Received: 25 January 2019 | Revised: 5 September 2019 | Accepted: 31 October 2019

DOI: 10.1111/jfr3.12580

ORIGINAL ARTICLE

CIWEM Chartered Institution of Water and Environmental Management Journal of Flood Risk Management WILEY

Sensitivity of flood frequency analysis to data record, statistical model, and parameter estimation methods: An evaluation over the contiguous United States

Lanxin Hu[1] | Efthymios I. Nikolopoulos[1,2] | Francesco Marra[3] | Emmanouil N. Anagnostou[1]

[1]Department of Civil and Environmental Engineering, University of Connecticut, Storrs, Connecticut

[2]Department of Mechanical and Civil Engineering, Florida Institute of Technology, Melbourne, Florida

[3]Institute of Earth Sciences, Hebrew University of Jerusalem, Jerusalem, Israel

Correspondence
Efthymios I. Nikolopoulos, Department of Mechanical and Civil Engineering, Florida Institute of Technology, Melbourne, Florida
Email: enikolopoulos@fit.edu

Funding information
Eversource Energy Center, University of Connecticut; Israel Science Foundation, Grant/Award Number: 1069/18

ABSTRACT

The current statistical methods applied in flood frequency analysis require long data records to obtain reliable estimates, particularly for long return periods. Moreover, the choice of the statistical model and the parameter estimation procedure may introduce uncertainty in the estimates. In this work, we investigate the sensitivity of flood frequency analysis to various sample sizes, statistical models, and parameter estimation methods over six major hydrological regions in the contiguous United States. Results show that flood frequency estimates based on annual maximum series approach convergence to the reference values (estimates derived from 70 years record) in terms of median for 35-year or longer records. However, the uncertainty remains significant and a record of 35 years (20 years) is associated with ~50% (100%) larger uncertainty on the estimated 100-year flood. The generalised extreme value distribution combined with maximum likelihood estimation method is associated with the largest uncertainty, while the log-Pearson type III exhibits comparable bias and smaller uncertainty. Application of the partial duration series approach to 20-year records shows no significant advantage. Our findings suggest that the hydroclimatic characteristics of the catchments exhibit limited impact on the uncertainty.

KEYWORDS

AMS, distribution, flood frequency, parameter estimation, PDS, sample size, sensitivity

1 | INTRODUCTION

Reliable and robust estimation of magnitude and frequency of floods are fundamental for infrastructure design, risk-assessment, and decision-making. Traditional engineering practice ties together flood risk assessment and at-site flood frequency analysis (FFA), allowing one to estimate flood magnitudes at given gauged locations for return periods beyond the available data record (Nagy, Mohssen, & Hughey, 2017; Rahman et al., 2013).

FFA consists of the identification of a statistical distribution that is able to model the probability of exceedance of extreme flood peaks. Traditionally, both the statistical distribution class and the parameters describing it is derived from past data records. The standard approach is based on the annual maximum series (AMS), which involves the statistical modelling of the largest peak flows observed in each year, a series of data that can be easily and unambiguously extracted (Coles, Bawa, Trenner, & Dorazio, 2001). Oftentimes, the generalised extreme value (GEV) distribution is selected as a statistical model, following the results of extreme value theorem: the maxima values of large samples ($n \to \infty$) of independent and identically distributed random variables can only converge to this distribution (Fisher & Tippett, 1928; Gnedenko, 1943). However, some of these hypotheses are rarely verified, so that other distributions are considered (Sangal & Biswas, 1970). Particularly, log Pearson type III (LP3) was extensively used for projects sponsored (IACWD, 1982) and is widely applied in practice in the United States due to its accurate estimation on magnitude and frequency of flood flows (Stedinger & Griffis, 2008; Subramanya, 2013; Wallis & Wood, 1985).

However, in the case of limited data records (usually 10 to 20 years), only few data points are available for the estimation of the model parameters, causing important uncertainties in the parameters estimation (Keast & Ellison, 2013). To overcome this limitation and extend the available information, the partial duration series (PDS) approach, also known as peaks over threshold (PoT) can be adopted (Cunnane, 1973; Madsen, Rasmussen, & Rosbjerg, 1997). This approach makes use of the flood peaks exceeding a sufficiently high threshold, thus including more than one flood per year on average (Armstrong, Collins, & Snyder, 2012). Extreme value theory, again, but with some additional hypotheses, provides indications on the statistical model to be used for such approach, the Generalised Pareto distribution (GPD).

The choices of statistical distribution and of the parameter estimation method used to model extreme values affect the FFA estimates, and have been thoroughly explored in past studies (Ahilan, Amp, Sullivan, & Bruen, 2012; Baratti et al., 2012; Bobée, Cavadias, Ashkar, Bernier, & Rasmussen, 1993; Haddad & Rahman, 2010; Haktanir & Horlacher, 1993; Kidson & Richards, 2005; Meshgi & Khalili, 2008; Michele & Rosso, 2001; Villarini & Smith, 2010; Wilson et al., 2011). Log Pearson type III (LP3), GEV distribution, Extreme Value type I (EV1), GPD are frequently adopted. In general, there has been no theoretical consensus about a globally accepted probability distribution for FFA across various sites and the choice of the distribution has been mostly based on national guidelines (Wilson et al., 2011). In this work, we examine the performance of GEV, owing to the classic theoretical background, and LP3, is the most widely applied in contiguous United States (CONUS). Additionally, various estimation methods are used for the identification of the distribution parameters from a given data sample. The most common estimation methods in flood frequency analysis are based either on the method of moments (MOM) or on the maximum likelihood estimates (MLE). Other authors have proposed the use of Probability Weighted Moments (PWM) (Greenwood, Landwehr, Matalas, & Wallis, 1979), and L-moments (Gubareva, 2011; Haddad & Rahman, 2010; Vivekanandan, 2015). For example, in the case of the United States, the guidelines for determining flood frequency according to *Bulletin 17B* (IACWD, 1982) (hereafter 17B), of the advisory committee on water information, recommended weighted moments estimators for use with LP3 distribution.

Both the AMS and PDS approaches have been widely studied. Cunnane (1973) showed that, for the same range of return periods, PDS estimation has smaller sampling variance than the AMS if the PDS contains 1.65 N flood peaks, where N is the number of years of record length. Other studies also considered the parameter estimation methodologies, showing that on the basis of a regional average estimate, PDS provides the most efficient estimation for heavy-tailed distributions (higher probability of particularly high values) when MLE is used, and AMS when MOM is used, whereas in the case of MOM and PWM, PDS is superior for light-tailed (lower probability of particularly high values) (Madsen et al., 1997). In general, the idea that PDS outperforms AMS in the case of limited data records was confirmed, even if most studies used the full available sample sizes directly to conduct FFA (Bezak, Brilly, & Šraj, 2014; Nagy et al., 2017). Recent studies, however, challenged this idea showing no significant improvement provided by the use of PDS with short data records, particularly when automatic threshold selection methods need to be adopted (Marra, Nikolopoulos, Anagnostou, & Morin, 2018; Schlögl & Laaha, 2017).

The most common problem encountered in at-site FFA is the limited length of observations (Rahman et al., 2013). Given the statistical basis of flood frequency estimates, this will lead to inappropriate choice of the extreme value probability distributions and to inaccurate estimation of its parameters (Archer, Leesch, & Harwood, 2007; Claps & Laio, 2003; Cunnane, 1985). Nevertheless, in many cases in practice, FFA is carried out on relatively short data records, which typically do not exceed 50 years, rarely reach 100 years, and often are limited to 20–30 years (Bhat et al., 2018; Opere, Mkhandi, & Willems, 2006; Villarini, Smith, Serinaldi, Ntelekos, & Schwarz, 2012). Relatively short records may be applicable in low-flow design, but are far from the requirements for high-flow design (typically over 100 years) (Tallaksen, 2000). Such limitations result in larger uncertainties in the estimation of the longer return-period quantiles (Cunnane, 1988; Ribatet, Sauquet, Grésillon, & Ouarda, 2006; Tallaksen, 2000).

As first indicated by Benson (1963) and later applied by NERC (1975), reliable quantile estimates can be obtained only for return periods that do not exceed by a factor of two the flow record length. Douglas, Vogel, and Kroll (2000) indicated that, in typical conditions, a minimum of 30 years of data with no gaps is deemed satisfactory. It has been demonstrated that, as the range of interest is represented by the upper tail of the distribution (longer return period), theoretical models that satisfactorily fit the central area (i.e., return period <10 years) could give biased estimates if extrapolated to much longer return periods (Landwehr, Matalas, & Wallis, 1978; Landwehr, Matalas, & Wallis, 1980). For certain distributions, such as LP3, the skewness used is very sensitive to sample size (IACWD, 1982). Chowdhury Jahir and Stedinger Jery (1991), Reis and Stedinger (2005), and Griffis and Stedinger (2007) emphasised that the available range of record is important to obtain accurate weighted regional skew because skewness estimator can be unstable due to limited sample size (IACWD, 1982).

Despite the abundance of literature available on FFA, a systematic comparison analysis that evaluates the sensitivity of results to extreme value approach, statistical model, and parameter estimation method under short record conditions and diverse hydroclimatic settings is still lacking. The objective of this study is to present a comprehensive sensitivity analysis of statistical models, which are used in flood frequency analysis. The sensitivity of the models to parameter estimation methods and data record length is studied. The work aims to improve our understanding of the interplay between these different factors affecting FFA and investigate these aspects as a function of different climate characteristics. CONUS is chosen for this study because of the sufficient number of gauges across different hydrological regions and diverse hydroclimatic conditions.

The study consists of two parts. First, the impact of sample size and the combination of different statistical distributions and parameter estimation methods on FFA based on the AMS approach are examined. Then, the ability of the PDS approach to improve flood frequency estimation based on shorter record (20-year) records, in comparison with AMS, is tested. The article is organised as follows: in section 2 we present the necessary information of the study area, available data and their use in the context of this study. In section 3 we describe the methodology used for flood frequency analyses with different approaches and their corresponding statistical models and parameter estimation procedures, as well as the measurement of bias. Section 4 presents and discusses the results and section 5 summarises the main conclusions of this study.

2 | STUDY AREA AND DATA

The study herein is based on historical annual maximum flows and continuous streamflow observations over the Contiguous United States (CONUS) extracted from the U.S. Geological Survey (USGS) database. The basin scales vary from 1.52 to 25,791 km^2, with a median value of 178 km^2.

In addition to the quality check from USGS, the gauge data were further examined following a set of criteria provided by the GAGES-II data set (Falcone, 2011) and detailed below. For a station to be included, it should have at least a record of annual maximum peak flows longer than 70 years (for AMS). Of these stations with continuous streamflow record of at least 20 years where used for PDS. Second, the selected drainage, of selected basins, should be natural or minimally disturbed by anthropogenic modification, such as regulation, urbanisation, or land use change. Third, since the FFA approaches examined are based on statistical stationarity, no statistically significant trend should be observed in the data. The stationarity of streamflow data were examined based on the Philips-Perron test (Phillips & Perron, 1988) and Mann–Kendall test (Kendall, 1975; Salmi, 2002). The gauging stations that did not meet both these criteria at 5% significance level were excluded from the analysis. The final data set (Figure 1) that includes stations that passed the three quality check criteria consists of 299 gauges with at least 70-year record for the AMS approach. From those, 220 gauges included also a record of 20 years of continuous streamflow for the PDS approach.

The available data set was further divided into hydroclimatic regions according to the watershed characteristics and Köppen climate classification, respectively. To ensure, an adequate representation of the within-region variability and to obtain robust results for the examined regions, we included in our analysis only the regions with at least 15 stations distributed across each region, and excluded regions with only few available gauges. Specifically, six main regions were considered due to their relatively large number of gauges and because they were representative of different climate groups. The climate group and corresponding region names, number of gauges for AMS and PDS and climate class are presented in Table 1. Three regions (01, 02, and 03) are from the eastern, two (05,07) are in the Central, and one region (17) is from the western US.

3 | METHODOLOGY

The analysis framework conducted in this article is presented in Figure 2. It is divided into two parts. The first part consists of AMS modelling, which is composed of three elements including: (a) a range of sample size from available annual maximum flow data, (b) different distributions, and (c) different methods for estimating the

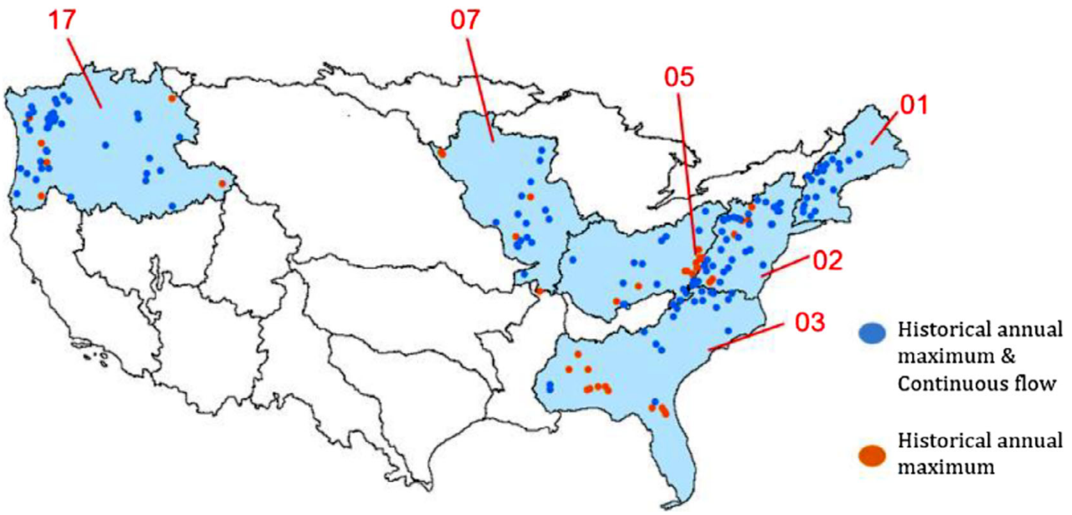

FIGURE 1 Selected gauges and hydrologic regions across CONUS. CONUS, contiguous United States

TABLE 1 Characteristic of selected hydrologic regions

Hydrological unit map code	Number of gauges (AMS)	Number of gauges (PDS)	Region	Climate characteristic
01	15	15	New England	Dfb
02	43	38	Mid-Atlantic	Dfa and Dfab
03	29	17	South Atlantic-Gulf	Cfb
05	25	16	Ohio	Dfa
07	19	15	Upper Mississippi	Dfa and Dfb
17	44	38	Pacific Northwest	Csb

Abbreviations: AMS, annual maximum series; PDS, partial duration series.

parameters of the distributions examined, all of which are built as several combinations for further sensitivity analyses. Our aim is to establish how these three elements affect flood frequency estimates at different exceedance probability levels.

The second part tests the advantage of PDS on limited sample size, and is structured as a comparison of flood frequency estimates obtained using AMS and PDS on the most recent 20 years of data (the period 1997–2016). The procedure in the AMS approach is similar to the first part, but for consistency, is limited to 20 years of data extracted from annual maximum flows. The PDS approach is applied at the 20 years for which continuous streamflow data are available.

Three distributions and three parameter estimation methods are used resulting in six combinations in total: LP3-MLE, LP3-17B, GEV-MLE, GEV-LMOM, GPD-MLE, and GPD-LMOM. Specifically, GEV and LP3 are applied in AMS. The following sections provide more details in statistical metrics on the AMS and PDS methods for the sensitivity analyses.

3.1 | AMS modelling

Two theoretical distribution functions, LP3 and GEV, are used in AMS modelling. Three parameter estimation methods MLE, L-moments (LMOM), and 17B are used. MLE maximises the probability likelihood of the sample data (Haddad & Rahman, 2010; Martins & Stedinger, 2000). LMOM is largely used in hydrological applications owing to the lower sensitivity to outliers and to the better estimation of the tail heaviness in presence of short data records (Hosking, 1990). The method characterises a wider range of distributions than conventional moments, and has relatively small sampling variance, especially in comparison with the classical coefficients of skewness and kurtosis (Ahilan et al., 2012; Bobée & Rasmussen, 1995; Chowdhury Jahir & Stedinger Jery, 1991; Pearson, 1991; Seckin, Haktanir, & Yurtal, 2011). In the case of LP3 distribution, weighted moments estimators based on the logarithms of sample data are used instead of the classic LMOM, as recommended by Bulletin 17B. In addition, 17B uses a generalised estimate of the skew coefficient of

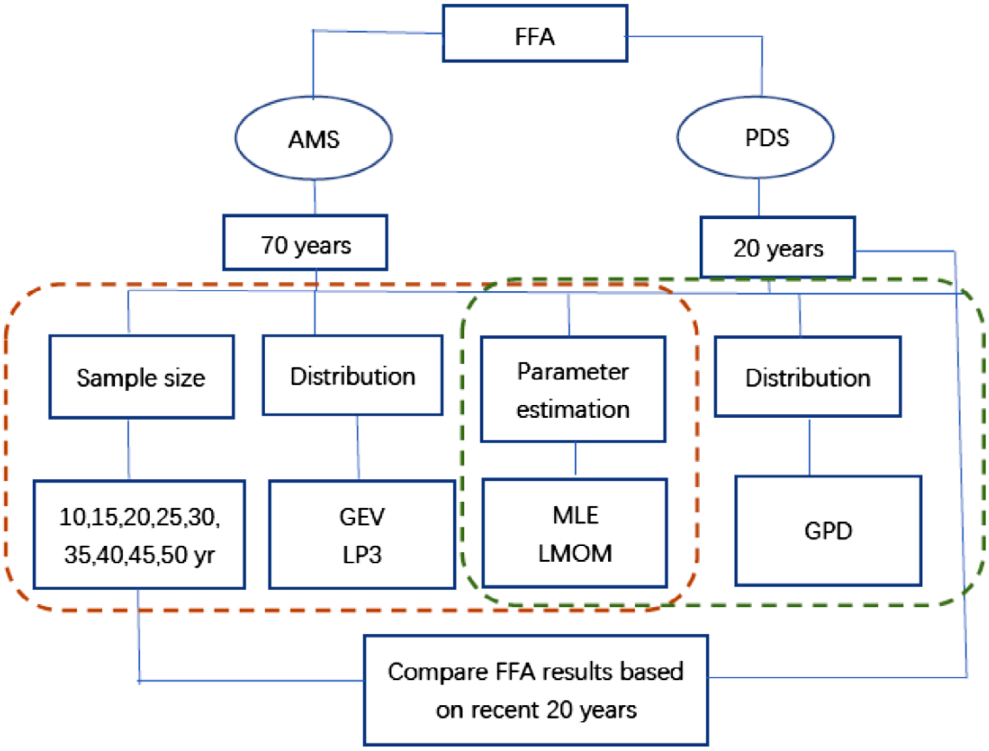

FIGURE 2 Methodology flowchart

the station record to reduce error and bias in the skewness estimation.

3.2 | PDS modelling

The PDS approach consists of modelling exceedances above a predefined threshold. PDS modelling is undertaken on continuous streamflow data.

Data used in the PDS approach should satisfy two requirements: independence of events and selection of an appropriate threshold, which will be further introduced in subsequent subsections. Under extreme value theory assumptions, exceedances are expected to follow a GPD distribution, which is described by three parameters (Bezak et al., 2014; Nagy et al., 2017; Pickands, 1975; Ribatet et al., 2006).

Parameters are estimated with MLE and LMOM technique, which are also used in AMS. LMOM allow for the estimation of the location parameter of the GPD distribution once the PDS is defined through the threshold (Hosking & Wallis, 1997). Therefore, we have GPD-MLE and GPD-LMOM for PDS modelling.

3.2.1 | Independence criteria

It is essential to ensure the independence of consecutive flood peak events. Studies have shown that independence criteria should be a function of catchments size and interval between separate flood events (Lang, Ouarda, & Bobée, 1999). In this study, we follow the recommendation of USWRC (1976) that the events should be separated by a time interval θ:

$$\theta > 5\,days + 2.59 * \log(A) \tag{1}$$

where A is the basin area in square kilometres.

3.2.2 | Threshold selection

Identification of the appropriate threshold is not a straightforward task and still remains an issue of uncertainty in PDS estimates (Langousis, Mamalakis, Puliga, & Deidda, 2016). An appropriate threshold should be high enough to represent the tail of the distribution and, at the same time, the retained data should provide an increased data sample with respect to the AMS (Caballero-Megido, Hillier, Wyncoll, Bosher, & Gouldby, 2018). High threshold level will reduce the number of events but increase the likelihood of independence and fulfil the theoretical requirements (Lang et al., 1999), while low threshold will more likely violate the theoretical requirements. Various approaches for threshold selection have been proposed (Bernardara, Mazas, Kergadallan, & Hamm, 2014; Bobée & Rasmussen, 1995; Lang et al., 1999; Tanaka, Takara, Snorrason, Finnsdottir, & Moss, 2002), but many of them have been quite subjective. In our study, we followed an objective method recently proposed by Solari, Egüen, Polo, and Losada (2017), which selects the threshold by minimising the complement of the *p*-value of the Anderson-Darling test.

3.3 | Sensitivity analysis and uncertainty quantification

Uncertainty in flood frequency estimates can be caused by model errors, which comprise inappropriate estimation of the population parameters owing to limited sampling, possible incorrect choice of the model, and possible non-optimal choice of the parameter estimation procedure. In this study, the impact of three factors is quantified: (a) short record length, (b) selection of statistical models, and (c) parameter estimation procedure. Particularly, the experiment is divided into two parts: first, the sensitivity of AMS models to the three above-mentioned factors is quantified; then, sample size effect is examined for AMS and PDS approaches focusing on 20-year records.

Uncertainty is quantified using the bootstrap approach suggested by Overeem, Buishand, and Holleman (2008), which comprises (a) the generation of synthetic records of specified length via random sampling with replacement of the full record of AMS or PDS and (b) the estimation of flood quantiles (e.g., for 5-, 10-, 50-, 100-year return period), from those synthetic records using different models and parameter estimation methods. The procedure is then repeated 10,000 times to create the bootstrap sample.

The impact of the three factors mentioned above is quantified using the relative difference (RD) defined as:

$$\mathrm{RD} = \frac{X_i - Xref}{Xref} \times 100\% \tag{2}$$

For a given flood quantile, estimated based on a given method (AMS or PDS) and parameter estimation technique, *Xi* is the flood estimate value estimated from synthetic records of length *i* and *X_ref* the reference quantile estimated from synthetic records of reference record length. The exact time period of data record in *X_ref* varied per station but all the selected stations covered the period 1947–2016. We considered two statistical indicators, the median and interquartile range (IQR). The RD of median and IQR allow us to quantify the impact of the factors examined on both bias and uncertainty in FFA estimates, respectively.

RD absolute values (that is, the magnitude of the RD is considered without considering its sign) are first examined as a measure of the degree of bias of a particular distribution/parameter estimation method/sample size; second, the sign of the RD is considered to indicate whether the reference is under- or over-estimated. More details are provided in the following sections.

3.3.1 | Sensitivity on AMS

The performance of different models (LP3-MLE, LP3-17B, GEV-MLE, and GEV-LMOM) in FFA estimation based on AMS for different sample sizes *S*(10-, 15-, 25-, 30-, 35-, 40-, 45-, 50-year) and return periods (5-, 10-, 50-, and 100-year return period), is quantified using the RD from bootstrap to quantify the variability between FFA result and reference.

In Equation (2), the reference *X_ref* corresponds to the 70 years AMS estimates. In sensitivity analyses, *X_ref* corresponds to the estimation based on each model with full record length (70 years) i.e., FFA estimates from each model at sample size *S* are compared against the FFA estimates of the same model at sample size equal to 70 years.

3.3.2 | AMS versus PDS

To evaluate AMS and PDS for short records, we compare their performance based on 20-year records. More specifically, flood frequency estimates on AMS with 20 years and continuous streamflow series on PDS are compared. The statistical models and parameter estimation methods in AMS are the same as in section 3.3.1. Since the true distribution of each sample size is not known, the reference for quantile estimates is based on the empirical distribution function from the full 70-year record, which is proposed as an optimal and fair reference to be more powerful than other tests of fit for a wide range of sample sizes (Ahmad, Sinclair, & Spurr, 1988). Therefore, *X_ref* in this section is the estimation based on the empirical distribution with a 70-year record. The classical Weibull formula is selected for the empirical estimates:

$$F_i = \frac{N - i + 1}{N + 1} \tag{3}$$

where F_i, *i, and N* are the empirical non-exceedance probability, the rank in descending order, and the number of extreme events (i.e., number of annual maxima), respectively.

4 | RESULTS AND DISCUSSION

4.1 | Sensitivity on AMS modelling

An example of FFA (USGS site number: 01013500) based on various distributions and parameter estimation methods

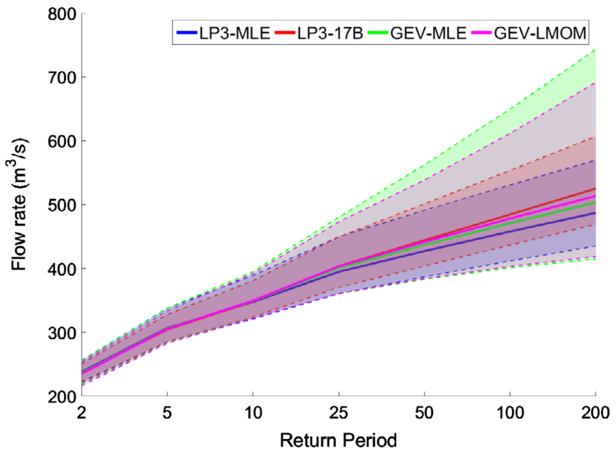

FIGURE 3 Example results of flood frequency estimates and 5th and 95th percentile (from bootstrapping, see section 3.3 for details) obtained for different distributions/parameter estimation methods. Results are shown for USGS gauge station 01013500. USGS, U.S. Geological Survey

is shown in Figure 3, where the resulting curves are reported together with the associated uncertainty bounds calculated as described in section 3.3. The confidence intervals are obtained from bootstrapping experiments. More specifically, we randomly sampled (with replacement) the original record (70 years) and each time, flood quantiles were estimated for each model/estimation method examined. The distribution of flood quantiles that resulted from 10,000 iterations was used to identify the 5th and 95th percentile that was used to represent uncertainty in our estimates. The four distributions based on full record length (70 years) begin to diverge at 25 years return period. Most noticeable differences occur in the upper bounds of the higher return period among the distributions. Compared to the FFA results for 2-year return period, the differences among the distributions increase nearly by a factor of 10, indicating an increase in uncertainty for higher return periods. LP3-based estimates have more than 60% less uncertainty than in GEV. LP3-MLE exhibits the lowest uncertainty, manifested as the narrowest shade in Figure 3. Conversely, GEV-MLE estimates are associated with the highest uncertainty because of the widest shade area. For parameter estimation methods, LMOM based estimates exhibit less uncertainty than MLE. The greatest difference between 17B and MLE based estimates is presented in the upper bounds, where the uncertainty of 17B based estimates is ~10% higher than the MLE for 200-year return period.

Similarly, a series of sensitivity analyses are carried out for each distribution and parameter estimation method (LP3-MLE, LP3-17B, GEV-MLE, GEV-LMOM) in the selected hydrological regions. Results on the RDs in median and IQR are summarised in Figures 4 and 5.

4.1.1 | Sample size

As record length increases and approaches the reference length, RDs exhibit a nonlinear rate of decrease, which highlights that the uncertainty is highly sensitive to sample size. RDs in the median reduce gradually and approach zero (i.e., no difference with reference) as sample size increases in the four models studied herein. This pattern is consistent in all regions examined and in different return periods, suggesting that the variability of sample size affects flood frequency estimates, and affects it dramatically at small sample sizes. Specifically, RDs in the median are about 10% for 20 years record length and 5% for 30 years record length. RDs in IQR show a noticeable decrease with increasing record length. They exhibit very high RDs ranging from 150 to 350% with 10 years record length, then dropping to less than 20% by 50 years record lengths. For the 10-year return period RDs are ~100% for 20 years record length and reduce to less than 50% for 35 years of record length. Results for the 100-year return period are similar but RDs are slightly higher.

4.1.2 | Distributions and parameter estimation methods

In terms of RDs based on medians compared to reference, LP3-17B in most regions exhibit overestimation, while the other three distributions, LP3-MLE, GEV-MLE, and GEV-LMOM always exhibit underestimation. In addition, LP3-17B presents no obvious decreasing trend along with increasing sample sizes in shorter return periods, showing its stability in various sample sizes for low quantiles. The results for RDs in the median for LP3-17B exhibits a regional dependence that is different from other methods. Specifically results for some regions are associated with overestimation, which is different from the rest of the methods examined. This essentially further highlights the regional dependence of FFA to estimation method. Potential attribution for this regional dependence can be related to the use of regional skewness coefficients that are involved in the 17B approach.

GEV-MLE and GEV-LMOM show a relatively higher decreasing trend in RDs for the sample size range from 10 to 30 years of record length, demonstrating that it is more sensitive to the smaller sample sizes. After 20 years, the impact of the record length is weak. More specifically, the RDs among four models are very small and after 30 years of record length the RDs are negligible and approach zero.

In terms of RDs based on IQR, the four models demonstrate similar results in the range of 160 to 200% for the 10-year return period. GEV-MLE exhibits high

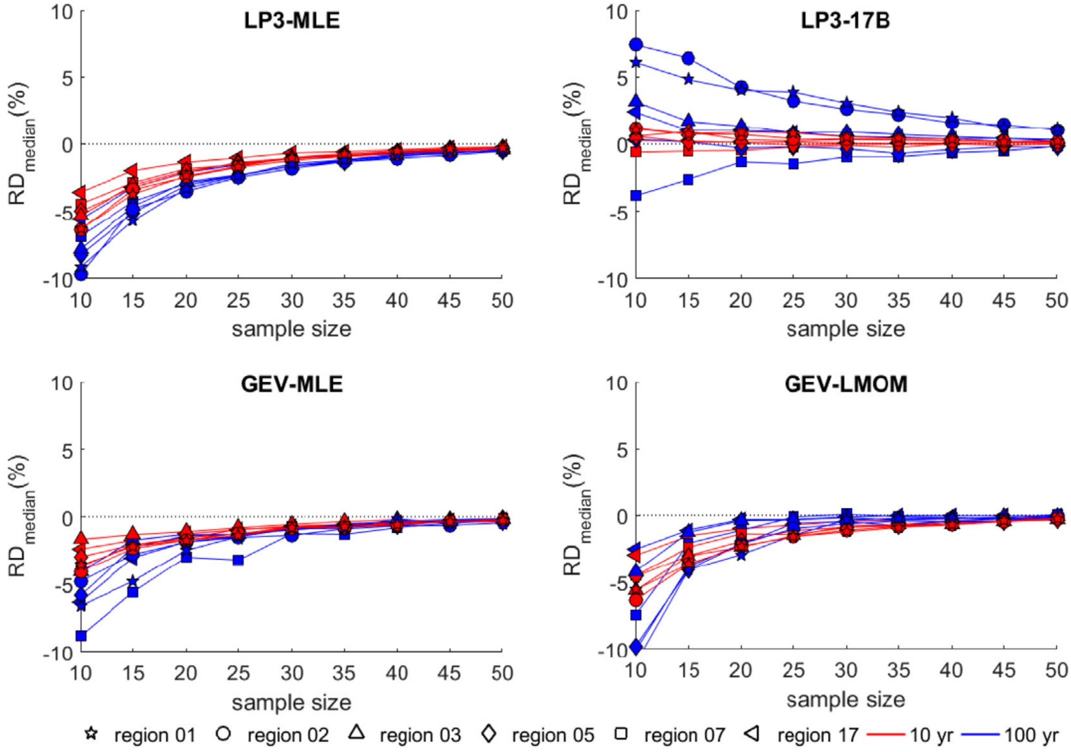

FIGURE 4 RDs_median of 10- and 100-year return period quantiles of annual maximum streamflow derived from nine sample sizes (10-, 15-, 20-, 25-, 30-, 35-, 40-, 45-, 50-year) using AMS among six regions (01, 02, 03, 05, 07, 17). AMS, annual maximum series

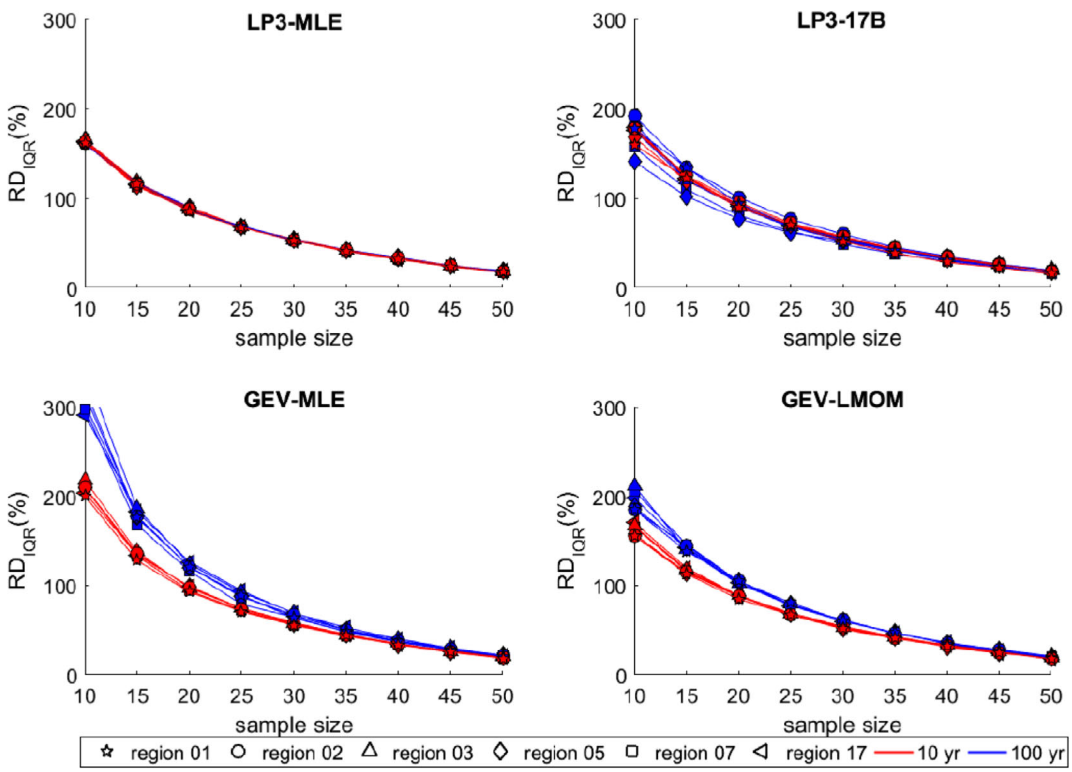

FIGURE 5 RDs_IQR of 10- and 100-year return period quantiles of annual maximum streamflow derived from nine sample sizes (10-,15-,20-,25-,30-,35-,40-,45-,50-year) using AMS among six regions(01, 02, 03, 05, 07, 17). AMS, annual maximum series

sensitivity to sample size and maintains a consistent behaviour for all regions. The value of RDs, however, are large for small sample sizes (300% for the 100-year return period and 200% for the 10-year return period). In general, RDs of GEV-MLE are approximately 30% higher than those in the other distributions when using smaller sample sizes. GEV-LMOM shows much less RDs than does GEV-MLE, which shows that GEV-LMOM has less uncertainty than GEV-MLE and thus highlights the impact of the parameter estimation method on the uncertainty of FFA estimates. In the 10-year and 100-year return period, GEV-MLE and GEV-LMOM both exhibit strong dependence of the results on the quantiles, while LP3-MLE yields stable predictions because there are no regional differences in the RDs within any of the return periods, showing it has no distinct dependence on the quantiles. The comparison between the 10-year and the 100-year return periods gave no significant divergence in short return periods for LP3-17B.

4.2 | Sensitivity on PDS

Analysis of flood frequency estimates based on the AMS approach (section 4.1) demonstrated that, even with 20 years of data, we are still dealing with significant uncertainty in FFA estimates that can be even 100% higher than the reference (i.e., estimates from 70 years of observations). An alternative method, PDS, allows us to partially overcome issues with short record length. In this section, we present a comparative analysis between the AMS and PDS approaches for 20 years of continuous streamflow observations. Evaluation of the flood frequency estimates from the two methods is performed against the empirical quantiles derived from the 70-year AMS record, which is considered the reference. Evaluation results for six models, four AMS models and two PDS models are presented in Figure 6 and Figure 7 for different return periods (5-, 10-, 25-, 50-year) and regions. Note that, since our reference record is 70 years long, we could not estimate empirical quantiles for the100-year return period and thus we present results for up to the 50-year return period.

In terms of RDs based on median, except for region 02 and 05, GPD-based estimates exhibit less variability for the 5-year return period. For the 10-year return period, only region 03 and 05 exhibit smaller variability with GPD-based estimates, while GPD-based estimates in the other regions did not show obvious advantage. In particular, GPD in region 02 presents more uncertainty.

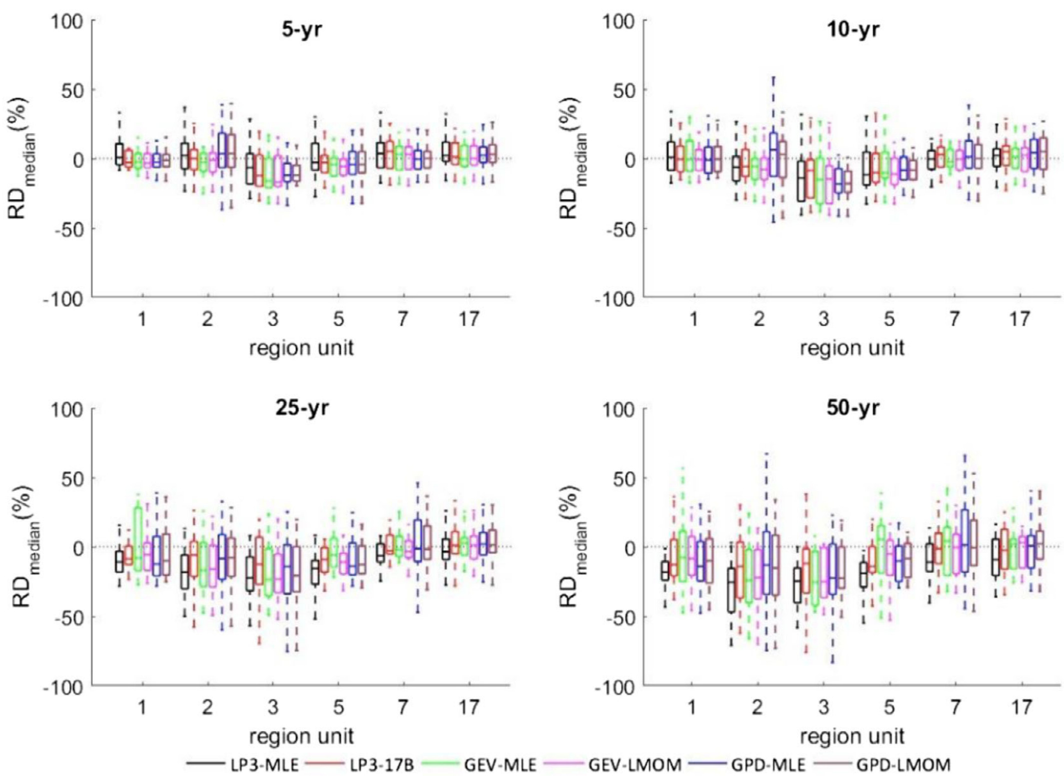

FIGURE 6 RDs_median of the 5-,10-,25- and 50-year return period quantiles derived from 20 years AMS and PDS in six regions (01, 02, 03, 05, 07, 17). AMS, annual maximum series; PDS, partial duration series

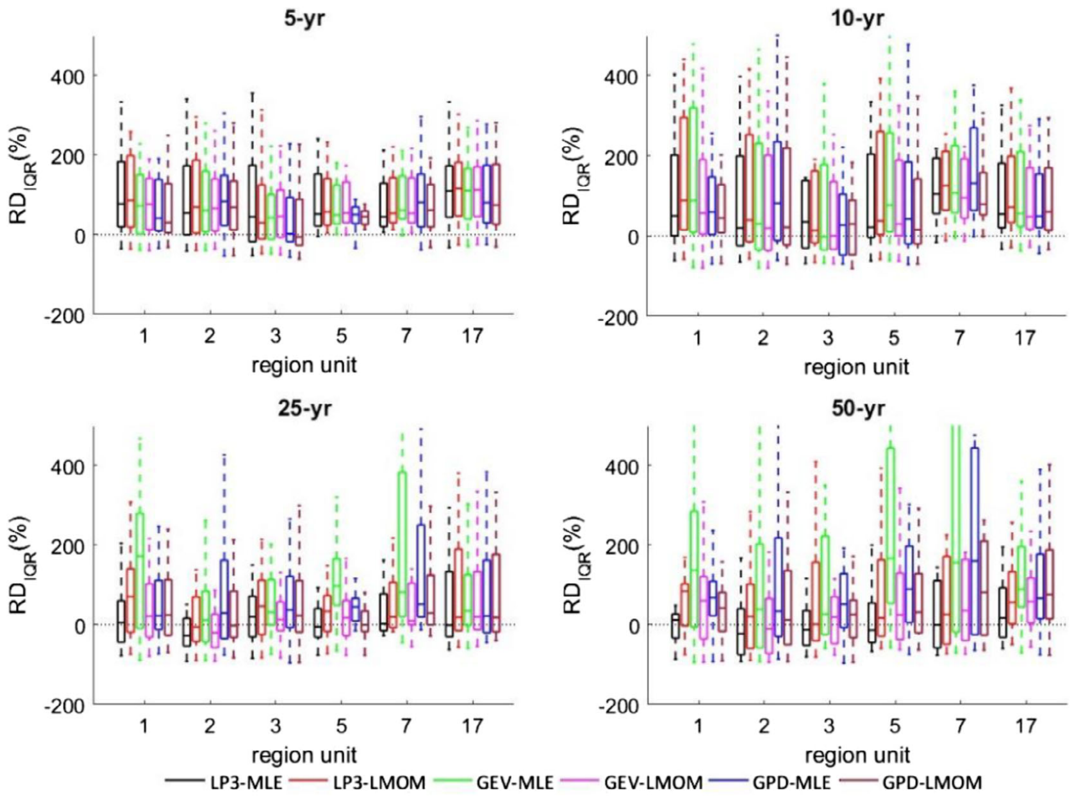

FIGURE 7 RDs_IQR of the 5-,10-,25- and 50-year return period quantiles derived from 20 years AMS and PDS in six regions (01, 02, 03, 05, 07, 17). AMS, annual maximum series; PDS, partial duration series

For the 25- and 50-year return periods, GPD-based estimates exhibit large variability except for of region17. Although GPD-MLE and GPD-LMOM exhibit good results with the least bias in region 03 for the 5- and 10-year return periods, this pattern is not consistent for all regions.

For the RDs based on IQR, GPD-based estimates present the lowest RDs and magnitude of uncertainty in region 05 while RDs among all distributions/parameter estimation methods in other regions are around 200% in the 5-year return period. It is notable that GPD-LMOM presents high uncertainty in region 07 for the 50-year return period. In other regions, GPD-MLE and GPD-LMOM did not show obvious advantages since they have similar uncertainties with distributions in AMS. For the 25-year return period, the results for region 01, 03, 05, and 17 are acceptable due to the RDs of IQR are close to those in AMS modelling but the results for region 02 and 07 are not.

Overall, flood frequency estimates based on PDS approach do not exhibit a consistent advantage over the equivalent AMS-based estimates for 20 years of record. This suggests that a priori acceptance of the common idea that PDS outperforms AMS in extracting extreme value information from short data records are not validated in this experiment. This was also been observed in other recent studies by Schlögl and Laaha (2017) and Marra et al. (2018).

5 | CONCLUSIONS

This study presents a comprehensive framework for sensitivity of FFA across CONUS, with respect to sample size, statistical model, and parameter estimation method. Results are presented for six major hydrologic regions and consistently show that all the factors examined correspond to significant sources uncertainty in flood frequency estimates. Record length has a major control in FFA and particularly affects the uncertainty of the estimates. As an example for the AMS approach, a record of around 35 years is associated with 50% higher uncertainty than the 70-year reference. FFA estimates and associated uncertainty are also dependent on the choice of distribution. GEV-MLE had the highest uncertainty, while LP3-MLE and LP3-17B exhibited the lowest among the distributions examined. This has important practical applications and suggests that when using short records to estimate high return periods (100 years or greater), careful consideration of the model/method used and quantification of the uncertainty associated with FFA estimates is imperative.

Comparison between AMS and PDS approach with 20 years of record length did not reveal a clear winner since the RDs in median and IQR did not exhibit consistent improvement across regions for any of the methods. Nevertheless, it should be mentioned that the findings in this study are not meant to defy other studies that may have shown superior performance of PDS relative to AMS (see for example Bezak et al., 2014). However, our results highlight clearly that in practice, identification of the most appropriate method is not a straightforward task and depends on a number of factors. As shown, the magnitude of uncertainties varies with parent distribution, model type, and sample size for a particular region. The results obtained from this study indicate the magnitude of uncertainty that can be expected in the quantile estimates and thus can hopefully provide some guidance to relevant application over CONUS.

Clearly a major limitation of current at-site FFA approaches relates to the fact that continuous observational records are relatively short (less than or equal to 20 years) and the fact that both AMS and PDS make only partial use of these data sets. An existing approach that used to deal with limited record is the Regional Flood Frequency Analysis (RFFA) (Kohnová & Szolgay, 2003; Mckerchar & Pearson, 1990; Mediero & Jiménez, 2007). While RFFA offers a potential solution, it has to be highlighted that a fundamental prerequisite for its application is the identification of statistically homogeneous regions, which is far from being a straightforward task for applications at the scale of CONUS, where areas with large variability in climate and watershed characteristics (both natural and anthropogenic) are observed. Alternatively, future research for improving at-site FFA could therefore focus on two possible directions that aim to (a) generate long-term hydrologic information by combining reanalysis data set with distributed hydrological models (Cea & Fraga, 2018; Hardesty, Shen, Nikolopoulos, & Anagnostou, 2018) and (b) investigate new statistical approaches (Marani & Ignaccolo, 2015) that overcome the limitations from shortage of data set and obtain more reliable assessment of high quantiles.

ACKNOWLEDGEMENTS

We would like to acknowledge the Eversource Energy Center at the University of Connecticut for the support of this research. F.M. was funded by the Israel Science Foundation [grant no. 1069/18].

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available by United States Geological Survey. These data were derived from the following resources available in the public domain: https://waterdata.usgs.gov/nwis/sw

ORCID

Efthymios I. Nikolopoulos https://orcid.org/0000-0002-5206-1249

REFERENCES

Ahilan, S., Amp, A. P. O. S., Sullivan, J. J., & Bruen, M. (2012). Influences on flood frequency distributions in irish river catchments. *Hydrology and Earth System Sciences*, *16*, 1137–1150.

Ahmad, M. I., Sinclair, C., & Spurr, B. (1988). Assessment of flood frequency models using empirical distribution function statistics. *Water Resources Research*, *24*, 1323–1328.

Archer, D. R., Leesch, F., & Harwood, K. (2007). Assessment of severity of the extreme river Tyne flood in January 2005 using gauged and historical information. *Hydrological Sciences Journal*, *52*, 992–1003.

Armstrong, W. H., Collins, M. J., & Snyder, N. P. (2012). Increased frequency of low-magnitude floods in new England 1. *JAWRA Journal of the American Water Resources Association*, *48*, 306–320.

Baratti, E., Montanari, A., Castellarin, A., Salinas, J. L., Viglione, A., & Bezzi, A. (2012). Estimating the flood frequency distribution at seasonal and annual time scales. *Hydrology and Earth System Sciences*, *16*, 4651–4660.

Benson, M. A. (1963). *Factors influencing the occurrence of floods in a humid region of diverse terrain*. Geological Survey: US Department of the Interior.

Bernardara, P., Mazas, F., Kergadallan, X., & Hamm, L. (2014). A two-step framework for over-threshold modelling of environmental extremes. *Natural Hazards and Earth System Sciences*, *14*, 635–647.

Bezak, N., Brilly, M., & Šraj, M. (2014). Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis. *Hydrological Sciences Journal*, *59*, 959–977.

Bhat, M. S., Alam, A., Ahmad, B., Kotlia, B. S., Farooq, H., Taloor, A. K., & Ahmad, S. (2018). Flood frequency analysis of river jhelum in kashmir basin. *Quaternary International*, *XXX*, XXX–XXX.

Bobée, B., Cavadias, G., Ashkar, F., Bernier, J., & Rasmussen, P. (1993). Towards a systematic approach to comparing distributions used in flood frequency analysis. *Journal of Hydrology*, *142*, 121–136.

Bobée, B., & Rasmussen, P. F. (1995). Recent advances in flood frequency analysis. *Reviews of Geophysics*, *33*, 1111–1116.

Caballero-Megido, C., Hillier, J., Wyncoll, D., Bosher, L., & Gouldby, B. (2018). Technical NOTE: Comparison of methods for threshold selection for extreme sea levels. *Journal of Flood Risk Management*, *11*, 127–140.

Cea, L., & Fraga, I. (2018). Incorporating antecedent moisture conditions and intraevent variability of rainfall on flood frequency analysis in poorly gauged basins. *Water Resources Research*, *54*, 8774–8791.

Chowdhury Jahir, U., & Stedinger Jery, R. (1991). Confidence interval for design floods with estimated skew coefficient. *Journal of Hydraulic Engineering*, *117*, 811–831.

Claps, P., & Laio, F. (2003). Can continuous streamflow data support flood frequency analysis? An alternative to the partial duration series approach. *Water Resources Research*, *39*(8), 1216.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208, p. 208). London: Springer.

Cunnane, C. (1973). A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of Hydrology*, *18*, 257–271.

Cunnane, C. (1985). Factors affecting choice of distribution for flood series. *Hydrological Sciences Journal*, *30*, 25–36.

Cunnane, C. (1988). Methods And Merits Of Regional Flood Frequency Analysis. *Journal of Hydrology*, *100*, 269–290.

Douglas, E., Vogel, R., & Kroll, C. (2000). Trends in floods and low flows in the United States: Impact of spatial correlation. *Journal of Hydrology*, *240*, 90–105.

Falcone, J. A. 2011. Gages-II: Geospatial attributes of gages for evaluating streamflow. US Geological survey.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*, 180–190.

Gnedenko, B. (1943). Sur la distribution LIMITE du TERME maximum D'UNE SERIE ALEATOIRE. *Annals of Mathematics*, *44*, 423–453.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., & Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions EXPRESSABLE in inverse form. *Water Resources Research*, *15*, 1049–1054.

Griffis, V., & Stedinger, J. (2007). Log-PEARSON type 3 distribution and its application in flood frequency analysis. Ii: Parameter estimation methods. *Journal of Hydrologic Engineering*, *12*, 492–500.

Gubareva, T. S. (2011). Types of probability distributions in the evaluation of extreme floods. *Water Resources*, *38*, 962–971.

Haddad, K., & Rahman, A. (2010). Selection of the best fit flood frequency distribution and parameter estimation procedure: A CASE study for Tasmania in Australia. *Stochastic Environmental Research and Risk Assessment*, *25*, 415–428.

Haktanir, T., & Horlacher, H. B. (1993). EValuation of various distributions for flood frequency analysis. *Hydrological Sciences Journal*, *38*, 15–32.

Hardesty, S., Shen, X., Nikolopoulos, E., & Anagnostou, E. (2018). A numerical framework for evaluating flood inundation hazard under different dam operation scenarios—A case study in Naugatuck river. *Water*, *10*, 1798.

Hosking, J. R. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, *52*, 105–124.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis: An approach based on L-MOMENTS*. CAMBRIDGE: Cambridge University Press.

IACWD, 1982. Guidelines for determining flood flow frequency, interagency committee on water data, bulletin 17-B. hydrology subcommittee, March 1982 (revised and corrected), 28 p. and appendices.

Keast, D., & Ellison, J. (2013). Magnitude frequency analysis of small floods using the annual and partial series. *Water*, *5*, 1816–1829.

Kendall, M. 1975. Multivariate Analysis, Charles Griffin & Company Ltd.

Kidson, R., & Richards, K. S. (2005). Flood frequency analysis: Assumptions and alternatives. *Progress in Physical Geography: Earth and Environment*, *29*, 392–410.

Kohnová, S., & Szolgay, J. (2003). Regional estimation of the index flood and the standard deviation of the summer floods in the Tatry mountains. *Journal of Hydrology and Hydromechanics*, *51*, 241–255.

Landwehr, J. M., Matalas, N., & Wallis, J. (1980). Quantile estimation with more or less floodlike distributions. *Water Resources Research*, *16*, 547–555.

Landwehr, J. M., Matalas, N., & Wallis, J. R. (1978). Some comparisons of flood statistics in real and log space. *Water Resources Research*, *14*, 902–920.

Lang, M., Ouarda, T., & Bobée, B. (1999). Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, *225*, 103–117.

Langousis, A., Mamalakis, A., Puliga, M., & Deidda, R. (2016). Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resources Research*, *52*, 2659–2681.

Madsen, H., Rasmussen, P. F., & Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. *Water Resources Research*, *33*, 747–757.

Marani, M., & Ignaccolo, M. (2015). *A METASTATISTICAL APPROACH TO RAINFALL EXTREMES*. A METASTATISTICAL approach to rainfall extremes. *Advances in Water Resources*, *79*, 121–126.

Marra, F., Nikolopoulos, E. I., Anagnostou, E. N., & Morin, E. (2018). Metastatistical extreme value analysis of hourly rainfall from short records: Estimation of high quantiles and impact of measurement errors. *Advances in Water Resources*, *117*, 27–39.

Martins, E. S., & Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, *36*, 737–744.

Mckerchar, A. I., & Pearson, G. P. (1990). Maps of flood statistics for regional flood frequency analysis in New Zealand. *Hydrological Sciences Journal*, *35*, 609–621.

Mediero, L. & Jiménez, A. Regional analysis for frequency estimation of annual flood peaks in ungauged basins of Spain. Geophysical Research Abstracts, 2007.

Meshgi, A., & Khalili, D. (2008). Comprehensive evaluation of regional flood frequency analysis by L- and LH-moments. II. development of lh-moments parameters for the generalized pareto and generalized logistic distributions. *Stochastic Environmental Research and Risk Assessment*, *23*, 137–152.

Michele, C. D., & Rosso, R. (2001). Uncertainty assessment of regionalized flood frequency estimates. *Journal of Hydrologic Engineering*, *6*, 453–459.

Nagy, B. K., Mohssen, M., & Hughey, K. F. D. (2017). Flood frequency analysis for a braided river catchment in New Zealand: Comparing annual maximum and partial duration series with varying record lengths. *Journal of Hydrology*, *547*, 365–374.

Opere, A. O., Mkhandi, S., & Willems, P. (2006). At site flood frequency analysis for the Nile equatorial basins. *Physics and Chemistry of the Earth, Parts A/B/C*, *31*, 919–927.

Overeem, A., Buishand, A., & Holleman, I. (2008). Rainfall depth-duration-frequency curves and their uncertainties. *Journal of Hydrology*, *348*, 124–134.

Pearson, C. P. (1991). New Zealand regional flood frequency analysis using L-moments. *Journal of Hydrology (New Zealand)*, *30*, 53–64.

Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*, 335–346.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, *3*, 119–131.

Rahman, A. S., Rahman, A., Zaman, M. A., Haddad, K., Ahsan, A., & Imteaz, M. (2013). A study on selection of probability distributions for at-site flood frequency analysis in Australia. *Natural Hazards*, *69*, 1803–1813.

Reis, D. S., & Stedinger, J. R. (2005). Bayesian mcmc flood frequency analysis with historical information. *Journal of Hydrology*, *313*, 97–116.

Ribatet, M., Sauquet, E., Grésillon, J.-M., & Ouarda, T. B. M. J. (2006). A regional Bayesian pot model for flood frequency analysis. *Stochastic Environmental Research and Risk Assessment*, *21*, 327–339.

Salmi, T. 2002. Detecting trends of annual values of atmospheric pollutants by the mann-kendall test and sen's slope estimates-the excel template application makesens, Llmatieteen laitos.

Sangal, B., & Biswas, A. K. (1970). The 3-parameter lognormal distribution and its applications in hydrology. *Water Resources Research*, *6*, 505–515.

Schlögl, M., & Laaha, G. (2017). Extreme weather exposure identification for road networks – A comparative assessment of statistical methods. *Natural Hazards and Earth System Sciences*, *17*, 515–531.

Seckin, N., Haktanir, T., & Yurtal, R. (2011). Flood frequency analysis of Turkey using L-moments method. *Hydrological Processes*, *25*, 3499–3505.

Solari, S., Egüen, M., Polo, M. J., & Losada, M. A. (2017). Peaks over threshold (pot): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, *53*, 2833–2849.

Stedinger, J. R. & Griffis, V. W. 2008. Flood frequency analysis in the united states: Time to update. American Society of Civil Engineers.

Subramanya, K. 2013. Engineering hydrology, 4E, TATA McGraw-Hill Education.

Tallaksen, L. M. (2000). Streamflow drought frequency analysis. In *Drought and drought mitigation in Europe* (pp. 103–117). Dordrecht: Springer.

Tanaka, S., Takara, K., Snorrason, A., Finnsdottir, H., & Moss, E. (2002). A study on threshold selection in pot analysis of extreme floods. *IAHS Publication*, *2002*, 299–306.

USWRC. (1976). *Guidelines for determining flood flow frequency. United States Water Resources Council Bull* (Vol. 17, p. 73). Hydrol comm: Washington, D.C.

Villarini, G., & Smith, J. A. (2010). Flood peak distributions for the eastern United States. *Water Resources Research*, *46*, 1–17.

Villarini, G., Smith, J. A., Serinaldi, F., Ntelekos, A. A., & Schwarz, U. (2012). Analyses of extreme flooding in Austria over the period 1951-2006. *International Journal of Climatology*, *32*, 1178–1192.

Vivekanandan, N. (2015). Flood frequency analysis using method of moments and L-moments of probability distributions. *Cogent Engineering*, *2*(1), 1018704.

Wallis, J. R., & Wood, E. F. (1985). Relative accuracy of log Pearson III procedures. *Journal of Hydraulic Engineering*, *111*, 1043–1056.

Wilson, D., Fleig, A. K., Lawrence, D., Hisdal, H., Pettersson, L. E. & Holmqvist, E. 2011. A review of NVE's flood frequency estimation procedures. *Norwegian Water Resources and Energy Directorate report*, 9, 2011.

How to cite this article: Hu L, Nikolopoulos EI, Marra F, Anagnostou EN. Sensitivity of flood frequency analysis to data record, statistical model, and parameter estimation methods: An evaluation over the contiguous United States. *J Flood Risk Management*. 2020;13:e12580. https://doi.org/10.1111/jfr3.12580