Metastatistical Extreme Value Distribution applied to floods across the continental United States

Arianna Miniussi[1,*][1], Marco Marani[1,2] and Gabriele Villarini[3]

[1] Department of Civil, Architectural, and Environmental Engineering, University of Padua, Padua, Italy

[2] Department of Civil and Environmental Engineering, Duke University, Durham, NC, USA

[3] IIHR-Hydroscience & Engineering, University of Iowa, Iowa City, Iowa, USA

Abstract

This study analyzes daily mean streamflow records from 5,311 U.S. Geological Survey stream gages in the continental United States and develops a Metastatistical Extreme Value Distribution (MEVD) tailored for flood frequency analysis. We compare the new tool with the Generalized Extreme Value (GEV) and Log-Pearson Type III (LP3) distributions and investigate the role of El Niño Southern Oscillation (ENSO) in the generation of floods. Hence, we formulate the MEVD in terms of mixture of distributions to describe the occurrence of flood peaks generated under different ENSO phases. We find that the MEVD outperforms GEV and LP3 distributions respectively in about 76% and 86% of the stations, with a significant improvement in the accuracy of quantiles corresponding to return periods much larger than the calibration sample size. The ENSO signature detected in the distributions of the daily peak flows does not necessarily improve the estimation of high return period flow values.

[1] Corresponding author.
E-mail address: arianna.miniussi@dicea.unipd.it (A. Miniussi)

* Current address: Catchment Hydrology Department, Helmholtz Centre for Environmental Research - UFZ, Halle (Saale), Germany

1. Introduction

Globally, in the period 1998-2017, floods have been the most frequent disaster (43.4% of the natural disasters) and have caused more than 140,000 deaths (representing 11% of the fatalities due to natural disasters of all types) (*Wallemacq and House*, 2018). Within this global context, during the 20th century floods in the United States were the number-one natural disaster in terms of the number of lives lost and property damage (*Perry*, 2000), the costliest (*Miller et al.*, 2008) and affected the largest number of people (*Stromberg*, 2007). They are also the second weather-related hazard in terms of fatalities in the United States, with 4,586 reported deaths between 1959 and 2005, mainly due to flash floods caused by heavy precipitation (*Ashley and Ashley*, 2008). Reliable flood frequency estimation methods are the basis to devise and implement strategies for the mitigation of these societal and economic impacts, with applications in a number of fields, from the design of hydraulic structures, to environmental management and planning, to flood insurance. A key concept is the design flood peak value, typically set in national regulations by specifying an average recurrence interval, or return period, T, associated with a probability of being exceeded in each year equal to $P=1/T$. The T-year flood, in turn, is estimated based on the analysis of past floods, requiring the selection of a probability distribution to perform this inference using a sample with size $S<<T$ (*Benson*, 1962).

Flood frequency analyses are commonly performed as a part of engineering and planning projects, but they too often represent a mere statistical fitting exercise, and do not attempt to incorporate a representation of the underlying physical processes. As noted by *Klemeš* (1988, 1993), the standard approach of extreme event probability estimation is moving towards a higher mathematical abstraction, renouncing any leveraging or understanding of the different flood-generating mechanisms at play. The hydrological-process information in the observations is thus often neglected, and the selection of an optimal statistical model through a goodness-of-fit metric remains the main focus of these types of analyses. The approach proposed here does not belong to the "standard" methodology on which flood frequency analyses are usually based. Three points of departure can indeed be highlighted. First, it is not a mere fit of a distribution on annual maxima but, by dealing with all the observations (see Section "Data and Methodology" for further details), the approach described here is closer to the "raw" observations. Second, the approach we propose for flood frequency analysis is not a mere "goodness of fit exercise", which is what Klemeš (1988, 1993) was stigmatizing. One of the important points highlighted here is that we need to address predictive uncertainty through a cross-validation procedure, rather than just use unexplained variance as a measure of uncertainty. Third, but not less important, the proposed approach does not just consider peak flows as random numbers, but attempts to base the statistical modelling on physical drivers.

The U.S. federal guidelines themselves (Bulletin 17-B (*IACWD*, 1982), and its updated version 17-C (*England et al.*, 2018)) recognize that the assumption under which stream gage records are generated by one single flood-generating mechanism may not always be realistic. They highlight the need to understand and more accurately identify these physical mechanisms and list the identification and treatment of mixed distributions to represent their diversity as a research and application priority.

Even though the idea of a more process-driven flood frequency analysis is not necessarily new (e.g. *Hirschboeck*, 1987), there has been a renewed interest in recent years in process-based formulations of extreme flood distributions (e.g., *Alila and Mtiraoui*, 2002; *Smith et al.* 2011; *Villarini and Slater*, 2017; *Barth et al.*, 2019). Flood-generating processes can quite naturally be analyzed using mixed distributions (e.g., *Alila and Mtiraoui*, 2002), but the determination of which flood peaks result from the different processes and of when the use of mixed distributions is beneficial remains an open problem (e.g., *Villarini and Slater*, 2017).

There are several different hydrological mechanisms that can drive the occurrence of flood events, including snowmelt, frontal systems, local convective processes, monsoons, and intense tropical cyclones (see *Villarini*, 2016; *Zhang et al.*, 2017). *Slater et al.* (2015) compared hydrologic and geomorphic drivers in flood hazard, while *Berghuijs et al.* (2016) analyzed the dominant flood generating mechanisms across the United States. *Barth et al.* (2017, 2019) investigated the role of atmospheric rivers in the generation of flood peaks across the western United States and suggested a weighted mixed population approach to perform a process-driven flood frequency analysis to reflect the differences in flood agents.

Within the context of mixed distributions, this study formulates a novel flood frequency distribution and uses it to investigate the role of El Niño Southern Oscillation (ENSO) in the generation of floods, and the detectability of its signature in observed records. ENSO is a major mode of variability of the coupled atmosphere-ocean system associated with episodes of above-normal (El Niño) and below-normal (La Niña) sea surface temperature in the tropical Pacific Ocean, with impacts on seasonal winds, rainfall, and temperature across the globe.

We propose here the use of a novel approach, the Metastatistical Extreme Value Distribution (MEVD), which can naturally incorporate mixed distributions to represent flood magnitudes generated by different mechanisms. The MEVD has been introduced by *Marani and Ignaccolo* (2015) and has been applied mostly to rainfall (*Zorzetto et al*., 2016; *Marra et al*., 2018; *Zorzetto and Marani*, 2019), for which it was shown to provide significantly smaller estimation uncertainty when compared to traditional approaches, especially when considering return periods that are larger than the sample size used for distribution estimation.

Currently, the MEVD has yet to be applied to flood magnitudes and flood frequency analysis. Here we provide the first such application and ask the following relevant questions: does the MEVD outperform the traditional Generalized Extreme Value (GEV) distribution in flood frequency analysis? Does the incorporation of mixed probability distributions representing different types of flood events associated with different ENSO phases improve the estimation of event magnitudes with high return periods?

To answer these questions, we apply the MEVD approach to daily records from stream gage stations across the continental United States (CONUS), examining the role played by mixtures of distributions associated with different ENSO phases. Results are compared and contrasted against those from the GEV distribution, providing qualitative and quantitative evaluations of their relative predictive performance. The GEV distribution is the natural reference benchmark for this comparison, as it stems from the widely-applied traditional Extreme Value Theory (*Smith*, 1987; *Katz et al.*, 2002). However, other methods are widely applied in the flood-frequency analysis practice. We thus extended the comparison of cross-validation performances to the Log-Pearson Type III (LP3), the distribution extensively used in flood frequency analyses in the US (*USWRC*, 1976).

The paper is organized as follows: the data and methodology used in the study are described in Section 2, followed by a Section 3 detailing the results of the analyses. Section 4 summarizes the main points of this study and concludes the paper.

2. Data and methodology

We analyzed daily records from 5,311 U.S. Geological Survey (USGS) stream gages across the continental United States (Figure 1, panel a). We focus on water years, defined to run between October 1 and September 30, and select only sites where at least 30 complete (i.e. with more than 330 daily observations/year) years of observation exist, and where no statistically significant trends are found (at the 5% level, based on the Mann-Kendall test; *Mann*, 1945, *Kendall*, 1975). The historical time series selected cover the period 1916-2017, with record lengths between 31 and 101 years (Figure 1, panels b and c; consult Supplementary Figure 1 for an overview of the spatial distribution of the lengths of the historical time series).

Flood frequency analysis requires the identification of independent events: here, we identify the largest flood peaks within blocks of length equal to $T = 10 \cdot days + log(A)$, where A is the drainage area in square miles (*Lang et al.*, 1999). Additionally, we discard the smallest discharge peak within any pair of consecutive peaks if the minimum flow between them does not drop below a threshold equal to 75% of the lower of the two (Water Resources Council, *USWRC*, 1976). This additional condition is necessary to eliminate secondary peaks occurring during recession periods of previous floods. The entire set of peak discharge values resulting from this selection process of uncorrelated events is here called the set of "ordinary events" to denote that it contains all the independent events that have occurred in the record, irrespective of their magnitudes.

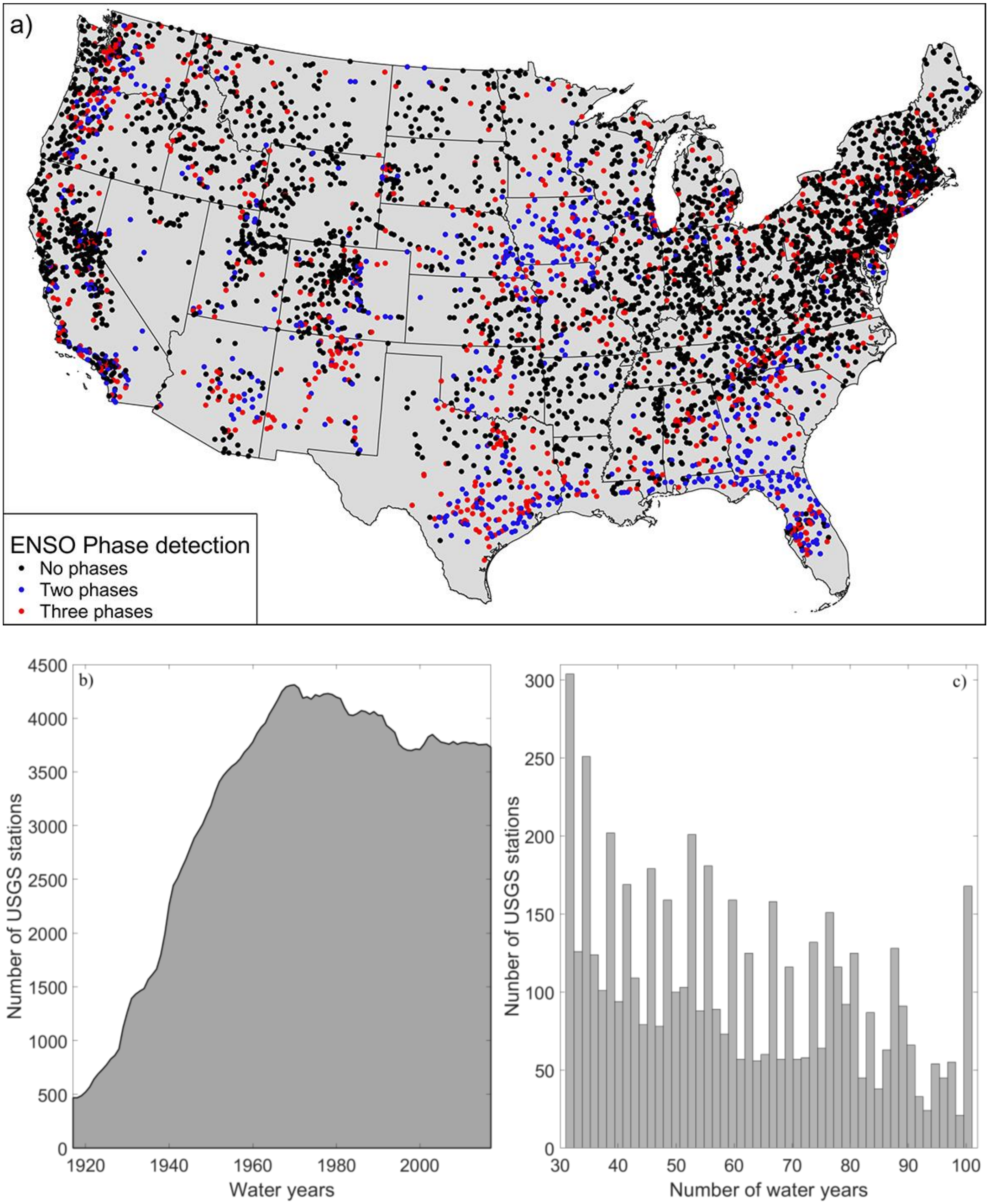

Figure 1 – Panel a: Spatial distribution of the selected stream gages. The colors refer to the statistical detectability of ENSO phases in the distribution of discharge peak values: black dots indicate stations where the frequency distributions of ordinary peak discharge values in different ENSO phases are indistinguishable from one another; stations for which three different phases are detected are displayed in red; in blue, stations where two out of three ENSO phases can be distinguished. Panel b shows the cumulative number of stations with data in each water year, while the histogram in panel c summarizes the number of stations in terms of the number of years on record.

2.1 MEVD approach

The Metastatistical Extreme Value Distribution, originally introduced by *Marani and Ignaccolo* (2015), is explicitly formulated on the basis of the probability distribution(s) of the ordinary values, from which the distribution of extremes (annual maxima) is then derived. Hence, the expression of the MEVD describing the extreme events is identified by estimating its parameters using the entire set of observed ordinary events. This is quite different from the assumptions at the basis of the traditional Extreme Value Theory (EVT), which focuses on fitting a distribution to the annual maxima or to relatively few values above a high threshold.

The MEVD approach treats as realizations of stochastic variables both the parameters of the distributions describing the ordinary events in each year, $F(x;\boldsymbol{\theta})$ (where $\boldsymbol{\theta}$ is the parameter vector), and the number, n, of yearly event occurrences. Under these premises, the MEVD of yearly maxima can be defined as:

$$\zeta(x) = \sum_{n=1}^{\infty} \int_{\Omega_{\boldsymbol{\theta}}} [F(x;\boldsymbol{\theta})]^n g(n,\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{1}$$

where $g(n,\boldsymbol{\theta})$ is the joint probability distribution of the random variables N and $\boldsymbol{\Theta}$ (discrete in n and continuous in $\boldsymbol{\theta}$) and Ω_{Θ} is the population of the parameters values. The ensemble average can be approximated by the sample average computed over all the years of the historical series (M), becoming

$$\zeta(x) = \frac{1}{M} \cdot \sum_{j=1}^{M} [F(x;\boldsymbol{\theta}_j)]^{n_j} \tag{2}$$

where $F(x;\boldsymbol{\theta}_j)$ is the cumulative distribution of ordinary values and n_j is the number of events in year j.

Here, we apply the MEVD to peak discharges and modify this approach to account for ordinary values belonging to different populations, corresponding, in the present case, to different ENSO phases. The cumulative distribution function $\zeta(x)$ of the mixed-MEVD can be written as follows:

$$\zeta(x) = \frac{1}{M}\sum_{j=1}^{M}\prod_{p=1}^{n_{ph}}\left[F_p\left(x;\boldsymbol{\theta}_j\right)\right]^{n_{j,p}} \tag{3}$$

where n_{ph} is the number of phases that induce statistically different distributions of the ordinary events and should therefore be considered separately; F_p is the yearly (or time window, when the low number of events/year requires parameter estimation to be performed of multi-year windows) cumulative distribution of the ordinary values in phase p; $n_{j,p}$ is the original yearly number of the peaks in phase p and year j; M is the number of years for which observations are available. Eq.(3) reduces to the original formulation in *Marani and Ignaccolo* (2015) when only one phase is present. The mixed MEVD formulation in Eq.(3) is the same as the approach proposed in *Marra et al.* (2019). Here, instead of using the further simplification (SMEV, *Marra et al.*, 2019) that removes the inter-annual variability in the statistics of the ordinary events (i.e. there is no dependence on j in Eq.(2)), we preserve the time variability of the distributions.

The first step in the application of the MEVD approach is identifying a suitable parametric distribution to represent the ordinary events. We evaluate three candidate distributions for the $F(x;\boldsymbol{\theta}_j)$ in Eq.(2): Weibull, Generalized Pareto, and Gamma distributions. We select the most suitable distribution on the basis of the skill score (see Section on the evaluation metrics for its definition), comparing the MEVD-estimated quantiles to the observed maxima. In the present analyses, the Gamma distribution was the best performing one (see Section 3 for further details).

Because the average number of flood events in a year is small for many stations (e.g., in 1243 of the 5311 analyzed stations the average number of peaks/year is smaller than or equal to 10; see Supplementary Figure 2), we explore estimating the parameters of the distribution using either five-year windows or the entire sample. Eq. (3) can be hence expressed in terms of the windows used for parameter estimation:

$$\zeta(x) = \frac{1}{M}\cdot\sum_{j=1}^{M}\prod_{p=1}^{n_{ph}}\left[F_p(x;\boldsymbol{\theta}_{k(j)})\right]^{n_{j,p}} \tag{4}$$

where *k(j)* is the index value identifying the window containing year *j*, $\boldsymbol{\theta}_{k(j)}$ is the parameter vector in the k^{th} window, p is the number of statistically different phases and M is the number of years on which the sample average is computed to approximate the ensemble average in Eq. (1). When considering window sizes longer than one year (*k* years in general, 5 and *S* years in the present analyses), the parameters of the ordinary distributions are the same for the *k* years used as calibration sample.

Unlike the application of the MEVD to the analysis of daily rainfall (*Zorzetto et al.*, 2016), the use of a yearly estimation window is not considered here due to the potential inaccuracies in the estimation of the parameters when few values are available. This is because the autocorrelation in discharge is much larger than in precipitation, leading to the need to use an inhibition window to identify independent flood events as previously described. This problem is further exacerbated when we stratify the data into different components of the mixture of distributions: the use of relatively long data windows (either 5-year windows or the whole calibration sample) for parameter fitting is thus necessary to make sure that, in periods in which multiple ENSO phases are present, there is a sufficient number of flood events in each phase to ensure a robust parameter estimation. After evaluating the number of peaks/year in stream gauges across the CONUS (see Supplementary Figure 2) we selected 5 years as the minimum window size, because a smaller one would have led to very few peaks/window in the driest areas, especially when ENSO phases are considered.

2.2 Fitting Procedure and Cross-Validation

2.2.1 Fitting Procedure

GEV fitting is performed on annual maxima using L-Moments (*Hosking*, 1990). *Zorzetto et al.* (2016) found that the cross-validation performance of the Peak-Over-Threshold GEV fits is indistinguishable from the performance of Maximum Likelihood or L-Moment GEV fits on annual maxima. Hence, we only present here results from the application of the latter. The parameters of the yearly Gamma distributions in the MEVD (Eq.(2)) are estimated on independent peaks from either 5-year windows or the whole sample via L-Moments (*Hosking*, 1990). LP3 fitting is performed on annual maxima using the method of moments

(*Griffis and Stedinger*, 2007). Low Floods have not been removed, because they still convey information about the frequency of events.

2.2.2 Evaluation Metrics

To identify the possible signature of ENSO phases in the distributions of ordinary flood peaks, we assign each event to one of the three ENSO phases based on the Extended Multivariate ENSO Index (https://www.esrl.noaa.gov/psd/enso/past_events.html). The phases are defined with a monthly time span by means of an index: -1 for El Niño, 1 for La Niña and 0 for the neutral phase. We then test whether the distributions of ordinary flood peaks for each phase are different from one another using the Kolmogorov-Smirnov test with the Bonferroni correction (*Bonferroni*, 1936) to account for multiple hypotheses testing (the three possible combinations among the phases, in this case). If the distributions of the peak magnitudes belonging to two separate ENSO phases are not statistically different at the 5% level, we combine all discharge peak values from both phases. Hence, we classify each time series in the dataset depending on whether: 1) three separate ENSO phases are distinguishable in the empirical distribution of ordinary flood peaks; 2) two separate ENSO phases are distinguishable (i.e. peaks from two of the phases were merged); 3) no ENSO phases are statistically different from each other in the set of ordinary events.

We estimate the values of the empirical cumulative frequency associated with observed peak discharge in the test sub-sample using the Weibull plotting position ($F_k = k/(L+1)$, where k denotes the k-th peak discharge value, Q_k, in an ascending order ranking). The estimated quantiles corresponding to each value F_k are computed using the three EV distributions by solving $DIST(Q^{DIST}_k) = F_k$, where *DIST* indicates the MEV, GEV and LP3 distributions.

We use two metrics to evaluate goodness-of-fit and estimation accuracy:

1. we compare estimated quantiles with observed ones through the computation of the Skill Score (SS $\in (-\infty; 1]$) (*Murphy and Winkler*, 1992; *Hashino et al.*, 2006), which provides a global metric of estimation accuracy:

$$SS(Q_{est}, Q_{obs}) = \rho^2_{Q_{est},Q_{obs}} - \left[\rho_{Q_{est},Q_{obs}} - \left(\sigma_{Q_{est}}/\sigma_{Q_{obs}}\right)\right]^2 - \left[\left(\mu_{Q_{est}} - \mu_{Q_{obs}}\right)/\sigma_{Q_{obs}}\right]^2 \qquad (5)$$

where $\rho_{Q_{est},Q_{obs}}$ is the correlation between the estimated values (Q_{est}) and the observations (Q_{obs}); $\sigma_{Q_{est}}$ and $\sigma_{Q_{obs}}$ ($\mu_{Q_{est}}$ and $\mu_{Q_{obs}}$) represent the standard deviation (mean) of the observations and estimations, respectively. The SS accounts for the potential skill (i.e., coefficient of determination) as well as conditional and unconditional biases. The SS is used both in the context of ordinary values fitting and of extreme values estimation evaluation. In the latter case, to provide a measure of the estimation of high quantiles, we compute the terms in the skill score definition (Eq. (5)) only on quantiles with return period $T_k=(1-F_k)^{-1} > S$, i.e. greater than the length of the dataset used for calibration. This reflects application needs, which target the estimation of extremes with return period much greater than the length of the observational time series available (estimation of quantiles with $T_k \leq S$ can be performed empirically, without the need to assume a specific probability distribution);

2. we compute the non-dimensional estimation error:

$$\varepsilon_j(S,T) = \left[Q_{est,j}(S,T) - Q_{obs,j}(S,T)\right]/Q_{obs,j}(S,T) \qquad (6)$$

for which we estimate a whole frequency distribution based on the $N_r = 1000$ Monte Carlo realizations. The Monte Carlo experiment allows us to eliminate any non-stationarity in the observational records, while, at the same time, it preserves the distribution of the flood peak values and their occurrence.

Over these realizations, we finally compute the Fractional Standard Error:

$$FSE(S,T) = \left[\frac{1}{N_r}\sum_{j=1}^{N_r} \varepsilon_j(S,T)^2\right]^{1/2} \qquad (7)$$

2.2.3 Cross-Validation

We quantify the uncertainty in estimating high quantiles associated with the use of the MEVD (in its single- or multi-phase versions) and of the GEV and LP3 distributions by means of a cross-validation procedure involving Monte Carlo simulations (with N_r = 1000 realizations for each station) as follows:

1. we randomly reshuffle the observational years on record keeping all the observations in their original year to preserve their yearly frequency distributions and the distribution of the number of events/year, to generate a realization without any systematic variability;

2. we divide the observational sample into two sub-samples obtained by randomly selecting S years from the original time series of length L_{tot}: this sub-sample is used for parameter estimation, while data in the remaining $L = L_{tot} - S$ years are used for testing;

3. in every realization, we compute both the SS and the FSE between the estimated and observed quantiles as described in the Section about the estimation metrics;

4. the whole procedure above is performed for different calibration sample sizes (S=10, 20, and 30 years), to evaluate how estimation uncertainty varies jointly with return period and calibration sample size.

3. Results

3.1 Ordinary Values

We start by selecting the most appropriate parametric distribution for ordinary peak discharge values based on the SS computed for yearly maxima.

The Gamma distribution provides the highest SS for most of the observational records analyzed (71% of the sites, Supplementary Figure 3). This is consistent with other studies in the literature (e.g., *Hann*, 1977; *Palynchk and Guo*, 2008; *Villarini and Strong*, 2014; *Slater and Villarini*, 2017). Hence, analyses were performed using the Gamma distribution at all CONUS sites, including the minority of sites where it was not the most skillful. This choice was made to obtain a method that would be simpler to apply and more homogeneous comparisons across the CONUS.

Another necessary step for the application of the MEVD towards the evaluation of the potential improvements associated with the use of a mixed ENSO-based MEVD, is to identify whether the distribution of ordinary flow peak values associated with different ENSO phases are different.

For most of the stations (3718 or 70% of the total), we did not detect statistically different distributions among the peaks occurred under different ENSO phases. At 883 sites (about 17%) the three phases are all different from one another. For the remaining 13% of the stations, we find common distributions to be shared by El Niño and the neutral phases or by La Niña and the neutral phase (Figure 1a). In most of the analyzed cases, different ENSO phases are detectable in ordinary peak discharge values in areas located in the eastern and southern United States, which are known to be more strongly affected by ENSO (e.g., *Emerton et al.*, 2017; *Mallakpour and Villarini*, 2017). We found detectable phases also in a group of stations in the U.S. Midwest, which are not usually areas that are affected by the effects of ENSO, but other processes might play a role.

3.2 Extreme Values Analysis

We now turn to the question of evaluating the predictive performance of the MEVD-Gamma formulations and compare the predictive performances of the two MEVD approaches, for the first time applied to flood frequency analysis, with those from the traditionally used GEV distribution and the Log-Pearson Type III distribution. The "optimal" MEVD-Gamma formulation compared below to traditional flood-frequency analysis approaches is, for each station, obtained by selecting the window size and the approach (single-component or multi-phase) that maximizes the Skill Score value. We will then quantify the potential benefits of including ENSO phases in extreme flood estimation.

When comparing the "optimal MEVD" (i.e. the MEVD formulation based on the number of phases that yielded the maximum SS value computed on yearly maxima) we find that MEVD outperforms the GEV applied to yearly maxima in ~78% of the stations based on the SS metric; Figure 2a shows the results from a S=10 years calibration period in terms of the relative difference between the SS from the optimal MEVD and the SS from the GEV distribution, divided by the absolute value of the SS for the GEV distribution here assumed as a reference. Regarding the comparison with the Log-Pearson Type III, Figure 2b shows that the MEVD outperforms the LP3 at a large number of stations (~86% of the stations), the same results being confirmed when looking at the FSE computed on the quantile corresponding to the maximum return period (i.e. the length of the test subsample + 1 year, Supplementary Figure 4). A calibration sample size of 10 years does not always allow a reliable the calculation of the third moment (skewness) when estimating LP3 parameters, hence leading to a low parameter estimation accuracy.

Based on these multiple comparisons, and on the lower performance of LP3, in the following we will focus on further comparative analyses between the predictive performances of the MEVD and GEV distribution.

We identify some areas where the performance of the GEV distribution is generally higher. They are characterized by a small number of independent events, and thus of uncorrelated flow peaks/year (e.g. the Rocky Mountains and southern California, where the number of independent peaks/year is frequently less than 10; see Supplementary Figure 2), usually in combination with short historical time series. The low number of flood events in these areas leads to a limited number of "ordinary peaks" that can be included in the MEVD in addition to yearly maxima. Moreover, previous findings showed that the GEV distribution does outperform the MEVD when the return period of interest is comparable with the calibration sample size (*Zorzetto et al.*, 2016).

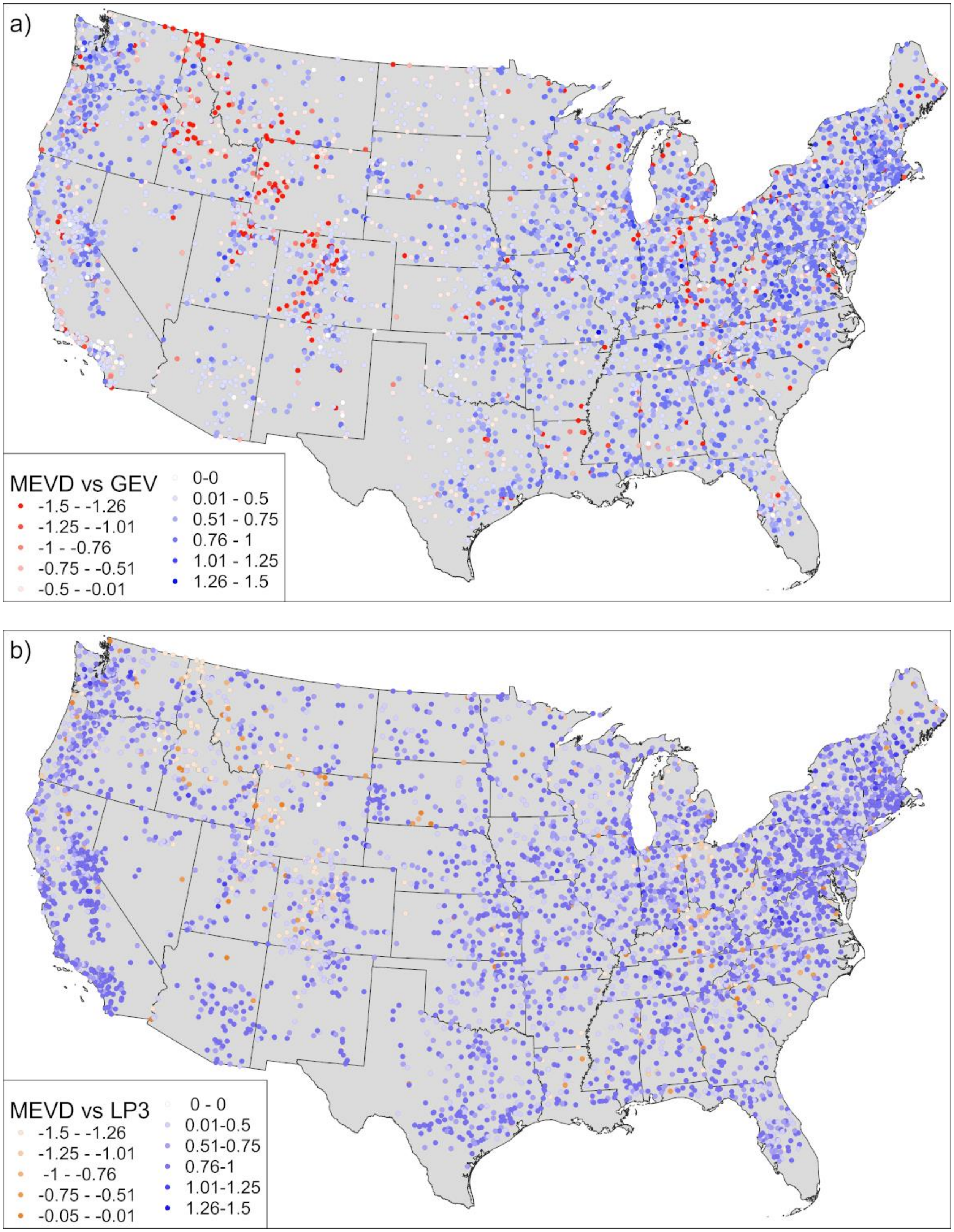

Figure 2 - Comparison between the performances of the optimal MEVD (i.e. the choice of window length and single-phase or multi-phase formulation yielding the highest SS) and that of the GEV (LP3) distribution expressed as $(SS_{MEVD} - SS_{GEV})/ |SS_{GEV}|$

($(SS_{MEVD} – SS_{LP3})/ |SS_{LP3}|$) values averaged over 1000 Monte Carlo realizations in panel a (b). Colors indicate which distribution provides the highest values of the SS for each station: shades of blue when the MEVD outperforms GEV (LP3), shades of red (orange) when the GEV (LP3) outperforms MEVD). White circles indicate stations for which SS values are equal to the first decimal digit.

In addition to characterize one method's predictive performance globally, we are also interested in focusing on the prediction accuracy for high return periods, being this the case for most practical applications. The return period associated with the maximum value in each test sub-series is estimated as $T_{max} = L_{tot} - S + 1$, where L_{tot} represents the length of the historical series: it is variable among the analyzed stations and ranges between 22 and 92 years. With reference to the range of the historical records available here (31-101 years of observations), the highest quantile for which the estimation error can be quantified using a calibration sample size of 10 years corresponds to a return period *T*=31(101)-10=22(92) years. When we look at $FSE(S=10\ years, T_{max})$ computed for the highest return period from the MEVD and GEV approaches, the MEVD outperforms the GEV distribution in about 76% of the analyzed stations (Figure 3). The information provided by Figure 3 is complemented for all return periods in Figure 4, where the FSE is plotted as a function of the ratio between return period and calibration sample size (*S*=10 years). For small values of the ratio between the return period (*T*) and the calibration sample size (*S*), the errors in the estimations computed with the two EV approaches are comparable both in terms of average value and uncertainty. When higher values of *T*/*S* are considered, the estimates provided by the traditional GEV distribution are less accurate than those provided by the MEVD approach, which shows a 30% improvement with respect to GEV estimates. Many engineering applications are almost exclusively focused on high return periods (i.e. return periods much larger than the span of observational time series): the uncertainty of the GEV-based estimates shows a steadily increasing trend of the FSE with increasing return period, while the MEVD estimation error stabilizes around a value of about 0.32 for high values of *T*/*S*. This result suggests that, when estimating quantiles corresponding to return periods much larger than values that have been observed, the estimation errors of the traditional EVT approach will become very large, and much larger than for the MEVD-based estimates. The results of the FSE computed with calibration sample sizes of 20 and 30 years are consistent with those from *S*=10 years (yet limited to smaller ratios of *T* over *S*) for both the GEV and LP3 distributions (second and third row in Figure 4).

The average number of peaks/year exhibits a large variation across space (from 3 to 31) suggesting that different analysis approaches may be differently effective in dry areas, where extremely few flood events are observed, and more humid areas, where the larger number of events/year makes available larger quantities of data. Also in consideration of the spatial pattern identified in Figure 2, it is thus interesting to analyze the possible dependence of the estimation performance associated with different EV approaches with respect to the number of floods/year. Figure 5 shows the FSE plotted as a function of T/S for two groups of stations representing two end-member cases: sites with less than 10 events/year and sites with more than 17 events/year (limits are defined in such a way that both groups include about 1000 stations). The advantage in the use of the MEVD approach instead of the traditional one is limited to higher values of T/S when few peaks are selected, while it always outperforms the GEV distribution when a greater number of peaks is available. This is linked to the fact that, for the same T/S, having a small number of peaks is not adding much information to the distribution of maxima, like it does when the number of peaks increases. However, the robustness of the MEVD with respect to the GEV distribution is confirmed for high ratios of return period over sample size.

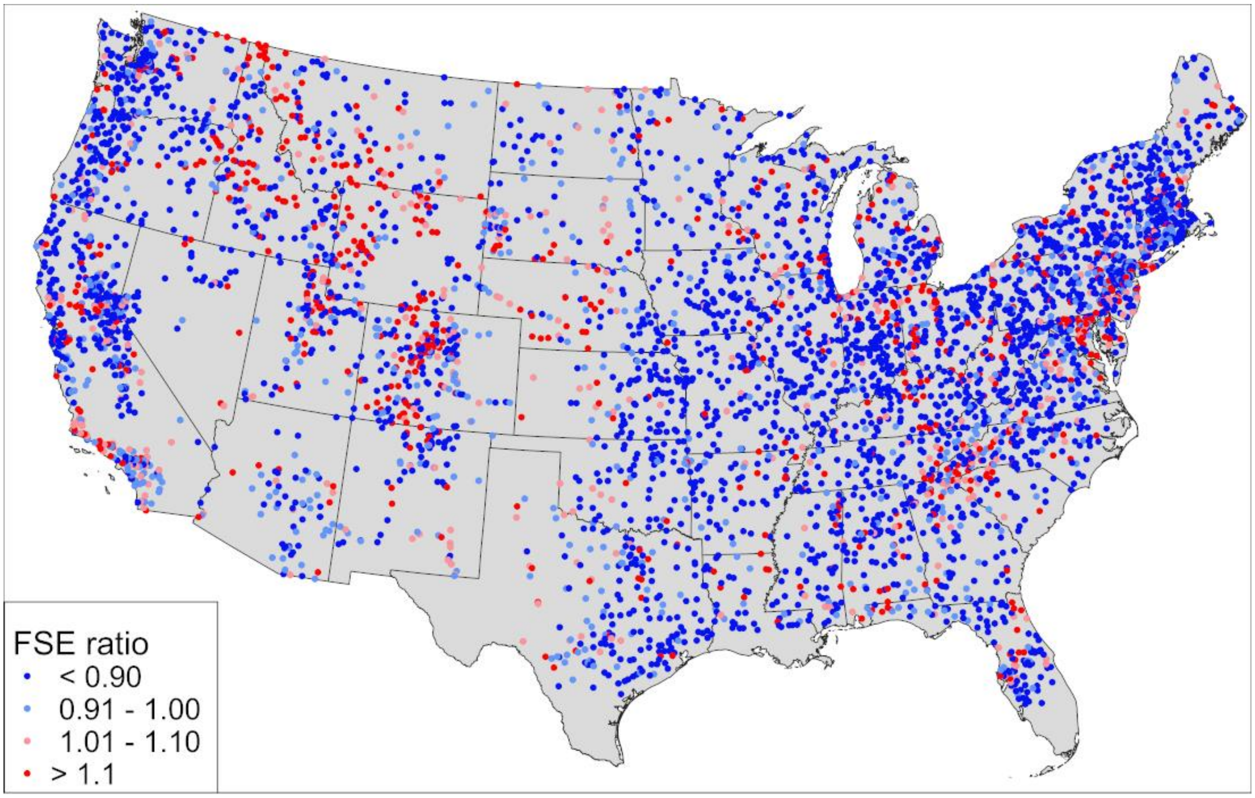

Figure 3 - Ratio between the Fractional Standard Error from the optimal MEVD (defined on the basis of the SS between the single-phase and multi-phase approach) and the FSE from the GEV distribution, computed for the highest return period at each station. Blue dots represent sites where the MEVD outperforms the GEV distribution (the ratio between the two FSEs is lower than one), while red dots indicate those stations in which the GEV distribution is providing a more accurate estimation. Shaded colors indicate small differences between the two approaches.

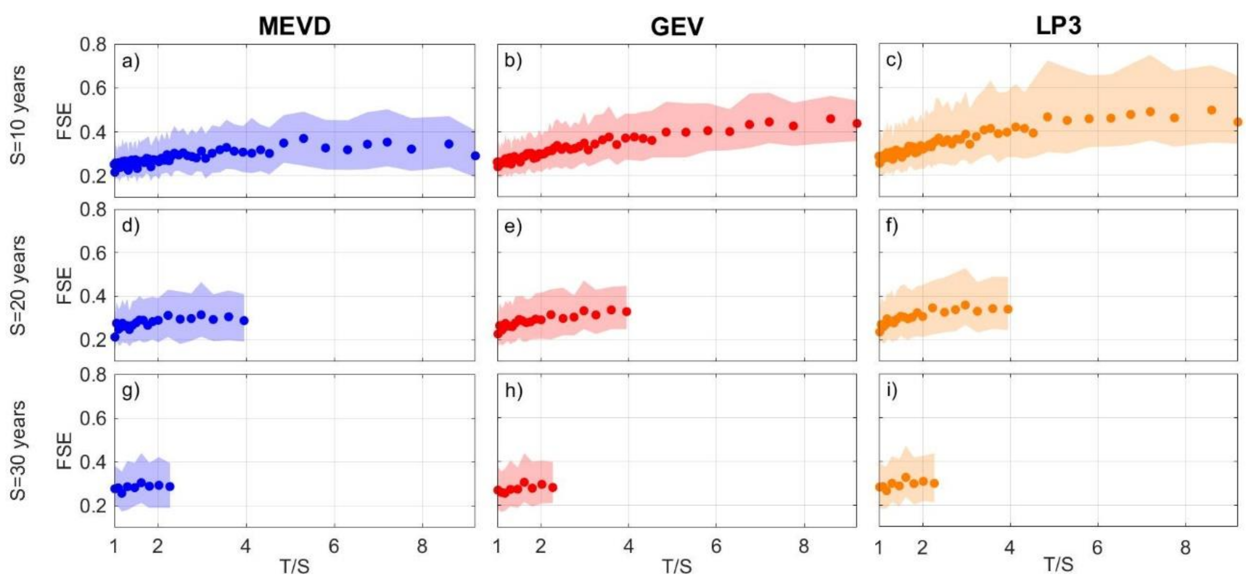

Figure 4 - Fractional Standard Error (FSE) for the MEVD (blue, a), d) and g)),the GEV distribution (red, b), e) and h)) and the Log-Pearson Type III (orange, c), f) and i)) averaged across all the stations for return periods greater than the calibration sample size, plotted as a function of the ratio between the return period (T) and the length of the calibration sample size (S=10, 20 and 30 years in the first, second and third row respectively). Dots represent the mean values computed across all stations and over T/S bins that include at least 300 values. Shaded areas indicate confidence intervals defined by the 25th and 75th percentiles.

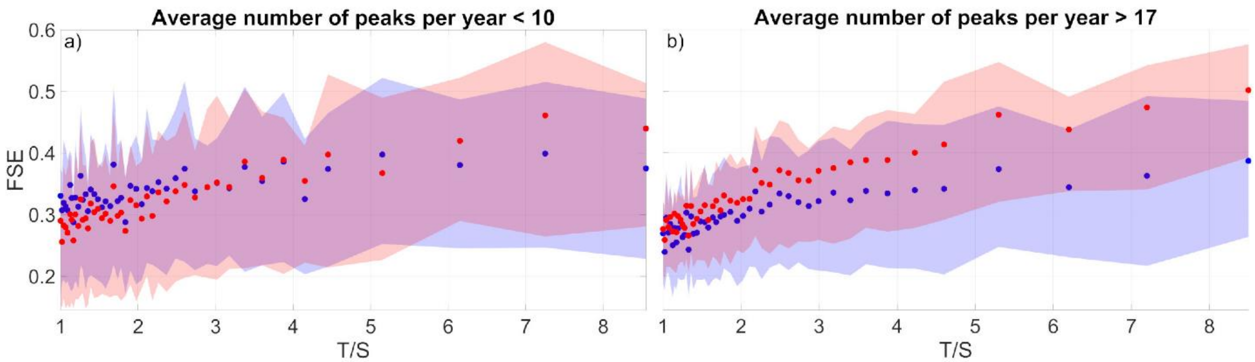

Figure 5 - Fractional Standard Error as a function of T/S for the MEVD (blue) and GEV distribution (red). Dots represent the mean of the FSE in each bin (the width of the bins is chosen such that they contain at least 100 values). Shaded areas are limited by the 25th and 75th percentiles. Panel a) refers to sites where, on average, less than 10 events/year occur. Panel b) shows results for sites where the average number of yearly floods is greater than 17.

We now focus on answering the second question, i.e. whether it is beneficial to adopt a mixed-distribution MEVD approach accounting for the different ENSO phases when estimating extreme flood magnitudes. In Figure 6 we show the performance of the optimal MEVD when the single-component approach and the mixed one are compared, on the SS basis (the performance of the two MEVD approaches is presented as the relative difference between the SS from the single-component MEVD and the one from the mixed approach, divided by the absolute value of the SS for the mixed MEVD here assumed as a reference). We find that including mixtures of distributions does not significantly improve the estimation: the number of sites where extremes are best described by a mixed MEVD is approximately equal to the number of sites where a single-population MEVD performs the best. Furthermore, in most cases the difference in skill scores between single-population and mixed MEVD estimates is negligible. This is consistent with the results obtained for rainfall by Marra et al. (2019), who found that introducing two populations to represent different rainfall-generating mechanisms in a relatively arid Mediterranean area did not yield improvements in the estimation of extremes. Moreover, whenever the mixed MEVD is selected, the value of the SS is generally comparable to what obtained using a single MEVD. The signal that we detected in the ordinary distributions is confirmed by the use of a mixed distribution for the estimation of extremes only along the eastern and south-eastern United States, which are known to be more strongly affected by ENSO.

These conclusions are corroborated by the analyses of the FSE values obtained from multi-phase and single-phase MEVD approaches (Supplementary Figure 7), which show negligible improvement in the estimation accuracy when adopting a mixed-distribution MEVD based on multiple ENSO phases.

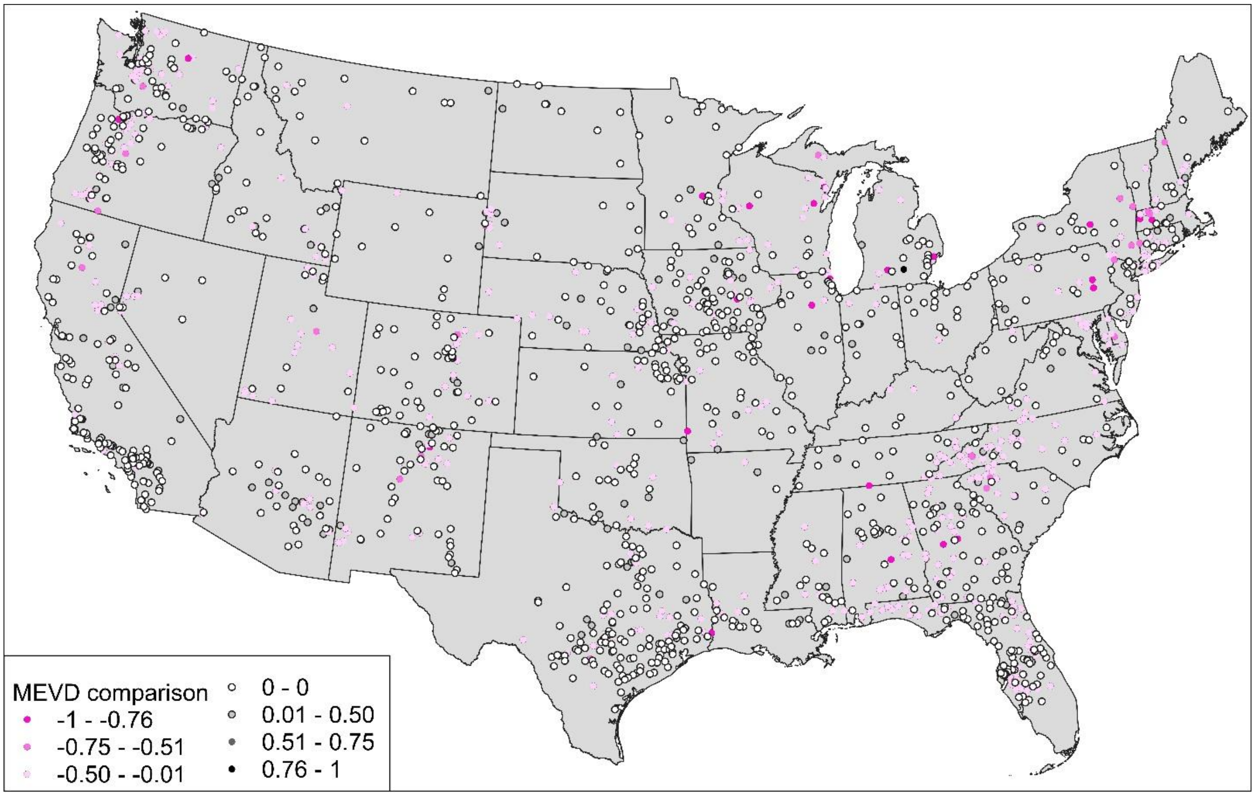

Figure 6 – Optimal MEVD formulation for stations in which different phases have been detected in the ordinary peak flow distribution. S=10 years is used for parameter estimation in the Monte-Carlo cross-validation evaluation of uncertainty. Magenta (black) circles represent stations for which the highest skill score is (not) obtained by including different ENSO phases. The relative performance of the two MEVD approaches has been evaluated through the ratio: $(SS_{non\text{-}mixed\text{-}MEVD} - SS_{mixed\text{-}MEVD}) / |SS_{mixed\text{-}MEVD}|$. *White circles represent the stations for which this ratio is equal to 0.*

With a more focused attention on the areas most influenced by ENSO (i.e. eastern and south-eastern US) and on those stations in which statistically different distributions were detected for the ordinary peaks, we look at the performance of the two MEVD approaches in absolute terms, to provide concrete indications of the estimation uncertainty at play. MEVD-estimates slightly under-estimate high-return period quantiles (see Figure 7) and, generally, the two approaches perform very similarly (consistently with the Skill Score values shown in Figure 6); however, in several cases the mixed-MEVD helps mitigating issues related to under-estimations of the quantiles compared to the observations (Figure 7).

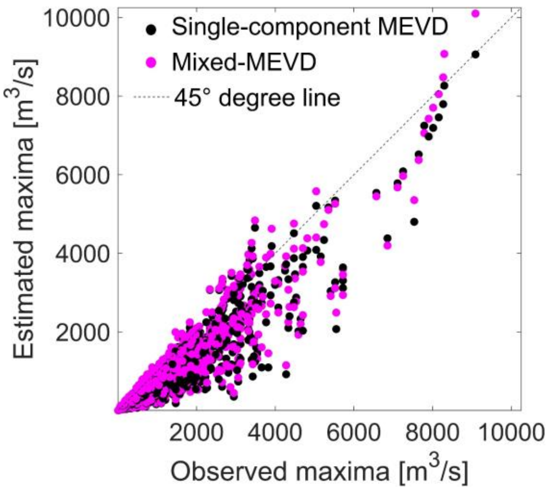

Figure 7 - Quantile-Quantile plot of the median values of the maxima estimated with the single- (mixed-) component MEVD in black (magenta) vs observed ones. These results are relative to the stations in the eastern and south-eastern United States where statistically significant differences in the ordinary-event distributions were detected.

4. Discussion and Conclusions

In this paper, we applied for the first time an adapted formulation of the Metastatistical Extreme Value Distribution (MEVD) to flood peaks observed at more than 5,000 USGS stream gages across the continental United States and proposed an alternative procedure to the standard flood frequency analysis. The major differences between the proposed and the standard flood frequency approach does not only lie in the distribution chosen to describe flood peaks, but also in the use of independent test samples to evaluate predictive performance within a Monte Carlo cross-validation approach. Goodness-of-fit evaluations, in fact, merely consist in the evaluation of the suitability of a distribution to describe a record of observations, without quantifying the skill of a method when presented with yet unoccurred extremes. A cross-validation approach instead allowed us to properly quantify the predictive uncertainty of MEVD-based estimates, both comparatively, with respect to traditional approaches, and in absolute terms.

Furthermore, we leveraged the ability of the MEVD of making use of all available independent observations to explore the potential benefits of a mixed-distribution approach, in which different ordinary-value distributions are used to describe different flood-generation mechanisms processes, with specific reference to ENSO phases.

Two fundamental steps were needed to develop a methodology tailored for flood frequency analysis. First, it was necessary to identify a robust method for the automatic and objective selection of independent flood peaks. Unlike previous MEVD applications, focused mainly on rainfall at the daily and sub-daily scale, the correlation in streamflow records is very significant, over large range of time scales. This circumstance required discarding potentially correlated flood peaks, and, as a consequence, the reduction of the number of events/year. The proposed approach performs analyses on data windows with length 5 years or higher to compensate for this effect, a particularly strong constraint when stratifying observed peaks based on different ENSO phases.

Second, a preliminary screening of candidate ordinary-value distributions was necessary. MEVD analyses of rainfall are based on the Weibull distribution (*Wilson and Toumi*, 2006; *Marani and Ignaccolo,* 2015), which was not found to be ideal in the case of flood peaks. Here, we compared the performance of three candidate distributions, eventually selecting the Gamma distribution as the best statistical model of flood peaks at the largest number of stations across the CONUS. This choice leads to a simpler MEVD-Gamma formulation, though applications seeking to optimize performance even further, may be based on an at-site selection of the optimal ordinary-value distribution.

The MEVD-Gamma approach outperforms the traditional GEV analysis of extreme flood peaks at about 76% of the stations, especially in the presence of records that are short with respect to the return period of interest. In fact, the MEVD displays the smallest Fractional Standard Error for small calibration sample sizes and high return periods, the case of greatest practical interest. We found similar results when comparing the MEVD with the widely-applied Log Pearson Type III. MEVD-based estimates of extreme peak flows outperform LP3 estimates at 86% of the stations across the CONUS. When the size of the calibration sample is increased (from S=10 years, to S=20, and to S=30 years) the three distributions considered here show comparable results for low values of T/S. The demonstrated superior predictive skill of the MEVD-Gamma approach to flood frequency analysis becomes especially relevant for engineering applications, where the estimation of exceptionally high design events is often required in the basis of short observational time series.

In a minority of cases the traditional GEV shows a reduced estimation uncertainty, in the presence of a low number of peaks/year and of short time series. In these stations, little additional information is available for the MEVD to exploit additionally to yearly maxima.

Considering ENSO as a factor potentially identifying different populations of flood peaks, we found that the estimation of high return period flows does not necessarily improve, even though the ENSO signature in the distributions of ordinary flood peaks was identified as statistically significant at several stream gages. We conclude that either the uncertainty intrinsic to extreme value estimation overwhelms the information contributed by ENSO phases or that just one of the ENSO phases detected in the distribution of the ordinary events effectively dominates the shape of the distributional tail and of the extreme values.

A final comment is in order. Even though the conclusion regarding the information that can be extracted from ENSO phases about extreme streamflows is negative, the ability of accounting for mixtures of distributions in the flood-peak MEVD formulation still has practical potential. In fact, several physical flood drivers can be identified (e.g., snowmelt, rain-on-snow, atmospheric rivers) whose role can be studied using the proposed approach. The introduction and formalization of a mixed-distribution MEVD for flood frequency analysis thus remains important, because it lends itself to applications to a variety of contexts in which different physical drivers can be defined, such as the North Atlantic Oscillation in western Europe (*Marani and Zanetti*, 2015) or the Arctic Oscillation in north eastern Europe (*Bartolini et al.*, 2009). Regarding ENSO, its detection in the distributions of the ordinary peaks is also still valuable: the detection or prediction of the occurrence of an ENSO phase justifies the use of MEVD parameters conditional to the known occurrence of that specific ENSO phase, with potential improvements over flood frequency analyses neglecting this information.

Acknowledgements

A.M. acknowledges the Foundation "Cassa di Risparmio di Padova e Rovigo" for funding her PhD under the project "Extreme Hydrologic Events" and the Foundation "Aldo Gini" for the grant in support of the period abroad at IIHR-Hydroscience & Engineering (University of Iowa). M.M. acknowledges support by the Venice Water Authority – Venice 2021 Project, Line 1.3 Modellazione numerica integrata del sistema bacino scolante-laguna-mare. G.V. acknowledges financial support by the USACE Institute for Water Resources and the National Science Foundation under CAREER Grant AGS-1349827.

References

Alila, Y., A. Mtiraoui, 2002. Implications of heterogeneous flood frequency distributions on traditional stream-discharge prediction techniques. Hydrol. Processes, 16 (5), 1065–1084. https://doi.org/10.1002/hyp.346

Ashley S. T, W.S. Ashley, 2008. Flood fatalities in the United States. Journal of Meteorology and Climatology, 47, 805-818. http://dx.doi.org/10.1175/2007JAMC1611.1

Bartolini, E., P. Claps, P. D'Odorico, 2009. Interannual variability of winter precipitation in the European Alps: Relations with the North Atlantic oscillation, Hydrol. Earth Syst. Sci., 13, 17–25. https://doi.org/10.5194/hess-13-17-2009

Barth, N. A., G. Villarini, M.A. Nayak, K. White, 2017. Mixed populations and annual flood frequency estimates in the western United States: The role of atmospheric rivers. Water Resour. Res., 53, 257-269. http://dx.doi.org/10.1002/2016WR019064

Barth, N., G. Villarini, K. White, 2019. Accounting for mixed populations in flood frequency analysis: Bulletin 17C perspective. Journal of Hydrologic Engineering. 24, 3. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001762

Benson, M.A., 1962. Evolution of methods for evaluating the occurrence of floods. U.S. Geol. Surv. Water Supply Pap., 1580-A. https://doi.org/10.3133/wsp1580A

Berghuijs, W.R., R.A. Woods, C.J. Hutton, M. Sivapalan, 2016. Dominant flood generating mechanisms across the United States. Geophys. Res. Lett., 43, 4382–4390. https://doi.org/10.1002/2016GL068070

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Emerton, R., H.L. Cloke, E.M. Stephens, E. Zsoter, S.J. Woolnough, F. Pappenberger, 2017. Complex picture for likelihood of ENSO-driven flood hazard. Nat. Commun., 8, 14796. doi:10.1038/ncomms14796

England, J.F., Jr., T.A. Cohn, B.A. Faber, J.R. Stedinger, W.O. Thomas, Jr., A.G. Veilleux, J.E. Kiang, R.R. Mason, Jr., 2018. "Guidelines for determining flood flow frequency – Bulletin 17C." Chap. B5 in USGS Techniques and Methods, Book 4. Reston, VA: USGS. https://doi.org/10.3133/tm4B5

Griffis, V.W. and J.R. Stedinger (2007). Log-Pearson type 3 distribution and its application in flood frequency analysis. II: Parameter estimation methods. Journal of Hydrologic Engineering, 12(5), 492-500. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(492)

Hann, C., 1977. Statistical Methods in Hydrology. The Iowa State University Press.

Hashino, T., A.A. Bradley, S.S. Schwartz, 2006. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrol and Earth System Sciences Discussions 3(2): 561–594. https://hal.archives-ouvertes.fr/hal-00305061

Hirschboeck, K.K., 1987. Hydroclimatically-defined mixed distributions in partial duration flood series, in: Singh, V.P. (Ed). In: Hydrologic Frequency Modeling: Proceedings of the International Symposium on Flood Frequency and Risk Analyses, 14-17 May 1986, Louisiana State University, Baton Rouge, U.S.A., Springer Netherlands, Dordrecht, pp. 199–212. https://doi.org/10.1007/978-94-009-3953-0_13

Hosking, J.R.M., 1990. L-Moments: Analysis and estimation of distributions using linear combinations of order statistics, Journal of the Royal Statistical Society. Series B (Methodological), 52, no.1, 105-124. https://www.jstor.org/stable/2345653

IACWD (Interagency Advisory Committee on Water Data), 1982. Guidelines for determining flood flow frequency: Hydrology subcommittee bulletin 17B. Reston, VA USGS

Kendall, M., G., 1975. Rank Correlation Methods. Charles Griffin, London

Katz, R. W., M.B. Parlange and P. Naveau (2002). Statistics of extremes in climatology and hydrology. Adv. Water Resour., 25 , 1287-1304. https://doi.org/10.1016/S0309-1708(02)00056-8

Klemeš, V., 1988. A hydrological perspective. Journal of Hydrology, 100, 3-28. https://doi.org/10.1016/0022-1694(88)90179-5

Klemeš, V., 1993. Probability of extreme hydrometeorological events; a different approach. In Z. W. Kundzewicz Extreme hydrological events: Precipitation, floods and droughts (pp. 167-176). Publication No. 213. Oxfordshire: IAHS.

Lang, M., T.B.M.J. Ouarda and B. Bobée, 1999. Towards operational guidelines for over-threshold modeling. Journal of Hydrology, 225, 103–117. https://doi.org/10.1016/S0022-1694(99)00167-5

Mallakpour, I. and G. Villarini, 2017. Analysis of changes in the magnitude, frequency, and seasonality of heavy precipitation over the contiguous USA. Theor Appl Climatol, 130, 345-363. doi:10.1007/s00704-016-1881-z

Mann, H., B., 1945. Non-parametric tests against trend. Econometrica, 13, 245-259. https://www.jstor.org/stable/1907187

Marani, M., M. Ignaccolo, 2015. A metastatistical approach to rainfall extremes. Adv in Water Res, 79, 121-126. https://doi.org/10.1016/j.advwatres.2015.03.001

Marani, M., S. Zanetti, 2015. Long-term oscillations in rainfall extremes in a 268 year daily time series. Water Resour. Res., 51, 639-647. https://doi.org/10.1002/2014WR015885

Marra, F., E.I. Nikolopoulos, E.N. Anagnostou, E. Morin, 2018. Metastatistical Extreme Value analysis of hourly rainfall from short records: Estimation of high quantiles and impact of measurement errors. Adv in Water Res, 117, 27-39. https://doi.org/10.1016/j.advwatres.2018.05.001

Marra, F., D. Zoccatelli, M. Armon, E. Morin, 2019. A simplified MEV formulation to model extremes emerging from multiple nonstationary underlying processes. Adv in Water Res, 127, 280-290. https://doi: 10.1016/j.advwatres.2019.04.002

Miller, S., R. Muir-Wood, A. Boissonnade, 2008. An exploration of trends in normalized weather-related catastrophe losses. Climate Extremes and Society, (Cambridge, UK: Cambridge University Press), pp. 225-247. https://doi.org/10.1017/CBO9780511535840.015

Murphy A.H., R.L. Winkler, 1992. Diagnostic verification of probability forecasts. Int J. of Forecasting 7(4): 435–455. https://doi.org/10.1016/0169-2070(92)90028-8

Palynchuk, B.,Y. Guo, 2008. Threshold analysis of rainstorm depth and duration statistics at Toronto, Canada. Journal of Hydrology, 348 (3-4), 535-545. https://doi:10.1016/j.jhydrol.2007.10.023

Perry, C.A., 2000. Significant floods in the United States during the 20th Century: USGS measures a century of floods. USGS Fact Sheet 024-00. Lawrence, KS: U.S. Geological Survey. https://doi.org/10.3133/fs02400

Slater, L.J., L.B. Singer, J.W. Kirchner, 2015. Hydrologic versus geomorphic drivers of trends in flood hazard. Geophys. Res. Lett. 42, 370-376. https://doi.org/10.1002/2014GL062482

Slater, L. J., G. Villarini, 2017. Evaluating the drivers of seasonal streamflow in the U.S. Midwest. Water, 9(9), 1-22. https://doi.org/10.3390/w9090695

Smith, J. A., Estimating the upper tail of flood frequency distributions (1987). Water Resour. Res., 23(8), 1657–1666. https://doi.org/10.1029/WR023i008p01657

Smith, J.A., G. Villarini, M.L. Baeck, 2011. Mixture distributions and the climatology of extreme rainfall and flooding in the Eastern US. Journal of Hydrometeorology, 12(2), 294-309. https://doi.org/10.1175/2010JHM1242.1

Stromberg, D., 2007. Natural disasters, economic development, and humanitarian aid. Journal of Economic Perspectives, 21(5): 199-222. https://www.jstor.org/stable/30033741

USWRC, 1976. Guidelines for determining flood flow frequency. United States Water Resources Council, Bull.17, Hydrol. Comm. Washington, DC

Villarini, G., 2016. On the seasonality of flooding across the continental United States. Adv in Water Res, 87, 80-91. http://dx.doi.org/10.1016/j.advwatres.2015.11.009

Villarini, G., L.J. Slater, 2017. Climatology of flooding in the United States. Oxford research encyclopedia of natural hazard science, Oxford University Press, Oxford, U.K.. https://doi.org/10.1093/acrefore/9780199389407.013.123

Villarini, G., A. Strong, 2014. Roles of climate and agricultural practices in discharge changes in an agricultural watershed in Iowa. Agriculture, Ecosystems and Environment, 188, 204-211. https://doi.org/10.1016/j.agee.2014.02.036

Wallemacq, P., R. House, 2018. Economic Losses, Poverty & Disasters 1998-2017. United Nations Office for Disaster Risk Reduction (UNDRR) and Centre for Research on the Epidemiology of Disasters (CRED). Retrieved from https://www.unisdr.org/we/inform/publications/61119

Wilson, P.S. and R. Toumi (2005). A fundamental probability distribution for heavy rainfall. Geophys. Res. Lett., 32, 1-4. https://doi.org/10.1029/2005GL022465

Zhang, M., X Chen, M Kumar, M Marani, M Goralczyk, (2017), Hurricanes and tropical storms: A necessary evil to ensure water supply?, Hydrological Processes 31 (24), 4414-4428. https://doi.org/10.1002/hyp.11371

Zorzetto, E., G. Botter, M. Marani, 2016. On the emergence of rainfall extremes from ordinary events. Geophys. Res. Lett., 43, 8076-8082. https://doi.org/10.1002/2016GL069445

Zorzetto, E., M. Marani, 2019. Downscaling of rainfall extremes from satellite observations. Water Resour. Res., 55, 156–174. https://doi.org/10.1029/2018WR022950