrological response which varies with the catchment water content and mimics the joint effect of different flow components (e.g., subsurface and surface runoff) (Basso et al., 2015).

The mechanistic-stochastic description of runoff generation processes summarized above yields an expression for the probability distribution of daily streamflows q (A.1, Botter et al. (2009)) which, in addition to λ and α, depends on the coefficient K and exponent a of a power law function (alike the storage-discharge relation mentioned earlier) used to describe hydrograph's recessions. It also provides a physically-grounded mathematical form for the probability distribution of ordinary peak flows (*sensu* Zorzetto et al. (2016), i.e., local flow peaks occurring as a result of streamflow-producing rainfall events) (A.2, Basso et al. (2016)) which depends on the same set of parameters. By further postulating independence of the peak flows, the physically-based probability distribution of flow maxima (i.e., maximum values in a specified time frame, such as a season or a year) finally emerges (A.3, Basso et al. (2016)).

These three probability distributions form a consistent and physically-grounded set of expressions to characterize the statistical properties of daily flows, ordinary peak flows and flow maxima. They describe their magnitudes, likelihoods of occurrence and the related flood hazard based on a set of four physically meaningful parameters (α, λ, a, K) which embody climatic and landscape attributes of the considered catchment. These probability distributions might therefore constitute a sound alternative to purely statistical distributions commonly used to characterize hydrological variables and their maxima.

2.2. Benchmarking PHEV against leading methods for flood hazard assessment

We benchmark the performance of PHEV! against two leading statistical models for flood hazard assessment, namely the Generalized Extreme Value (GEV) (Gnedenko, 1943; Coles, 2001) and Metastatistical Extreme Value (MEV) (Marani and Ignaccolo, 2015; Zorzetto et al., 2016) distributions. We ruled out the widely applied Log Pearson Type III distribution (England et al., 2019) as a suitable term of comparison in this study as a recent work showed it being outperformed by GEV and MEV distributions for estimating magnitude and frequency of river floods from short data samples (Miniussi, Marani and Villarini, 2020).

The GEV distribution is a standard tool (Kjeldsen and Bayliss, 2008; DWA, 2012) traditionally applied to samples of maxima (as per the block maxima approach) under the assumption that the number of events within each block tends towards infinity (Coles, 2001). It is also used to characterize the peaks above a high threshold (in the peak over threshold approach) under the hypotheses of a Poisson distributed number of peaks, whose magnitudes follow a Generalized Pareto distribution (Pickands, 1975).

The MEV distribution is instead a recently proposed method to estimate extremes by exploiting the features of ordinary events, which is currently gaining momentum

(Marra et al., 2018, 2019; Schellander et al., 2019; Zorzetto and Marani, 2019; Hosseini et al., 2020; Zorzetto and Marani, 2020; Miniussi, Marani and Villarini, 2020; Miniussi, Villarini and Marani, 2020; Miniussi and Marani, 2020; Marra et al., 2020). It relaxes the mentioned assumptions which lie at the heart of the GEV distribution and regards as random variables both the number of independent ordinary events occurring in the considered time interval and the parameters of the distribution used to describe their magnitudes, without any restriction on the distribution underlying them. Details on these two methods are respectively available in Coles (2001) and Zorzetto et al. (2016).

For the present analyses we used rainfall and streamflow time series of gauges unaffected by climatic and anthropogenic hydrograph alterations from the US Model Parameter Estimation Experiment (MOPEX) dataset (Schaake et al., 2006; Wang and Hejazi, 2011). We only retained catchments for which at least 30 years of observations are available, with less than 10% of missing data in each season (December-February, March-May, June-August, September-November). The resulting mininum, mean and maximum lengths of the analyzed records are 35, 52 and 55 years. We performed all computations at seasonal scale (Durrans et al., 2003; Baratti et al., 2012) to account for the seasonality of rainfall and runoff generation processes. Catchments where precipitation falls as snow (i.e., when average daily temperatures below zero degrees occur during precipitation) for more than 50% of a season were discarded to comply with key hypotheses of the theoretical framework underlying PHEV (Botter et al., 2013), which are violated where snow dynamics play a decisive role. This initial data screening resulted into 161 gauges (483 catchment-season case studies) retained for analysis, whose geographical locations are shown with grey dots in Figure 1G.

We further selected eight stations from different US hydro-climatic regions (red squares in Figure 1G and Table B1) with the aim of using them for a more in-depth assessment of the capability of the theoretical extreme value distributions (i.e., PHEV, GEV and MEV) to predict magnitudes of events with return periods much longer than the available length of observations. This reflects typical conditions in the practice, where the magnitude of events with rather long return periods (say, 100-1000 years) must be extrapolated on the basis of relatively short samples of observations. In light of the limited length of observed discharge time series typically available, and being the maximum empirical return period about equal to the duration of the record, in such situations observations cannot serve as a benchmark for evaluating performances in the estimation of the highest quantiles of the distribution.

For this reason, we utilized a stochastic rainfall generator and a parsimonious hydrological model (see details in Garbin et al. (2019)) to produce 1000 years long synthetic time series of daily rainfall and streamflow for the select catchments. The series were used as a reference and compared to empirical and statistical estimates derived from shorter samples extracted from the 1000 years long data series. The rainfall generator and hydrological model comply with the main physical assumptions of PHEV detailed at the beginning (being MEV and GEV purely statistical models which do not require specific assumptions on the underlying processes). In this way we created a controlled

experiment which enables analyzing the robustness of both empirical estimates and theoretical extreme value distributions to non-ergodic data samples only.

Furthermore, we employed a cross-validation procedure (Zorzetto et al., 2016; Miniussi, Marani and Villarini, 2020) to analyze the ability of the different methods to extrapolate information outside the range covered by the observations. This would not be possible in a customary goodness-of-fit approach, which provides no information on the capability of a method to predict events which are not included in the calibration sample. To do so, we first divided the available series of observed (synthetic) data into two complementary parts by randomly selecting from them (through resampling without substitution) 100 (1000) subsets of S years. These are the training sets used to calibrate the theoretical distributions, whereas the remaining parts constitute the validation samples. We used $S = 10$ years for the observations and repeated the process for $S = 10$, 20, 30 and 60 years for the long synthetic data series to explore how the different models perform with various sizes of the calibration sample.

We fitted the GEV distribution on the calibration sample of seasonal maxima by means of L-moments (Hosking, 1990). Likewise, we employed L-moments to estimate the parameters of a Gamma distribution here adopted to describe the whole set of ordinary flow peaks in the MEV approach. Further details on the calibration of GEV and MEV are available in (Miniussi, Marani and Villarini, 2020). Parameters of PHEV were instead either directly derived from daily rainfall and streamflow series (α, λ and a) or obtained through maximum likelihood calibration on the sample of seasonal maxima (K) (Basso et al., 2016). In particular, we computed α as the average amount of precipitation in rainy days, λ as the ratio between long-term average streamflow and α, and a as the median value of the recession exponents obtained by fitting a power law function to $dq/dt - q$ pairs observed for each hydrograph recession. The maximum likelihood approach here used to calibrate the parameter K is known to provide frail estimates when small calibration samples are available, as in this application. Although aware of the handicap thus imposed to the performance of PHEV, we were compelled to adopt this method due to the current unavailability of more suitable approaches (such as L-moments) for the novel physically-based extreme value distribution.

We finally used several metrics (e.g., skill score (Murphy and Winkler, 1992; Hashino et al., 2007), relative error (Abramowitz and Stegun, 1972), quantile-quantile plot (Barnett, 1975)) to evaluate and compare performances of PHEV, GEV and MEV when both observed data (for the MOPEX dataset) and long synthetic series (for the eight select basins) are analyzed. We introduce the significance of these metrics in the following, where results are discussed.

3. Results and Discussion

Figure 1A-C summarizes the performances of PHEV, GEV and MEV for predicting observed flood quantiles with return periods longer than the length of the calibration samples (i.e., 10 years) in all the case studies. Warmer colors indicate higher density of

the point clouds and thus mark the most frequent performances of the methods. PHEV (panel A) tends to slightly underestimate flood magnitudes in the record (i.e., the red-to-yellow core of the point cloud stands below the 45-degree line). Higher variability of its performances (i.e., larger spread of points in the figure) compared to GEV and MEV is manifest, especially in the case of high quantiles for which overestimation often occurs. On the contrary, both GEV (panel B) and MEV (panel C) correctly estimate lower flood magnitudes (i.e., the point clouds are centered on the 45-degree lines for low quantiles) but tend to systematically underestimate the larger floods. In the case of MEV, this behavior is likely due to using a Gamma distribution to describe ordinary events, in agreement with Villarini and Strong (2014) and Slater and Villarini (2017). In fact, such a light tailed distribution might underestimate ordinary peaks characterized by low probability, eventually resulting in underestimation of maxima with long return periods. Although the optimization of the MEV methodology is outside the scope of this work, we recognize that the choice of a site-specific distribution of ordinary peaks might reduce underestimation issues in the MEV approach (Miniussi, Marani and Villarini, 2020).

As for PHEV, previous studies scantily reported about the possible existence of a relation between recession coefficient K and peak flow and its capability to interfere with the characterization of daily flow distributions provided by the mechanistic-stochastic model underlying it (Basso et al., 2015; Ghosh et al., 2016). Was there such a relation, the parameters describing the hydrologic behavior of the catchment would vary along with the peak flow magnitude, thus precluding the use of a single effective K to portray the catchment hydrologic response and the resulting magnitude of floods. At the same time, such a relation would also suggest caution about the dependability of purely statistical flood estimates beyond the range of recorded magnitudes, as ostensibly good performances of these methods might indicate their higher flexibility to represent conditions which occurred in the observed time frame rather than robustness in estimating uncharted flood magnitudes (Kirchner, 2006).

We computed the recession coefficient by fitting a power law function with exponent set equal to a (i.e., the median exponent across all recessions in a case study) to $dq/dt - q$ pairs observed for each hydrograph recession (Biswal, 2021), and investigated the possible existence of a relation between K and peak flow in the MOPEX basins. Panels E and F display examples of catchments where we either did not or identified such a relation. We then classified all the case studies into two groups based on the slope of their K - peak flow relation, rendering them in panel D with either light or dark green markers if the relation was respectively weak (i.e., absolute value of the slope $\leq$ 0.4) or strong (i.e., absolute value of the slope $>$ 0.4). Notice that choosing alternative thresholds around this value does not substantially impact the key results of the study.

The performance of PHEV remarkably improves in case studies exhibiting weak relations between K and peak flow (i.e., the majority of light green markers lie just below the 45-degree line in panel D, whereas dark green markers mainly stand above it). The Two-sample Kolmogorov-Smirnov test for the probability distribution of the

relative estimation errors confirms that the two groups are significantly different in a statistical sense (p-value < 0.01). The variability of PHEV diminishes for the set of cases featuring weak K- peak flow relations (the interdecile and interquartile ranges of the relative error decrease by 25% and 48%) and, although PHEV tends to underestimate flood magnitudes in the range of observed quantiles, the bias of the estimate seems to be relatively stable for increasing quantiles compared to GEV and MEV.

The proposed analysis of hydrograph recession characteristics is thus an effective method to distinguish ex ante in what conditions PHEV shall be applied, namely in all cases (268 out of 483 in our dataset, which will be referred in the following analyses) when the recession coefficient K is fairly constant across recessions and thereby agrees with the mechanistic description of runoff generation processes underlying PHEV. Possible causes of the variability of K across events and basins are a much debated topic in recent years (Rupp et al., 2009; Shaw et al., 2013; Bart and Hope, 2014; Biswal and Nagesh Kumar, 2015; Patnaik et al., 2015; Dralle et al., 2018; Tashie et al., 2019, 2020). The geomorphological theory of recession flow curves (Biswal and Marani, 2010) suggests, on the basis of theoretical and observational arguments, that an enhanced variability of K is linked to the existence of manifold hydrologic storage units which are differentially activated across events, either because of their uneven distribution along river networks or as a result of variable degrees of saturation preceding the events (Biswal and Marani, 2014; Mutzner et al., 2013). This behavior is conceivably related to the existence of assorted triggers of runoff events and floods (Tarasova et al., 2020).

The contrasting behaviors of PHEV, GEV and MEV for the prediction of the highest quantiles are further analyzed in the insets of panels B, C and D of Figure 1, which display the median performances among case studies belonging to ten different hazard categories (i.e., the deciles of the observed flood quantiles). For PHEV (inset of panel D, light green hue) markers lie just below and parallel to the 45-degree line, indicating a slight tendency to underestimation that is independent of the flood magnitude. The markers instead diverge from the 45-degree lines in the insets of panels B and C (i.e., for GEV and MEV), indicating that the estimation errors associated to these distributions are low for small floods but increase with increasing quantiles.

To better visualize these traits and their implications for estimating magnitudes of events with long return periods, we plotted the median relative errors for each hazard class as a function of the observed quantiles (Figure 2A). We then extrapolated the performances of the different extreme distributions to higher quantiles by fitting regression lines using only half the observed errors, randomly extracted a hundred times from the whole set of them by means of resampling without substitution. This procedure yields an evaluation of the uncertainty of the extrapolations (shaded areas in panel A) and concurrently allows for inspecting their robustness against the observations themselves, avoiding potential errors due to the limited sample available. The performance of PHEV (light green shaded area) is lesser affected by increasing flood quantiles, whereas it ceaselessly deteriorate with the magnitude of floods for GEV and MEV (red and blue shaded areas). This finding is especially relevant for the estimation

of high quantiles associated to very long return periods. In this situation, errors of PHEV shall remain fairly limited while those of GEV and MEV are expected to further increase.

We provide a stricter assessment of the performances of the three extreme value approaches for very high quantiles in the remainder of Figure 2. There we summarize the analyses performed on 1000 years long synthetic time series of rainfall and streamflow for eight select catchments from different US hydro-climatic regions (red squares in Figure 1G), which exhibit weak relations between K and peak flow.

A compendium of the results is given in Figure 2B-G, which displays boxplots of the skill scores (panels B-C) and relative errors (panels D-G) of PHEV, GEV and MEV. We computed both the metrics in a predictive fashion, i.e., by only considering in the computation quantiles corresponding to return periods T longer than the sample size S used to calibrate the parameters of the distributions.

Figure 2B-C shows the skill scores of PHEV, GEV and MEV when results for all the calibration sample sizes are pooled together. The skill score is a global metric of estimation accuracy that varies between $-\infty$ and 1, the latter value indicating a perfect match between predicted quantiles and the validation sample. PHEV and GEV have on average similar values of median skill score, which are higher than for MEV (panel B). The latter result would likely improve if variations of the parameters of the ordinary peak distribution across time intervals would be allowed for MEV, thus exploiting its full potentiality (Miniussi and Marani, 2020). Although GEV appears to have the lowest variance across the distributions, panel C clarifies the reality and discloses the large number and extent of outliers among GEV estimates. A similar pattern is also highlighted by the relative errors between estimated and observed maxima, computed for each $T > S$ and then pooled in the boxplots of Figure 2D-G for all select river basins and every S = 10, 20, 30 and 60 years. A relative error equal to zero indicates a perfect match between quantiles predicted by the model and those estimated from the validation sample. Negative values of the median relative error for MEV (panels D and F) reflect its tendency to underestimate high quantiles, while limited error variances and number of outliers indicate stability of its performances. Notwithstanding comparable performances of GEV and PHEV for what concerns median errors, the latter exhibits lower error variance than the former in both the ranges of $\frac{T}{S}$ analyzed in the figure and its capability to provide estimates does not strongly deteriorate for $\frac{T}{S} > 10$, as also confirmed by the distribution of outliers (panel G).

The capabilities of the three distributions (and of an empirical estimation method, i.e., the Weibull plotting position) to reproduce the reference normalized flood magnitude-frequency curve (grey dots) obtained from a 1000 years long synthetic time series when a limited yet common in the hydrological practice calibration sample (S=20 years) is available are exemplified in panels H-M. We display the results for these two specific case studies to emphasize how the shape of the flood frequency curve might affect the results given by the three methods. Namely, we chose one case (panels H-J) where the flood frequency curve is rather predictable (i.e., the magnitude of the normalized

streamflow smoothly rises with increasing return period) and one case (panels K-M) characterized by the opposite behavior (i.e., the flood magnitude markedly increases above a certain return period).

The black lines and shaded areas span the breadth of empirical (i.e., the range spanned by the calibration samples randomly extracted 1000 times), PHEV (light green), GEV (red) and MEV (blue) estimates. The empirical method reveals its fickleness, as the related estimates highly depend on the specific set of flood events sampled in the time frame S even for short return periods (i.e., the uncertainty of the maximum associated to an assigned frequency is considerable for return periods as low as 10 years). The use of a theoretical extreme value distribution reduces this variability (i.e., the shaded areas span smaller ranges across the y-axes at short return periods). However, GEV estimates (panels I and L) are as variable for just slightly larger return periods (i.e., 50 years). Conversely, the uncertainty of MEV is limited (i.e., the blue shades in panels J and M are narrow) for the whole range of investigated return periods, but the method tends to underestimate magnitudes of rare events with return periods longer than a few hundred years, in particular when the flood frequency curve is characterized by the presence of an inflection point (panel M). PHEV (panels H and K) provides dependable estimates of flood magnitude-frequency curves which are less affected by the specific set of flood events available in the sample (i.e., the amplitude of the light green shade is as well limited). Albeit more uncertain than MEV, the pattern of PHEV estimates closely resembles that of the reference flood magnitude-frequency curve (grey dots) for return periods above a hundred years, especially when the curve trends upward. Remarkably, knowledge of the magnitude of events with return periods substantially longer than the available data series is required at most in the practice.

4. Conclusions and Outlook

The PHysically-based Extreme Value (PHEV) distribution of river flows emerges from a mechanistic-stochastic description of basin-scale soil moisture dynamics and runoff generation processes, which provides consistent analytical expressions of the probability distributions of daily flows, peak flows and flow maxima. It constitutes a physically-based alternative to purely statistical methods hitherto used to characterize extremes of hydrological variables, whose parameters are all but one measurable from daily rainfall and streamflow records.

A novel method relying on flow recession analyses and the identification of an inverse relation between recession coefficients and peak flows allows for the domain of applicability of PHEV to be defined a priori. In the absence of such relation PHEV guarantees remarkable performances which are comparable to those of two leading statistical models for flood hazard assessment. In particular, PHEV estimates entail lesser increasing bias with larger flood quantiles than it occurs for the other benchmark approaches. This feature ensures more reliable appraisal of the magnitude of rare extreme floods from relatively short data series, an especially valuable achievement

for the community of professionals involved in environmental hazards assessment.

Notwithstanding the highlighted perks of adopting a mechanistic-stochastic approach to estimate magnitudes and frequencies of floods from daily streamflow dynamics, work remains to be done. Stable calibration methods, such as probability weighted and L-moments, shall be developed for PHEV as done in the past decades for standard probability distributions now routinely used in extreme value analyses. The search for approaches to estimate values of the parameter K without resorting to calibration on observed maxima (thus achieving entirely calibration-free predictions of flood statistics) shall also continue. Given the apparent dependence of K on the wetness conditions of river basins (Shaw et al., 2013; Bart and Hope, 2014), satellite-derived water storage and soil moisture products are promising allies in this regard, as recently shown by Sharma et al. (2020) and Basso et al. (2021).

The mechanistic-stochastic conceptualization of runoff generation processes underlying PHEV might as well be further advanced to explicitly include the variability of recession coefficients in its mathematical expression. Seminal works in this sense (Botter, 2010; Basso et al., 2015) demonstrated visible improvements of high flow estimates when this variability is embedded in the mechanistic-stochastic description of daily flows. Using the physically-based distribution of peak flows underpinning PHEV to describe ordinary events in the MEV framework constitutes another promising research avenue whereby capitalizing on the assets of both methods, namely the mechanistic description of runoff generation processes embedded in PHEV and the possibility to explicitly account for the inter-annual variability of parameters provided by MEV.

Finally, our findings hint at the possibility to leverage the reliability of PHEV estimates for very long return periods to pinpoint inflection points of flood magnitude-frequency curves, whose occurrence entail abrupt increases of the magnitude of high flows, amplified hydrological hazard and a propensity of rivers to generate extreme floods. Addressing these aspects would advance the PHysically-based Extreme Value (PHEV) distribution of river flows towards ripeness for a standard application in the practice.

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. Hydro-climatic time series from the MOPEX data set are available at https://hydrology.nws.noaa.gov/pub/gcip/mopex/USData. Synthetic hydro-climatic time series for the select catchments used in this study to assess performances for very long return periods are available at: http://www.hydroshare.org/resource/abd449fbced146eb8d0ec8ad95754f1c.

Appendix A. Probability distributions of daily streamflows, ordinary peak flows and flow maxima

The expression of the probability distribution of daily streamflows q (Botter et al., 2009) is:

$$p(q) = C_1 q^{-a} e^{\frac{\lambda q^{1-a}}{K(1-a)} - \frac{q^{2-a}}{\alpha K(2-a)}} \tag{A.1}$$

where α and λ are average and frequency of runoff pulses, K and a are coefficient and exponent of power law hydrograph's recessions, and C_1 is a normalization constant.

The expression of the probability distribution of ordinary peak flows (Basso et al., 2016) is:

$$p_j(q) = C_2 q^{1-a} e^{\frac{\lambda q^{1-a}}{K(1-a)} - \frac{q^{2-a}}{\alpha K(2-a)}} \tag{A.2}$$

which depends on the same parameters of (A.1), where C_2 is also a normalization constant.

The expression of the physically-based probability distribution of flow maxima (i.e., maximum values in a specified time frame) (Basso et al., 2016) is:

$$p_M(q) = \lambda \tau e^{-\lambda \tau D_j(q)} p_j(q) \tag{A.3}$$

where $\tau[T]$ is the duration in days of the chosen time frame, $p_j(q)$ is the probability distribution of peak flows (A.2), and $D_j(q) = \int_q^\infty p_j(q)\,\mathrm{d}q$ is the exceedance cumulative probability (i.e., the duration curve) of peak flows.

Appendix B. Subset of river basins from different US hydro-climatic regions

Information concerning the eight select catchments from different US hydro-climatic regions, which are used in this study to assess performances of the theoretical extreme distributions for very long return periods ($O(10^3)$), are available in Table A1.

Table B1. Select catchments from different US hydro-climatic regions used to assess the performance of the analyzed methods for very long return periods ($O(10^3)$)

USGS ID	Longitude	Latitude	River name	State	Area [km^2]
12098500	-121.9486	47.1514	White River near Buckley	WA	1039
11501000	-121.8486	42.5847	Sprague River near Chiloquin	OR	4092
11025500	-116.8653	33.1069	Santa Ysabel Creek near Ramona	CA	290
08032000	-95.4306	31.8922	Neches River near Neches	TX	2966
05471500	-92.6586	41.3553	South Skunk River near Oskaloosa	IA	4235
03448000	-82.5925	35.5019	French Broad River at Bent Creek	NC	1751
03075500	-79.4256	39.4219	Youghiogheny River near Oakland	MD	347
01372500	-73.8731	41.6531	Wappinger Creek near Wappingers Falls	NY	469

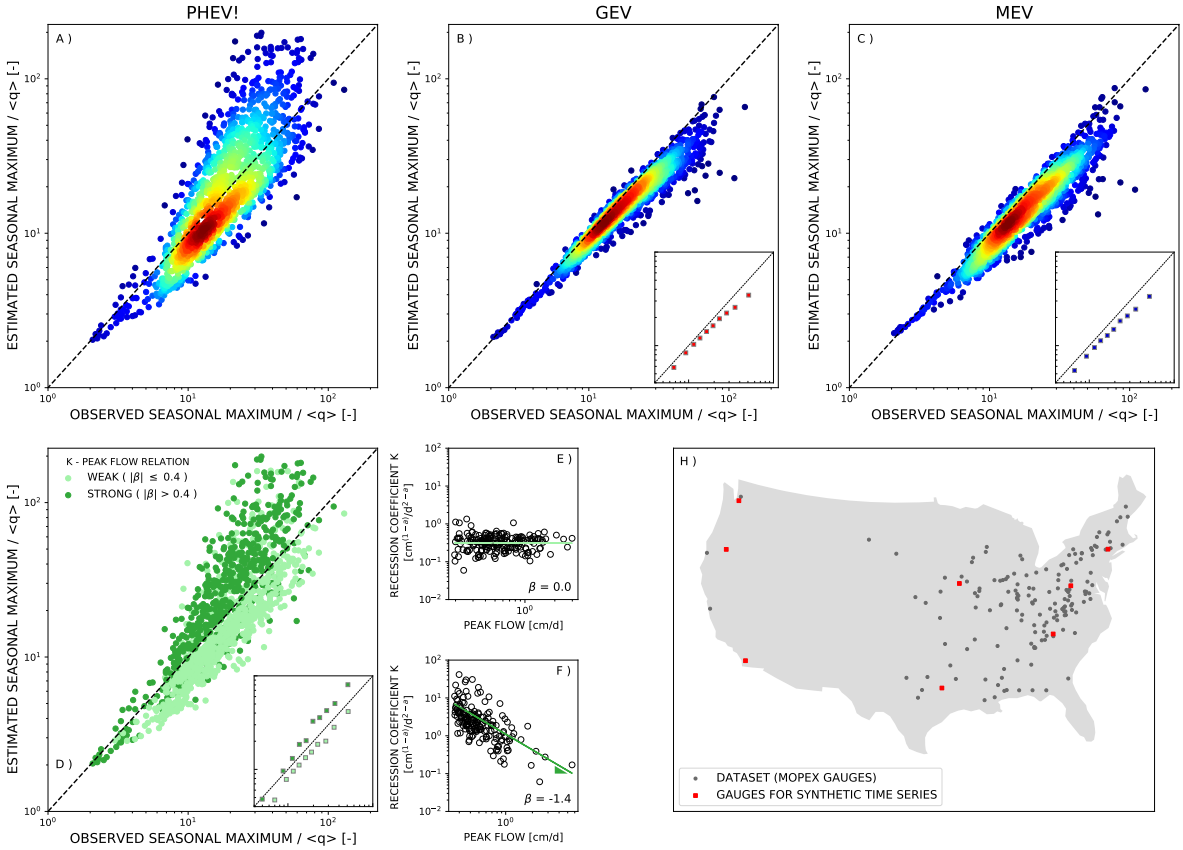

Figure 1. (A-D) Quantile-quantile plots of normalized flood magnitudes (i.e., seasonal maxima divided by the long term average daily flow, $<q>$, of the catchment) estimated for all catchments and seasons by means of Weibull plotting position of observations in the validation samples (horizontal axes) and either (vertical axes) (A and D) the PHysically-based Extreme Value distribution (PHEV), (B) the Generalized Extreme Value distribution (GEV) or (C) the Metastatistical Extreme Value distribution (MEV) calibrated on 10 years long samples. Different realizations of the calibration samples have been assembled by resampling without substitution the available observational series a hundred times (the remaining part of the data series constitutes each time the validation sample). Median results among all the realizations are plotted. Notice that only quantiles corresponding to return periods longer than the length of the calibration samples (i.e., $T_r > 10$ years) and until the second longest T_r which could be estimated from the validation sample (to avoid large sampling uncertainty inherent to the longest empirical return period) are shown. Warmer colors in panels (A) to (C) indicate higher density of points and thus mark the most frequent performances of the methods. Dark green markers in panel (D) identify case studies for which a strong relation between recession coefficient and ordinary peak flows was detected (see panel (F)). The current mechanistic description of the catchment hydrologic response embedded in PHEV does not consider such a relation and prevents it from providing satisfactory performances in these cases, which can be identified ex ante through hydrograph recession analyses. Light green markers in panel (D) instead indicate catchments where such a relation is weak (see panel (E)). In these cases the mechanistic conceptualization of hydrological processes underpinning PHEV is suitable and the physically-based extreme value distribution of river flows guarantees satisfactory performances. The insets of panels (B), (C) and (D) represent the median performances for all case studies in the figures belonging to ten different hazard categories (i.e., the deciles of the observed normalized flood quantiles). (E-F) Exemplary case studies displaying (E) weak (USGS ID: 03253500, spring) and (F) strong (USGS ID: 03504000, summer) relations between recession coefficients and ordinary peak flows. The recession coefficient is fairly constant for increasing peak flows in (E), thus supporting the adoption of an effective parameter K in these cases and justifying its efficacy for estimating flood quantiles (light green markers in (D)). Conversely, the recession coefficient markedly decreases (of several orders of magnitudes) for increasing values of the peak flows in (F). This determines declining performances of PHEV for increasing quantiles when an effective parameter K is adopted (dark green markers in (D)). (G) Location of gauges from the MOPEX dataset analyzed in this study (grey dots). Red markers display a subset of river basins from varied hydro-climatic regions that have been selected to perform analyses on 1000 years long synthetic time series (see Figure 2).

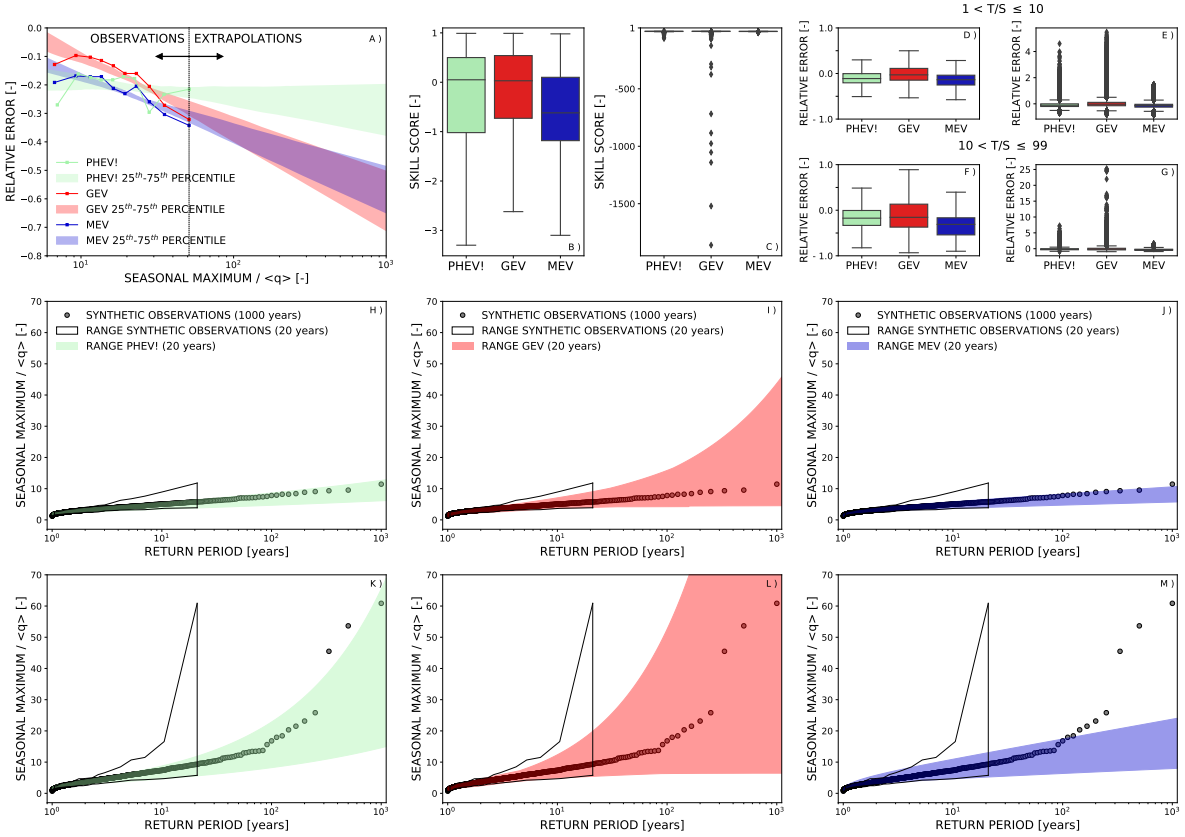

Figure 2. (A) Relative flood estimation errors for increasing quantiles resulting from the use of PHEV (light green), GEV (red) and MEV (blue) distributions. The left-hand side of the panel encompasses the range of quantiles for which errors associated to the use of different statistical distributions could be evaluated from observations. Extrapolations of these errors for higher quantiles (i.e., less frequent floods), whose estimation is often required in the practice, are displayed as shaded areas on the right-hand side of the plot. (B-G) Boxplots (without and with displayed outliers) summarizing performance metrics for all select case studies and every $S = 10$, 20, 30 and 60 years. Performance is assessed by comparing event magnitudes estimated from the $(1000 - S)$ years long series of synthetic observations with those computed by means of statistical distributions calibrated on shorter data series of length S assembled through resampling without substitution. Only return periods longer than the sample size S used for calibration are considered to compute performances. (B-C) Skill score of PHEV, GEV and MEV and (D-G) Relative error between quantiles predicted by the models and the validation sample. Panels (D) and (E) show errors for return periods between one and ten times the length of the calibration samples (i.e., errors for frequent floods), whereas panels (F) and (G) display errors for rare extreme events (i.e., return periods between 10 and 99 times the length of the calibration sample). (H-M) Normalized (i.e., seasonal maximum divided by the long term average daily flow, $< q >$) flood magnitude-frequency curves estimated by means of synthetic data series and three different statistical distributions. Results for the Youghiogheny River near Oakland, MD (USGS gauge 03075500) in the spring season (panels H, I, J) and the South Skunk River near Oskaloosa, IA (USGS gauge 05471500) in the summer season (panels K, L, M) are here displayed as exemplary case studies. Grey dots present the estimates obtained by means of Weibull plotting position of 1000 years long synthetic time series generated for the considered basins. Black solids lines and shaded areas enclose the ranges of normalized flood magnitude-frequency curves obtained when shorter data series of 20 years length (assembled by resampling without substitution the long series a thousand times) are used for estimation either by means of Weibull plotting position of the shorter series (black solid lines) or the support of (H and K) the PHysically-based Extreme Value distribution (PHEV, light green areas), (I and L) the Generalized Extreme Value distribution (GEV, red areas) and (J and M) the Metastatistical Extreme Value distribution (MEV, blue areas). The plots highlight superior reliability of PHEV for long return periods, especially when discontinuities in the flood magnitude-frequency curve exist.

References

Abramowitz M and Stegun I E 1972 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* New York.

Baratti E, Montanari A, Castellarin A, Salinas J L, Viglione A and Bezzi A 2012 *Hydrol. Earth Syst. Sci.* **16**, 4651–4660.

Barnett V 1975 *Appl. Stat.* **24**, 95 – 108.

Bart R and Hope A 2014 *J. Hydrol.* **519**, 205 – 213.

Barth N, Villarini G and White K 2019 *J. Hydrol. Eng.* **24**.

Basso S, Ghazanchaei Z and Tarasova L 2021 *Sci. Total Environ.* **756**, 143469.

Basso S, Schirmer M and Botter G 2015 *Adv. Water Resour.* **82**, 98 – 105.

Basso S, Schirmer M and Botter G 2016 *Geophys. Res. Lett.* **43**(17), 9070–9076.

Bevere L, Gloor M and Sobel A 2020 Natural catastrophes in times of economic accumulation and climate change Technical report Swiss Re Institute.

Biswal B 2021 *Adv. Water Resour.* **147**, 103822.

Biswal B and Marani M 2010 *Geophys. Res. Lett.* **37**(24). L24403.

Biswal B and Marani M 2014 *Adv. Water Resour.* **65**, 34 – 42.

Biswal B and Nagesh Kumar D 2015 *Adv. Water Resour.* **77**, 37 – 43.

Botter G 2010 *Water Resour. Res.* **46**(12).

Botter G, Basso S, Rodriguez-Iturbe I and Rinaldo A 2013 *Proc. Natl. Acad. Sci. USA* **110**(32), 12925–12930.

Botter G, Porporato A, Rodriguez-Iturbe I and Rinaldo A 2007 *Water Resour. Res.* **43**(2).

Botter G, Porporato A, Rodriguez-Iturbe I and Rinaldo A 2009 *Water Resour. Res.* **45**(10).

Coles S 2001 *An Introduction to Statistical Modeling of Extreme Values* London.

Dralle D N, Hahm W J, Rempe D M, Karst N J, Thompson S E and Dietrich W E 2018 *Hydrol. Process.* **32**(13), 1978–1992.

Durrans S R, Eiffe M A, Thomas Jr. W O and Goranflo H M 2003 *J. Hydrol. Eng.* **8**.

DWA 2012 Ermittlung von hochwasserwahrscheinlichkeiten Technical report Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall (DWA).

England J, Cohn T, Faber B, Stedinger J, Thomas W, Veilleux A, Kiang J and Mason R 2019 Guidelines for determining flood flow frequency—Bulletin 17C Technical report U.S. Geological Survey.

Garbin S, Celegon Alessi E, Fanton P and Botter G 2019 *R. Soc. Open Sci.* **6**(2), 181428.

Ghosh D K, Wang D and Zhu T 2016 *Adv. Water Resour.* **88**, 8–13.

Gnedenko B 1943 *Ann. Math.* **44**(3), 423–453.

Hashino T, Bradley A A and Schwartz S S 2007 *Hydrol. Earth Syst. Sci.* **11**(2), 939–950.

Hirpa F A, Salamon P, Beck H E, Lorini V, Alfieri L, Zsoter E and Dadson S J 2018 *J. Hydrol.* **566**, 595–606.

Hosking J 1990 *J. R. Stat. Soc. Ser. B* **52**(1), 105–124.

Hosseini S R, Scaioni M and Marani M 2020 *Geophys. Res. Lett.* **47**.

Katz R W, Parlange M and Naveau P 2002 *Adv. Water Resour.* **25**, 1287–1304.

Kirchner J 2006 *Water Resour. Res.* **42**(3).

Kjeldsen, T.R. J D and Bayliss A 2008 Improving the feh statistical procedures for flood frequency estimation Technical report European Environment Agency (EEA).

Klemeš V 1989 *in* 'Proceedings of the World Meteorological Organization Technical Conference Held in Geneva' World Meteorological Organization London p. 43–51.

Knijff J M V D, Younis J and Roo A P J D 2010 *Int. J. Geogr. Inf. Sci.* **24**(2), 189–212.

Kuffour B N O, Engdahl N B, Woodward C S, Condon L E, Kollet S and Maxwell R M 2020 *Geosci. Model Dev.* **13**(3), 1373–1397.

Laio F, Porporato A, Ridolfi L and Rodriguez-Iturbe I 2001 *Adv. Water Resour.* **24**(7), 707 – 723.

Marani M and Ignaccolo M 2015 *Adv. Water Resour.* **79**, 121–126.

Marra F, Borga M and Morin E 2020 *Geophys. Res. Lett.* **47**.

Marra F, Nikolopoulos E, Anagnostou E and Morin E 2018 *Adv. Water Resour.* **117**, 27–39.

Marra F, Zoccatelli D, Armon M and Morin E 2019 *Adv. Water Resour.* **127**, 280–290.

Maxwell R and Miller N 2005 *J. Hydrometeorol.* **6**, 233–247.

Metzger A, Marra F, Smith J A and Morin E 2020 *J. Hydrol.* **590**.

Miniussi A and Marani M 2020 *Water Resour. Res.* **56**(7).

Miniussi A, Marani M and Villarini G 2020 *Adv. Water Resour.* **136**.

Miniussi A, Villarini G and Marani M 2020 *Geophys. Res. Lett.* **47**.

Morrison J E and Smith J A 2002 *Water Resour. Res.* **38**(12).

Murphy A H and Winkler R L 1992 *J. Forecast.* **7**, 435–455.

Mutzner R, Bertuzzo E, Tarolli P, Weijs S V, Nicotina L, Ceola S, Tomasic N, Rodriguez-Iturbe I, Parlange M B and Rinaldo A 2013 *Water Resour. Res.* **49**(9), 5462–5472.

Patnaik S, Biswal B, Kumar D N and Sivakumar B 2015 *J. Hydrol.* **528**, 321 – 328.

Pickands J I 1975 *Ann. Stat.* **3**(1), 119–131.

Porporato A, Daly E and Rodriguez-Iturbe I 2004 *Am. Nat.* **164**, 625 – 632.

Rupp D E, Schmidt J, Woods R A and Bidwell V J 2009 *J. Hydrol.* **377**(1), 143–154.

Schaake J, Duan Q, Andréassian V, Franks S, Hall A and Leavesley G 2006 *J. Hydrol.* **320**(1), 1 – 2. The model parameter estimation experiment.

Schellander H, Lieb A and Hell T 2019 *Earth Space Sci.* p. 2019EA000557.

Sharma D, Patnaik S, Biswal B and Reager J T 2020 *Geosciences* **10**(10).

Shaw S B, McHardy T M and Riha S J 2013 *Water Resour. Res.* **49**(9), 6022–6028.

Slater L and Villarini G 2017 *Water* **9**(9), 2019EA000557.

Tarasova L, Basso S and Merz R 2020 *Geophys. Res. Lett.* **47**(22), e2020GL090547.

Tashie A, Pavelsky T and Emanuel R E 2020 *Water Resour. Res.* **56**(3), e2019WR026425.

Tashie A, Scaife C I and Band L E 2019 *Hydrol. Process.* **33**(19), 2561–2575.

The Economist 2017 *The Economist* .

Villarini G and Strong A 2014 *Agric. Ecosyst. Environ.* **188**, 2014–2011.

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

Wang D and Hejazi M 2011 *Water Resour. Res.* **47**.

Zorzetto E, Botter G and Marani M 2016 *Geophys. Res. Lett.* **43**, 8076–8082.

Zorzetto E and Marani M 2019 *Water Resour. Res.* **55**, 156–174.

Zorzetto E and Marani M 2020 *Adv. Water Res.* **135**, 103483.