AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR028631

Special Section:
Advancing process representation in hydrologic models: Integrating new concepts, knowledge, and data

Improvements to Flood Frequency Analysis on Alluvial Rivers Using Paleoflood Data

Joeri B. Reinders[1] and Samuel E. Muñoz[1,2]

[1]Department of Marine and Environmental Sciences, Marine Science Center, Northeastern University, Nahant, MA, USA, [2]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA

Key Points:

- We run a series of computational experiments to evaluate the influence of paleoflood data on flood frequency analysis for alluvial rivers
- Incorporating more than one paleoflood event into flood frequency analyses improves extreme flood probability estimates
- Large uncertainties in paleoflood discharge estimates can reduce the accuracy of flood frequency analyses

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. Reinders,
reinders.j@northeastern.edu

Citation:

Reinders, J. B., & Muñoz, S. E. (2021). Improvements to flood frequency analysis on alluvial rivers using paleoflood data. *Water Resources Research*, *57*, e2020WR028631. https://doi.org/10.1029/2020WR028631

Received 20 AUG 2020
Accepted 29 MAR 2021

Abstract Hydrologists and engineers routinely use flood frequency analyses to compute flood probabilities for mitigation, infrastructure planning, and emergency management. Conventional flood frequency analyses—in which annual discharge maxima from a stream gage are fit to a statistical probability distribution—often encounter large uncertainties when estimating extreme flood levels. Most gage records span relatively short periods of time (<100 years), and thus the most extreme and infrequently occurring flood events tend to be poorly represented in instrumental data sets. Here, we demonstrate how a new generation of paleoflood records derived from floodplain sediments can be used to improve the accuracy and precision of extreme flood probability estimates along alluvial rivers. We use a series of simulation experiments to show that incorporating large numbers of paleoflood events in flood frequency analyses can significantly reduce the uncertainty of extreme flood estimates when the paleoflood data are sufficiently accurate and precise. Our results illustrate that robust paleoflood records can improve the shape parameter of flood frequency distributions, which determine the thickness of distribution tails, when as many as 50 paleoflood events are incorporated. We conclude by demonstrating how an alluvial paleoflood data set reduces uncertainty in a flood frequency analysis for a gage on the lower Mississippi River. Finally, we provide recommendations for how to incorporate paleoflood information into flood frequency analysis to improve the accuracy of extreme flood probabilities.

Plain Language Summary Water resource managers and engineers rely on estimates of extreme flood probabilities for flood mitigation, infrastructure planning, and emergency management. However, computing accurate estimates of extreme flood probabilities can be challenging due to the brevity and irregularity of instrumental river discharge measurements. Here, we evaluate how paleoflood data—information about floods that occurred prior to instrumental measurements derived from geological archives—can improve the accuracy of extreme flood probability estimates on low-gradient rivers. Our results show that paleoflood events can help to improve the accuracy of extreme flood probabilities and we provide a series of recommendations for how this data can be integrated into current flood probability estimation methods.

1. Introduction

Severe flooding ranks among the costliest and most frequently occurring natural disasters and affects communities across the United States (Mallakpour & Villarini, 2015; NOAA, 2020; A. B. Smith, 2020). Flood mitigation and emergency management depend on estimates of extreme flood probabilities [e.g., the 100-year flood (Q_{100}); a flood with a 1% chance of occurrence each year] to prepare for large floods effectively. Traditionally, the probability of extreme floods are estimated using a flood frequency analysis, where annual hydrologic maxima from a stream gage are fit to a probability distribution to compute the likelihood of a given flood level at the gage site (Kidson & Richards, 2005). These conventional flood frequency analyses, however, face a major drawback related to the brevity of instrumental records. Instrumental discharge records are often short—especially for periods in which rivers were minimally influenced by human activities—and may not include large and infrequent events (Benito, Brázdil, et al., 2015; Cohn & Stedinger, 1987). As such, flood frequency analyses performed with short instrumental records may underestimate or overestimate the probabilities of extreme flooding (Hosking & Wallis, 1986a; Merz & Bloschl, 2008a).

Improving probability estimates of hydrologic extremes can be accomplished by including information into flood frequency analyses based on hydrological reasoning that builds on conventional statistical methods

(Klemeš, 1993). Described as "flood frequency hydrology," this approach encompasses the addition of information in three categories: temporal, spatial, and causal (Merz & Bloschl, 2008a). Temporal information includes data related to flood history before the instrumental collection of streamflow data, while spatial information includes data derived from neighboring regions or watersheds, and causal information describes knowledge of the processes that generate flooding (Merz & Bloschl, 2008a). For example, paleoflood data derived from slackwater sediments in bedrock canyons represent temporal information that have been integrated into and improved flood frequency analyses (Baker, 1987, 2008; Benito, Brázdil, et al., 2015; Frances et al., 1994; Macdonald et al., 2014). In recent years, the use of paleoflood data has been expanded to include other archive types and geographic settings, including lake sediments, tree-rings, speleothems, and historical archives (Wilhelm et al., 2020).

Despite the potential of paleoflood data to improve estimates of extreme flood probabilities (Madsen et al., 2013) and the diversity of methods that describe how to incorporate them into flood frequency analysis—including the "Bulletin 17C" guidelines for the United States (England, Cohn, et al., 2019)—paleoflood archives are not consistently incorporated into flood hazard estimates in many countries outside the US, particularly for large alluvial rivers (Castellarin et al., 2012; Madsen et al., 2013). Advances in paleoflood hydrology now allow for the reconstruction of past floods on alluvial rivers by using coarse-grained flood deposits that accumulate in floodplain depressions, including oxbow lakes (Fuller et al., 2018; Jones, 2012; Leigh, 2018; Munoz et al., 2015, 2018; Toonen, 2012; Toonen et al., 2015, 2020). Oxbow lakes, which form when a meandering river cuts itself off, are common in floodplains and provide a source for hydrologic temporal information for alluvial rivers around the world, including places that do not have long systematic records (Toonen, Munoz, et al., 2020). However, these alluvial paleoflood records differ from most paleo stage indicators (PSI), like slack water deposits, in terms of the number of paleoflood events recorded, the paleoflood record length, and the uncertainty of discharge estimates. It is thus difficult to determine their additional value to flood frequency analysis, as many common practices in paleoflood hydrology flood frequency analyses may not apply or have not yet been assessed in this context (Harden, O'Connor, et al., 2011).

Here, we aim to advance the use of hydrological reasoning using temporal information in the form of paleoflood data to improve the biases of flood frequency analyses resulting from short instrumental records for alluvial rivers. We evaluate the ability of oxbow lake derived paleoflood records to improve extreme flood estimates under a range of scenarios. We run a series of numerical experiments in which flood frequency analyses are performed on simulated instrumental discharge and paleoflood data with different characteristics including record length, number of events, and uncertainty structure. Our results show that including large numbers of paleoflood events can increase the accuracy of frequency analysis estimates, but that uncertainties over the paleoflood magnitude can in some cases reduce the effectiveness of flood frequency analyses. We illustrate our recommendations by incorporating paleoflood data into a flood frequency analysis for the lower Mississippi River at Vicksburg (USGS gage 07289000). We conclude with recommendations describing best practices for how to incorporate oxbow lake derived temporal information into flood frequency analysis.

2. Background

2.1. Paleoflood Data in Flood Frequency Analysis

Adding paleoflood and historical flood data to systematic annual maxima data is a technique which reduces sample bias of short instrumental streamflow records (Benito, Lang, et al., 2004; Hosking & Wallis, 1986b; MacDonald et al., 2014). Incorporating paleoflood data into frequency analysis requires an extended version of the standard distribution fitting algorithms applied in analyses with only systematic instrumental data (Frances et al., 1994; Stedinger & Cohn, 1986). For example, Bulletin 17C recommends the Expected Moment Algorithm (EMA), which is an adjusted version of a Method of Moment procedure which computes distribution parameters from a sample's moments (Blainey et al., 2002; England et al., 2003, 2019). Other scholars attempted to estimate distribution parameters from the L-moments of a sample via the Partial Probability Weighted Moments (PPWM) algorithm, an extended systematic Probability Weighted Moments (PWM) method (Greenwood, 1979; Q. J. Wang, 1990). We settled for a Maximum Likelihood Estimator (MLE) algorithm which performs consistently well with the incorporation of paleoflood data, different lengths of instrumental data, and multiple extreme value distribution models (Blainey et al., 2002; Frances

et al., 1994; Payrastre et al., 2011; Stedinger & Cohn, 1986; Strupczewski et al., 2017). An MLE computes the likelihood that a sample is generated from a given set of distribution parameters via the *a priori* determined distribution probability density function (Stedinger & Cohn, 1986).

Aside from the Frequentist methods described above, a number of Bayesian flood frequency methods have been available to further reduce the uncertainty of return level estimates. Combined with Markov Chain Monte Carlo (MCMC) algorithms, Bayesian methods have successfully been applied to instrumental discharge records around the world in both a parametric (Lam et al., 2017; Reis & Stedinger, 2005), and nonparametric setting (O'Connell, 2005). In addition, existing Bayesian algorithms offer the advantages of allowing for the incorporation of multiple types of paleo-data, such as paleohydrologic bounds and 2D model output (Kuczera, 1999; O'Connell et al., 2002). Both parametric and nonparametric Bayesian algorithms reduce the biases of traditional flood frequency analysis, however, conventional flood frequency procedures continue to be primarily based on Frequentist statistics (Castellarin et al., 2012; Madsen et al., 2013). Therefore, we decided to focus our study on reducing biases in flood frequency analysis using Frequentist, instead of Bayesian, means.

The increased availability of paleoflood data has prompted interest in quantifying their added value to flood frequency analyses. Factors that determine the value of paleoflood data include the number of paleofloods, the paleoflood record length, the length of the systematic record, the shape of the flood frequency distribution, and the flood quantile of interest (Blainey et al., 2002). In simulation experiments where paleofloods are selected because they exceed a given threshold, the magnitude of that threshold also affects the information value of paleoflood data (Blainey et al., 2002; Frances et al., 1994)—but this primarily reflects the number of paleofloods in the record (Guo & Cunnane, 1991). Hosking and Wallis (1986a, 1986b) demonstrated that including only one paleoflood significantly improved the accuracy of Q_{100} estimates, specifically when instrumental gage records are short. Other authors reiterate this result and show that adding even more events can further improves Q_{100} estimates (England, Salas, & Jarrett, 2003; Guo & Cunnane, 1991; Strupczewski et al., 2014). However, this improvement is conditional on the length of the paleoflood record—such that short records with an increasing amount of paleoflood events can eventually degrade Q_{100} estimates (Guo & Cunnane, 1991)—and the type of distribution fitting algorithm (England, Salas, & Jarrett, 2003).

Increasing the length of the paleoflood record can substantially improve the accuracy of Q_{100} estimates, yet this is dependent on other factors, such as the type of distribution fitting algorithm (Blainey et al., 2002; England, Salas, & Jarrett, 2003; Stedinger & Cohn, 1986) and systematic record length (Strupczewski et al., 2014). For example, the Historical Weighting Moments parameter estimation method, recommended in Bulletin 17B (an earlier version of Bulletin 17C), only marginal improves the Q_{100} estimates as record lengths increase (England, Salas, & Jarrett, 2003; Stedinger & Cohn, 1986), but in combination with EMA and MLE algorithms longer paleoflood records do improve Q_{100} estimates (Blainey et al., 2002; England, Salas, & Jarrett, 2003). When only systematic records are available, the more extreme the flood quantile of interest, the larger the error of the quantile estimate (Blainey et al., 2002; England, Salas, & Jarrett, 2003). This pattern is similar when paleofloods are included in the analysis, although dampened for more extreme quantiles (Blainey et al., 2002; England, Salas, & Jarrett, 2003). In addition, Frances et al. (1994) show that the statistical gain of paleoflood data to improve flood frequency analysis increases for values up to Q_{10}, but then stagnates. This means that adding paleofloods does not provide a relative advantage for estimating, for example, a Q_{100} event over a Q_{1000} event.

Uncertainty about the properties of the paleoflood archive can also lead to biased Q_T estimates. We can distinguish two important types of error inherent to paleoflood data: (1) uncertainty regarding the paleoflood magnitude and (2) dating error of the paleoflood. Hosking and Wallis (1986b) demonstrate, using a simulation experiment, that errors larger than 25% over the paleoflood magnitude can reduce the effects of incorporating paleoflood data into flood frequency analyses. However, Blainey et al. (2002) show that paleoflood data with log-normally distributed 2σ errors of ±30% only marginally reduces the uncertainty of a flood frequency analysis; their analysis also included a ±40-year dating error, typical for the normally distributed uncertainty of radiocarbon dates (Blainey et al., 2002). Taken alone, this dating error only reduced uncertainty of the Q_{100} estimates by 1.7% in combination with the best possible paleoflood scenario (Blainey et al., 2002). Strupczewski et al. (2014), however, demonstrate that uncertainty about the length of the paleoflood record, as a result of chronological uncertainty, can reduce the value of paleoflood data to

flood frequency analysis. Although there has been a large body of work addressing paleoflood archives role in flood frequency analysis through simulation experiments, our aim here is to explore aspects that have received less attention but are important when considering how best to incorporate paleoflood data into flood frequency analysis on alluvial rivers.

2.2. Oxbow Lake Sediments as Paleoflood Record

Recent advances in paleoflood hydrology have expanded the use of paleoflood archives beyond bedrock canyons to encompass a range of hydrologic settings, including alluvial rivers (Wilhelm et al., 2019). In alluvial settings, paleoflood records depend on the deposition of distinctive sediments during flood events in floodplain depressions, including oxbow lakes (Toonen, Munoz, et al., 2020); these alluvial paleoflood archives have distinct characteristics from other archives that have traditionally been incorporated into flood frequency analysis (Ely & Baker, 1985; Levish et al., 2003; Thorndycraft et al., 2005). In alluvial river valleys, coarse-grained sediments deposited in oxbow lakes during overbank floods serve as flood magnitude proxy (Toonen, Munoz, et al., 2020), where high-magnitude flood events result in the mobilization and deposition of coarser-grained material (Fuller et al., 2019; Munoz, Giosan, Therrell, et al., 2018; Toonen, Winkels, et al., 2015). These alluvial paleoflood hydrology techniques can generate multi-centennial flood records with over 20 paleoflood events, but with event magnitude uncertainties of >10%. Moreover, these sediment records are relatively short, encompassing centuries to millennia, as oxbow lakes fill in with both river sediments and eroded floodplain material (Toonen, Kleinhans, & Cohen, 2012). Consequently, the ratio between the record length and amount of paleofloods is much higher compared to other paleoflood archives.

Unsystematic flood data—including both historic and paleoflood events—have been grouped into four classes: (1) floods with a known magnitude, (2) floods with an unknown magnitude but below some magnitude threshold, (3) floods with an unknown magnitude but above some magnitude threshold, and (4) floods with a magnitude between some range (Blainey et al., 2002; England, Salas, & Jarrett, 2003). Prior work addressing the value of paleoflood data in flood frequency analyses primarily focuses on paleoflood records from slackwater deposits (Hosking & Wallis, 1986b; Stedinger & Cohn, 1986). Slackwater sediments—flood deposits preserved at tributary mouths and rock shelters along bedrock canyons—typically provide precise estimates of past flood and registration threshold stage through hydraulic models, as topography remains relatively stable over time (Baker, 2008; Kochel & Baker, 1982; O'Connor et al., 2014). In contrast to bedrock canyons, alluvial floodplains are geomorphically dynamic and the processes involved in sediment deposition during overbank flow are complex and sensitive to subtle topographic variation, sediment provenance, and human modifications to the floodplain and channel (Asselman & Middelkoop, 1995; Munoz, Giosan, Blusztajn, et al., 2019), so reconstructing a paleoflood magnitude or proxy registration threshold in these settings is more difficult than in bedrock canyons.

To estimate paleoflood magnitudes from oxbow lake sediments, a linear regression is used to relate sediment texture of historic floods to the measured peak discharge of those floods at a nearby gage; this statistical model is then used to estimate the magnitude of prehistoric floods (Fuller et al., 2019; Leigh, 2018; Toonen, Winkels, et al., 2015). While this approach has been successful for several alluvial rivers (Leigh, 2018; Munoz, Giosan, Therrell, et al., 2018; Toonen, Winkels, et al., 2015), it cannot provide a proxy registration threshold, and results in discharge estimates with errors of 10%–30%. Here, we decided to focus our simulation experiments on floods with a known magnitude as this most closely reflects the characteristics of alluvial paleoflood archives. Consequently, our MLE algorithm does not include binomial observations, ranges, or multiple proxy registration thresholds even though these techniques are widely used in flood frequency analysis for other paleoflood archives (England, Salas, & Jarrett, 2003; Frances et al., 1994).

Previous work on the Rhine River showed that this type of integration of paleoflood records can significantly improve the precision of return levels by reducing confidence intervals of the return level plot, especially in combination with advanced statistical methods such as bootstrapping and MCMC algorithms (Bomers et al., 2019; Toonen, Winkels, et al., 2015). Here, we evaluate the ability paleoflood records to also improve the accuracy of extreme flood estimate using a series of simulation experiments that represent the characteristics of alluvial paleoflood records. Specifically, we evaluate the effect of the amount of paleofloods and the uncertainty of the paleoflood estimates on flood frequency analysis - two characteristics where alluvial paleoflood archives distinct from slack water deposits. We examine the effects of these characteristics on Q_T

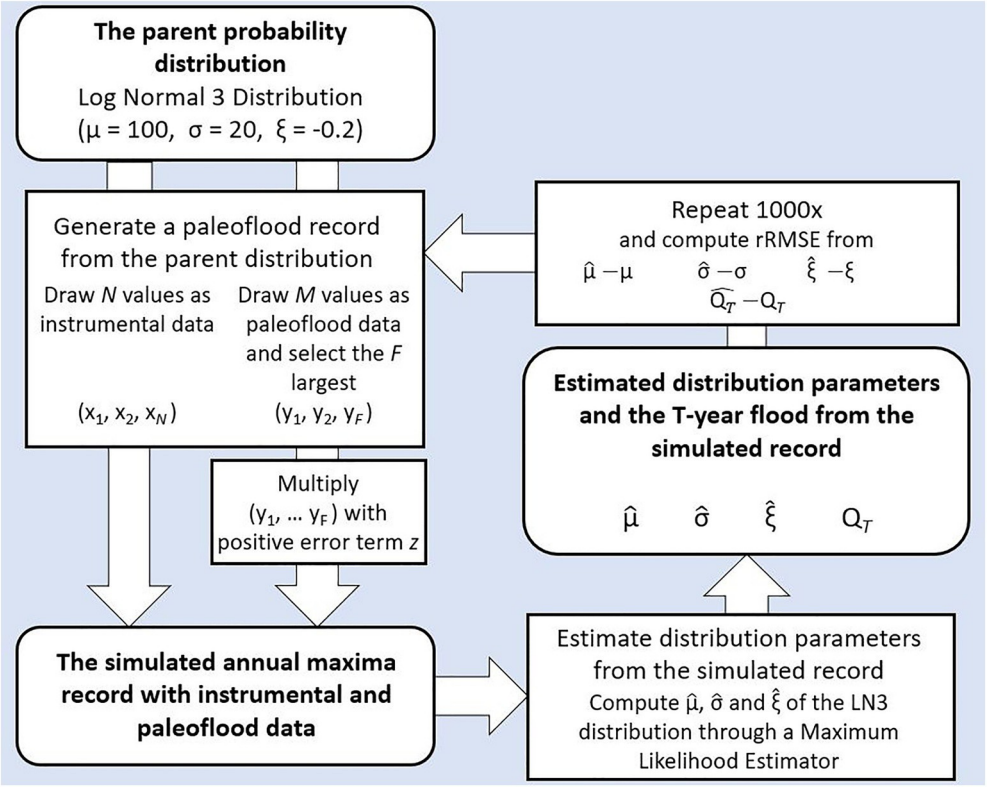

Figure 1. Schematic overview of the simulation experiments implemented in this study to evaluate the efficacy of paleoflood records in flood frequency analysis.

values, as well as the estimates of specific distribution parameters including location, scale, and shape. The structure of our simulation experiments is similar to those of Stedinger and Cohn (1986), Hosking and Wallis (1986b), and Strupczewski et al. (2017), where we fit a parametric distribution to synthetic (paleo)flood records drawn from a predefined distribution Our first set of experiments examines the effect of increasing the number of paleoflood events on distribution parameters, Q_T estimates, while a second set of experiments focuses on the effect of uncertainty structures (i.e., over- and under-estimation) over the magnitude of paleoflood discharges (Figure 1).

3. Data and Methods

3.1. Simulation Procedure

Our approach for assessing the utility of paleoflood events in flood frequency analyses involves estimating the distribution parameters of a simulated record with both instrumental data (x_1, x_2, ..., x_N), and paleoflood data (y_1, y_2,...,y_F) from a (pre)historical period M (Figure 2). We chose to draw data from a three-parameter Log Normal (LN3) distribution (Equation 1), because it fits the average statistical properties of annual maxima records throughout the United States (Vogel & Fennessey, 1993; Vogel & Wilson, 1996) (Figure S1). The LN3 distribution has a location, scale, and shape parameter (Hosking & Wallis, 1997). The location parameter (μ) sets the center of the distribution, the scale parameter (α) describes its variance, and the shape parameter (ξ) describes the asymmetry of the distribution and thus impacts tail thickness and extreme values. The parent distribution has a location parameter of 100 and a scale parameter of 20. We picked these values based on the ratio between μ and α of the peak flow distribution of the Mississippi River at Vicksburg gauge station (USGS gage 07289000). We performed the analyses with a shape parameter of −0.2, similar to the Mississippi at Vicksburg, and an additional experiment with thicker tails for which the shape parameter is −0.6.

$$f(x) = \frac{e^{\frac{y-}{2}}}{\sqrt{}}, y = \begin{cases} -^{-1}\log\left\{1 - \left(\frac{x-}{}\right)\right\}, \neq 0 \\ \frac{x-}{}, = 0 \end{cases} \quad (1)$$

We construct records by generating N instrumental years and M historical years (y_1, y_2, ..., y_M). Only the F largest flood magnitudes from $y_{1:M}$ are selected as the paleoflood record and, if part of the experiment, multiplied by some Gaussian error term z. To estimate the location, scale, and shape parameters of the sample, an LN3 distribution is fit to the sample distribution with an MLE algorithm (Equation 2) similar to Frances et al. (1994) and Stedinger and Cohn (1986). In Equation 2, x_t is the systematic record, y_t the paleoflood events, h the length of the paleoflood record, k the number of paleoflood events, and X_0 the perception threshold. We take the magnitude of the lowest paleoflood as X_0 because it is difficult to establish a real perception threshold for these environments as established above. We ran 1,000 simulations for every scenario in which N, M, F, and the distribution from which z is drawn are fixed, but $x_{1:N}$ and $y_{1:M}$ vary. The Root Mean Square Error (RMSE) is computed from the 1000 simulations to statistically summarize the deviation from the true parameter values, Q_{10}, Q_{50}, Q_{100}, Q_{500}, Q_{1000}, Q_{10000}, and $Q_{1000000}$ (Strupczewski et al., 2017). The RMSE is the square root of the average squared difference between the estimated and true parameters—we divide it by the average estimate of the 1,000 simulations to get the relative RMSE (rRMSE).

$$L(_t, _t, _t) = \prod_{t=1}^{n} f_x(x_t) \binom{h}{k} F_x(X_0)^{(h-k)} \prod_{t=1}^{k} f_x(y_t) \quad (2)$$

Unlike previous simulation experiments with paleoflood data, our simulation set up only involves the selection of the F largest floods, also known as type 2 censoring (Blainey et al., 2002). Type 1 censoring implies that the paleoflood record results from the selection of all floods above some fixed threshold, in simulation experiments often Q_T. It is important to distinguish between the two as they result in different behavior of the paleoflood record length to the paleoflood events amount ratio (M:F) as paleoflood records get longer - which will affect the accuracy of the flood frequency analysis. For type 1 censoring, this ratio remains stable because more paleofloods will be observed as events as more cross the threshold in a longer record. For type 2 censoring, this ratio is not stable as the amount of paleoflood events remains the same (F) even though the record gets longer. Previous simulation experiments have shown that when the ratio is high (e.g., a short record with many events), it reduces the accuracy of the flood frequency analysis (Guo & Cunnane, 1991; O'Connor et al., 2014). As described above, oxbow lake paleoflood archives typically have high M:F ratios, which can be reconstructed flexibly with type 2 censoring (or type 1 censoring with a low threshold). Therefore, we decide to focus on type 2 censoring as it better resembles the characteristics of oxbow lake paleoflood archives.

3.2. Experiments

First, we run a set of experiments to evaluate the combined effects of the number of paleoflood events and instrumental record lengths. Initially, the number of paleoflood events is increased from 0 to 8 and from 10 to 50, while the length of the (pre)historical period is fixed at 300 years. Next, the length of the (pre)historical period varies from 0 to 1,000 years by increments of 200 years while the number of paleoflood events was fixed at 10. For every combination of paleoflood events and historical period length, described above, 10 simulations are run with increasing instrumental years from 10 to 100. After every scenario the rRMSEs are computed for all Q_T values and the three distribution parameters. A second set of experiments evaluates the role of uncertainty in the paleoflood discharge data on the Q_{100} levels and distribution parameters. Three uncertainty structures of the paleoflood discharge magnitudes—nondirectional random error, structural underestimation, and structural overestimation—are simulated by multiplying the floods $y_{1:F}$ with an error term z drawn from a normal distribution D as $N(\mu,\sigma)$. For the nondirectional random error scenario, D has a mean 1 and an increasing standard deviation from 0.1 to 0.4. Meaning that the paleoflood discharge is as likely to be both lesser or greater in magnitude than the true paleoflood discharge, but with increasing de-

viation as σ gets larger. For under or overestimation, σ is fixed (0.1) but μ is respectively changing from 1 to 0.8 or 1 to 1.2. Meaning that the paleoflood discharge can still be larger or smaller than the true paleoflood discharge but the odds of the paleoflood becoming larger increases as the mean (μ) moves up and vice versa. To summarize, this means that generated paleoflood discharge is multiplied by a factor around 1 and the error becomes greater as μ moves away from 1, or σ becomes larger. We assume the range of these error values based on the estimate uncertainty of paleoflood data retrieved from the Mississippi River (Munoz, Giosan, Therrell, et al., 2018). Here we want to explicitly study the effect of error associated with paleoflood events, however, we recognize that systematic records also have error associated with them which is addressed by other authors (Blainey et al., 2002; Kuczera, 1996).

We estimate the Q_T values and the three distribution parameters for each type of uncertainty and with increasing lengths of instrumental data, like the previous experiment. Finally, different uncertainty structures and the number of paleoflood events are combined to test the effect of more paleoflood events with error on the accuracy of Q_{100} estimates. To do this, we generate paleoflood data with z from a fixed distribution $N(\mu,\sigma)$, for each type of uncertainty, and increased the number of paleoflood events from 0 to 10 for every scenario. All simulation experiments are performed for both the parent distribution with a shape of −0.2 and −0.6.

4. Results

4.1. Paleoflood Record Length

Our simulation experiments show that the rRMSE of $Q_{100,}$ and all three distribution parameters reduces as more paleofloods are included, but the decrease of rRMSE diminishes as more paleoflood events are added and as instrumental record lengths are increased (Figure 2).The reduction of the Q_{100} rRMSE stagnates after the addition of 10 paleofloods (Figures 2a and 3b). Our simulations also show that the rRMSE of Q_{100} becomes insensitive to the instrumental record length with the inclusion of more than six paleoflood events, as after that the Q_{100} rRMSE does not improve further (Figures 2a and 2b). These findings support that an analysis with more than six paleofloods is most beneficial when instrumental records are short, as the Q_{100} rRMSE of analyses that exclude paleoflood data exponentially decreases as instrumental record length increases. For example, there is a 75% reduction in the Q_{100} rRMSE when six paleofloods are added to a short (20 years) instrumental record relative to an analysis with no paleofloods; the same analysis with a longer (40 years) instrumental record reduces the Q_{100} rRMSE by 60% (Figure 2a). Even in the case of relatively long instrumental records (100 years), the rRMSE is still reduced by 45% when six paleoflood events are included in the analysis, although the rRMSE reduction is moderate (0.024) when the instrumental record is this long.

Our results also document a relatively increasing reduction of rRMSE for Q_{500} and larger return periods compared to Q_{100} rRMSE when paleoflood data are included in frequency analysis, reflecting the difficulty of estimating extreme flood levels using only instrumental data (Figure 3). We also find the opposite trends for return levels smaller than Q_{100}, where improvements with the addition of paleoflood data are less pronounced (Figure 3). Adding more paleoflood events is least helpful to estimate smaller return periods, but for very large return levels (i.e., $Q_{1000000}$) there is a substantial difference in rRMSE between using 10 or 50 paleoflood events (Figure 3b). Increasing the length of time from which paleofloods are retrieved only reduces the Q_{100} rRMSE notably when multiple paleofloods are included (Figure S2). However, including many paleofloods from a relatively short period (e.g., 40 floods over 200 years) results in a larger Q_{100} rRMSE relative to an analysis that does not include paleoflood data (Figure S3).

The improvements to flood frequency analyses with the inclusion of paleoflood data are primarily due to reduced rRMSE of the shape parameter, which describes the tail end of distribution (Figure 2). For example, when two paleoflood events are added to 20 years of instrumental data, the rRMSE of the shape parameter is reduced by 0.41 (33%) relative to an analysis without paleoflood data. The effect of more paleoflood flood on the rRMSE of the shape parameter is minimal, such that the rRMSE in the scenario described above (i.e., 20 years of instrumental data) is only marginally improved (0.47 rRMSE; 39%) if six paleoflood events are included instead of two (Figure 2b). Only when instrumental records are shorter than 40 years and more than 10 paleofloods are included will the shape parameter rRMSE further decrease with additional

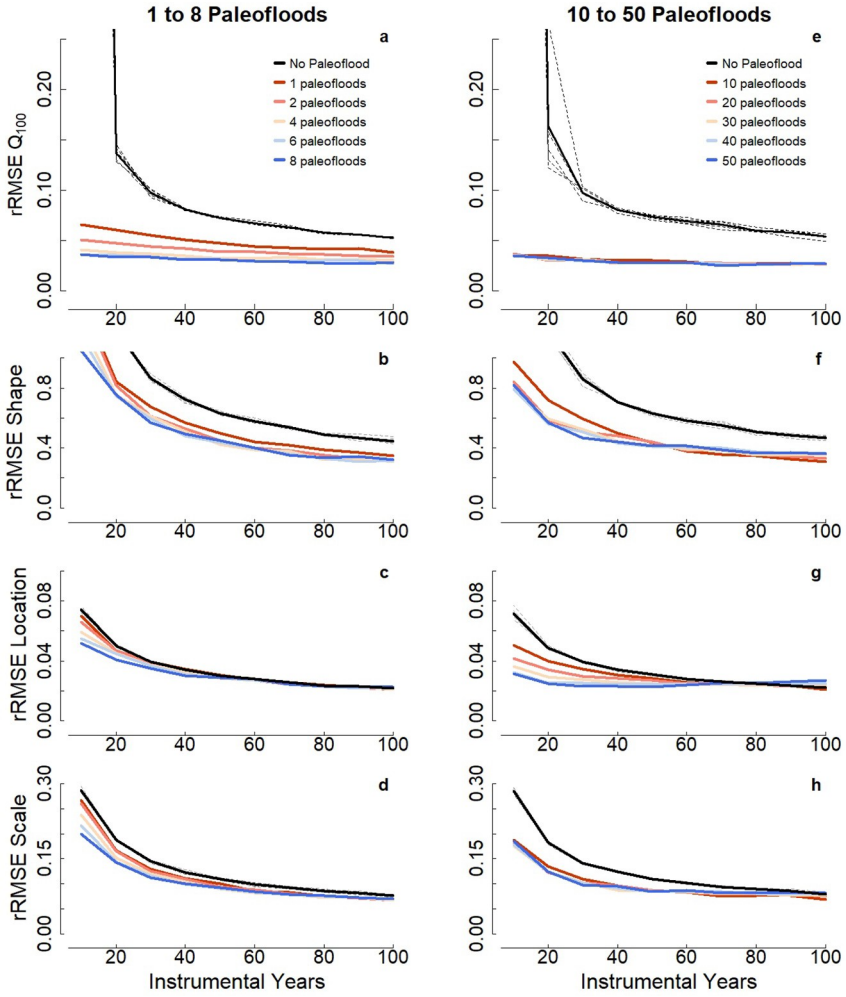

Figure 2. The relative Root Means Square Error (rRMSE) of the estimated 100-years-flood (Q_{100}) and the three parameters of the Log Normal 3 distribution ($\xi = -0.2$), from flood frequency analyses with different numbers of paleoflood events (over a 300-year period) as a function of the gaged instrumental record length. The striped black lines denote simulations with only instrumental data and the solid black line is their average. Panels (a–d) simulations with 1–8 paleoflood events; panels (e–h): simulations with 10–50 paleoflood events.

paleofloods (Figure 2f). In contrast to the shape parameter, the inclusion of paleoflood data exerts only a marginal influence on the rRMSE of the location parameter (Figure 2c). For example, with 40 years of instrumental data and two paleoflood events, the location rRMSE is reduced by only 3% relative to an analysis that does not include paleofloods, though the relative reduction becomes larger when more than 10 paleoflood events are included (e.g., 65% reduction in rRMSE [0.05] with 20 paleoflood events). We also note that when instrumental data sets are >40 years in length, the inclusion of >10 paleoflood events worsen estimates of the scale and location parameters relative to an analysis that includes only instrumental data (Figures 2g and 2h).

We also evaluated the influence the parent distribution to find that a distribution with thicker tails (shape parameter of −0.06) reduces the accuracy of Q_{100} estimates in scenarios with <8 paleofloods included, but more paleoflood do not further improve the results (Figure 4). The inclusion of paleoflood data improves the accuracy of the shape parameter for a parent distribution with thicker tails compared to a distribution for thin tails. For example, an analysis with six paleofloods and 20 years of instrumental from a parent distribution with a shape parameter of −0.6 results 64% lower rRMSE compared to the same analysis with a parent distribution with a shape parameter of −0.2. The location and shape parameter are less sensitive to changes in tail thickness (Figure 4c). Together, our results demonstrate that the inclusion of paleoflood data

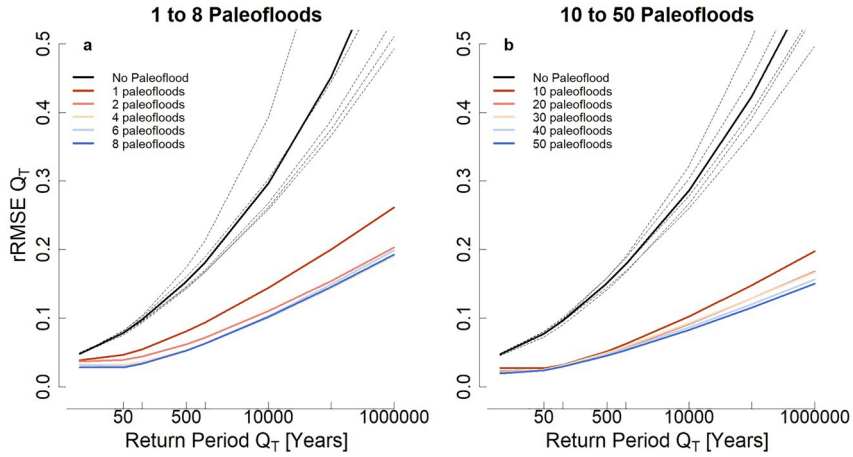

Figure 3. The relative root means square error (rRMSE) of different flood quantiles, ranging from the 10-year flood (Q_{10}) to the 1,000,000-year flood ($Q_{1000000}$), from flood frequency analyses with different numbers of paleoflood events (over a 300-year period). The striped black lines denote simulations with only instrumental data and the solid black line is their average. Panel (a) simulations with 1–8 paleoflood events; panel (b) simulations with 10–50 paleoflood events.

increases the accuracy of extreme flood estimates by improving the estimation of the shape parameter, but that the use of >10 paleoflood events can reduce the accuracy of the location and scale parameters.

4.2. Paleoflood Record Uncertainty

Our experiments examined three uncertainty structures in the paleoflood data—random, negative, and positive biases—and generally found that paleoflood data with large errors of any type result in poorer or estimates as good as an analysis without paleoflood data (Figure 5). Normally distributed 1σ errors of > 0.2 (i.e., where the ratio between the mean and standard deviation is larger than 0.2) result in location (Figure 5b) and scale (Figure 5c) parameter estimates with higher rRMSE than when no paleoflood data are included. The shape parameter is most resilient to random error, as paleoflood estimates with 1σ errors < 0.3 perform better than analyses based only on instrumental data (Figure 5a). For the shape parameter, paleoflood data with 1σ errors < 0.4 improve the analysis when instrumental records are <60 years in length. We also found that structural underestimation and overestimation of paleoflood discharges has a similar impact on the rRMSE of location and scale parameters (Figure 5). In general, our results show that the addition of paleoflood events with 1σ errors > 0.2 do not improve—and sometimes worsen—the estimates of Q_{100} flood levels in a frequency analysis when instrumental records are >20 years in length.

Finally, our analyses show that the accuracy of extreme flood levels reduces as a large number of uncertain paleoflood events is included in a frequency analysis—even if the paleoflood errors are relatively small—due to compounding of individual errors (Figure 6). For example, the inclusion of two paleoflood events with random 1σ errors of 0.2 and 80 years of instrumental data results in larger rRMSE of Q_{100} than if only instrumental data are used and including eight paleoflood events with these errors doubles the rRMSE of Q_{100} (Figure 6a). We observe a similar effect for structural underestimation of paleoflood errors, though structural overestimation is less affected by the number of paleoflood events—especially when instrumental records are short (Figure 6b). The inclusion of paleoflood events is most beneficial when instrumental records are <20 years in length, even when more than two paleoflood events with errors >20% are included. In short, we show that large errors in paleoflood magnitude estimates generally reduce the effectiveness of flood frequency analyses, and that the inclusion of a large number of paleoflood events compounds their uncertainty in a flood frequency analysis.

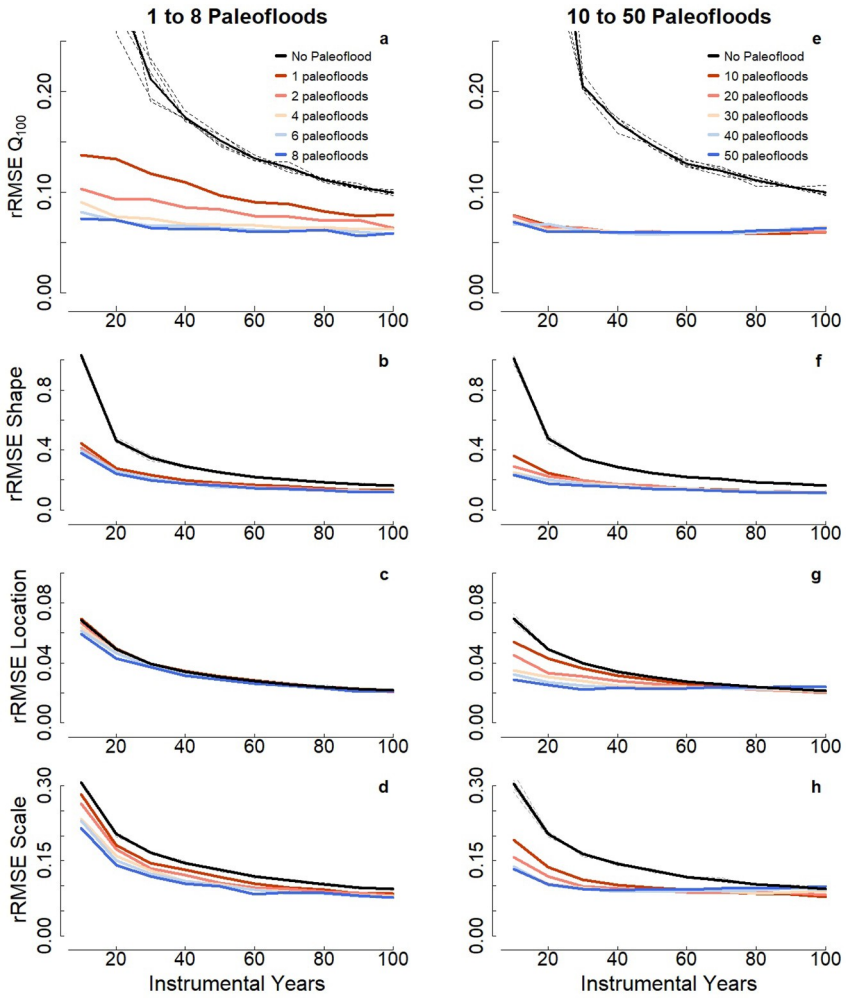

Figure 4. The relative root means square error (rRMSE) of the estimated 100-year-flood (Q_{100}) and the three parameters of the Log Normal 3 distribution ($\xi = -0.6$) from flood frequency analyses with different numbers of paleoflood events (over a 300-year period) as a function of the gaged instrumental record length. The striped black lines denote simulations with only instrumental data and the solid black line is their average. Panels (a–d): simulations with 1–8 paleoflood events; panels (e–h): simulations with 10–50 paleoflood events.

4.3. Robustness Test

To evaluate the robustness of our results, we run the same simulation experiments for a GEV distribution model. Different watersheds within the United States will have different underlying flood frequency distributions, as the country is hydrological diverse with several flood generating mechanisms (England, Salas, & Jarrett, 2003; Merz & Blöschl, 2003). In this context, a robust estimator can be defined as one which returns reliable results under a broad set of scenarios (England, Salas, & Jarrett, 2003; Stedinger et al., 1993). Here, we compare the LN3 simulation results to simulations from the GEV distribution, another common three-parameter distribution. For example, Salinas et al. (2014) identified the GEV distribution as a pan-European flood distribution, but the GEV distribution also characterizes the statistical properties of peak flow in rivers across the United States (Vogel & Wilson, 1996).

Similar to the LN3 distribution, the GEV distribution has a location, scale, and shape parameter (Hosking & Wallis, 1997). The GEV distribution summarizes three different extreme value distributions: the Gumbel distribution (GEV-type I) which has a shape parameter of 0, the Fréchet distribution (GEV-type II) which has a positive shape parameter, and the Weibull distribution (GEV-type III) which has a negative shape parameter. Here, larger shape parameters are associated with fatter tails. We assess the results for two sets of

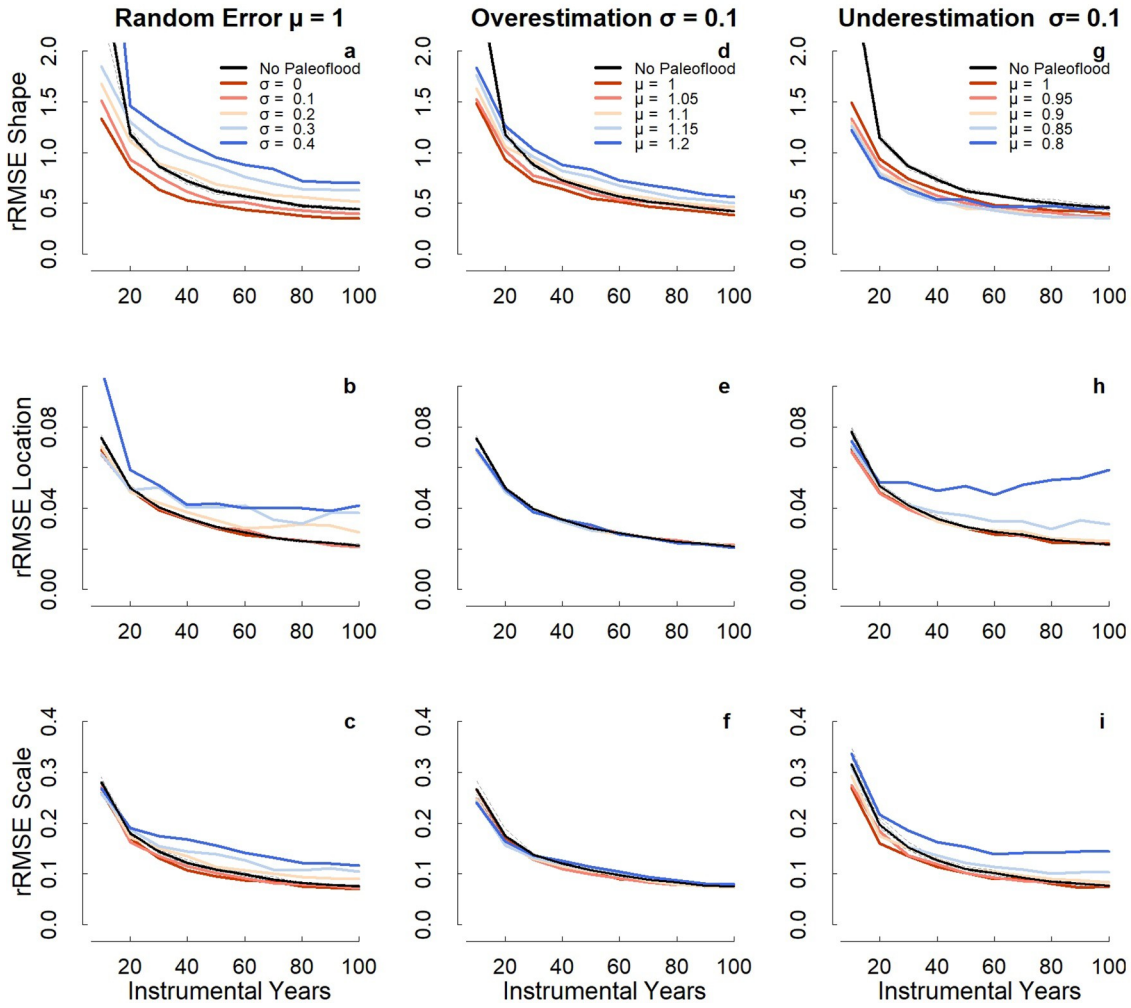

Figure 5. The relative root mean square error (rRMSE) of the estimated 100-year-flood (Q_{100}) and the three parameters of the Log Normal 3 (LN3) distribution ($\xi = -0.2$) from flood frequency analyses with one paleoflood event (over a 300-year period) as a function of the gaged instrumental record length. The striped black lines denote simulations with only instrumental data and the solid black line is their average. Panels (a–c): simulations with 1 paleoflood that is multiplied with z from a Normal distribution with mean of 1 and standard deviation denoted by the colored lines; panels (d–f): simulations with 1 paleoflood event that is multiplied with z from a Normal distribution with a standard deviation of 0.1 and a mean denoted by the colored lines; panels (g–i): simulations with one paleoflood event that is multiplied with z from a Normal distribution with a standard deviation of 0.1 and a mean denoted by the colored lines.

parameters, one that represents gauge data from the Mississippi River at Vicksburg [location: 100; scale: 20; shape: −0.1] and a distribution with fatter tails [location: 100; scale: 20; shape: 0.3].

Generally, simulations with a GEV distribution result in higher absolute rRMSE's compared to those with LN3 simulations, however the way rRMSE changes with different numbers of paleoflood events, different instrumental record lengths or differences in estimated flood quantiles remains similar (Figure 7). For example, the rRMSE values for no-error flood frequency analyses with LN3 and GEV distributions are alike, except for scenarios without paleoflood data, for which the GEV distributions shows larger rRMSE values (Figure 7). The same is true for other analyses performed in this study (Figures S9–S13).

4.4. Application to Mississippi River at Vicksburg

We applied the insights described above to perform an adjusted flood frequency analysis for the lower Mississippi River at Vicksburg (USGS gage 07289000) using paleoflood records derived from nearby oxbow lake sediments developed by Munoz, Giosan, Therrell, et al. (2018) (Figure 8). The instrumental record for the

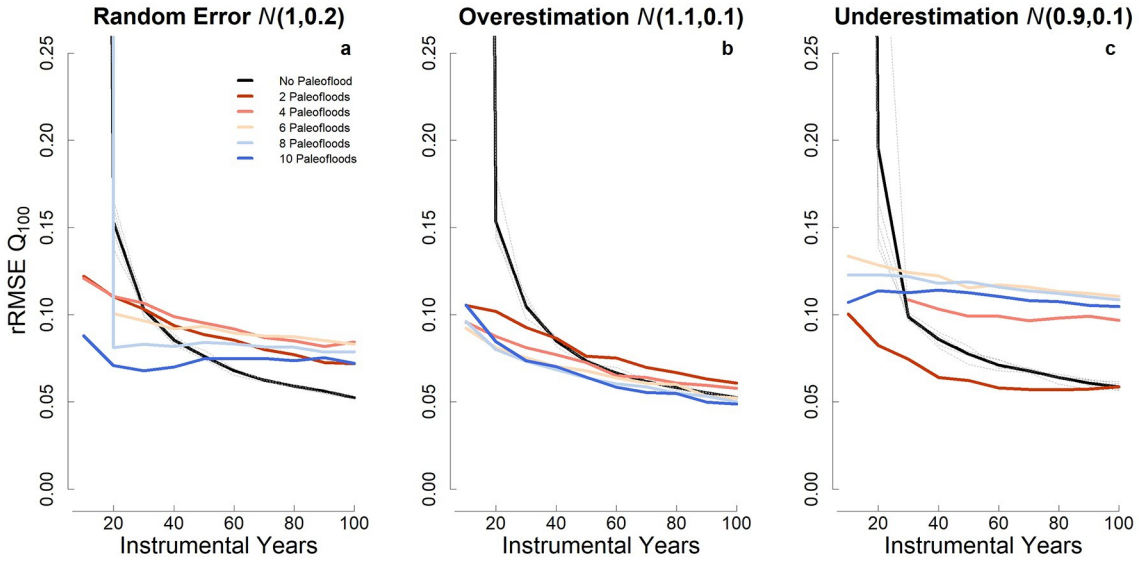

Figure 6. The relative root mean square error (rRMSE) of the estimated 100-year-flood (Q_{100}) and the three parameters of the Log Normal 3 (LN3) distribution ($\xi = -0.2$) from flood frequency analyses with 2–10 paleoflood events (over a 300-year period) as a function of the gaged instrumental record length. Panel (a) simulations with paleofloods multiplied by a random error term from a Normal distribution with a mean of 1 and standard deviation of 0.2; panel (b) simulations with paleofloods multiplied by an overestimation error term from a Normal distribution with a mean of 1.1 and standard deviation of 0.1; panel (c) simulations with paleofloods multiplied by an underestimation error term from a Normal distribution with a mean of 0.9 and standard deviation of 0.1.

Mississippi River at Vicksburg includes continuous annual peak discharge data spanning 128 years, from 1887 to 2015. However, human activities within the Mississippi River basin—including efforts to improve navigation and mitigate flooding—have profoundly influenced the hydrology of the lower Mississippi River over the last century (Munoz, Giosan, Therrell, et al., 2018; Pinter et al., 2008; Remo et al., 2009; L. M. Smith & Winkley, 1996). Levees and other engineering structures have been constructed along the lower Mississippi River since the eighteenth century, but, but a severe flood in 1927 resulted in the failure of the existing levee system and triggered more comprehensive legislation that federalized and significantly expanded flood mitigation and navigation infrastructure across the lower Mississippi River through the Flood Control Act of 1928 (Camillo, 2012). These activities, together with other upstream changes to the Mississippi River and its basin (e.g., land use change, dams), have profoundly influenced river stages, discharge, sediment loads, and channel morphology (Meade & Moody, 2010; Mossa, 1996; Munoz, Giosan, Therrell, et al., 2018; Pinter et al., 2008; Remo et al., 2009; B. Wang & Xu, 2018). The influence of anthropogenic climate change on the discharge of the lower Mississippi River remains unclear (Tao et al., 2014; van der Wiel et al., 2018), but may also contribute to nonstationarity of flood peaks beginning in the twentieth century. Together, geomorphic and hydroclimatic changes to the Mississippi River challenge the assumption of stationarity throughout the entire instrumental period. Here, we use a cutoff of date of 1928 to designate the end of "unregulated" flows of the lower Mississippi River because these predates the major investments in infrastructure associated with the MR&T project. Under this assumption, the instrumental period of unregulated peak flows for the Mississippi River at Vicksburg is reduced to 41 years that we supplement with existing paleoflood data.

We supplemented the instrumental data with paleoflood records derived from oxbow lake sediments along the lower Mississippi described by Munoz, Giosan, Therrell, et al. (2018). This paleoflood data set is derived from sedimentary records collected from three oxbow lakes along the Lower Mississippi River using an approach similar to other alluvial paleoflood records (Munoz, Gruley, et al., 2015; Toonen et al, 2015, 2020) and includes 45 paleoflood events with associated age and peak discharge estimates that span the last ~500 years. Peak discharges of the paleofloods were estimated and calibrated via a linear regression to the Vicksburg gage using the approach of Toonen, Winkels, et al. (2015). The estimated peak discharges of paleofloods had a 1σ error of ~0.1, so we included only the three largest paleoflood events, dated to ca. 1609 CE, 1620 and 1625, to avoid compounding errors and reducing the effectiveness of the paleoflood data (Figure 6). The historic

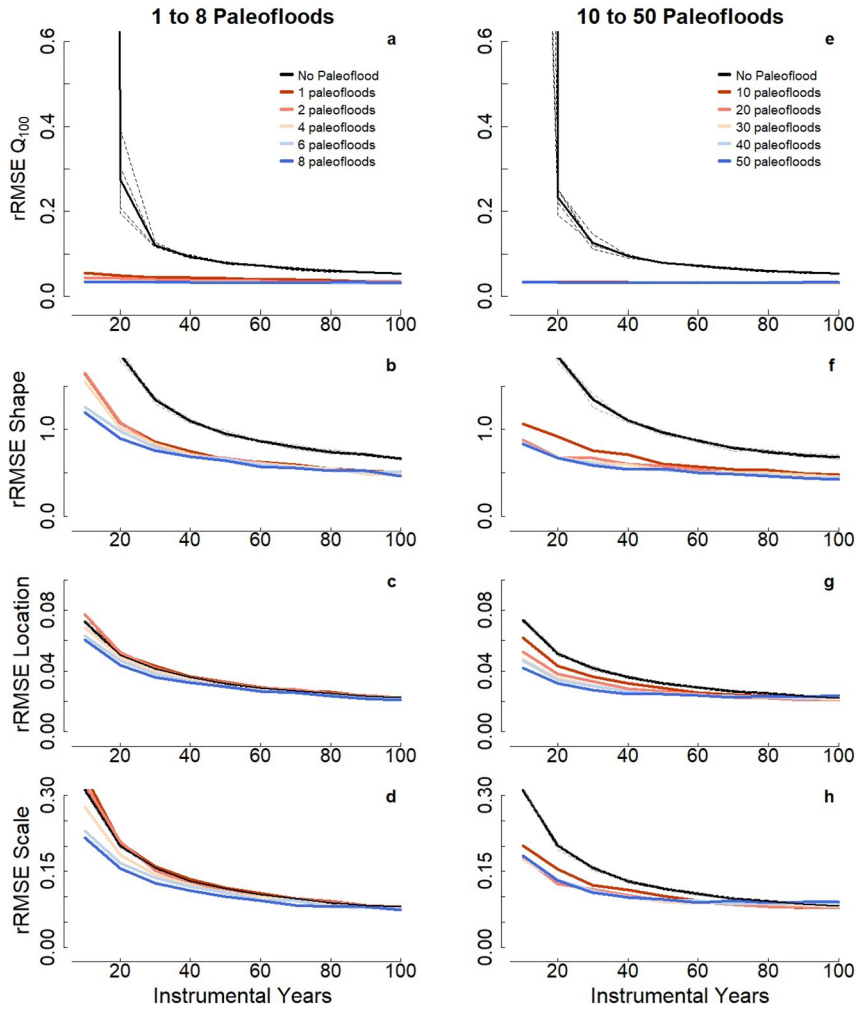

Figure 7. The relative root means square error (rRMSE) of the estimated 100-year-flood (Q_{100}) and the three parameters of the GEV distribution ($\xi = -0.1$), from flood frequency analyses with different numbers of paleoflood events (over a 300-year period) as a function of the gaged instrumental record length. The striped black lines denote simulations with only instrumental data and the solid black line is their average. Panels (a–d): simulations with 1–8 paleoflood events; panels (e–h): simulations with 10–50 paleoflood events.

period was computed via the formula described in Schendel and Thongwichian (2017) based on an uniform distribution, which counters the bias of using a too short period in which we observe the flood events if we only took the length of the available data as our historic period. The lowest paleoflood discharge level was assumed as the proxy registration threshold.

Our analysis shows that the inclusion of three paleofloods significantly reduces the confidence intervals of extreme flood levels (i.e., Q_{100} and Q_{500}), while the mean estimate of these flood levels only marginally changes with the inclusion of paleofloods (Figure 8b). Confidence interval width for Q_{100} and Q_{500} reduce by 53.7% and 62.2%, respectively, with the inclusion of paleofloods data, while the mean Q_{100} and Q_{500} estimates decrease by 8.7% and 11.9%, respectively, with the inclusion of paleoflood data. The substantial improvement in the precision of extreme flood levels with the inclusion of paleofloods results in Q_{100} and Q_{500} estimates and confidence intervals that do not exceed the Project Design Flood—a hypothetical maximum discharge that existing MR&T infrastructure are designed to withstand (Camillo, 2012).

Paleoflood event discharges are often more extreme than the values measured over the instrumental period (Baker, 2008), and thereby serve to increase the return levels of Q_{100} and Q_{500} when included in a frequency analysis. In the case of the lower Mississippi River, the most extreme paleoflood discharge is significantly

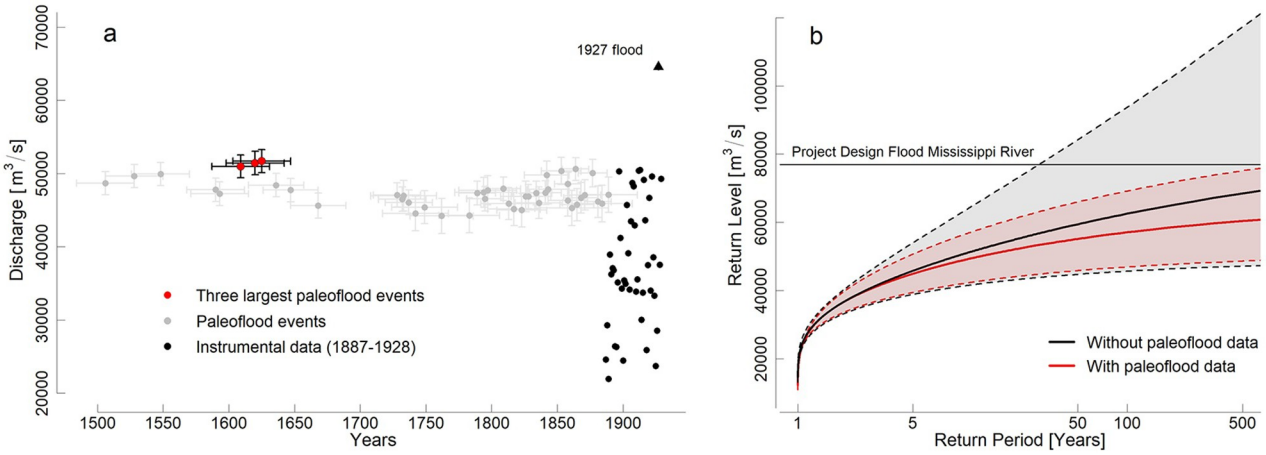

Figure 8. Application of alluvial paleoflood data to a flood frequency analysis for the Mississippi River at Vicksburg (USGS gage 07289000): (a) peak annual discharges for the Mississippi River at Vicksburg for the "unregulated" period (i.e., prior to 1928) from gage data (black; 1927 flood as triangle) and paleoflood data (gray = all paleofloods; red = three largest paleofloods selected for frequency analysis) from Munoz, Giosan, Therrell, et al. (2018); (b) flood levels associated with different flood return periods for a flood frequency analysis with a GEV distribution using only instrumental data (black line, shading is 95% confidence interval) and using 3 paleoflood events to supplement the instrumental record (red line, shading is 95% confidence interval).

lower (25%; 12,854 m^3/s) than the largest unregulated (pre-1928) instrumental datapoint (the 1927 flood) (Figure 8a). The MLE algorithm used to conduct the flood frequency analysis assumes a threshold such that all paleoflood events greater than the lowest paleoflood discharge are recorded. As a result, the shape parameter of our flood frequency analysis with paleofloods approaches the shape parameter of the analysis without paleofloods such that extreme return levels are similar between the two analyses. In other cases, where paleoflood magnitudes exceed those in the instrumental record, the inclusion of paleofloods may significantly alter the shape parameter and estimates of extreme flood level. It should be noted that there are multiple techniques that could make the flood frequency analysis more accurate, for instance including historical data with a different registration threshold. Here, however, our goal was strictly aimed at demonstrating the influence of including paleofloods in a flood frequency analysis.

5. Discussion

The main objective of this study is to develop approaches for including hydrological information into flood frequency analysis procedures for alluvial rivers. To achieve this, we examined how to include oxbow lake derived paleoflood data into flood distribution parameter estimations to improve the accuracy of extreme flood estimates. Our work provides insights into the use of these paleoflood archives from alluvial floodplain to aid return level estimates; specifically, the effect of the number of paleoflood events and their uncertainty.

Our findings demonstrate that including oxbow lake paleoflood records with more than two events can improve the estimation of extreme flood probabilities (Figure 2a), although we also show the limits of these improvements when instrumental data records are >40 years in length. Prior work exploring the value of paleoflood evidence for flood frequency analysis by Stedinger and Cohn (1986) and Strupczewski et al. (2014, 2017) found a similar reduction of the Q_{100} rRMSE with the inclusion of one or two paleoflood events. We extend these analyses by examining the effect of up to 50 events in light of new paleoflood records with high *M:F* ratios and show that incorporating 10 paleofloods can result in a Q_{100} rRMSE reduction of over 50%. Our results also add new insights by showing the influence of paleoflood data on the individual distribution parameter estimates. The reduction of the Q_{100} rRMSE resulting from the inclusion of paleoflood data reflects the improved accuracy of shape parameter estimates (Figure 2b), whereas the inclusion of too many paleofloods can skew the location and scale parameter estimates (Figures 2g and 2h). The shape parameter is dependent on the extreme values in an annual peak flow record, and these extremes are often poorly represented in instrumental data sets but well represented in a paleoflood record. In contrast, the location and scale parameters reflect average conditions that are well represented in the instrumental

data, such that the addition of paleoflood events could bias the location parameter toward higher values (Figure 2g).

Paleoflood events further reduce the rRMSE of the shape parameter estimates for flood frequency distributions with thick tails (e.g., $\xi = -0.6$), although the percentual decrease of rRSME with the inclusion of paleofloods remains comparable to thinner tailed distributions. These results show that oxbow lake paleoflood archives can be particular useful for rivers where the distribution of hydrologic maxima has thick tails (i.e., where extreme floods are more common), especially if we have interest in the shape parameter of the distribution. This further develops the argument that the underlying flood frequency distribution determines the value of paleoflood data in flood frequency analysis (England, Salas, & Jarrett, 2003). In addition, our findings show that adding paleofloods to an analysis benefit in particularly more extreme flood levels ($>Q_{100}$). This contributes to findings from Frances et al. (1994) who showed that a fixed number of paleofloods do not provide additional statistical gain to more extreme flood quantiles compared to, for example, Q_{100}; however every additional paleoflood does reduce the rRMSE of extreme flood quantiles more than that of Q_{100}.

With our analyses, we also show that the utility of paleoflood data in flood frequency analyses is conditional on the level of uncertainty of the inferred paleoflood discharge (Figure 5). One paleoflood with 1σ error up to 0.2 improves the estimates of all the distribution parameters, but these errors compound if more floods are included. For example, when two paleofloods with a 1σ random error of 0.2 are included in a frequency analysis alongside 80 years of instrumental data, the inclusion of paleofloods results in Q_{100} rRMSE estimates that are 30% larger than if no paleofloods are included (Figure 6a). We note, however, that our simulations show that the influence of paleoflood errors on the flood level estimates is highly sensitive to the number of instrumental years and the number of paleofloods included (Figure 6). We also found that systematic over and underestimation of the paleoflood discharges affect parameter estimates differently than nondirectional random error (Figures 5b and 5c). It should be noted, though, that in actual paleoflood estimates all these types of error naturally coexist, behave in multivariate ways, and are likely not normally distributed. Here, we advocate for adding the error of the paleoflood magnitude estimates as an important factor for determining the value of paleoflood data—particularly paleoflood magnitudes inferred from oxbow lake sediments, as they can have discharge magnitude errors of 10%–30%.

As opposed to previous studies, our simulation experiments explore the interaction between distribution parameters and paleoflood data, which paves the way for new approaches to flood frequency analyses, also outside of the alluvial setting. As noted above, the location and scale parameter are best estimated using only instrumental records, particularly when they are >40 years in length (Figures 2c and 2d). Yet, the shape parameter improves significantly with >20 paleoflood events, even though this does not improve, or even worsens, the estimation of the location and scale parameter (Figures 2b and 2f). Thus, using the location and scale parameter estimates from instrumental records only, but including paleoflood data to estimate the shape parameter, will be more beneficial than estimating all parameters using only instrumental or paleoflood data alone.

The application of our findings to the Mississippi River illustrate how the insights gained through our theoretical simulations experiments can both increase precision (i.e., reducing the width of confidence intervals) and accuracy (i.e., changing the shape of the return level plot) of a flood frequency analysis on a real river. In this example, we used paleoflood data to improve the accuracy and precision of unregulated extreme flood estimates. This provides one approach to address a common problem in water resources management, where the period of instrumentation overlaps with significant human impacts to the river and basin that can generate nonstationarities in the annual maxima (Milly et al., 2008). Constraining a flood frequency analysis of unregulated flows with paleoflood data provides a baseline that is useful for comparison to recent and projected hydrologic changes, and in the calibration and validation of hydrologic models over a wider range of observations. An alternative solution is to use a nonstationary flood frequency model, in which distribution parameters are dependent on time (Katz et al., 2013) or temperature (Cheng et al., 2014). The application of nonstationary models in a paleoflood context holds promise as a means to integrate all available data, although the discontinuous nature of paleoflood data, together with uncertainties in their ages and magnitude estimates, are challenges that are yet to be overcome.

Finally, we note that paleoflood and historical flood data are useful beyond their application in flood frequency analysis—even in cases where the flood estimates are qualitative or are associated with large uncertainties. For example, changes in the frequency of paleoflood events provide a means to identify the role of climate variability in mediating flood hazard or place recent floods in a longer-term context (e.g., Fuller et al., 2019; Harden, O'Connor, & Driscoll, 2015; Toonen et al., 2017; Knox, 2000; Munoz, Giosan, Therrell, et al., 2018). Our study focuses on the use of paleoflood data to improve flood frequency analysis, and we build on prior work to describe a set of best practices around their application in this context.

6. Conclusions and Recommendations

Our analyses provide a set of recommendations for the use of hydrologic information to improve flood frequency analysis. These recommendations are as follows:

1. The use of paleoflood data in flood frequency analysis in alluvial settings is most effective when instrumental discharge records are short (<40 years in length); estimates of extreme floods do not meaningfully improve when >10 paleoflood events are included in the frequency analysis. The inclusion of more than 10 paleoflood does help for estimating flood quantiles larger than Q_{500}.
2. It is not recommended to use multiple paleoflood data in alluvial settings if the instrumental records are >60 years in length, the (pre)historical period is short, and paleoflood records have discharge estimates errors >20%. If the paleoflood discharge error is <20%, it is recommended to not use more than two paleofloods. One can further reduce these errors by including other flood frequency techniques such as binomial observations, hydrologic bounds, or ranged magnitude discharges.

Data Availability Statement

Data in this study are available through the United States Geological Survey (USGS) Water Data for the Nation (https://waterdata.usgs.gov/nwis). The R-code used for the analyses in this article are available from the Zenodo open respiratory (http://doi.org/10.5281/zenodo.4587899).

U.S. Geological Survey (USGS), 2020.

Acknowledgments
The authors thank Willem Toonen, Paul Hudson, Ed Beighley, Auroop Ganguly, Dick Bailey for valuable discussion and comments on this work. In addition, we thank John England and two anonymous reviewers, who provided thorough feedback which significantly improved this study. This project was supported by grants from the U.S. National Science Foundation (EAR-1804107 and EAR-1833200).

References

Asselman, N. E. M., & Middelkoop, H. (1995). Floodplain sedimentation: quantities, patterns and processes. *Earth Surface Processes and Landforms*, *20*(6), 481–499. https://doi.org/10.1002/esp.3290200602

Baker, V. R. (1987). Paleoflood hydrology and extraordinary flood events. *Journal of Hydrology*, *96*(1–4), 79–99. https://doi.org/10.1016/0022-1694(87)90145-4

Baker, V. R. (2008). Paleoflood hydrology: Origin, progress, prospects. *Geomorphology*, *101*(1–2), 1–13. https://doi.org/10.1016/j.geomorph.2008.05.016

Benito, G., Brázdil, R., Herget, J., & Machado, M. J. (2015). Quantitative historical hydrology in Europe. *Hydrology and Earth System Sciences*, *19*(8), 3517–3539. https://doi.org/10.5194/hess-19-3517-2015

Benito, G., Lang, M., Barriendos, M., Llasat, M. C., Francés, F., Ouarda, T. ., & Bobée, B. (2004). Use of systematic, palaeoflood and historical data for the improvement of flood risk estimation. Review of scientific methods. *Natural Hazards*, *31*(3), 623–643.

Blainey, J. B., Webb, R. H., Moss, M. E., & Baker, V. R. (2002). Bias and information content of paleoflood data in flood-frequency analysis. *Ancient Floods, Modern Hazards: Principles and Applications of Paleoflood Hydrology*, *5*, 161–174.

Bomers, A., Schielen, R. M. J., & Hulscher, S. J. M. H. (2019). Decreasing uncertainty in flood frequency analyses by including historic flood events in an efficient bootstrap approach. *Natural Hazards and Earth System Sciences*, *19*(8), 1895–1908. https://doi.org/10.5194/nhess-19-1895-2019

Camillo, C. A. (2012). *Divine providence: The 2011 flood in the Mississippi River and tributaries project*.

Castellarin, A., Kohnová, S., Gaál, L., Fleig, A., Salinas, J. L., Toumazis, A., et al. (2012). *A review of applied-statistical methods for flood-frequency analysis in Europe*. The Centre for Ecology & Hydrology (FloodFreq COST Action ES0901). Retrieved from http://nora.nerc.ac.uk/id/eprint/19286/

Cheng, L., AghaKouchak, A., Gilleland, E., & Katz, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Climatic Change*, *127*(2), 353–369. https://doi.org/10.1007/s10584-014-1254-5

Cohn, T. A., & Stedinger, J. R. (1987). Use of historical information in a maximum-likelihood framework. *Journal of Hydrology*, *96*(1–4), 215–223. https://doi.org/10.1016/0022-1694(87)90154-5

Ely, L. L., & Baker, V. R. (1985). Reconstructing paleoflood hydrology with slackwater deposits: Verde River, Arizona. *Physical Geography*, *6*(2), 103–126. https://doi.org/10.1080/02723646.1985.10642266

England, J. F., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas, W. O. J., Veilleux, A. G., et al. (2019). Guidelines for determining flood flow frequency bulletin 17C. In *U.S. Geological Survey techniques and methods, Book 4 chap. B5*. US Geological Survey. https://doi.org/10.3133/tm4B5

England, J. F., Salas, J. D., & Jarrett, R. D. (2003). Comparisons of two moments-based estimators that utilize historical and paleoflood data for the log Pearson type III distribution. *Water Resource Research*, *39*(9), 1243. https://doi.org/10.1029/2002wr001791

Frances, F., Salas, J. D., & Boes, D. C. (1994). Flood Frequency Analysis with systematic and historical or paleoflood data based on the two-parameter general extreme value models. *Water Resource Research*, *30*(6), 1653–1664. https://doi.org/10.1029/94WR00154

Fuller, I. C., Macklin, M. G., Toonen, W. H., Turner, J., & Norton, K. (2019). A 2000 year record of palaeofloods in a volcanically-reset catchment: Whanganui River, New Zealand. *Global and Planetary Change*, *181*, 102981. https://doi.org/10.1016/j.gloplacha.2019.102981

Greenwood, J. A., Landwehrau, J. M., Matalasau, N. C., & Wallisau, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resource Research*, *15*(5), 1049–1054. https://doi.org/10.1002/9781118445112.stat00803

Guo, S. L., & Cunnane, C. (1991). Evaluation of the usefulness of historical and palaeological floods in quantile estimation. *Journal of Hydrology*, *129*(1–4), 245–262. https://doi.org/10.1016/0022-1694(91)90053-K

Harden, T. M., O'Connor, J. E., & Driscoll, D. G. (2015). Late Holocene flood probabilities in the Black Hills, South Dakota with emphasis on the Medieval Climate Anomaly. *Catena*, *130*, 62–68. https://doi.org/10.1016/j.catena.2014.10.002

Harden, T. M., O'Connor, J. E., Driscoll, D. G., & Stamm, J. F. (2011). Flood-frequency analyses from paleoflood investigations for Spring, Rapid, Boxelder, and Elk Creeks, Black Hills, western South Dakota: *U.S. Geological Survey Scientific Investigations Report 2011–5131*, 136.

Hosking, J. R. M., & Wallis, J. R. (1986a). Paleoflood hydrology and flood frequency analysis. *Water Resource Research*, *22*(4), 543–550. https://doi.org/10.1029/WR022i004p00543

Hosking, J. R. M., & Wallis, J. R. (1986b). The value of historical data in flood frequency analysis. *Water Resource Research*, *22*(11), 1606–1612. https://doi.org/10.1029/WR022i011p01606

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

Jones, A. F., Macklin, M. G., & Brewer, P. A. (2012). A geochemical record of flooding on the upper River Severn, UK, during the last 3750 years. *Geomorphology*, *179*, 89–105. https://doi.org/10.1016/j.geomorph.2012.08.003

Katz, R. W. (2013). Statistical methods for nonstationary extremes. In *Extremes in a changing climate* (pp. 15–37). Springer.

Kidson, R., & Richards, K. S. (2005). Flood frequency analysis: Assumptions and alternatives. *Progress in Physical Geography: Earth and Environment*, *29*(3), 392–410. https://doi.org/10.1191/0309133305pp454ra

Klemeš, V. (1993). *Probability of extreme hydrometeorological events — A different approach*, In Extreme hydrological events: Precipitation, floods and droughts (Vol. *213*, pp. 167–176). IAHS Publications.

Knox, J. C. (2000). Sensitivity of modern and Holocene floods to climate change. *Quaternary Science Reviews*, *19*(1–5), 439–457. https://doi.org/10.1016/S0277-3791(99)00074-8

Kochel, R. C., & Baker, V. R. (1982). Paleoflood hydrology. *Science*, *215*(4531), 353–361. https://doi.org/10.1126/science.215.4531.353

Kuczera, G. (1996). Correlated rating curve error in flood frequency inference. *Water Resources Research*, *32*(7), 2119–2127. https://doi.org/10.1029/96WR00804

Kuczera, G. (1999). Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. *Water Resources Research*, *35*(5), 1551–1557. https://doi.org/10.1029/1999WR900012

Lam, D., Thompson, C., Croke, J., Sharma, A., & Macklin, M. (2017). Reducing uncertainty with flood frequency analysis: The contribution of paleoflood and historical flood information. *Water Resources Research*, *53*, 2312–2327. https://doi.org/10.1002/2016WR019959

Leigh, D. S. (2018). Vertical accretion sand proxies of gaged floods along the upper Little Tennessee River, Blue Ridge Mountains, USA. *Sedimentary Geology*, *364*, 342–350. https://doi.org/10.1016/j.sedgeo.2017.09.007

Levish, D. R., England, J. F., Jr, Klawon, J. E., & O'Connell, D. R. H. (2003). *Flood hazard analysis for Seminoe and Glendo dams, Kendrick and North Platte projects, Wyoming, Final Report*. Bureau of Reclamation.

MacDonald, N., Kjeldsen, T. R., Prosdocimi, I., & Sangster, H. (2014). Reassessing flood frequency for the Sussex Ouse, Lewes: The inclusion of historical flood information since AD 1650. *Natural Hazards and Earth System Sciences*, *14*(10), 2817–2828. https://doi.org/10.5194/nhess-14-2817-2014

Madsen, H., Lawrence, D., Lang, M., Martinkova, M., & Kjeldsen, T. R. (2013). A review of applied methods in europe for flood-frequency analysis in a changing environment. *The*Centre for Ecology & Hydrology (FloodFreq COST Action ES0901). Retrieved from: https://www.wmo.int/pages/prog/hwrp/publications/Floodfreq_report.pdf

Mallakpour, I., & Villarini, G. (2015). The changing nature of flooding across the central United States. *Nature Climate Change*, *5*(3), 250–254. https://doi.org/10.1038/nclimate2516

Meade, R. H., & Moody, J. A. (2010). Causes for the decline of suspended-sediment discharge in the Mississippi River system, 1940–2007. *Hydrological Processes*, *24*(1), 35–49. https://doi.org/10.1002/hyp.7477

Merz, R., & Blöschl, G. (2003). A process typology of regional floods. *Water Resources Research*, *39*(12), 1340. https://doi.org/10.1029/2002WR001952

Merz, R., & Blöschl, G. (2008). Flood frequency hydrology: 1. Temporal, spatial, and causal expansion of information. *Water Resources Research*, *44*(8), W08432. https://doi.org/10.1029/2007WR006744

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity is dead: Whither water management? *Science*, *319*, 573. https://doi.org/10.1126/science.1151915

Mossa, J. (1996). Sediment dynamics in the lowermost Mississippi River. *Engineering Geology*, *45*(1–4), 457–479. https://doi.org/10.1016/S0013-7952(96)00026-9

Munoz, S. E., Giosan, L., Blusztajn, J., Rankin, C., & Stinchcomb, G. E. (2019). Radiogenic fingerprinting reveals anthropogenic and buffering controls on sediment dynamics of the Mississippi River system. *Geology*, *47*(3), 271–274. https://doi.org/10.1130/G45194.1

Munoz, S. E., Giosan, L., Therrell, M. D., Remo, J. W. F., Shen, Z., Sullivan, R. M., et al. (2018). Climatic control of Mississippi River flood hazard amplified by river engineering. *Nature*, *556*(7699), 95–98. https://doi.org/10.1038/nature26145

Munoz, S. E., Gruley, K. E., Massie, A., Fike, D. A., Schroeder, S., & Williams, J. W. (2015). Cahokia's emergence and decline coincided with shifts of flood frequency on the Mississippi River. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(20), 6319–6324. https://doi.org/10.1073/pnas.1501904112

NOAA National Centers for Environmental Information (NCEI). (2020). *U.S. Billion-Dollar weather and climate disasters*. Rerieved from https://www.ncdc.noaa.gov/billions/

O'Connell, D. R. (2005). Nonparametric Bayesian flood frequency estimation. *Journal of Hydrology*, *313*(1–2), 79–96. https://doi.org/10.1016/j.jhydrol.2005.02.005

O'Connell, D. R. H., Ostenaa, D. A., Levish, D. R., & Klinger, R. E. (2002). Bayesian flood frequency analysis with paleohydrologic bound data. *Water Resources Research*, *38*(5), 16-1–16-3. https://doi.org/10.1029/2000WR000028

O'Connor, J. E., Atwater, B. F., Cohn, T. A., Cronin, T. M., Keith, M. K., Smith, C. G., & Mason, R. R. (2014). Assessing inundation hazards to nuclear powerplant sites using geologically extended histories of riverine floods, tsunamis, and storm surges. *U.S. Geological Survey Scientific Investigations Report 2014–5207*, 66. http://dx.doi.org/10.3133/sir20145207

Payrastre, O., Gaume, E., & Andrieu, H. (2011). Usefulness of historical information for flood frequency analyses: Developments based on a case study. *Water Resources Research*, *47*, W08511. https://doi.org/10.1029/2010WR009812

Pinter, N., Jemberie, A. A., Remo, J. W., Heine, R. A., & Ickes, B. S. (2008). Flood trends and river engineering on the Mississippi River system. *Geophysical Research Letters*, *35*, L23404. https://doi.org/10.1029/2008GL035987

Reis, D. S., Jr, & Stedinger, J. R. (2005). Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology*, *313*(1–2), 97–116. https://doi.org/10.1016/j.jhydrol.2005.02.028

Remo, J. W., Pinter, N., & Heine, R. (2009). The use of retro-and scenario-modeling to assess effects of 100+ years river of engineering and land-cover change on Middle and Lower Mississippi River flood stages. *Journal of Hydrology*, *376*(3–4), 403–416. https://doi.org/10.1016/j.jhydrol.2009.07.049

Salinas, J. L., Castellarin, A., Viglione, A., Kohnová, S., & Kjeldsen, T. R. (2014). Regional parent flood frequency distributions in Europe - Part 1: Is the GEV model suitable as a pan-European parent? *Hydrology and Earth System Sciences*, *18*(11), 4381–4389. https://doi.org/10.5194/hess-18-4381-2014

Schendel, T., & Thongwichian, R. (2017). Considering historical flood events in flood frequency analysis: Is it worth the effort? *Advances in Water Resources*, *105*, 144–153. https://doi.org/10.1016/j.advwatres.2017.05.002

Smith, A. B. (2020). *2010-2019: A landmark decade of U.S. Billion-dollar weather and climate disasters. Beyond the data*. Retrieved from https://www.climate.gov/news-features/blogs/beyond-data/2010-2019-landmark-decade-us-billion-dollar-weather-and-climate

Smith, L. M., & Winkley, B. R. (1996). The response of the Lower Mississippi River to river engineering. *Engineering Geology*, *45*(1–4), 433–455. https://doi.org/10.1016/S0013-7952(96)00025-7

Stedinger, J. R., & Cohn, T. A. (1986). Flood frequency analysis with historical and paleoflood information. *Water Resources Research*, *22*(5), 785–793. https://doi.org/10.1029/WR022i005p00785

Stedinger, J. R., Vogel, R. M., & Foufoula-Georgiou, E. (1993). Frequency analysis of extreme events. In D. R. Maidment, (Ed.), *Handbook of hydrology* (pp. 18.1–18.66). McGraw-Hill.

Strupczewski, W. G., Kochanek, K., & Bogdanowicz, E. (2014). Flood frequency analysis supported by the largest historical flood. *Natural Hazards and Earth System Sciences*, *14*(6), 1543–1551. https://doi.org/10.5194/nhess-14-1543-2014

Strupczewski, W. G., Kochanek, K., & Bogdanowicz, E. (2017). Historical floods in flood frequency analysis: Is this game worth the candle? *Journal of Hydrology*, *554*, 800–816. https://doi.org/10.1016/j.jhydrol.2017.09.034

Tao, B., Tian, H., Ren, W., Yang, J., Yang, Q., He, R., et al. (2014). Increasing Mississippi river discharge throughout the 21st century influenced by changes in climate, land use, and atmospheric CO_2. *Geophysical Research Letters*, *41*, 4978–4986. https://doi.org/10.1002/2014GL060361

Thorndycraft, V. R., Benito, G., Rico, M., Sopeña, A., Sánchez-Moya, Y., & Casas, A. (2005). A long-term flood discharge record derived from slackwater flood deposits of the Llobregat River, NE Spain. *Journal of Hydrology*, *313*(1–2), 16–31. https://doi.org/10.1016/j.jhydrol.2005.02.003

Toonen, W. H. J., Kleinhans, M. G., & Cohen, K. M. (2012). Sedimentary architecture of abandoned channel fills. *Earth Surface Processes and Landforms*, *37*(4), 459–472. https://doi.org/10.1002/esp.3189

Toonen, W. H. J., Winkels, T. G., Cohen, K. M., Prins, M. A., & Middelkoop, H. (2015). Lower Rhine historical flood magnitudes of the last 450 years reproduced from grain-size measurements of flood deposits using End Member Modelling. *Catena*, *130*, 69–81. https://doi.org/10.1016/j.catena.2014.12.004

Toonen, W. H., Munoz, S. E., Cohen, K. M., & Macklin, M. G. (2020). High-resolution sedimentary paleoflood records in alluvial river environments: A review of recent methodological advances and application to flood hazard assessment. In *Palaeohydrology* (pp. 213–228). Springer.

U.S. Geological Survey (USGS). (2020). *USGS water data for the nation*. Retrieved from https://waterdata.usgs.gov/nwis/

Van der Wiel, K., Kapnick, S. B., Vecchi, G. A., Smith, J. A., Milly, P. C. D., & Jia, L. (2018). 100-year Lower Mississippi floods in a global climate model: Characteristics and future changes. *Journal of Hydrometeorology*, *19*(10), 1547–1563. https://doi.org/10.1175/ jhm-d-18-0018.1

Vogel, R. M., & Fennessey, N. M. (1993). L moment diagrams should replace product moment diagrams. *Water Resources Research*, *29*(6), 1745–1752. https://doi.org/10.1029/93WR00341

Vogel, R. M., & Wilson, I. (1996). Probability distribution of annual maximum, mean and minimum streamflows in the United States. *Journal of Hydrologic Engineering*, *1*(4), 69–76. https://doi.org/10.1061/(asce)1084-0699(1996)1:2(69)

Wang, B., & Xu, Y. J. (2018). Decadal-scale riverbed deformation and sand budget of the last 500 km of the Mississippi river: Insights into natural and river engineering effects on a large alluvial river. *Journal of Geophysical Research: Earth Surface*, *123*, 874–890. https://doi.org/10.1029/2017JF004542

Wang, Q. J. (1990). Unbiased estimation of probability weighted moments and partial probability weighted moments from systematic and historical flood information and their application to estimating the GEV distribution. *Journal of Hydrology*, *120*(1–4), 115–124. https://doi.org/10.1016/0022-1694(90)90145-N

Wilhelm, B., Ballesteros Cánovas, J. A., Macdonald, N., Toonen, W. H., Baker, V., Barriendos, M., et al. (2019). Interpreting historical, botanical, and geological evidence to aid preparations for future floods. *Wiley Interdisciplinary Reviews: Water*, *6*(1), e1318. https://doi.org/10.1002/wat2.1318

References From the Supporting Information

Peel, M. C., Wang, Q. J., Vogel, R. M., & McMahon, T. A. (2001). The utility of L-moment ratio diagrams for selecting a regional probability distribution. *Hydrological Sciences Journal*, *46*(1), 147–155. https://doi.org/10.1080/02626660109492806