Underestimation of Frequent Floods when Using Annual Maxima for Frequency Analysis: Drivers, Spatial Variability, and Possible Solutions

F. Dell'Aira[1], A. Cancelliere[2], and C. I. Meier[1]

[1]University of Memphis, Department of Civil Engineering, Memphis, Tennessee, USA.
[2]University of Catania, Department of Civil Engineering and Architecture, Catania, Italy.

Corresponding author: Francesco Dell'Aira (fdllaira@memphis.edu)

Author Contributions:

Funding acquisition and Supervision: C. I. Meier; Conceptualization: F. Dell'Aira and A. Cancelliere; Methodology, Software, and Data Curation: F. Dell'Aira; Writing – review & editing: F. Dell'Aira and C.I. Meier

Key Points:

- The underestimation of frequent floods due to frequency analysis with annual maxima displays a strong spatial pattern and can be severe.
- The variability in underestimation is explained by specific distributional attributes reflecting climate and basin characteristics.
- We propose an equation to correct frequent floods predicted with annual maxima, but it is not valid in basins with mixed flood populations.

Abstract

Because they are conceptually unable to consider events at the sub-annual scale, probabilistic flood analyses based on annual maxima (AM) underestimate the actual frequency of frequent floods (with return periods under 5 years), so that peaks-over-threshold (POT) approaches should be preferred. While this has been acknowledged for decades, frequent floods are still estimated too often using AM, probably because the procedure is simpler, and AM series are longer and easier to obtain. However, the negative bias incurred when performing flood frequency with AM can be severe. This affects fields such as river restoration, stream ecology, and fluvial geomorphology, which require a correct characterization of frequent floods. Using hundreds of U.S. watersheds with natural flow regimes, across different climatic and geomorphic conditions, we systematically study the variability in how AM frequency analyses underestimate frequent floods, finding clear spatial patterns. Exploiting the duality between the Generalized Extreme Value and the Generalized Pareto distributions (used for modeling AM and POT, respectively), we identify the drivers of frequent-flood underestimation, studying the influence of the distributions' shapes. In turn, with the support of an optimal feature-selection technique, we determine the physical drivers explaining underestimation, from a wide spectrum of basin descriptors, investigating their linkages with the distributional characteristics that affect underestimation. A theoretical relationship is derived to infer the underestimation rate, allowing for post-hoc correction of AM-predicted frequent floods, without the need to perform POT frequency analyses. However, this approach underperforms at sites with mixed flood populations.

Plain Language Summary

Engineers and river scientists perform probabilistic analyses of floods to describe how frequently a given peak discharge is equaled or exceeded at a river location. The two approaches for selecting the peak values to be analyzed yield different flood predictions: annual maxima (AM), which takes only the maximum discharge from each single year in the record, tends to underestimate flood frequency (or overestimate the average time between events) as compared to peaks-over-threshold (POT), which includes all floods above a threshold. Even though this bias becomes significant when estimating frequent floods (those that occur on average at least once every 5 years), which play crucial roles in stream restoration, river ecology, and fluvial geomorphology, many still prefer AM over POT. This work studies frequent flood underestimation by AM at hundreds of U.S. basins, showing that its severity is strongly site-dependent and influenced by the climate: higher underestimation rates should be expected in arid and semi-arid regions. A theoretical correction approach is proposed to adjust the magnitude of frequent floods predicted with AM. An investigation into its limits of applicability finds poorer performances for basins where major floods happen anytime in the year, due to the occurrence of different flood-generating mechanisms.

1 Introduction

State-of-the-art methods for flood frequency analysis (FFA) use either annual maxima (AM) or peaks-over-threshold (POT) series (Pan et al., 2022). AM consider the largest event for each water year (starting on October 1, in the U.S.; Barth et al., 2017), while POT (also known as "exceedances" or partial duration series – PDS; Bezak et al., 2014) correspond to all the independent peaks extracted from the continuous hydrograph, that exceed a suitably defined threshold (Coles, 2001). The two methods predict average interarrival times (AITs) between two floods larger than a certain magnitude (also referred to as "events") which are conceptually different from each other (Wang & Holmes, 2020). AM-FFA produces what is conventionally referred to as return period R, i.e., the average number of years with no events before a year with at least one event. Mathematically, the domain of R is $(1,\infty)$ years, which implies that AM-FFA cannot consider events potentially occurring multiple times annually (Wang & Holmes, 2020). In contrast, POT-FFA predicts an Average Recurrence Interval (ARI) with domain $(0,\infty)$ years, thus also covering more frequent events, at the sub-annual scale (Wang & Holmes, 2020). This conceptual difference implies that, given the AM and POT series of peak flows observed at a given river section, the two methods will predict different AIT values between consecutive events of the same flood magnitude, independently of any sampling variability effect. Since their difference is negligible for large floods (Langbein, 1949; Wang & Holmes, 2020), the two methods have often been used almost interchangeably in many FFA applications (Adamowski, 2000; Bezak et al., 2014; Karim et al., 2017; Madsen et al., 1997; Metzger et al., 2020; Norheim, 2018; Ouarda et al., 2006). On the other hand, if the analysis focuses on frequent floods (FFs; i.e., events with R not larger than 5 years), the conceptual difference between R and ARI may translate into significantly different numerical values of the AITs predicted by the two methods, for the same flood magnitude (Ball et al., 2019; Karim et al., 2017; Wyżga, 1995). R predicted for a given FF by AM-FFA is larger than its corresponding ARI from POT-FFA (Langbein, 1949). Under the assumption that the annual number of exceedances follows a Poisson distribution (Wang & Holmes, 2020), Eq. (1) provides Langbein's relationship between R and ARI for a given flood of magnitude Q (Langbein, 1949).

$$\frac{1}{R(Q)} = 1 - \exp\left(-\frac{1}{ARI(Q)}\right) \tag{1}$$

The AIT between two FF events estimated from R may not reflect the real, higher frequency of occurrence of that FF, because using yearly time-blocks for sampling extreme events cannot accurately capture the interarrival time of frequent peaks, that may occur more than once per year. For such frequent events, the ARI from POT-FFA represents a better and conceptually more appropriate estimate of the actual AIT between two occurrences than the return period R (Ball et al., 2019; Karim et al., 2017; Wyżga, 1995).

Overestimating the AIT of a given FF event when using AM-FFA is equivalent to underestimating its frequency. If the focus of the analysis is identifying the flood peak magnitude Q that can occur with a given average frequency or AIT (e.g., once every 2 years), performing AM-FFA would result in underestimating Q.

There are multiple issues connected to the use of AM series for FFA. Two well-known drawbacks are: 1) a limited number of peak values, as compared to POT, for the same flow record (Bezak et al., 2014; Caires, 2009; Cunnane, 1973; Pan et al., 2022; Prosdocimi, 2018; Robson & Reed, 1999; Tavares & Da Silva, 1983), and 2) the risk of including low peaks from dry years in the analysis (Cohn et al., 2013; England et al., 2019; Lamontagne et al., 2016; Plavšić et al., 2016), which may arise from different hydrological mechanisms than the other peaks, and therefore come from a different population, violating the necessary assumption of i.i.d. events (Klemeš, 2000). Because of the first issue, national guidelines from different countries recommend minimum record lengths for performing AM-FFA (Robson & Reed, 1999; England et al., 2019). As to the low outliers, often referred to as "potentially influential low flows" (PILFs; Cohn et al., 2013), U.S. Bulletin 17C (England et al., 2019) recommends their preliminary removal from AM series using the "multiple Grubbs-Beck test" (MGBT; Cohn et al., 2013).

FF-underestimation is another well-known issue of AM-FFA (Langbein, 1949), but apparently it has not received the same attention in the hydrologic community. In the U.S., e.g., AM- has often been preferred over POT-FFA by many governmental agencies (e.g., Feaster et al., 2014; Kennedy & Paretti, 2014; Law & Tasker, 2003; Southard, 2010, Virginia Department of Transportation, 2021), even when predicting FFs with return periods as low as, e.g., 2 years, or even less. The use of state-of-the-art techniques (such as MGBT) may remove small peaks from AM series, which would be automatically ignored in the corresponding POT series. However, this does not resolve the issue of FF underestimation, which is not due to the presence of PILFs in the AM, but rather to the fundamental conceptual difference between R and ARI.

There is a number of reasons why AM-FFA still enjoys greater popularity, such as: 1) wider availability of AM series as compared to POT (Norheim, 2018; Prosdocimi et al., 2014); 2) greater simplicity since, in contrast with POT-, AM-FFA does not require applying independence criteria between subsequent flood events, nor selecting a threshold for defining extreme events (Pan et al., 2022); 3) the range of quantiles affected by FF underestimation is irrelevant in many engineering applications, which focus on more extreme, higher return-period floods.

With reference to the latter issue, major civil engineering works subject to risk of flooding, such as bridges (Benedict & Knight, 2021), storm sewers (Sun et al., 2011), dam-drainage systems (Khaddor et al., 2021), levees (Huang et al., 2015), and other hydraulic structures for river flood control (Cipollini et al., 2021; Lendering et al., 2019; Scussolini et al., 2016) are all designed to withstand relatively extreme events, with large return periods (Ponce, 1989; Rasekh et al., 2010; Sayers et al., 2013), depending on their strategic importance and the threat that their failure would pose to human lives and properties (Cipollini et al., 2021; Lendering et al., 2019; Morrison et al., 2018; Shah et al., 2018; Tung, 2005; Vogel & Castellarin, 2017). However, there are many other applications where accurate prediction of frequent floods is critical. Regular, low-magnitude floods play a more relevant role than extreme (but rarer) inundation events in a series of river-related phenomena such as changes in fluvial morphology (Death et al., 2015; Harvey et al., 1979; Wolman & Miller, 1960), sediment transport (Markus & Demissie, 2006), and dynamics of the stream ecosystem (Bendix & Hupp, 2000; Johnson et al., 1995; Meier, 2008), which are all crucial aspects in river restoration projects and river science (Wohl et al., 2015). Much research on fluvial geomorphology focuses on the role of FFs (e.g., with return periods between one and two years), which have been shown to simultaneously perform sufficient geomorphic work as well as occur frequently enough, so they tend to determine the channel's shape (Death et al., 2015; Harvey et al., 1979; Wolman & Miller, 1960). In river ecology, FFs affect the dynamic interactions between main channel and floodplain, with major impacts on the extension of the habitats cyclically available to the aquatic biota (Johnson et al., 1995; Wohl et al., 2015) and riparian vegetation (Bendix & Hupp, 2000; Meier, 2008; Wohl et al., 2015).

Biased FF predictions from using AM may negatively affect scientific and practical progress in these fields. For instance, a river restored based on the wrong design discharge may end up flooding as much as three times more often than per-design conditions, which could cause channel unravelling; similarly, environmental studies on river ecology and geomorphology, where FFs play a dominant role, may be based on wrong assumptions.

So far, FF-underestimation has been systematically investigated only in a few regions of the world, such as Poland (Wyżga, 1995) and Australia (Karim et al., 2017; Keast & Ellison, 2013). These authors suggest that the degree of FF underestimation is site-dependent, with the variability attributed to differences in the climate and resulting flow regime across catchments, as rivers with flashy behavior, typical of dry regions, experienced greater FF underestimation from AM-FFA than rivers in more humid regions, characterized by more stable flows (Karim et al., 2017; Wyżga, 1995). Some authors (e.g., Keast & Ellison, 2013; Page & McElroy, 1981) pointed out that Langbein's equation (Eq. 1) may misestimate the actual difference between R and ARI of FFs in some Australian basins. Keats & Ellison (2013) suggested that this could be due to a violation of the assumption of event independence, when extracting the PDS from long-term hydrographs.

Our work stems from the idea that a systematic comparison of FF predictions from AM and POT series at a wider, e.g., continental scale, can provide a deeper insight into the phenomenon of FF underestimation by AM-FFA and its climatic dependence. We choose the continental U.S. (CONUS) as study region, since FF underestimation by AM-FFA has not been systematically investigated there, yet.

Exploiting the duality between AM- and POT-FFA, valid under certain hypotheses (Wang & Holmes, 2020), we derive a theoretical expression of the underestimation of the T-year quantile from AM-FFA, as compared to POT-FFA, as a function of T and the parameters of the distribution of AM. This equation can be used directly to correct AM-based estimates of FF quantiles, without needing to perform POT-FFA.

It is known from decades of regionalization studies (e.g., Adamowski, 2000; Burn, 1997; Castellarin et al., 2001; Dalrymple, 1960; Hosking & Wallis, 1997; Hosking et al., 1985; Laio et al., 2011; Lun et al., 2021; Madsen et al., 1997; Metzger et al., 2020; Smith et al., 2015; Stedinger & Lu, 1995; Zaman et al., 2012; Zrinji & Burn, 1994) that there exist linkages between watershed characteristics and the parameters of flood distributions. Hence, the relationship between FF underestimation and the characteristics of the flood distribution can also be used to identify basin and climatic attributes that, through their influence on flood distributions, contribute to the spatial variability in FF underestimation. We deploy a wrapper method for feature selection (Babatunde et al., 2014; Huang et al., 2007) for this purpose. The list of candidate watershed characteristics encompasses a wide spectrum of information, including topographic, geomorphic, land-cover, and climatic descriptors, as well as hydrologic signatures (Addor et al., 2017). We also consider indicators of flow seasonality, derived from peak series using circular statistics (Villarini, 2016).

The large scale of this study also allows us to obtain a clearer view of the regions (and their characteristics) where Langbein's equation misestimates the difference between R and ARI of FFs.

To sum up, the contributions of this work can be synthesized as follows: 1) rigorously frame the phenomenon of FF underestimation from AM, allowing for a better understanding of the physical and theoretical drivers of variability in underestimation across varied sites; 2) provide a framework for preliminary estimates of the degree of FF underestimation by AM-FFA in a given basin, based only on its location and physical characteristics; 3) contribute a rigorous, theoretical method to correct AM-based estimates of FFs without performing POT-FFA in the first place, thus overcoming the common difficulties that POT-FFA typically entails; and 4) investigate where and why this theoretical method, as well as Langbein's equation, are (or are not) valid in real-world applications, possibly revisiting the explanation provided by Keast & Ellison (2013).

2 Methodology

Figure 1 shows a conceptual map of the methodological steps of this work, which are described in more detail in the following subsections.

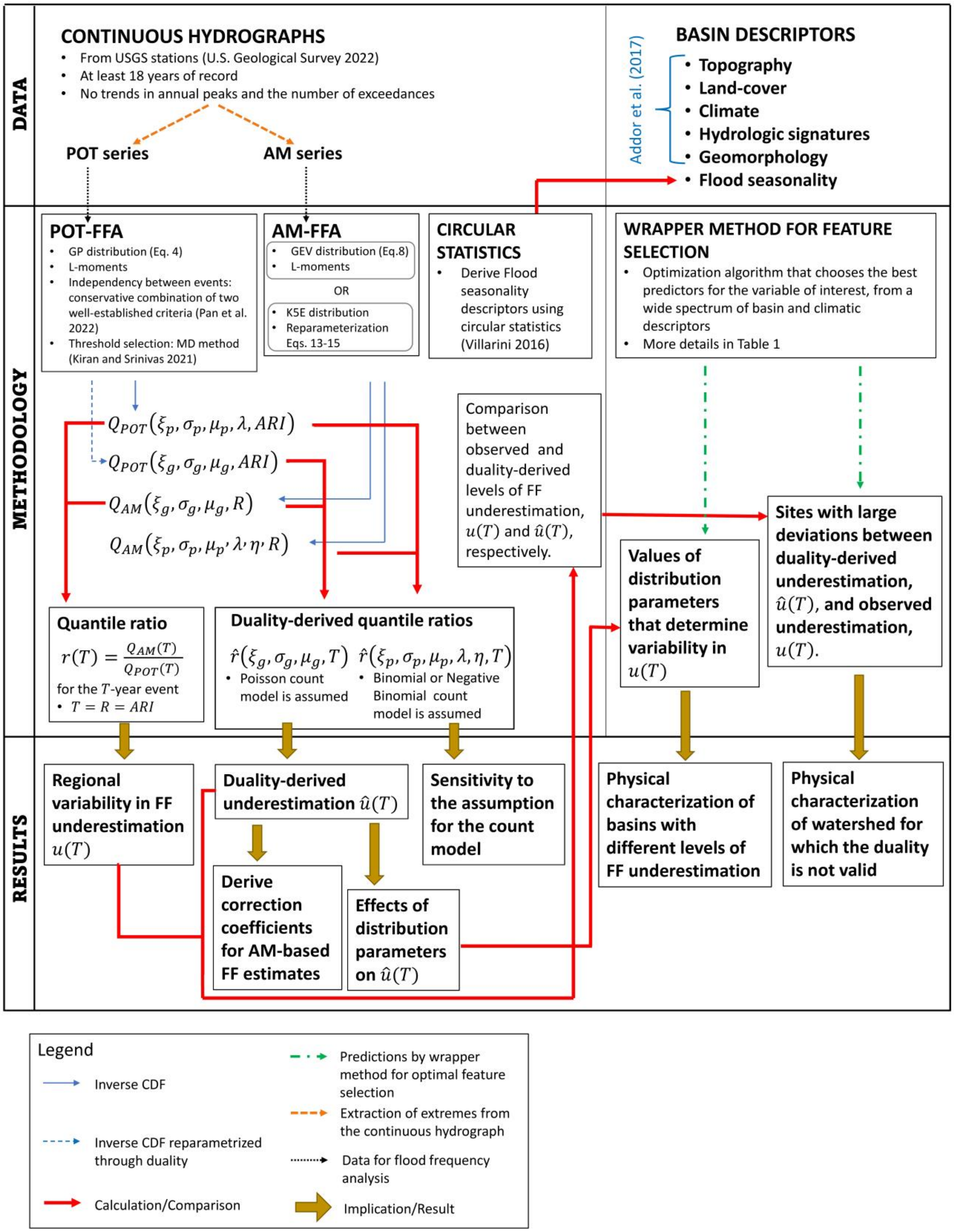


Figure 1. Scheme of the methodological steps of this work.

2.1 Quantile ratio as measure of the underestimation of FFs from AM-FFA

One way of comparing FF estimates from AM- and POT-FFA is to consider the same flood magnitude Q and assess the difference between its R and ARI values (e.g., Langbein, 1949; Wang & Holmes, 2020). Here we consider instead the same value of AIT T, for both R and ARI (i.e., $T = R = ARI$), and estimate the corresponding quantiles $Q_{AM}(T)$ and $Q_{POT}(T)$, using the inverse cumulative distribution functions (inverse CDFs) of the AM and POT distributions, respectively, where the probability is expressed in terms of T. In this way, the quantile ratio $r(T)$ of $Q_{AM}(T)$ to $Q_{POT}(T)$ (Eq. 2) readily provides a measure of the underestimation of the T-year flood by AM-FFA, given that the actual frequency of FFs is better reflected by the POT-based quantile estimate (Karim et al., 2017).

$$r(T) = \frac{Q_{AM}(T)}{Q_{POT}(T)} \tag{1}$$

Ignoring any effects of sampling variability, $Q_{AM}(T)$ is expected to be smaller than $Q_{POT}(T)$ for small Ts and become closer to $Q_{POT}(T)$ for increasing T, as reflected in Langbein's equation. Hence, $0 < r(T) \leq 1$, approximately. From $r(T)$, the percentage of underestimation due to using AM-FFA is obtained as $u(T) = [1 - r(T)] \times 100\%$.

2.2 Duality-based quantile ratio

Under the hypotheses of i.i.d. POTs distributed as $G_{POT}(Q)$ and number m of exceedances per year distributed as $P(m)$, the distribution of AM, $F_{AM}(Q)$ is univocally determined by the total probability theorem (Eq. 3; Önöz & Bayazit, 2001). $F_{AM}(Q)$ is referred to as the derived distribution of AM (Meier et al., 2016).

$$F_{AM}(Q \leq Q_T) = \sum_{k=0}^{\infty} P(m = k)[G_{POT}(Q_T)]^k \tag{2}$$

Furthermore, the parameters of the two distributions $G_{POT}(Q)$ and $F_{AM}(Q)$ can be related to each other by a set of reparameterization equations (Madsen et al., 1997; Prosdocimi & Kjeldsen, 2022; Wang & Holmes, 2020). This property is termed the "duality" between the AM and POT distributions (Coles, 2001; Wang & Holmes, 2020) and can be exploited to rewrite $r(T)$ as an expression of the parameters of a single distribution, either that for the AMs or the POTs. It is worth noticing that, for small quantiles, the choice of the distributions for AM and POTs is expected to have only a minor impact on $r(T)$ and $u(T)$.

We use the Generalized Pareto (GP) distribution (Equation 4), with shape ξ_p, scale σ_p, and location μ_p, to model the magnitude of exceedances.

$$P_{GP}(Q \leq Q_T, Q > \mu_p) = \begin{cases} 1 - \left(1 + \xi_p \dfrac{Q_T - \mu_p}{\sigma_p}\right)^{-\frac{1}{\xi_p}}, & for\ \xi_p \neq 0 \\ 1 - \exp\left(-\dfrac{Q_T - \mu_p}{\sigma_p}\right), & for\ \xi_p = 0 \end{cases} \tag{3}$$

GP is defined on the set $\{Q_T: Q_T > \mu_p\}$ when $\xi_p \geq 0$, and $\left\{Q_T: \mu_p < Q_T < \mu_p - \frac{\sigma_p}{\xi_p}\right\}$ when $\xi_p < 0$. For $\xi_p = 0$, GP degenerates to a shifted exponential distribution (Coles, 2001). For the AM, we consider the derived distributions from the total probability theorem (Eq. 3) for three alternative count models of m, namely the Poisson (PSN), Negative Binomial (NEG), and Binomial (BIN) distributions, given by Eqs. (5), (6), and (7), respectively. While the PSN is the most popular (Bezak et al., 2014; Pan et al., 2022; Wang & Holmes, 2020), some authors (e.g., Bezak et al., 2014; Bhunya et al., 2013; Önöz & Bayazit, 2001) have proposed NEG and BIN as alternative models to deal with cases of over- or under-dispersion, respectively.

$$P_{PSN}(m = k, \lambda\,) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{4}$$

The parameter λ of PSN represents the expected value $E(m)$ of the number m of exceedances (Önöz & Bayazit, 2001).

$$P_{NEG}(m = k, \alpha, \gamma) = \binom{\gamma + k - 1}{k}\alpha^k(1-\alpha)^\gamma \tag{5}$$

The parameters α and γ of NEG can be derived from $E(m) = \lambda = \frac{\alpha\gamma}{1-\alpha}$, and the variance of the number of exceedances, $Var(m) = V = \frac{\alpha\gamma}{(1-\alpha)^2}$, as $\alpha = 1 - \frac{\lambda}{V}$ and $\gamma = \lambda\frac{1-\alpha}{\alpha} = \frac{\lambda^2}{V-\lambda}$ (Bhunya et al., 2013).

$$P_{BIN}(m = k, \beta, \delta) = \binom{\delta}{k}\beta^k(1-\beta)^{\delta-k} \tag{6}$$

The parameters β and δ of the BIN can be derived from $E(m) = \lambda = \beta\delta$ and $Var(m) = V = \beta\delta(1-\beta)$ as $\delta = \frac{\lambda^2}{\lambda - V}$ and $\beta = \frac{\lambda}{\delta} = \frac{\lambda - V}{\lambda}$ (Önöz & Bayazit, 2001).

The derived distribution of AM assuming GP and PSN for magnitude and number of exceedances, respectively, is the Generalized Extreme Value (GEV) distribution (Eq. 8), with shape, scale, and location parameters (ξ_g, σ_g, and μ_g, respectively) assuming values in the ranges $-\infty < \xi_g < +\infty$, $\sigma_g > 0$, and $-\infty < \mu_g < +\infty$, respectively (Coles, 2001).

$$P_{GEV}(Q \le Q_T) = \begin{cases} \exp\left[-\left(1 + \xi_g\dfrac{Q_T - \mu_g}{\sigma_g}\right)^{-\frac{1}{\xi_g}}\right], & for\ \xi_g \ne 0 \\ \exp\left[-\exp\left(-\dfrac{Q_T - \mu_g}{\sigma_g})\right)\right], & for\ \xi_g = 0 \end{cases} \tag{7}$$

The GEV has a lower (upper) bound equal to $\mu_g - \frac{\sigma_g}{\xi_g}$ for $\xi_g > 0$ ($\xi_g < 0$), while it is unbounded for $\xi_g = 0$ (Coles, 2001).

Reparameterization Eqs. (9), (10), and (11) provide the relationships between the parameters of the GP and GEV distributions (Wang & Holmes, 2020).

$$\xi_p = \xi_g = \xi \tag{8}$$

$$\lambda = \begin{cases} \left(1 - \xi\dfrac{\mu_g - \mu_p}{\sigma_g}\right)^{-\frac{1}{\xi}}, & for\ \xi \ne 0 \\ \exp\left(\dfrac{\mu_g - \mu_p}{\sigma_g}\right), & for\ \xi = 0 \end{cases} \tag{9}$$

$$\sigma_p = \sigma_g - \xi(\mu_g - \mu_p) \tag{10}$$

Alternatively, when the NEG or BIN are considered to model the number of exceedances, the derived distribution is a 5-parameter extension of the 4-parameter Kappa (KPP) family (K5E; Eq. 12; Eastoe & Tawn, 2010), of which the GEV is a member (Hosking, 1994).

$$P_{K5E}(Q \leq Q_T) = \begin{cases} \left[1 - \eta \left(1 + \xi_k \dfrac{Q_T - \mu_k}{\sigma_k}\right)^{-\frac{1}{\xi_k}}\right]^{\frac{\lambda}{\eta}}, & for\ \xi_k \neq 0 \\ \left[1 - \eta \exp\left(-\dfrac{Q_T - \mu_k}{\sigma_k}\right)\right]^{\frac{\lambda}{\eta}}, & for\ \xi_k = 0 \end{cases} \tag{11}$$

The special KPP case (Hosking, 1994) occurs when $\lambda = 1$. Reparameterization equations for the GP-K5E duality are given by Eqs. (13), (14), and (15).

$$\mu_k = \mu_p\ , \quad \sigma_k = \sigma_p\ , \quad \xi_k = \xi_p \tag{12}$$

$$\beta = \frac{\alpha}{\alpha - 1} = \frac{\lambda - V}{\lambda} = \eta \tag{13}$$

$$\delta = -\gamma = \frac{\lambda^2}{\lambda - V} = \frac{\lambda}{\eta} \tag{14}$$

Eqs. (16), (17), and (18) represent the inverse CDFs of the GEV, GP, and K5E distributions, respectively, where the probability is suitably expressed in terms of R or ARI, by considering $G_{POT}(Q_{POT}(ARI)) = 1 - 1/(\lambda\, ARI)$ and $F_{AM}(Q_{AM}(R)) = 1 - 1/R$, respectively.

$$Q_{AM}(\xi_g, \sigma_g, \mu_g, R) = \begin{cases} \mu_g + \dfrac{\sigma_g}{\xi_g}\left\{\left[\ln\left(\dfrac{R}{R-1}\right)\right]^{-\xi_g} - 1\right\}, for\ \xi_g \neq 0 \\ \mu_g - \sigma_g \ln\left[\ln\left(\dfrac{R}{R-1}\right)\right], & for\ \xi_g = 0 \end{cases} \tag{15}$$

$$Q_{POT}(\xi_p, \sigma_p, \mu_p, \lambda, ARI) = \begin{cases} \mu_p + \dfrac{\sigma_p}{\xi_p}\left[(\lambda\, ARI)^{\xi_p} - 1\right], & for\ \xi_p \neq 0 \\ \mu_p + \sigma_p \ln(\lambda\, ARI), & for\ \xi_p = 0 \end{cases} \tag{16}$$

$$Q_{AM}(\xi_p, \sigma_p, \mu_p, \lambda, \eta, R) = \begin{cases} \mu_p + \dfrac{\sigma_p}{\xi_p}\left\{\left[\dfrac{1}{\eta}\left(1 - \left(1 - \dfrac{1}{R}\right)^{\frac{\eta}{\lambda}}\right)\right]^{-\xi_p} - 1\right\}, for\ \xi_p \neq 0 \\ \mu_p - \sigma_p \ln\left\{\dfrac{1}{\eta}\left[1 - \left(1 - \dfrac{1}{R}\right)^{\frac{\eta}{\lambda}}\right]\right\}, & for\ \xi_p = 0 \end{cases} \tag{17}$$

Under the assumption of a PSN count model, replacing Eqs. (9), (10), and (11) into Eqs. (16) and (17) leads to Eq. (19) and (20), respectively, where the AM-quantile is expressed in terms of the parameters of the distribution of POTs, and vice versa.

$$Q_{AM}(\xi_p, \sigma_p, \mu_p, \lambda, R) = \begin{cases} \mu_p + \dfrac{\sigma_p}{\xi_p}\left[\left(\dfrac{1}{\lambda}\ln\dfrac{R}{R-1}\right)^{-\xi_p} - 1\right], & for\ \xi_p \neq 0 \\ \mu_p - \sigma_p \ln\left[\dfrac{1}{\lambda}\ln\left(\dfrac{R}{R-1}\right)\right], & for\ \xi_p = 0 \end{cases} \tag{18}$$

$$Q_{POT}\left(\xi_g,\sigma_g,\mu_g,ARI\right)=\begin{cases}\mu_g+\dfrac{\sigma_g}{\xi_g}\left(ARI^{\xi_g}-1\right), & for\ \xi_g\neq 0\\ \mu_g+\sigma_g\ln(ARI), & for\ \xi_g=0\end{cases} \tag{19}$$

Eqs. (19) and (20) can be used to obtain expressions of the quantile ratio $\hat{r}\left(\xi_g,\sigma_g,\mu_g,T\right)$ and $\hat{r}\left(\xi_p,\sigma_p,\mu_p,\lambda,T\right)$ for the T-year event [Eqs. (21) and (22), respectively], as functions of the parameters of a single distribution, either the GEV or the GP, respectively. For convenience, expressions are derived for the general case $\xi_g\neq 0$.

$$\hat{r}\left(\xi_g,\sigma_g,\mu_g,T\right)=\frac{Q_{AM}\left(\xi_g,\sigma_g,\mu_g,T\right)}{Q_{POT}\left(\xi_g,\sigma_g,\mu_g,T\right)}=\frac{1+\dfrac{1}{\xi_g}\dfrac{\sigma_g}{\mu_g}\left[\left(\ln\dfrac{T}{T-1}\right)^{-\xi_g}-1\right]}{1+\dfrac{1}{\xi_g}\dfrac{\sigma_g}{\mu_g}\left[T^{\xi_g}-1\right]} \tag{20}$$

$$\hat{r}\left(\xi_p,\sigma_p,\mu_p,\lambda,T\right)=\frac{Q_{AM,T}\left(\xi_p,\sigma_p,\mu_p,\lambda,T\right)}{Q_{POT,T}\left(\xi_p,\sigma_p,\mu_p,\lambda,T\right)}=\frac{1+\dfrac{1}{\xi_p}\dfrac{\sigma_p}{\mu_p}\left[\left(\dfrac{1}{\lambda}\ln\dfrac{T}{T-1}\right)^{-\xi_p}-1\right]}{1+\dfrac{1}{\xi_p}\dfrac{\sigma_p}{\mu_p}\left[(\lambda T)^{\xi_p}-1\right]} \tag{21}$$

Eqs. 21 and 22 are valid under the assumption of PSN count model for m.

Alternatively, if a NEG or BIN count model is assumed, a duality-based expression of the quantile ratio can be obtained using the parameters of the K5E (Eq. 23) from the ratio of Eqs. (18) and (17).

$$\hat{r}\left(\xi_p,\sigma_p,\mu_p,\lambda,\eta,T\right)=\frac{1+\dfrac{1}{\xi_p}\dfrac{\sigma_p}{\mu_p}\left\{\left[\dfrac{1}{\eta}\left(1-\left(1-\dfrac{1}{T}\right)^{\frac{\eta}{\lambda}}\right)\right]^{-\xi_p}-1\right\}}{1+\dfrac{1}{\xi_p}\dfrac{\sigma_p}{\mu_p}\left[(\lambda T)^{\xi_p}-1\right]} \tag{22}$$

In either case, the duality-based underestimation $\hat{u}(T)$ is obtained from the corresponding duality-based quantile ratio $\hat{r}(T)$ as $\hat{u}(T)=[1-\hat{r}(T)]\times 100\%$.

2.3 Annual maxima and peaks-over-threshold analyses

We use L-moments (LM-method, Hosking, 1990) for fitting GP on the POT series and GEV on the AM series, and Eqs. (13), (14), and (15) to derive K5E parameters from the duality with the GP. λ and V are equaled to the sample mean and variance of m extracted from the POT series, respectively. The fitting of the GEV and GP on the AM and POT series, respectively, is assessed through the Kolmogorov-Smirnov goodness-of-fit (GOF) test (Kottegoda & Rosso, 1997).

Independent peaks are identified considering a conservative combination of two popular independence criteria (both reported in Pan et al., 2022): one proposed in Bulletin 17 (U.S. Water Resources Council, 1976) and the other recommended by Cunnane (1979) and in Volume 3 of the UK Flood Estimation Handbook (Robson & Reed, 1999).

Regarding threshold selection, we adopted the Mahalanobis distance (MD)-based method by Kiran & Srinivas (2021), since it has been shown to outperform many other recent automated threshold-selection criteria in the literature (Kiran & Srinivas, 2021). To test the sensitivity of our results, we also considered the method by Solari et al. (2017), which uses a GOF test to evaluate the various samples of exceedances obtained with the moving threshold, instead of L-moments.

2.4 Feature selection for identifying optimal predictors

A wrapper method for optimal feature selection (Babatunde et al., 2014; Huang et al., 2007) is used to characterize watersheds with different levels of FF underestimation, as well as those basins where the duality (Eq. 21) is not valid. It couples an optimization algorithm with a learning machine, where the latter is trained to map input into output variables while the former determines the optimal predictors among a wide spectrum of basin attributes, based on the performance of the learning machine (Huang et al., 2007; Babatunde et al., 2014). In this work, these attributes are either mapped into distributional characteristics that affect FF underestimation or are used to classify catchments where Eq. (21) is not valid.

The Non-Dominated Sorting Genetic Algorithm II (NSGA-II; Deb et al., 2002) is used as an optimizer, considering three conflicting objectives on the predictive power and the number of optimal features (see Table 1). The property of genetic algorithms (GAs) of dealing with a multitude of candidate solutions ("population") spread over the solution space (Simpson et al., 1994), makes NSGA-II particularly effective in avoiding local optima, quite typical in feature-selection problems (Huang et al., 2007), and suitable for identifying synergies among groups of two or more explanatory variables which could be irrelevant individually, but may display high explanatory power when combined with others (Taormina & Chou, 2015).

We deploy an ensemble of 12 multi-layer perceptrons (MLPs; Hornik et al., 1989), trained independently from each other, as the learning machine. Ensemble predictions average out any bias from the single training instances (Aggarwal et al., 2018), hence an unbiased assessment of each set of candidate input variables is obtained. The evaluation is based on the average performance on test basins in a five-fold validation framework (i.e., five iterative splits of the full dataset into training and test sets, with 80%-20% proportion, respectively).

Table 1 summarizes the tuning parameters of the optimization routine and the training/assessment of the learning machine in the two cases of 1) training a regressor model to map basin attributes into distributional characteristics that affect FF underestimation, and 2) training a binary classifier to identify watersheds where the duality-derived expression of the quantile ratio (Eq. 21) is not valid. Note that the output of a MLP classifier can be interpreted as the predicted probability of a positive case (i.e., basin where Eq. 21 is not valid).

For both optimization routines, the number of features N_f is minimized while maximizing the performance of the learning machine, as measured by the other two objectives. Therefore, the optimal Pareto fronts (Figures 6a and 8a) contain multiple trade-off solutions (see, e.g., Dell'Aira et al., 2021) with different performances and numbers N_f of optimal features, each corresponding to different subsets of basin attributes. These are broken down into a heatmap (Figures 6b and 8b), with basin attributes and N_f on the two axes, and the color gradient displaying the relative frequency at which each attribute is considered in alternative optimal solutions with same N_f. Key basin attributes are those that are used more frequently in alternative optimal solutions with same N_f, as well as those used in the most parsimonious solutions (i.e., with the smallest N_f values).

The way key basin attributes affect the target variable is studied using bivariate partial-dependence plots (PDPs; Figures 7 and 9), which show the marginal effect of a pair of predictors on the output of the learning machine, averaging out the effects of all the other input features (Auret & Aldrich, 2012).

Table 1. Parameter tuning of the optimization routine and learning machine training coupled in the wrapper method for feature selection. Depending on the target variable, regression or classification models are trained on candidate optimal features.

	Model/Algorithm /Hyperparameter	Regressor (sigma/mu)	Classifier (anomaly detection)	Notes/References
Feature Optimization	Optimization algorithm	NSGA-II	NSGA-II	(Deb et al. 2002)
	Population size N_{pop}	1000	1000	Survival of the fittest individuals to keep population size constant.
	Number of generations N_{gen}	≥ 200	≥ 200	Stop after 20 generations with no significant changes in the optimal population (but not before 200).
	Crossover – mutation probabilities	0.85-0.15	0.85-0.15	Probability of occurrence of one of the two genetic operators for each pair of parent individuals (Simpson et al. 1994).
	Objective 1	N_f	N_f	Number of input features.
	Objective 2	MAE	ROC_AUC	Mean Absolute Error, in the interval $[0; +\infty[$; Receiver Operating Characteristic (ROC) area under the curve, in the interval $[0.5;\ 1]$. Optimal value of ROC_AUC is 1, below 0.5 the classifier performs worse than a random classifier (Fernández et al. 2018)
	Objective 3	IQR	PR_AUC	Width of the interquartile range of errors, in the interval $[0; +\infty[$ (ideal value is 0); Precision-Recall area under the curve, in the interval $[r; 1]$, where r is the proportion of actual positives over the total number of cases; optimal value is 1; below r is worse than a trivial classifier that marks all instances as positive (Fernández et al. 2018)
Learning Machine Training	Model	MLP	MLP	Multi-layer perceptron (Hornik et al. 1989)
	Ensemble size	12	12	Ensemble predictions average out any bias related to the single training instance (Aggarwal 2018).
	Number of hidden layers	3	3	
	Activation	Sigmoid	Sigmoid	
	Hidden units per layer	30	100	
	Number of epochs	~500	~20000	Early stop after 100 – 10000 epochs with no performance improvement on the validation set.

	Optimizer	Adam	SGD	Adaptive Moment Estimation – Stochastic Gradient Descent (Fatima 2020; Landro et al. 2020)
	Performance metric	MAE	Binary cross-entropy	Mean Absolute Error (Wang et al. 2020)
	Oversampling (For imbalanced datasets)	-	SMOTE with Tomek links	Technique to artificially generate other positive examples in an otherwise imbalanced dataset (Fernández et al. 2018). Our original dataset only has ~20% of positive examples, i.e., basins where duality apparently is not valid, which would hamper the training, resulting in a classifier that only predicts negative labels.
	Cost-sensitive learning (For imbalanced datasets)	-	Weights 1:5	Error on positive examples (false negative) are weighted 5 times more than errors on negative examples (false positive) in the error function minimized during the training, to reduce chances of misclassification of the positive examples (Fernández et al. 2018).

2.5 Descriptors of flood seasonality as a proxy for the presence of mixed populations

The date in which a flood occurs contains information about its generating mechanism: for instance, snowmelt floods are concentrated in the springtime in the U.S., while flooding events during the summertime may be caused by, e.g., convective storms, tropical cyclones, or the monsoon phenomenon, depending on the specific geographical location (Villarini, 2016).

Villarini (2016) used circular statistics to analyze the seasonality of AM. The annual peak in year k^{th} is represented by a vector $\boldsymbol{z}_{AM,k}$ of unit length and direction ϑ_k in the complex plane (Eq. 24, with $i = \sqrt{-1}$), where ϑ_k corresponds to the time of year when the flood occurred (in radians, see Figure 3a-b). The vectorial sum $\bar{\boldsymbol{z}}_{AM}$ of vectors $\boldsymbol{z}_{AM,k}$ across multiple years (Eq. 25) represents the long-term average seasonality, which offers insights into the seasonal patterns of flood occurrence and the mechanisms that may drive them. Its module $|\bar{\boldsymbol{z}}_{AM}|$, in the interval [0; 1], is a measure of the strength of seasonality (Eq. 26). For example, if the annual vectors $\boldsymbol{z}_{AM,k}$ are clustered in the same season, $|\bar{\boldsymbol{z}}_{AM}|$ will be close to 1 (Figure 2b), indicating that one single dominant mechanism is responsible for the largest floods, whereas if vectors $\boldsymbol{z}_{AM,k}$ are spread out more widely, $|\bar{\boldsymbol{z}}_{AM}|$ would be close to 0 (Figure 2a), suggesting the presence of multiple major mechanisms operating at different times of the year. The direction of $\boldsymbol{z}_{AM,k}$, $\bar{\vartheta}_{AM}$ (Eq. 27, defined by parts to fit ϑ_k in the interval [0; 2π]) indicates the time of the year (in radians) when annual peaks are mostly concentrated.

$$\boldsymbol{z}_{AM,k} = \cos\vartheta_k + i\sin\vartheta_k \tag{23}$$

$$\bar{\boldsymbol{z}}_{AM} = \frac{1}{N_{years}}\sum_{k=1}^{N_{years}}\cos\vartheta_k + i\frac{1}{N_{years}}\sum_{k=1}^{N_{years}}\sin\vartheta_k = \bar{z}_{AM,x} + i\bar{z}_{AM,y} \tag{24}$$

$$|\bar{\mathbf{z}}_{AM}| = \sqrt{{\bar{z}_{AM,x}}^2 + {\bar{z}_{AM,y}}^2} \tag{25}$$

$$\bar{\vartheta}_{AM} = \begin{cases} \arctan\frac{\bar{z}_{AM,y}}{\bar{z}_{AM,x}}\ , & if\ \bar{z}_{AM,x} > 0 \\ \arctan\frac{\bar{z}_{AM,y}}{\bar{z}_{AM,x}} + \pi\ , & if\ \bar{z}_{AM,x} < 0\ and\ \bar{z}_{AM,y} \geq 0 \\ \arctan\frac{\bar{z}_{AM,y}}{\bar{z}_{AM,x}} - \pi\ , & if\ \bar{z}_{AM,x} < 0\ and\ \bar{z}_{AM,y} < 0 \end{cases} \tag{26}$$

In Eqs. 25–27, $\bar{z}_{AM,x}$ and $\bar{z}_{AM,y}$ represent the real and imaginary component of $\bar{\mathbf{z}}_{AM}$, respectively, while N_{years} the number of years of record.

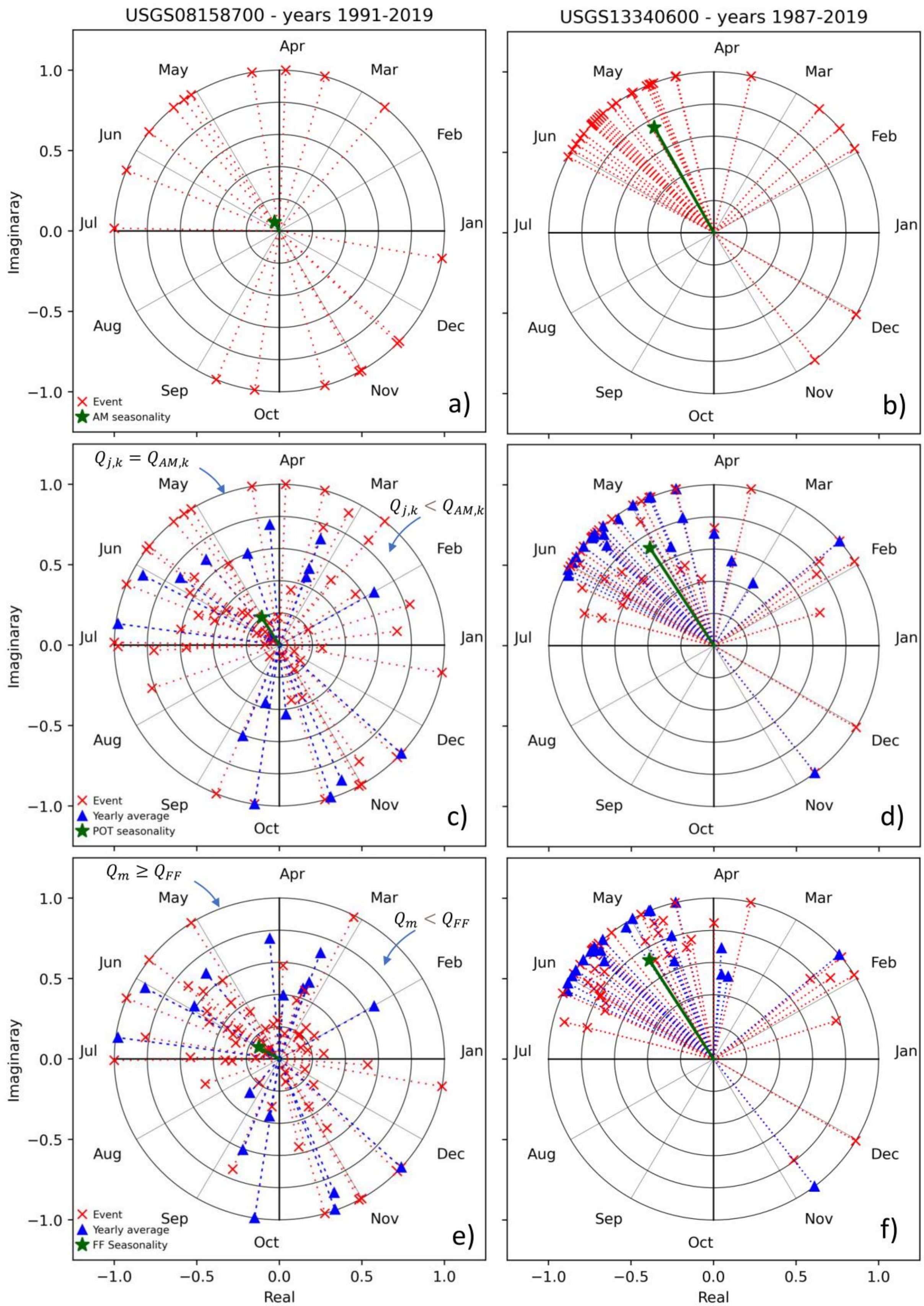

USGS08158700 - years 1991-2019
USGS13340600 - years 1987-2019
Apr
Mar
Feb
Jan
Dec
Nov
Oct
Sep
Aug
Jul
Jun
May
Imaginaray
Real
Event
AM seasonality
Yearly average
POT seasonality
FF Seasonality
$Q_{j,k} = Q_{AM,k}$
$Q_{j,k} < Q_{AM,k}$
$Q_m \geq Q_{FF}$
$Q_m < Q_{FF}$
a)
b)
c)
d)
e)
f)

Figure 2. Descriptors of a-b) AM-, c-d) POT-, and e-f) FF-seasonality, for two USGS stations. The column on the left shows a case of low seasonality, as peaks are evenly spread within the year. The case on the right shows a basin with high flood seasonality, concentrated in the Spring season (May and June).

This approach can provide insights into the climatic mechanisms behind AM floods, but it does not account for other relevant floods that may occur in the same or in other seasons. Hence, we introduce two other measures of flood seasonality, namely the POT seasonality and the FF seasonality, both derived from the POT series. In the former (Figure 2c-d), each POT from the k^{th} year is scaled by the corresponding AM, so that all floods contribute to the overall seasonality proportionally to their relative magnitude, as compared to the annual peak (Eq. 28). $N_{POT,k}$ is the number of POTs in the k^{th} year, while $Q_{AM,k}$ is its annual flood. $Q_{j,k}$ is the j^{th} POT in year k^{th}, while $\vartheta_{j,k}$ is the time of the year when it occurred.

$$\bar{\boldsymbol{z}}_{POT} = \frac{\sum_{k=1}^{N_{years}} \sum_{j=1}^{N_{POT,k}} \frac{Q_{j,k}}{Q_{AM,k}} \cos \vartheta_{j,k}}{\sum_{k=1}^{N_{years}} \sum_{j=1}^{N_{POT,k}} \frac{Q_{j,k}}{Q_{AM,k}}} + i \frac{\sum_{k=1}^{N_{years}} \sum_{j=1}^{N_{POT,k}} \frac{Q_{j,k}}{Q_{AM,k}} \sin \vartheta_{j,k}}{\sum_{k=1}^{N_{years}} \sum_{j=1}^{N_{POT,k}} \frac{Q_{j,k}}{Q_{AM,k}}} = \bar{z}_{POT,x} + i\bar{z}_{POT,y} \tag{27}$$

POT seasonality indices $|\bar{\boldsymbol{z}}_{POT}|$ and $\bar{\vartheta}_{POT}$ are obtained replacing the components $\bar{z}_{POT,x}$ and $\bar{z}_{POT,y}$ of $\bar{\boldsymbol{z}}_{POT}$ in Eqs. 26 and 27, respectively.

For the FF seasonality (Figure 2e-f), all POTs equal or larger than the reference frequent flood Q_{FF}, considered herein as the 2-year event, $Q_{FF} = Q_{POT}(T = 2)$, are represented by a unit vector (like the annual maxima in Eq. 24) and therefore weighted in the same way, while smaller events in the PDS are suitably scaled by Q_{FF} so as to decrease their importance in the overall seasonality (Eq. 29). In this way, information about both major, but rare events, as well as frequent smaller floods, is incorporated into the seasonality indices, and the choice of Q_{FF} defines the lower bound for the range of flood sizes considered with full weight.

$$\bar{\boldsymbol{z}}_{FF} = \frac{\sum_{m=1}^{N_{POT}} \frac{Q_m}{\max\,[Q_m, Q_{FF}]} \cos \vartheta_m}{\sum_{j=1}^{N_{POT}} \frac{Q_m}{\max\,[Q_m, Q_{FF}]}} + i \frac{\sum_{m=1}^{N_{POT}} \frac{Q_m}{\max\,[Q_m, Q_{FF}]} \sin \vartheta_m}{\sum_{j=1}^{N_{POT}} \frac{Q_m}{\max\,[Q_m, Q_{FF}]}} = \bar{z}_{FF,x} + i\bar{z}_{FF,y} \tag{28}$$

N_{POT} is the number of POTs in the PDS, Q_m is the m^{th} POT of the series, and ϑ_m is the time of the year when it occurred. FF seasonality indices $|\bar{\boldsymbol{z}}_{FF}|$ and $\bar{\vartheta}_{FF}$ are obtained replacing the components $\bar{z}_{FF,x}$ and $\bar{z}_{FF,y}$ of $\bar{\boldsymbol{z}}_{FF}$ in Eqs. 26 and 27, respectively.

Unlike the AM seasonality, which only provides insight into the drivers of the largest flood each year, the POT seasonality is expected to incorporate information about all major floods in the record. In addition, the FF seasonality looks at the seasonal distribution of all peaks in the POT series, but emphasizing those that are larger than the reference frequent flood Q_{FF}.

These descriptors are intended to complement each other; from their comparison, one should infer information on the seasonality of events across a range of magnitudes, possibly with different patterns of seasonality.

3 Case study basins

We considered the subset of CONUS watersheds from the well-known CAMELS dataset (Addor et al., 2017) for which continuous flow data (U.S. Geological Survey, 2022) are available for a minimum of 18 water years. Years with incomplete records were discarded if the gap was larger than 20%.

Basins with trends in the AM series or in the number of POTs per year, detected through the Mann-Kendall test (Bayazit, 2015), were excluded from the analysis, to ensure stationarity. Potentially influential low flows (PILFs; Cohn et al., 2013) were removed from the AM series using the MGBT algorithm recommended in the USGS Bulletin 17C (England et al., 2019), adopting the R package "MGBT" by Asquith et al. (2021). The methodology assumes that the log-transformed AM follow a normal distribution, and a statistical test is used to identify any low outliers in the series (Cohn et al., 2013).

The final dataset contains 432 basins with minimum human impacts, whose attributes encompass topographic, geomorphologic, climatic, and land-cover (LC) information, as well as hydrologic signatures (Addor et al., 2017), supplemented with the flood seasonality indices described in Section 2.5.

In what follows, basin attributes from the CAMELS dataset are referenced using the same names as in Addor et al. (2017). The only exception is the dominant land-cover (dom_land_cover) categorical variable, from which we derive as many binary variables as the number of categories. Each binary variable takes the value of 1 for basins with that specific dominant land-cover, 0 otherwise. These binary variables are indicated by the "LC-" prefix, such as, e.g.: $LC - Mixed\ Forests$, $LC - Decid. BL\ Forest$ (deciduous broadleaf), $LC - Croplands$, etc.

The same approach was adopted for the dominant geologic class ($geol\_class\_1st$) categorical variable, but none of the resulting binary variables were optimal predictors according to the wrapper method.

4 Results and discussion

4.1 Observed and duality-derived underestimation

Underestimation of the T-year event from AM-, with respect to POT-FFA, is obtained from Eq. (2) considering the quantiles $Q_{AM}(T)$ and $Q_{POT}(T)$ from the GEV and GP distributions, respectively. The Kolmogorov-Smirnov GOF test (Kottegoda & Rosso, 1997) indicates a good fit of these two distributions to the AM and POT series, respectively, for all the considered basins. Also, the estimated FF quantiles display no sensitivity to the PDS threshold-selection method. Figure 3a shows the observed spatial pattern of underestimation $u(T)$ of FFs from AM-FFA, as compared to POT-FFA (Eq. 2).

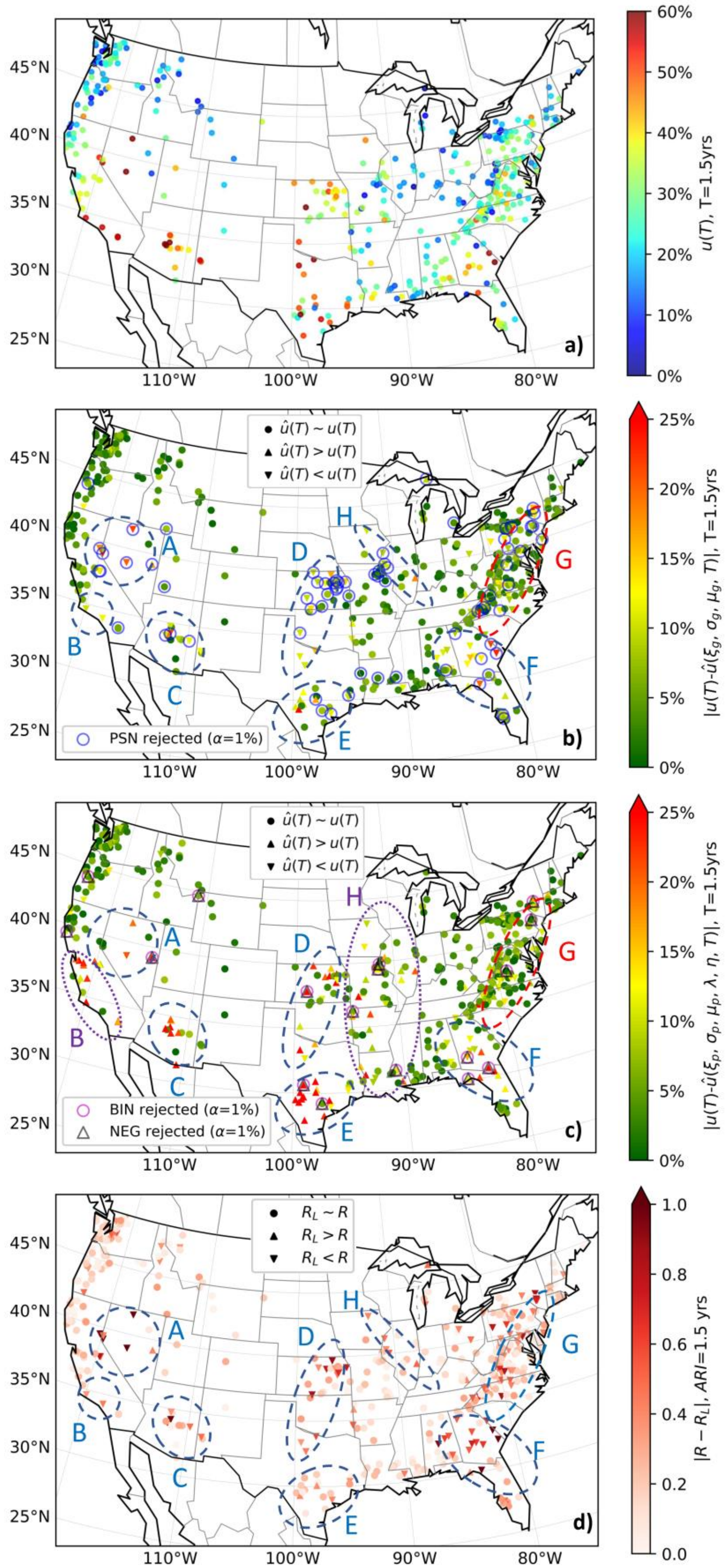

a)
b)
c)
d)
$u(T)$, T=1.5yrs
$\hat{u}(T) \sim u(T)$
$\hat{u}(T) > u(T)$
$\hat{u}(T) < u(T)$
PSN rejected (α=1%)
$|u(T)-\hat{u}(\xi_g, \sigma_g, \mu_g, T)|$, T=1.5yrs
BIN rejected (α=1%)
NEG rejected (α=1%)
$|u(T)-\hat{u}(\xi_p, \sigma_p, \mu_p, \lambda, \eta, T)|$, T=1.5yrs
$R_L \sim R$
$R_L > R$
$R_L < R$
$|R - R_L|$, ARI=1.5 yrs
A
B
C
D
E
F
G
H
110°W
100°W
90°W
80°W
25°N
30°N
35°N
40°N
45°N

Figure 3. a) Spatial pattern of observed FF underestimation $u(T)$ for the 1.5-year event; b) absolute deviation between the observed and duality-derived underestimation, $|u(T) - \hat{u}(T)|$ for the 1.5-year event under the assumption of PSN count model (Eq. 21); catchments where the deviation exceeds 8% are marked with triangles, pointing either up or down depending on the sign of the deviation; c) same as b), but with BIN or NEG count model (Eq. 23); d) Absolute deviation of Langbein-estimated return period R_L from the GEV-derived R of the POT-quantile with ARI of 1.5 years; catchments where the deviation exceeds 0.35 years are marked with triangles, pointing either up or down depending on the sign of the deviation. In b), c), and d), clusters A-H of basins with large deviations are circled.

There is a clear spatial structure in the degree of underestimation of the 1.5-year quantile. Minima of 10-20% are observed in many northern and some north-central states. A band of minima runs from the states of Washington and Oregon, on the west coast, moving south-east down to Colorado and northern New Mexico. There are few observations in the north-central U.S., due to the lack of CAMELS stations with sufficiently long flow records for the Dakotas, Nebraska, and Minnesota. Moving east, minima are also observed in the strip of territory starting from New England, on the east coast, and in states around the Great Lakes, down to Kentucky and West Virginia, on the west side of the Appalachian range.

The Appalachian range is a clear dividing line on the map, as the land east of it, to the Atlantic coast, is characterized by rates of FF underestimation of about 25-30%, with some peaks up to 45%, in contrast with the 10-20%, that prevails on its western side.

Moving down to Florida, most watersheds still present a 25-30% underestimation, but there are also a few basins characterized by 50% or more, resulting in greater heterogeneity overall. A discontinuity along the eastern coast can be observed, north of Florida, with a few basins between Georgia and South Carolina (i.e., the Savannah River) showcasing underestimation rates of 60%, and a few others in the south of Georgia with rates of 45-50%. These are higher values of underestimation than those generally observed for other watersheds close to the Atlantic, all clustered in the same region. Interestingly, this area also represents a singularity from a climatic perspective, with respect to the rest of the eastern coast. More precisely, it is the only region not significantly affected by precipitation from tropical cyclones (Villarini & Smith, 2010).

Basins close to the Gulf coast and in the hinterland above it, up to Tennessee and Arkansas, present a gradual increase in the rate of underestimation, moving from east to west. A similar east-west pattern is also observed when moving from south of the Great Lakes to Kansas, through Illinois and Missouri. This spatial trend of increasing underestimation reaches a local peak in the central and south-central U.S. (i.e., Texas, Oklahoma, and Kansas), where some of the highest underestimation rates in the country are observed, above 40%, although with great heterogeneity due to the occasional presence of catchments with only moderate underestimation, particularly in coastal Texas. High values of underestimation are also observed in the south-west US (New Mexico, Arizona, central and southern California).

Along the west coast of California, moving from south to north, a gradual decrease of the degree of underestimation occurs, from 50-55% to 30%. The underestimation further decreases northward, reaching minima of 10-20% along the coasts of Oregon and Washington.

Figures 3b and 3c show the absolute deviations of duality-derived underestimation $\hat{u}(T)$ (Eqs. 21 and 23) from the observed $u(T)$. Filled circular markers indicate deviations not larger than 8%, showing that very accurate predictions of underestimation can be derived from the duality in those basins. Slight differences are imputable to sample variability in parameter estimation.

About 20% of the basins are categorized as anomalous, i.e., the deviation exceeds 8%; these are marked with a triangle, pointing either up or down to indicate that $\hat{u}(T) > u(T)$ or $\hat{u}(T) < u(T)$, respectively. Figure 3b shows deviations when the PSN count model is adopted (Eq. 21) while Figure 3c depicts results for BIN or NEG count models (Eq. 23). It is interesting to note that anomalous catchments with high deviations are strongly clustered in space. The most striking groups are in the south-central and south-western U.S., as well as in the area that encompasses northern Florida and southern Georgia, with deviations above 10-15% (Figures 3b-c). Another cluster of anomalous deviations, although not as strong, is observed along the Appalachian range. All these clusters are observed independently of the assumption on the distribution of the yearly number of exceedances, since a similar structure is observed for both a PSN (Figure 3b) and BIN or NEG (Figure 3c) count model. However, in the latter case, some clusters are more widespread (e.g., clusters B and E in Figure 3c), and more anomalous watersheds are observed, including in regions that are not affected when considering a PSN count model (e.g., group H in Figure 3c).

A BIN or NEG count model results in overall larger deviations. For instance, for the clusters observed in Kansas and Texas, the duality-derived estimates under BIN (or NEG) count model often predict levels of underestimation 25% larger than the observed (Figure 3c), while under the assumption of a PSN count model absolute deviations typically do not exceed 10% (Figure 3b). There are a few catchments in the Sierra Nevada/Great Basin and at the boundary between Georgia and South Carolina where an opposite behavior is observed, i.e., deviations are larger when a PSN count model is considered. It is worth noting that in these few watersheds the hypothesis that the number of exceedances is PSN distributed can be rejected, based on the Chi-square GOF test (Kottegoda & Rosso, 1997; open circles in Figure 3b), but not the hypotheses of a BIN or NEG count model (open circles and open triangles in Figure 3c, respectively).

However, from a broader perspective, looking at the distribution of the number of exceedances is not decisive to choose the most accurate expression for duality-based predictions of underestimation. The point biserial correlation between watersheds where, e.g., PSN can be rejected and catchments where $\left|\hat{u}\left(\xi_g, \sigma_g, \mu_g, T\right) - u(T)\right|$ exceeds 8% is a modest 0.40. The p-value of the Chi-square test and the magnitude of the deviation are also poorly correlated. Furthermore, there is some overlap between basins where all the three considered count models should be rejected, and interestingly all these watersheds belong to the clusters of basins with large deviations (Figures 3b-c). Thus, the reasons for the poor performance of the duality-based predictions of underestimation must be identified by looking at what other theoretical assumptions are violated in regions with large deviations, beyond what distribution best fits the series of the annual count of exceedances. We propose an explanation in Subsection 4.4.

So far, all comparisons between observed and theoretical underestimation have focused on the 1.5-year event, taken as a representative FF. Figure 4 shows the observed and duality-derived underestimation for other FF quantiles, considering values of T up to 5 years. As expected, the underestimation tends to decrease for larger Ts, indicating that predictions from AM and POT-FFA converge for increasingly less frequent floods. This is consistent with Langbein's equation (Eq. 1), which predicts smaller differences between R and ARI for increasing T (Wang & Holmes, 2020).

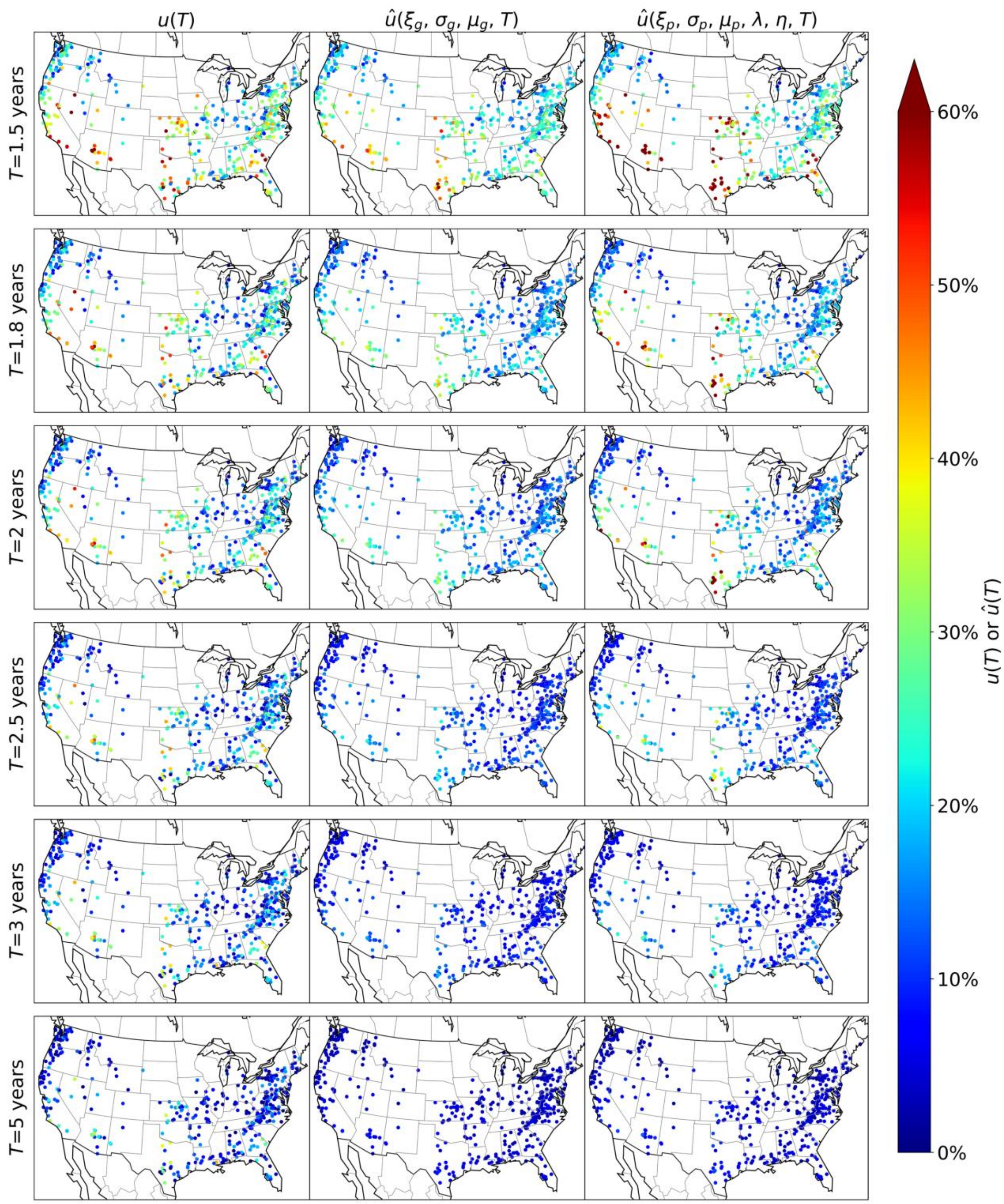


Figure 4. Spatial distribution of the observed underestimation $u(T)$, and duality-derived underestimations $\hat{u}(\xi_g, \sigma_g,\ \mu_g, T)$, and $\hat{u}(\xi_p, \sigma_p,\ \mu_p, \lambda, \eta, T)$, for a range of Ts.

Relatively higher underestimation rates are still observed in south-central and south-western U.S., independently of T, so that the spatial structure of $u(T)$ is preserved. Both duality-derived underestimations, $\hat{u}(\xi_g, \sigma_g, \mu_g, T)$ and $\hat{u}(\xi_p, \sigma_p, \mu_p, \lambda, \eta, T)$, match the observed underestimation, overall.

4.2 **Duality as a tool to adjust AM-based FF estimates**

From a practical standpoint, Eq. (21) can be used to estimate the amount of FF underestimation by AM-FFA without the need to compare the AM-based estimate of the T-year event to the corresponding POT value. Like Langbein's equation (Eq. 1), Eq. (21) is valid under the assumption of a PSN count model for the annual number of exceedances. The duality-derived quantile ratio $\hat{r}\left(\xi_g, \sigma_g, \mu_g, T\right)$ can be regarded as a correction coefficient of the T-year quantile obtained from AM-FFA, function of T and the GEV parameters. The corrected, duality-based T-year quantile $Q^*_{POT}(T)$ can be obtained from the AM-based quantile using Eq. (30), without the need to perform POT-FFA.

$$Q^*_{POT}(T) = \frac{1}{\hat{r}\left(\xi_g, \sigma_g, \mu_g, T\right)} Q_{AM}(\xi_g, \sigma_g, \mu_g, T) \qquad (29)$$

Although the accuracy of the predicted underestimation is lower in some regions (Figure 3b), the errors between observed and predicted underestimation are relatively small, so that overall the bias from applying an "incorrect correction" will still be much smaller than considering the AM-based FF quantile without any correction.

It is preferable to use the GP-GEV duality and resulting $\hat{r}\left(\xi_g, \sigma_g, \mu_g, T\right)$ (Eq. 21) instead of the GP-K5E duality with its correction coefficient $\hat{r}(\xi_p, \sigma_p, \mu_p, \lambda, \eta, T)$ (Eq. 23), because 1) the GP-GEV duality leads to smaller absolute errors than the GP-K5E duality, overall (Figure 3b-c); and 2) the GEV is a commonly used 3-parameter distribution, in contrast to the 5-parameter K5E required for computing $\hat{r}(\xi_p, \sigma_p, \mu_p, \lambda, \eta, T)$.

4.3 **Theoretical and physical drivers of underestimation**

Eq. (21) also affords to study the effects that GEV parameters have on the level of underestimation $\hat{u}(\xi_g, \sigma_g, \mu_g, T)$. Figure 5 maps $\hat{u}(\xi_g, \sigma_g, \mu_g, T)$ in the σ_g-μ_g and σ_g/μ_g-ξ_g planes, for two values of T.

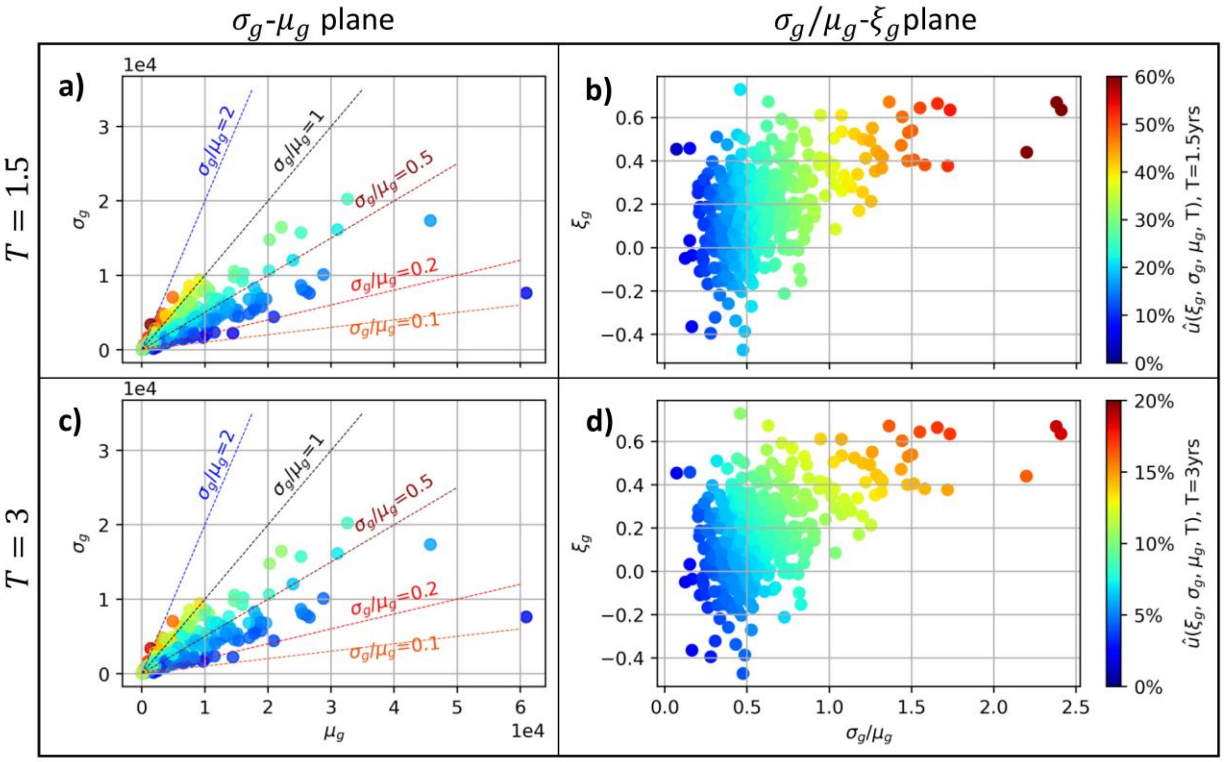
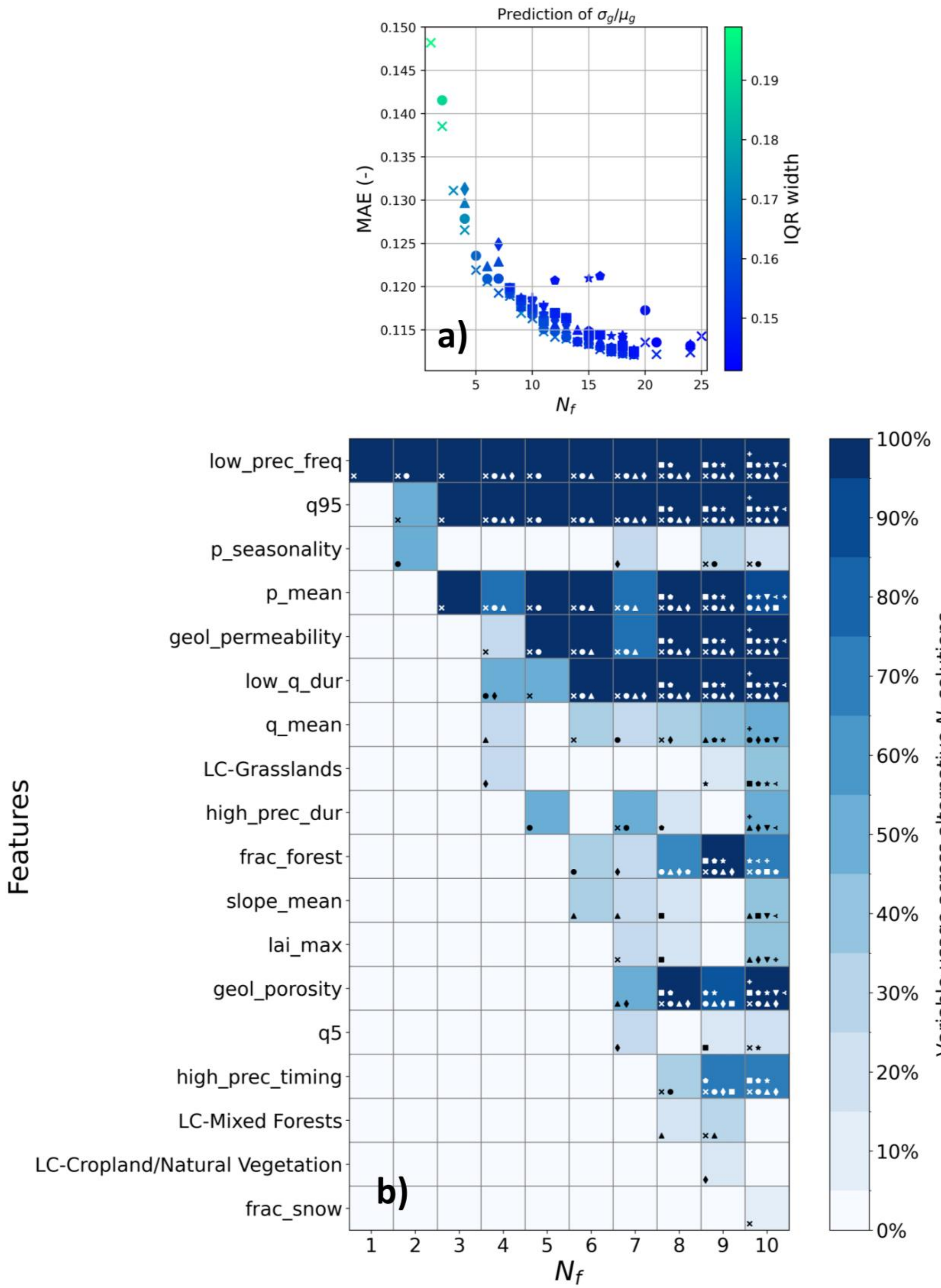


Figure 5. Case study basins mapped in the $\mu_g - \sigma_g$ and $\xi_g - \sigma_g/\mu_g$ planes. The color gradient shows the computed underestimation $\hat{u}(\xi_g, \sigma_g, \mu_g, T)$ for $T = 1.5$ (a and b) and $T = 3$ years (c and d).

For small Ts (e.g., 1.5 years; Figure 5a-b) the scale-location ratio σ_g/μ_g is the main control over $\hat{u}(\xi_g, \sigma_g, \mu_g, T)$, with larger σ_g/μ_g values associated to greater underestimation while the effect of the shape ξ_g is negligible. For larger Ts (e.g., 3 years; Figure 5c-d) though, the shape parameter also contributes to the amount of underestimation, with larger (positive) ξ_g associated to larger $\hat{u}(\xi_g, \sigma_g, \mu_g, T)$, for a given σ_g/μ_g ratio. Hence, the convergence of quantiles estimated from AM- and POT-FFA for increasing Ts is slower at sites with larger shape parameters. For our case study, the largest σ_g/μ_g values are all paired with large ξ_g values, indicating that U.S. basins most affected by FF underestimation tend to experience this issue for a wider range of Ts, as compared to watersheds with modest underestimation (also see Figure 4).

Optimal predictors for σ_g/μ_g show that arid climates are associated to larger σ_g/μ_g ratios (Figures 6-7), and therefore greater FF underestimation.


Prediction of σ_g/μ_g
MAE (-)
0.150
0.145
0.140
0.135
0.130
0.125
0.120
0.115
a)
5
10
15
20
25
N_f
IQR width
0.19
0.18
0.17
0.16
0.15
Features
low_prec_freq
q95
p_seasonality
p_mean
geol_permeability
low_q_dur
q_mean
LC-Grasslands
high_prec_dur
frac_forest
slope_mean
lai_max
geol_porosity
q5
high_prec_timing
LC-Mixed Forests
LC-Cropland/Natural Vegetation
frac_snow
b)
1
2
3
4
5
6
7
8
9
10
N_f
Variable usage across alternative N_f solutions
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%

576

Figure 6. a) Pareto front of optimal trade-off solutions among three competing objectives to minimize: number of predictors N_f, mean absolute error MAE, and width of the interquartile range IQR. Each point corresponds to a learning machine trained to estimate σ_g/μ_g with a set of N_f basin characteristics; b) heatmap with the frequency of usage of variables as optimal predictors in different solutions with same N_f; solutions with N_f up to 10 are considered (see online Supporting information for the full heatmap).

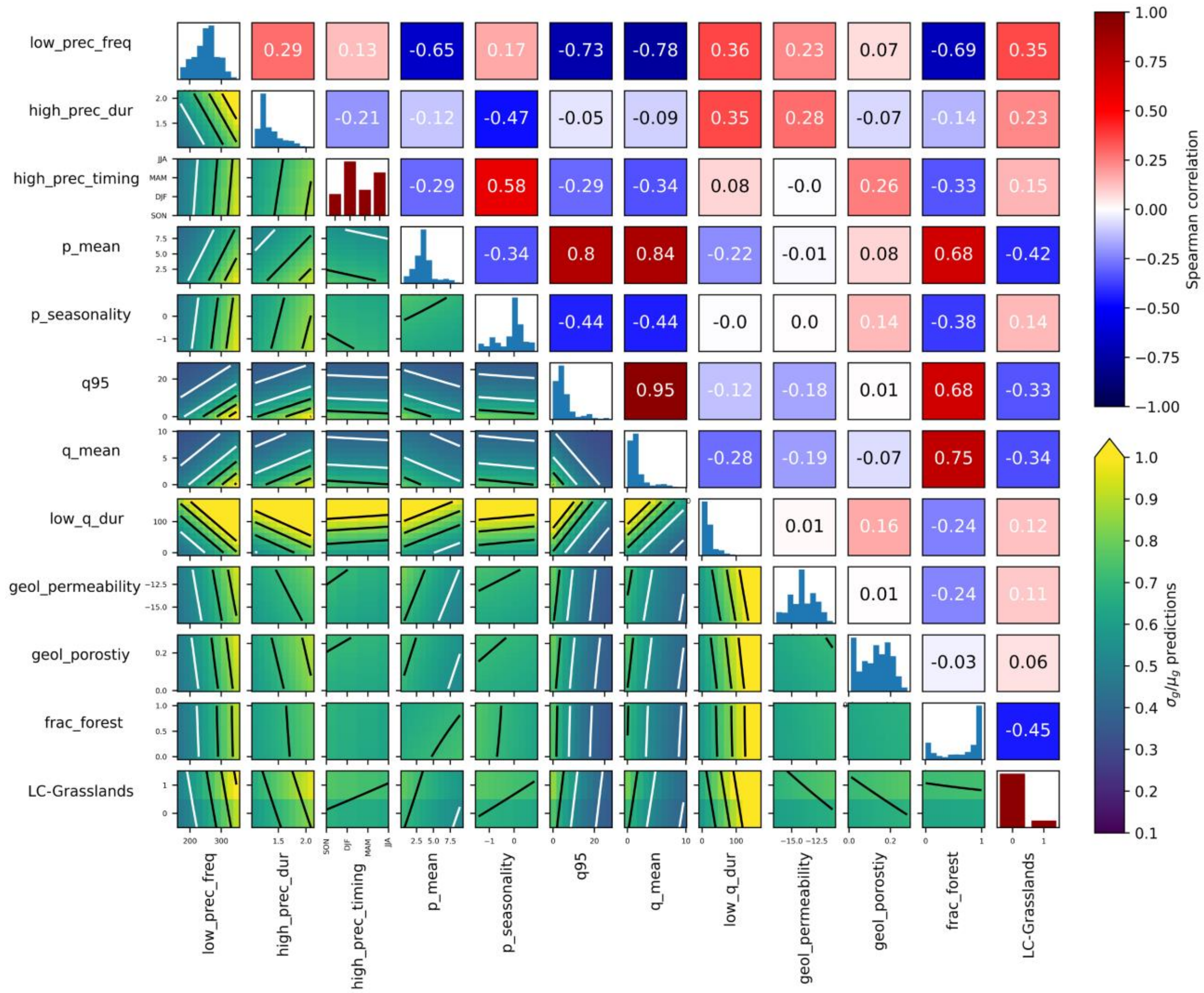


Figure 7. Lower triangular matrix: bivariate partial-dependence plots that show the relationships between key basin attributes and σ_g/μ_g values. Upper triangular matrix: Spearman correlation of key basin attributes. Diagonal: frequency distribution of the key basin attributes; brown histograms are used for categorical and binary variables, while continuous variables are in blue.

The number of dry days per year (low_prec_freq; Addor et al., 2017) has the strongest predictive power on σ_g/μ_g, as it is used in every optimal solution, including that for $N_f = 1$ (Figure 6).

Similar observations have been made for Europe (Lun et al., 2021), where the aridity index was identified as the main control on the coefficient of variation of annual maxima (CV_{AM}). Although the wrapper method never considered the aridity index ($aridity$ in the CAMELS dataset) among the optimal predictors of σ_g/μ_g, our findings are equivalent to those of Lun et al. (2021) for Europe, because of the strong positive correlation between CV_{AM} and σ_g/μ_g (Pearson correlation $\rho_P(CV_{AM}, \sigma_g/\mu_g) = 0.85$ and Spearman correlation $\rho_S(CV_{AM}, \sigma_g/\mu_g) = 0.87$, for our dataset), as well as between low_prec_freq and the aridity index [$\rho_P(aridity, low\_prec\_freq) = 0.74$, $\rho_S(aridity, low\_prec\_freq) = 0.82$].

Optimal solutions shown by the heatmap in Figure 6b, and PDPs in Figure 7 help identify other basin characteristics that have a strong control over σ_g/μ_g and, in turn, on FF underestimation by AM-FFA. The type of climate and the size of the river have the strongest influence. E.g., watersheds with low mean daily precipitation values (p_mean) are characterized by larger values of the σ_g/μ_g ratio. This is enhanced at locations that also experience long periods with low flows each year (high low_q_dur). Low average precipitation is typical of an arid climate [$\rho_S(aridity, p\_mean) = -0.65$], and low values of p_mean may be associated to large numbers of dry days (low_prec_freq). Having long periods with low flows in this kind of climate may be a sign of intermittent, flashy behavior, reflected by larger σ_g/μ_g ratios. This explains the prediction of higher σ_g/μ_g values at locations that display both signs of an arid climate (e.g., large values of low_prec_freq and small values of p_mean) as well as persistent low flows.

Large σ_g/μ_g values are also predicted at basins with long durations of high precipitation events (i.e., with large $high\_prec\_dur$ values) and persistent low flows during the year (large low_q_dur values). These are typical in regions dominated by synoptic-scale weather systems (Addor et al., 2017), where annual precipitation and flow cycles display strong seasonality, with maxima concentrated in winter and minima in summer. Watersheds from these locations are characterized by $high\_prec\_timing$ in the December-February (DJF) period, and negative $p\_seasonality$ values, which both indicate that precipitation events occur predominantly in winter. The range of variability of floods (reflected by σ_g) associated to this kind of climate can be wide, which explains why the learning machine predicts large σ_g/μ_g values when the precipitation cycle displays strong winter seasonality concurrent with large average duration of high precipitation events ($high\_prec\_dur$).

Indicators of river size relative to basin area, such as the mean and the 95-percentile of daily flow per unit area (i.e., q_mean and $q95$, respectively) also represent strong controls over the σ_g/μ_g ratio. This was expected, as rivers with greater flows have larger μ_g and are generally more stable (Dell'Aira et al., 2022), resulting in narrower ranges of variability (therefore, smaller σ_g), and consequentially smaller σ_g/μ_g values.

Basin attributes of secondary importance for predicting σ_g/μ_g include vegetation land-cover (e.g., $LC\_Grassland$; $frac\_forest$) and geomorphic information (e.g., $geol\_permeability$; $geol\_porosity$). The former may be regarded as a proxy for the type of climate, while the latter may affect the hydrologic response of basins. It is worth noting that, ceteris paribus, the learning machine assigns larger σ_g/μ_g values to watersheds with grassland dominant land-cover ($LC_{Grassland} = 1$), more frequent in arid and semi-arid regions (Addor et al., 2017), than catchments with other dominant land-cover types. This suggests that the learning machine is exploiting the relationships between climate and vegetation type.

Results on the optimal predictors for the shape parameter ξ_g show similar dependencies to those for σ_g/μ_g, with large ξ_g values associated to dry regions and small, negative values to humid areas. This is in agreement with previous research (e.g., Metzger et al., 2020; Villarini & Smith, 2013), matching our observations (Figure 4) that basins in arid and semi-arid regions show slower rates of convergence of quantiles estimated from AM- and POT-FFA, for increasing T. A large, positive ξ_g results in a GEV-PDF without upper bound, which may better describe the flashy behavior of rivers in arid catchments, in contrast with the more stable flows in humid regions. The list of basin attributes that affect ξ_g includes variables highly correlated to the aridity index, such as $runoff\_ratio$ [$\rho_S(aridity, runoff\_ratio) = -0.81$] and low_prec_dur [$\rho_S(aridity, low\_prec\_dur) = 0.77$], as well as information on the type of vegetation, which is a proxy for the type of climate. Heatmaps and PDPs of the optimal basin attributes related to the variability in ξ_g do not add any additional insight; therefore, they are not published in this work.

4.4 **Validity of the duality-based quantile ratio and Langbein's equation**

We concluded in Section 4.1 that the validity limits of Eq. (21) are not determined by the violation of the hypothesis of a PSN count model. We speculate here that the clusters of anomalous basins with large $|u(T) - \hat{u}(\xi_g, \sigma_g, \mu_g, T)|$ deviations (for T equal to 1.5 years) can be explained by the occurrence of mixed flood populations. Clusters (indicated with letters A–G in Figure 4b-c) are found in regions where the presence of mixed populations is well-acknowledged. E.g., a large proportion of flood events in the Sierra Nevada (western part of cluster A), coastal California (B), and central Arizona (C) is generated by atmospheric rivers (ARs), resulting in strongly heterogeneous populations (Barth et al., 2017, 2019; Villarini, 2016). In the Sierra, orographically enhanced precipitation in the November-April period and snowfall in winter (with consequent snowmelt floods in April-July) contribute further flood-generating mechanisms (Barth et al., 2017; U.S. Water Resources Council, 1976). In the Great Basin (eastern part of cluster A), snowmelt, frontal storms, and convective precipitation may generate major floods in the springtime, winter, and summer months, respectively (Burkham, 1988). In Arizona and New Mexico (cluster C), floods in the summer period may be caused by a variety of different processes, such as convective events (some of these connected to the North American monsoon activity, depending on the region) and eastern North Pacific tropical cyclones (Barth et al., 2017; Villarini, 2016). Coastal Texas, northern Florida/southern Georgia, and the Appalachian range (clusters E, F, and G) present the lowest AM seasonality within the U.S. (Villarini, 2016), indicating that AM may be observed in a different season each year, in turn suggesting the presence of multiple flood-generating mechanisms. Tropical cyclones and extratropical systems (TCs and ETSs, respectively), as well as organized warm-season convective systems (OWSCS) represent some of the possible drivers in those regions (Villarini, 2016; Villarini & Smith, 2010, 2013; Villarini et al., 2014). Further heterogeneity is introduced by the sensitivity of TC-generated floods to the phases of the North Atlantic and El Niño-Southern oscillations (Villarini et al., 2014), which may introduce variability in the characteristics of the flood population across years. Bulletin 17B (U.S. Interagency Advisory Committee on Water Data, 1982) already recommends separating TC-generated floods from other peaks of the series, for FFA applications. Southern Georgia displays the largest deviations observed for cluster F, which may be due to the presence of ETS-generated floods in early spring (Villarini & Smith, 2010), which apparently represent most AM events in this region, as suggested by the AM seasonality concentrated in the March-April period (Villarini, 2016). This indicates the presence of one dominant AM generating mechanism, related to the occurrence of ETSs, even though the POT series may come from heterogeneous flood populations, as exceedances come from a variety of different generating mechanisms. These conditions may lead to a more severe violation of the assumption of identically distributed events because it implies that one generating mechanism produces peaks that are systematically larger than the events produced by other mechanisms, exacerbating bimodality in the flood population. A similar explanation can be provided for clusters D and H, where the medium-to-strong AM seasonality is concentrated in the May-June period (Villarini, 2016), concurrently with North Atlantic low-level jets (NALLJs; Weaver et al., 2012). This indicates that NALLJs represent the dominant AM generating mechanism for that region, therefore introducing maxima that come from a notably different distribution as compared to not-NALLJs induced floods.

Results from optimal feature analysis support our hypothesis that mixed populations affect the validity of Eq. (21), as the key basin attributes for watershed classification are all related to the seasonality of flood and precipitation, as well as the type of vegetation, which in turn can be related to the type of climate. Clusters of anomalous basins are identified considering a threshold $\left|u(T) - \hat{u}\left(\xi_g, \sigma_g, \mu_g, T\right)\right| > 8\%$.

Figures 8-9 show that flow elasticity (i.e., sensitivity) to changes in precipitation ($stream\_elas$), the runoff to precipitation ratio ($runoff\_ratio$), as well as measures of flow and precipitation seasonality ($high\_prec\_timing$, low_prec_timing, $|\bar{\mathbf{z}}_{\boldsymbol{FF}}|$, $|\bar{\mathbf{z}}_{\boldsymbol{AM}}|$, and $\bar{\theta}_{AM}$) all play an important role in affecting the probability of having a basin with large $|u(T) - \hat{u}(\xi_g, \sigma_g, \mu_g, T)|$ deviations.

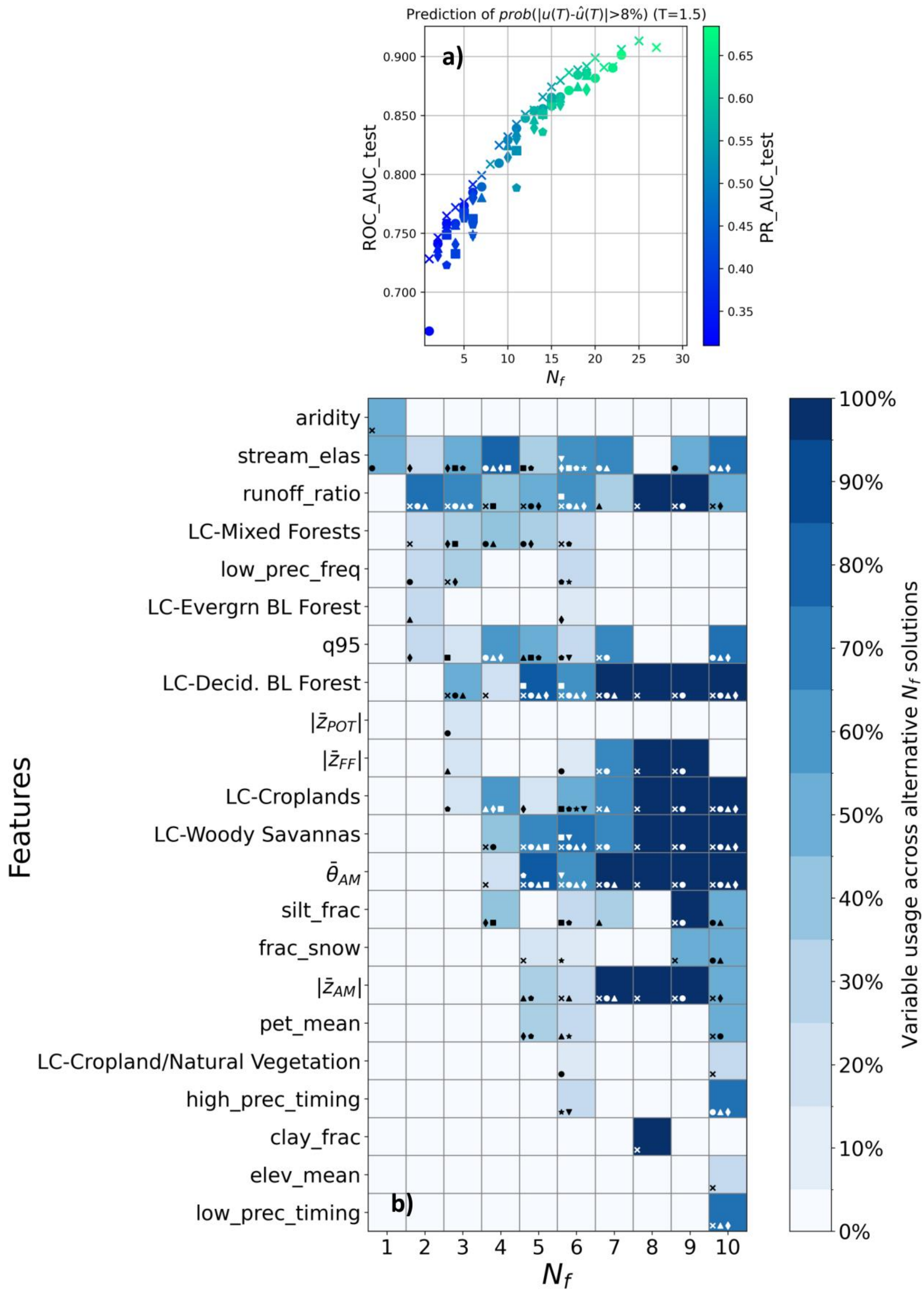

Prediction of prob(|u(T)-û(T)|>8%) (T=1.5)
a)
ROC_AUC_test
PR_AUC_test
N_f
b)
Features
aridity
stream_elas
runoff_ratio
LC-Mixed Forests
low_prec_freq
LC-Evergrn BL Forest
q95
LC-Decid. BL Forest
$|\bar{z}_{POT}|$
$|\bar{z}_{FF}|$
LC-Croplands
LC-Woody Savannas
$\bar{\theta}_{AM}$
silt_frac
frac_snow
$|\bar{z}_{AM}|$
pet_mean
LC-Cropland/Natural Vegetation
high_prec_timing
clay_frac
elev_mean
low_prec_timing
Variable usage across alternative N_f solutions
N_f

Figure 8. a) Pareto front of optimal trade-off solutions among three competing objectives to minimize: number of predictors N_f, ROC area under the curve (ROC_AUC), and Precision-Recall area under the curve (PR_AUC). Each point corresponds to a learning machine trained to classify basins with large deviations $\left|u(T) - \hat{u}\left(\xi_g, \sigma_g, \mu_g, T\right)\right|$, hence, where the duality-derived Eq. (21) is not valid, using a set of N_f basin characteristics; b) heatmap with the frequency of usage of variables as optimal predictors in different solutions with same N_f; solutions with N_f up to 10 are considered (see online Supporting information for the full heatmap).

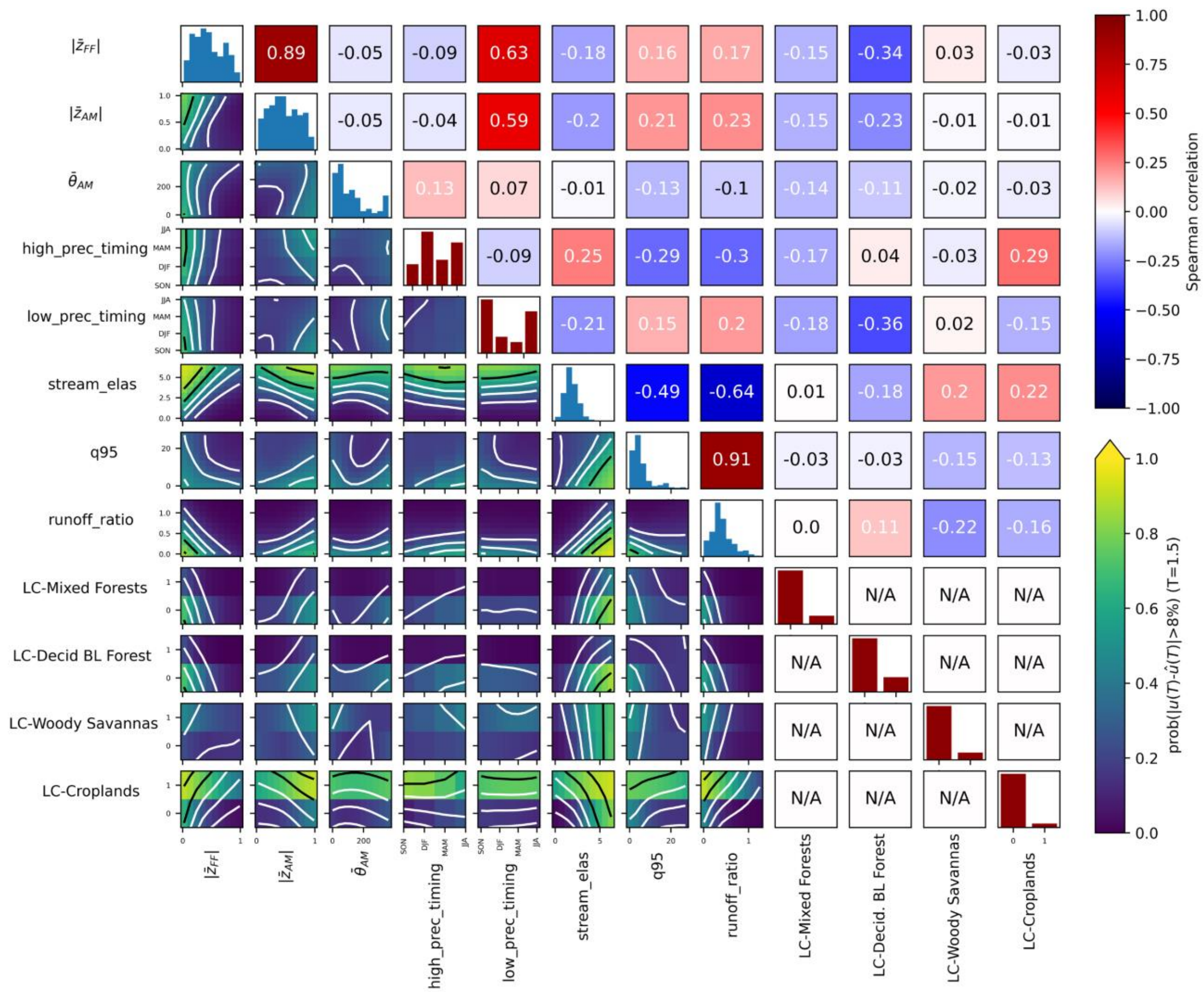


Figure 9. Lower triangular matrix: PDPs that show the relationships between key basin attributes and the probability that the deviation between observed and duality-derived underestimation is >8%. This is considered as an empirical indicator that the GP-GEV duality (Eq. 21) is not valid at a given watershed. Upper triangular matrix: Spearman correlation of key basin attributes; correlation of binary variables is not computed because they are different categories of the same categorical variable. Diagonal: frequency distribution of the key basin attributes; brown histograms are used for categorical and binary variables, while blue for continuous variables.

A low FF seasonality (i.e., $|\bar{\boldsymbol{z}}_{\boldsymbol{FF}}|$ close to 0) is generally associated with a high probability of having an anomalous watershed. Strong AM seasonality ($|\bar{\boldsymbol{z}}_{\boldsymbol{AM}}|$ close to 1) concurrent with low FF seasonality leads to higher chances of an anomaly, while if both AM and FFs present strong seasonality, the duality equation (Eq. 21) should give a good estimate of the underestimation. A strong AM seasonality coupled with a strong FF seasonality indicates that peaks tend to occur all in the same period of the year, resulting in a homogenous flood population. In contrast, a strong AM seasonality paired to a low FF seasonality suggests that there is one generating mechanism that often results in the largest annual event to occur in the same season, across multiple years, but other types of floods are also present in the peak series.

Another sign of large $\left|u(T) - \hat{u}\left(\xi_g, \sigma_g, \mu_g, T\right)\right|$ deviations is a high discharge-precipitation elasticity ($stream\_elas$). This is because a strong flood sensitivity to precipitation may result in greater changes of the characteristics of flood distributions across years with different amounts and time distributions of rainfall. The highest values of flow elasticity in the U.S. are observed in arid and semiarid regions (Sankarasubramanian et al., 2001).

Regarding the effect of dominant land-cover type, mixed or deciduous-broadleaf-forest catchments, common along the Appalachian range, are less likely to be classified as anomalous than savanna basins, more typical in parts of Texas and California. This reflects the fact that there are both regular and anomalous watersheds in the eastern U.S., characterized either by mixed or deciduous-forest land-cover, while basins in savanna regions present large $\left|u(T) - \hat{u}\left(\xi_g, \sigma_g, \mu_g, T\right)\right|$ deviations more systematically. Cropland LC-dominated watersheds, widespread in central U.S. (including in dry climate regions such as Texas and Kansas), get a high probability of displaying large deviations if the $runoff\_ratio$ is low (typical of dry climates) or $stream\_elas$ is high. This may be interpreted as a way of identifying the types of basins observed in Kansas and northern Texas (clusters E and part of D, respectively) by cross-checking multiple characteristics typical of those regions.

To conclude, the highly non-linear relationships shown in Figure 9 between key basin attributes and the probability of large deviation $\left|u(T) - \hat{u}\left(\xi_g, \sigma_g, \mu_g, T\right)\right|$ all have a quite straightforward interpretation if the hypothesis of the influence of mixed populations is deemed correct. Or, at least, they are not in conflict with each other. This may be regarded as an empirical, a posteriori proof in support of this hypothesis.

Langbein's equation is valid under the same two assumptions required by Eq. (21): i.i.d. peaks and that the number of exceedances is PSN distributed. Considering thresholds of 8% for the $\left|u(T) - \hat{u}\left(\xi_g, \sigma_g, \mu_g, T\right)\right|$ deviations and 0.35 years for the errors in the Langbein-estimated return period R_L (Figure 3d), the point biserial correlation between watersheds where Eq. (21) and Eq. (1) produce large errors is 0.78. This high correlation suggests that regions where the two theoretical equations are not perfectly valid are overall the same (also compare Figures 4b-d); in both cases, the most likely explanation is the occurrence of mixed populations, violating the assumption of identically distributed events. In practice, for design purposes or any other case where one needs to know the flood magnitude for a given frequency, Eq. (21) should be used together with AM-FFA, as it allows for directly correcting the AM-based flood estimate. In contrast, Langbein's equation is more useful in a verification framework, i.e., when one is interested in assessing the actual frequency of the design flood computed from AM-FFA.

5 Conclusions

Frequent flood underestimation by AM-FFA is a well-known phenomenon in engineering practice but is poorly understood from a theoretical standpoint. Probably this is one of the reasons why the issue has been systematically overlooked, with many practitioners across many disciplines using AM to predict FFs such as the 2-year quantile, or even more frequent floods.

This work considers a large sample of watersheds with minimum human impact to show that the level of FF underestimation can vary widely depending on the GEV parameters, and in turn the type of climate, the size of the river, and other basin characteristics that affect the distributional characteristics of AM. The scale-location ratio is the main control over the amount of underestimation, for a given average interarrival time T, while the shape parameter determines how quickly AM- and POT-estimated quantiles converge, for increasing T.

We propose a practical relationship, derived from the theoretical duality between the GEV and GP distributions, that can be used to correct AM-based estimates of FFs, considering that their actual frequency is better reflected by the ARI predicted by POT-FFA. However, we were able to characterize some regions in the U.S. where this useful tool underperforms, as does the well-known Langbein's equation, misestimating the gap between R (from AM-) and ARI (from POT-FFA). We conclude that the poor performance of both approaches is imputable to the occurrence of mixed flood populations. In these regions, the negative bias introduced by using AM-FFA can reach up to 60% for a T of 1.5 years. Such levels of underestimation of frequent flood magnitude are of practical concern for a range of river science and engineering fields, so that the use of POT should be mandatory in these cases.

Acknowledgments

This research was supported by the Tennessee Department of Transportation (TDOT) through the project "Updating equations for peak flow estimation in urban creeks and streams of Tennessee", started on September 18, 2020 (State Project Number: RES2020-23). This research was also supported through high performance computing (HPC) resources at The University of Memphis, Memphis, TN.

Data availability statement

Flow series used to perform the analyses in this work are available from the USGS website at the link *http://dx.doi.org/10.5066/F7P55KJN* (U.S. Geological Survey 2022). In order to avoid provisional data, flow series up to 2020 water year have been considered. Basin characteristics come from the CAMELS dataset (Addor et al. 2017), available at the link *https://ral.ucar.edu/solutions/products/camels*. Our results supporting the findings of this work are accessible through the link: *https://livememphis-my.sharepoint.com/:f:/g/personal/fdllaira_memphis_edu/Eo02kp_GJvpGlDF9ulhmtaQBjlSTmfAahRoLjFtZz-fmaQ?e=zNQWEI* (this is a temporary link for the peer review process; we will upload the results in a public repository before publication; please use the password "4peer-review_only"). Neural network training has been performed by the open-source Python library Tensorflow (Abadi et al., 2016). Figures in this work have been produced by means of the open-source Python library Matplotlib (Hunter, 2007) and its Basemap toolkit (*https://basemaptutorial.readthedocs.io/en/latest/external_resources.html*).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Adamowski, K. (2000). Regional analysis of annual maximum and partial duration flood data by nonparametric and L-moment methods. *Journal of Hydrology*, 229(3-4), 219-231.

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: catchment attributes and meteorology for large sample studies. *Hydrology and Earth System Sciences*, 21 (10), 5293 5313.

Aggarwal, C. C. (2018). *Neural networks and deep learning*. Springer.

Asquith, W.H., England, J.F., and Herrmann, G.R., 2021, MGBT—Multiple Grubbs–Beck low-outlier test: U.S. Geological Survey software release, R package, Reston, Va., *https://doi.org/10.5066/P9CW9EF0*.

Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, 35, 27-42.

Babatunde, O. H., Armstrong, L., Leng, J., & Diepeveen, D. (2014). A genetic algorithm-based feature selection. *International Journal of Electronics Communication and Computer Engineering*, 5(4), 899-905.

Ball, J., Babister, M., Nathan, R., Weeks,W., Weinmann, E., Retallick, M., Testoni, I. (2019) *Australian Rainfall and Runoff: A Guide to Flood Estimation*; Commonwealth of Australia (Geoscience Australia).

Barth, N. A., G. Villarini, M. A. Nayak, and K. White (2017), Mixed populations and annual flood frequency estimates in the western United States: The role of atmospheric rivers, *Water Resources Research*, 53, 257–269.

Barth, N. A., Villarini, G., & White, K. (2019). Accounting for mixed populations in flood frequency analysis: Bulletin 17C perspective. *Journal of Hydrologic Engineering*, 24(3), 04019002.

Bayazit, M. (2015). Nonstationarity of hydrological records and recent trends in trend analysis: a state-of-the-art review. *Environmental Processes*, 2, 527-542.

Bendix, J., & Hupp, C. R. (2000). Hydrological and geomorphological impacts on riparian plant communities. *Hydrological Processes*, 14(16-17), 2977-2990.

Benedict, S. T., & Knight, T. P. (2021). Benefits of compiling and analyzing hydraulic-design data for bridges. *Transportation Research Record*, 2675(11), 1073-1081.

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(5), 959-977.

Bhunya, P. K., Berndtsson, R., Jain, S. K., & Kumar, R. (2013). Flood analysis using negative binomial and Generalized Pareto models in partial duration series (PDS). *Journal of Hydrology*, 497, 121-132.

Burkham, D. E. (1988). *Methods for delineating flood-prone areas in the Great Basin of Nevada and adjacent states*. U.S. Geological Survey Water-Supply Paper 2316.

Burn, D. H. (1997). Catchment similarity for regional flood frequency analysis using seasonality measures. *Journal of Hydrology*, 202(1-4), 212-230.

Caires, S. (2009). *A Comparative Simulation Study of the Annual Maxima and the Peaks-Over-Threshold Methods*. Deltares Report 1200264-002 for Rijkswaterstaat, Waterdienst.

Castellarin, A., Burn, D. H., & Brath, A. (2001). Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. *Journal of Hydrology*, 241(3-4), 270-285.

Cipollini, S., Fiori, A., & Volpi, E. (2021). Structure-based framework for the design and risk assessment of hydraulic structures, with application to offline flood detention basins. *Journal of Hydrology*, 600, 126527.

Cohn, T. A., England, J. F., Berenbrock, C. E., Mason, R. R., Stedinger, J. R., & Lamontagne, J. R. (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. *Water Resources Research*, 49(8), 5047-5058.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.

Cunnane, C. (1973). A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of Hydrology*, 18(3-4), 257-271.

Cunnane, C. 1979. A note on the Poisson assumption in partial duration series models. *Water Resources Research* 15:489-494.

Dalrymple, T. (1960). *Flood-Frequency Analyses, Manual of Hydrology: Part 3. Flood-Flow Techniques*. U.S. Government Printing Office, Washington.

Death, R. G., Fuller, I. C., & Macklin, M. G. (2015). Resetting the river template: The potential for climate-related extreme floods to transform river geomorphology and ecology. *Freshwater Biology*, 60(12), 2477-2496.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2): 182–197.

Dell'Aira, F., Cancelliere, A., Creaco, E., & Pezzinga, G. (2021). Novel comprehensive approach for phasing design and rehabilitation of water distribution networks. *Journal of Water Resources Planning and Management*, 147(3), 04021001.

Dell'Aira, F., Chy, T. J., Goebel, T. H., & Meier, C. I. (2022). Inferring hydrological properties of the rainfall-runoff conversion process through artificial neural network modeling. In *Proc., World Environmental and Water Resources Congress 2022 (ASCE)* (pp. 1264-1278).

Eastoe, E. F., & Tawn, J. A. (2010). Statistical models for overdispersion in the frequency of peaks over threshold data for a flow series. *Water Resources Research*, 46(2).

England, J. F., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas, W. O., Veilleux, A. G., Kiang, J. E. & Mason, R. R. (2019). *Guidelines for Determining Flood Flow Frequency – Bulletin 17C (Ver. 1.1, May 2019)*. US Geological Survey Techniques and Methods, Book 4, Chap. B5. U.S. Geological Survey.

Fatima, N. (2020). Enhancing performance of a deep neural network: A comparative analysis of optimization algorithms. ADCAIJ: *Advances in Distributed Computing and Artificial Intelligence Journal*, 9(2), 79-90.

Feaster, T. D., Gotvald, A. J., & Weaver, J. C. (2014). Methods for Estimating the Magnitude and Frequency of Floods for Urban and Small, Rural Streams in Georgia, South Carolina, and North Carolina, 2011. *Report No. 2014-5030. United States. Dept. of the Interior (https://rosap.ntl.bts.gov/view/dot/27190).*

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer Cham.

Harvey A.M., Hitchcock D.H. & Hughes D.J. (1979) Event frequency and morphological adjustment of fluvial systems in upland Britain. In: *Adjustments of the Fluvial System* (Eds D.D. Rhodes & G.P. Williams), pp. 139–167. Kendall Hunt, Dubuque, IA.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT press.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.

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(1), 105-124.

Hosking, J. R. M. (1994). The four-parameter kappa distribution. *IBM Journal of Research and Development*, 38(3), 251-258.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional Frequency Analysis. An Approach Based on L-moments*. Cambridge University Press.

Hosking, J. R. M., Wallis, J. R., & Wood, E. F. (1985). An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. *Hydrological Sciences Journal*, 30(1), 85-109.

Huang, J., Cai, Y., & Xu, X. (2007). A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recognition Letters*, 28(13), 1825-1844.

Huang, W. C., Yu, H. W., & Weng, M. C. (2015). Levee reliability analyses for various flood return periods–a case study in southern Taiwan. *Natural Hazards and Earth System Sciences*, 15(4), 919-930.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

Johnson, B. L., Richardson, W. B., & Naimo, T. J. (1995). Past, present, and future concepts in large river ecology. *BioScience*, 45(3), 134-141.

Karim, F., Hasan, M., & Marvanek, S. (2017). Evaluating annual maximum and partial duration series for estimating frequency of small magnitude floods. *Water*, 9(7), 481.

Keast, D., & Ellison, J. (2013). Magnitude frequency analysis of small floods using the annual and partial series. *Water*, 5(4), 1816-1829.

Kennedy, J. R., & Paretti, N. V. (2014). Evaluation of the Magnitude and Frequency of Floods in Urban Watersheds in Phoenix and Tucson, Arizona. *U.S. Geological Survey Scientific Investigations Report 2014-5121 (https://doi.org/10.3133/sir20145121)*.

Khaddor, I., Achab, M., Soumali, M. R., Benjbara, A., & Alaoui, A. H. (2021). The impact of the construction of a dam on flood management. *Civil Engineering Journal*, 7(2), 343-356.

Kiran, K. G., & Srinivas, V. V. (2021). A Mahalanobis distance-based automatic threshold selection method for Peaks Over Threshold model. *Water Resources Research*, 57, e2020WR027534.

Klemeš, V. (2000). Tall tales about tails of hydrological distributions. I. *Journal of Hydrologic Engineering*, 5(3), 227-231.

Kottegoda, N. T., & Rosso, R. (1997). Statistics, Probability, and Reliability for Civil and Environmental Engineers (Vol. 735). *New York: McGraw-Hill*.

Laio, F., Ganora, D., Claps, P., & Galeati, G. (2011). Spatially smooth regional estimation of the flood frequency curve (with uncertainty). *Journal of Hydrology*, 408(1-2), 67-77.

Lamontagne, J. R., Stedinger, J. R., Yu, X., Whealton, C. A., & Xu, Z. (2016). Robust flood frequency analysis: Performance of EMA with multiple Grubbs-Beck outlier tests. *Water Resources Research*, 52(4), 3068-3084.

Landro, N., Gallo, I., & La Grassa, R. (2020). Mixing Adam and SGD: a combined optimization method. arXiv preprint arXiv:2011.08042.

Langbein, W.B., 1949. Annual floods and the partial-duration flood series. *Transactions, American Geophysical Union*, 30 (6), 879–881. (*doi:10.1029/TR030i006p00879*).

Law, G. S. and Tasker, G. D. (2003). Flood-Frequency Prediction Methods for Unregulated Streams of Tennessee, 2000. *U.S. Geological Survey Water-Resources Investigations Report 03-4176* (*https://pubs.usgs.gov/wri/wri034176/PDF/wrir034176.pdf*).

Lendering, K. T., Sebastian, A., Jonkman, S. N., & Kok, M. (2019). Framework for assessing the performance of flood adaptation innovations using a risk-based approach. *Journal of Flood Risk Management*, 12(S2), e12485.

Lun, D., Viglione, A., Bertola, M., Komma, J., Parajka, J., Valent, P., & Blöschl, G. (2021). Characteristics and process controls of statistical flood moments in Europe–a data-based analysis. *Hydrology and Earth System Sciences*, 25(10), 5535-5560.

Madsen, H., Pearson, C. P., & Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 2. Regional modeling. *Water Resources Research*, 33(4), 759-769.

Markus, M., & Demissie, M. (2006). Predictability of annual sediment loads based on flood events. *Journal of Hydrologic Engineering*, 11(4), 354-361.

Meier, C.I. (2008). Cottonwood Establishment in a Gravel-Bed River. Dissertation, the University of Montana.

Meier, C. I., Moraga, J. S., Pranzini, G., & Molnar, P. (2016). Describing the interannual variability of precipitation with the derived distribution approach: effects of record length and resolution. *Hydrology and Earth System Sciences*, 20(10), 4177-4190.

Metzger, A., Marra, F., Smith, J. A., & Morin, E. (2020). Flood frequency estimation and uncertainty in arid/semi-arid regions. *Journal of Hydrology*, 590, 125254.

Morrison, A., Westbrook, C. J., & Noble, B. F. (2018). A review of the flood risk management governance and resilience literature. *Journal of Flood Risk Management*, 11(3), 291-304.

Norheim, S. (2018). Flood frequency analysis: Comparing annual maximum series and peak over threshold: A case study for Norway (Master's thesis, Høgskulen på Vestlandet).

Önöz, B., & Bayazit, M. (2001). Effect of the occurrence process of the peaks over threshold on the flood estimates. *Journal of Hydrology*, 244(1-2), 86-96.

Ouarda, T. B., Cunderlik, J. M., St-Hilaire, A., Barbet, M., Bruneau, P., & Bobée, B. (2006). Data-based comparison of seasonality-based regional flood frequency methods. *Journal of Hydrology*, 330(1-2), 329-339.

Page, K. J., & McElroy, L. (1981). Comparison of annual and partial duration series floods on the Murrumbidgee River. *JAWRA Journal of the American Water Resources Association*, 17(2), 286-289.

Pan, X., Rahman, A., Haddad, K., & Ouarda, T. B. (2022). Peaks-over-threshold model in flood frequency analysis: a scoping review. *Stochastic Environmental Research and Risk Assessment*, 1-17.

Plavšić, J., Mihailović, V., & Blagojević, B. (2014). Assessment of methods for outlier detection and treatment in flood frequency analysis. In Proceedings of the Mediterranean Meeting on Monitoring, modelling and early warning of extreme events triggered by heavy rainfalls. PON 01_01503-MED-FRIEND project University of Calabria, Cosenza (Italy), June 26th-28th, 2014 (pp. 181-192).

Ponce, V. M. (1989). *Engineering Hydrology: Principles and Practices*. Prentice Hall Englewood Cliffs, NJ.

Prosdocimi, I. (2018). German tanks and historical records: the estimation of the time coverage of ungauged extreme events. Stochastic environmental research and risk assessment, 32(3), 607-622.

Prosdocimi, I., & Kjeldsen, T. R. (2022). Exploring links between annual Maxima and Peaks over Threshold: A Study for the UK Records. Abstract (H42E-1323) presented at 2022 AGU Fall Meeting, 12-16 Dec.

Prosdocimi, I., Kjeldsen, T. R., & Svensson, C. (2014). Non-stationarity in annual and seasonal series of peak flow and precipitation in the UK. *Natural Hazards and Earth System Sciences*, 14(5), 1125-1144.

Rasekh, A., Afshar, A., & Afshar, M. H. (2010). Risk-cost optimization of hydraulic structures: methodology and case study. *Water Resources Management*, 24(11), 2833-2851.

Robson, A. J . and Reed, D. W. (1999) Flood Estimation Handbook: Volume 3. *Statistical Procedures for Flood Frequency Estimation*. Centre for Ecology & Hydrology.

Sankarasubramanian, A., Vogel, R. M., & Limbrunner, J. F. (2001). Climate elasticity of streamflow in the United States. *Water Resources Research*, 37(6), 1771-1781.

Sayers, P., Yuanyuan, L., Galloway, G., Penning-Rowsell, E., Shen, F., Wen, K., & Le Quesne, T. (2013). *Flood Risk Management: A strategic approach.* Paris, UNESCO.

Scussolini, P., Aerts, J. C., Jongman, B., Bouwer, L. M., Winsemius, H. C., de Moel, H., & Ward, P. J. (2016). FLOPROS: an evolving global database of flood protection standards. *Natural Hazards and Earth System Sciences*, 16(5), 1049-1061.

Shah, M. A. R., Rahman, A., & Chowdhury, S. H. (2018). Challenges for achieving sustainable flood risk management. *Journal of Flood Risk Management*, 11, S352-S358.

Simpson, A. R., Dandy, G. C., & Murphy, L. J. (1994). Genetic algorithms compared to other techniques for pipe optimization. *Journal of Water Resources Planning and Management*, 120(4), 423-443.

Smith, A., Sampson, C., & Bates, P. (2015). Regional flood frequency analysis at the global scale. *Water Resources Research*, 51(1), 539-553.

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(4), 2833-2849.

Southard, R.E., 2010, Estimating the Magnitude and Frequency of Floods in Urban Basins in Missouri. *U.S. Geological Survey Scientific Investigations Report 2010-5073 (https://doi.org/10.3133/sir20105073).*

Stedinger, J. R., & Lu, L. H. (1995). Appraisal of regional and index flood quantile estimators. *Stochastic Hydrology and Hydraulics*, 9(1), 49-75.

Sun, S. A., Djordjević, S., & Khu, S. T. (2011). A general framework for flood risk-based storm sewer network design. *Urban Water Journal*, 8(1), 13-27.

Taormina, R., & Chau, K. W. (2015). Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. *Journal of Hydrology*, 529, 1617-1632.

Tavares, L. V., & Da Silva, J. E. (1983). Partial duration series method revisited. *Journal of Hydrology*, 64(1-4), 1-14.

Tung, Y. K. (2005). Flood defense systems design by risk-based approaches. *Water International*, 30(1), 50-57.

U.S. Geological Survey (2022). National Water Information System data available on the World Wide Web (USGS Water Data for the Nation), accessed July 11, 2022, at URL *http://dx.doi.org/10.5066/F7P55KJN*.

U.S. Interagency Advisory Committee on Water Data, 1982, Guidelines for Determining Flood Flow Frequency, Bulletin No. 17B: *U.S. Office of Water Data Coordination, Reston, Virginia.*

U.S. Water Resources Council, 1976, Guidelines for Determining Flood Flow Frequency, Bulletin No. 17: *U.S. Water Resources Council, Subcommittee on Hydrology, Washington, D.C.*

Villarini, G. (2016). On the seasonality of flooding across the continental United States. *Advances in Water Resources*, 87, 80-91.

Villarini, G., Goska, R., Smith, J. A., & Vecchi, G. A. (2014). North Atlantic tropical cyclones and US flooding. *Bulletin of the American Meteorological Society*, 95(9), 1381-1388.

Villarini, G., & Smith, J. A. (2010). Flood peak distributions for the eastern United States. *Water Resources Research*, 46(6).

Villarini, G., & Smith, J. A. (2013). Flooding in Texas: Examination of temporal changes and impacts of tropical cyclones. *JAWRA Journal of the American Water Resources Association*, 49(4), 825-837.

Virginia Department of Transportation (2021). *Drainage Manual*. (https://www.virginiadot.org/business/locdes/hydra-drainage-manual.asp)

Vogel, R. M., & Castellarin, A. (2017). Risk, reliability, and return periods and hydrologic design. In: *Handbook of Applied Hydrology*; Singh, VP, Ed.; McGraw-Hill Book Company: New York, NY, USA.

Wang, C. H., & Holmes, J. D. (2020). Exceedance rate, exceedance probability, and the duality of GEV and GPD for extreme hazard analysis. *Natural Hazards*, 102, 1305-1321.

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2020). A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 1-26.

Weaver, S. J., Baxter, S., & Kumar, A. (2012). Climatic role of North American low-level jets on US regional tornado activity. *Journal of Climate*, 25(19), 6666-6683.

Wohl, E., Lane, S. N., & Wilcox, A. C. (2015). The science and practice of river restoration. *Water Resources Research*, 51(8), 5974-5997.

Wolman, M. G., & Miller, J. P. (1960). Magnitude and frequency of forces in geomorphic processes. *The Journal of Geology*, 68(1), 54-74.

Wyżga, B. (1995). Evaluating the occurrence of low magnitude floods: a study of the reliability of the annual maximum series method. *Geografiska Annaler*. Series A. Physical geography, 23-33.

Zaman, M. A., Rahman, A., & Haddad, K. (2012). Regional flood frequency analysis in arid regions: A case study for Australia. *Journal of Hydrology*, 475, 74-83.

Zrinji, Z., & Burn, D. H. (1994). Flood frequency analysis for ungauged sites using a region of influence approach. *Journal of Hydrology*, 153(1-4), 1-21.