Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

Mixture Gumbel models for extreme series including infrequent phenomena

Thomas Rodding Kjeldsen, Hyunjun Ahn, Ilaria Prosdocimi & Jun-Haeng Heo

To cite this article: Thomas Rodding Kjeldsen, Hyunjun Ahn, Ilaria Prosdocimi & Jun-Haeng Heo (2018) Mixture Gumbel models for extreme series including infrequent phenomena, Hydrological Sciences Journal, 63:13-14, 1927-1940, DOI: 10.1080/02626667.2018.1546956

To link to this article: https://doi.org/10.1080/02626667.2018.1546956

HYDROLOGICAL SCIENCES JOURNAL
2018, VOL. 63, NOS. 13–14, 1927–1940
https://doi.org/10.1080/02626667.2018.1546956

Mixture Gumbel models for extreme series including infrequent phenomena

Thomas Rodding Kjeldsen[a], Hyunjun Ahn[b], Ilaria Prosdocimi[c] and Jun-Haeng Heo[b]

[a]Department of Architecture and Civil Engineering, University of Bath, Bath, UK; [b]Department of Civil and Environmental Engineering, Yonsei University, Seoul, South Korea; [c]Department of Mathematical Sciences, University of Bath, Bath, UK

ABSTRACT

A Gumbel mixture distribution is proposed for modelling extreme events from two different mechanisms: one phenomenon occurring annually and one occurring infrequently. A new Monte Carlo simulation procedure is presented and used to assess the consequence of fitting traditional Gumbel or GEV models to annual maximum series originating from two different populations. The results show that mixture models are preferred to single-population models when the two populations are very different. The Gumbel mixture model was applied to annual maximum 24-hour rainfall data from 64 South Korean raingauges, split into events generated by typhoon and non-typhoon rainfall. The results show that the use of a mixture model provides a more accurate description of the observed data than the Gumbel distribution, but is comparable to the GEV model. The theoretical and practical results highlight the need for more robust methods for identifying the underlying populations before mixture models can be recommended.

ARTICLE HISTORY
Received 8 May 2018
Accepted 10 October 2018

EDITOR
R. Woods

ASSOCIATE EDITOR
E. Volpi

KEYWORDS
flood frequency estimation; mixture distribution; typhoon; South Korea

1 Introduction

Flood frequency analysis is typically conducted by fitting a two-, three- or four-parameter statistical distribution to an annual maximum series of rainfall or flow events from which a design event of a return period T is estimated as the $(1 - 1/T)$ quantile in the chosen distribution. However, progress in hydrological science, data availability and statistical methods, as well as the emergence of new challenges related to environmental change and the prediction of very extreme events, has prompted some researchers to advocate a break from traditional methods as outlined above and a move towards a process-based approach to design flood estimation (e.g. Alila and Mtiraoui, 2002, Merz and Blöschl 2008, Barth *et al.* 2017). This leads naturally to the exploration of statistical models representing data originating from different underlying processes. Indeed, there are multiple examples in the literature of mixture-type distributions being used for frequency analysis of observed series of hydrological extremes. For example, Waylen and Woo (1982) used the product of two Gumbel distributions to represent annual maximum floods originating from a mixture of rainfall and snowmelt events. Rossi *et al.* (1984) presented the two-component extreme value (TCEV) distribution to describe highly skewed flood data from 39 Italian catchments. Calenda *et al.* (2009) proposed a mixture model for an annual maximum series on the River Tiber in Rome consisting of a normal distribution for ordinary floods and a Gumbel distribution for "exceptional floods", classifying the two types of events based on an event magnitude threshold value. Villarini and Smith (2010) argued that floods in the eastern United States arise from distinctly different weather systems influencing the tail behaviour of the observed series. Murphy (2001) analysed streamflow data from Massachusetts, USA, and defined three distinct types of events: ordinary, tropical cyclone and ice-jam release. He found that the existence of the different types of floods caused lack of fit when applying the standard flood frequency estimation technique described in Bulletin 17B. Yoon *et al.* (2013) used a mixture model to represent annual maximum rainfall events in South Korea, consisting of a mixture of events generated by typhoon and non-typhoon weather systems. A common theme in all these studies is that they advocate the use of mixture distributions based on case studies of observed records and hydrometeorological considerations.

However, mixture models typically involve the estimation of more parameters than single population models, which can be problematic in terms of model performance. Cunnane (1985) argued that it is not satisfactory to arbitrary postulate the existence of two or more populations as a way of achieving a more flexible model. These concerns were also raised by

CONTACT Thomas Rodding Kjeldsen t.r.kjeldsen@bath.ac.uk
Supplemental data for this article can be accessed here.

Rossi *et al.* (1984), who warned against the use of the TCEV distribution for at-site analysis due to the high degree of uncertainty associated with design quantiles derived from four-parameter distributions. Similarly, Serago and Vogel (2018) highlighted that earlier studies into flood frequency methods found that parameter parsimonious models generally have lower mean square error (MSE) than more complex models. The *Flood Studies Report* (NERC, 1975) did not specifically pursue mixture distributions with British flood data but highlighted potential practical issues associated with the definition of distinctly different event types. Despite the widespread interest in mixture models as a way of progressing flood hydrology and the reservations expressed above, no scientific assessment of the robustness of mixture distributions when applied to hydrological extremes has yet been conducted. In the light of this knowledge gap, the aim of this study is to investigate how useful mixture distributions are in describing annual maximum series originating from mixture populations. In particular, given the sampling uncertainty involved in the estimation of model parameters in at-site analysis, how different should the two populations be before it is advisable to consider a mixture distribution over more traditional single population models such as Gumbel or GEV? First, the assumptions behind mixture models are reviewed, and a new Monte Carlo simulation procedure proposed. Next, the consequences of choosing a single population model are examined when, in reality, floods are generated by a mix of two different processes. Finally, the impact of using mixture models is investigated by considering 24-hour annual maximum series of rainfall events from South Korea, classified as either typhoon or non-typhoon events.

2 Frequency models of mixed extreme phenomena

In this section, we introduce and discuss the assumptions underpinning mixture models, with emphasis on the occurrence of infrequent events, and describe procedures for estimating the model parameters from annual maximum series.

2.1 Model description

Consider two different flood generating processes, each consistently resulting in a peak event within each water year, for example, annual maximum floods resulting from snowmelt events or rainfall, as considered in Waylen and Woo (1982). The magnitude of the annual maximum event resulting from each mechanism is considered as independent random variables X_1 and X_2, with cumulative distribution functions (cdf) given as F_1 and F_2, respectively. The annual maximum event, X, is then defined as the maximum value of X_1 and X_2, and the cdf given as:

$$\begin{aligned} F_X(x) &= P\{\max(X_1, X_2) \le x\} \\ &= P\{X_1 \le x \cap X_2 \le x\} = F_1(x)F_2(x) \end{aligned} \quad (1)$$

where the last equals sign implies that the annual maximum values generated by each of the two processes are independent. If F_1 and F_2 are two-parameter Gumbel distributions, then Equation (1) becomes the TCEV distribution as defined by Rossi *et al.* (1984). Fitting of this model requires two distinct series of maximum events to be identified, and a distribution fitted to each series. Using a TCEV distribution is therefore only reasonable if each of the two processes generates an event in every year, e.g. at least one rainfall and one snowmelt flood event in each year.

However, it is not always the case that both processes generate events every year, as some processes might represent a more infrequent phenomenon that does not occur every year at every location. Examples of such infrequent events might include tropical storms or ice-jam release (Murphy 2001, Macdonald *et al.* 2006, Lindenschmidt *et al.* 2018), atmospheric rivers (Lavers *et al.* 2011, Barth *et al.* 2017) or typhoons (Yoon *et al.* 2013, Shin *et al.* 2015). In such cases, a cdf of the mixture population can be derived from the law of total probability. Consider an annual maximum series where the value in each year is generated by either Process 1 or Process 2. Whether the annual maximum for a given year is generated by Process 1 or Process 2 is the result of a random experiment that can have only two outcomes: the annual maximum value in this year is generated by Process 1 (event E_1) or it is generated by Process 2 (event E_2), where the terminology "event" refers to the outcome of a random trial. In each year, the event E_1, defined as the annual maximum value being generated by Process 1, has a probability of occurring $P\{E_1\} = 1 - \omega$. The complementary event, E_2, that the annual maximum value is a result of Process 2, has a probability $P\{E_2\} = \omega$. It is assumed here that Type 2 events are less frequent than Type 1 events, i.e. $P\{E_2\}<P\{E_1\}$, which is the same as $\omega<0.50$. It is also assumed that only Type 1 or Type 2 exists, i.e. $P\{E_2\} + P\{E_1\} = 1$. The cdf of the annual maximum series X can now be defined according to the law of total probability:

$$\begin{aligned} F_X(x) &= P\{X_1 \le x|E_1\}P\{E_1\} \\ &\quad + P\{X_2 \le x|E_2\}P\{E_2\} \\ &= G_1(x)(1-\omega) + G_2(x)\omega \end{aligned} \quad (2)$$

where the probability distributions, G_1 and G_2, are defined as conditional distributions, and thus different from the unconditional distributions, F_1 and F_2, used in the first approach (Equation (1)). The corresponding probability density function is defined as:

$$f_X(x) = g_1(x)(1-\omega) + g_2(x)\omega \tag{3}$$

where g_1 and g_2 are the corresponding conditional probability distribution functions of the Type 1 and Type 2 events. Using unconditional distributions with the second approach is not correct and can result in misleading results. For example, Yoon *et al.* (2013) appear to have used unconditional distributions with a mixture-model framework to represent annual maximum daily rainfall in South Korea, classified as originating from typhoon and convective rainfall. Due to the inclusion of years where no or only minor impact of typhoons on rainfall was recorded, their study came to the surprising, and counterintuitive, conclusion that design rainfall estimates would be higher if typhoon events were not considered in the analysis; this despite the largest events on record being caused by typhoon events. The relaxation of the requirement for the mixture model in Equation (2) that a Type 2 event occurs in each and every year is a key difference with the TCEV model in Equation (1), which assumes that both Type 1 and Type 2 events occur in each year.

2.2 Parameter estimation

In this study it is assumed that the conditional distributions in the mixture model (Equation (2)) are two-parameter Gumbel distributions, each with a cdf defined as:

$$G_i(x) = \exp\left(-\exp\left[\frac{x-\mu_i}{\alpha_i}\right]\right), i = 1,\ 2 \tag{4}$$

where μ and α are the location and shape parameters, respectively. The model parameters for each of the two flood processes can be estimated using a variety of methods, including the method of moments, method of L-moments or likelihood-based methods. The idea behind a mixture model is that the annual maximum series consists of two or more distinctly different types of events. Therefore, it is argued here that first step should involve separating the annual maximum series into two sub-samples based on classification of each event as being either Type 1 or Type 2. Next, the conditional distributions G_1 and G_2 should be fitted to each of the two sub-samples separately. Finally, the two conditional distributions are combined through Equation (2). In this study the method of L-moments is used for estimation of the location and scale parameter for each process, i.e.:

$$\alpha = \frac{\lambda_2}{\ln 2} \tag{5a}$$

$$\mu = \lambda_1 - \gamma\alpha \tag{5b}$$

where λ_1 and λ_2 are the first- and second-order L-moments, respectively, and $\gamma = 0.5772$ is Euler's constant (Hosking and Wallis 1997). It is possible to define the likelihood function for the combined mixture distribution in Equation (2) and find the combination of five parameters $(\mu_1, \alpha_1, \mu_2, \alpha_2, \omega)$ that maximizes the likelihood function. However, this approach is akin to fitting a five-parameter distribution directly to the data without considering the different types of events, not unlike fitting any other five-parameter distribution to the data, such as a Wakeby distribution (Park *et al.* 2001).

According to Equation (2), the weight parameter ω in the mixture model is the probability that in a - given year the annual maximum (AM) event is a Type 2 event, $P\{E_2\}$, where again it is assumed that Type 2 events are less frequent than Type 1 events, i.e. ω<0.50. A naive estimator of the weight parameter is the ratio between the number of observed Type 2 events and the total number of events, i.e.:

$$\hat{\omega} = \frac{\text{Number of Type 2 AM events}}{\text{Total number of AM events}} \tag{6}$$

However, a closer inspection of how Type 1 and 2 events are generated in the annual maximum series is required to ensure that the correct model features are captured by the estimator. For a Type 2 event to become the annual maximum event (an E_2 event), two conditions must be satisfied: (i) an event of Type 2 must have occurred (this event is denoted F_2 with associated probability $P\{F_2\}$), and (ii) the magnitude of the Type 2 event must exceed the magnitude of the annual maximum Type 1 event occurring in the same year ($X_2 > X_1$). Thus, assuming that the event F_2 and the magnitude of the Type 2 event are independent, the probability $P\{E_2\}$ is given as:

$$\begin{aligned} P\{E_2\} &= P\{F_2 \cap X_2 - X_1 > 0\} \\ &= P\{F_2\}P\{X_2 - X_1 > 0\} \end{aligned} \tag{7}$$

As specified above, Type 2 events are more infrequent but typically larger than Type 1 events, i.e. $P\{E_2\}$ <0.5, $P\{X_2 > X_1\} > 0.5$ and $0 \le P\{F_2\} \le 1$. Evaluation of Equation (7) requires the statistical distribution of the difference between two Gumbel distributed random variables $Z = X_2 - X_1$. Details of the derivation of the probability distribution function (pdf) for Z is shown in the Appendix and the solution is given as:

$$f_Z(z) = \alpha_1^{-1}\exp\left(\frac{\mu_2}{\alpha_2} + \frac{z+\mu_1}{\alpha_1}\right)\Omega(\mu_1, \alpha_1, \mu_2, \alpha_2, z) \tag{8}$$

where $\mu_1, \alpha_1, \mu_2, \alpha_2$ are the location and scale parameters of the two Gumbel distributions representing Type 1 and Type 2 events, and Ω is a function defined as:

$$\Omega(\mu_1, \alpha_1, \mu_2, \alpha_2, z) = \int_0^\infty u^{\frac{\alpha_2}{\alpha_1}}\exp\left(-\exp\left(\frac{\mu_2}{\alpha_2}\right)u - \exp\left(\frac{\mu_1+z}{\alpha_1}\right)u^{\frac{\alpha_2}{\alpha_1}}\right)\mathrm{d}u \tag{9}$$

The function Ω is evaluated numerically and, finally, numerical integration of the pdf provides the required estimate of the cumulative probability, $P\{Z > 0\}$.

The introduction of Equation (7) has implications for model specification and design of a Monte Carlo simulation procedure, as outlined in the next section.

3 Monte Carlo simulation study of model choice

Calculating design flood events requires a choice on the flood frequency distribution to be used, typically a two-, three- or four-parameter distribution such as the Gumbel, GEV or kappa distributions (e.g. Salinas *et al.* 2014, Kjeldsen *et al.* 2017). It is well known that the choice of distribution requires a trade-off between ability to fit the data and parameter parsimonious models to constrain the uncertainty of the design events (e.g. Laio *et al.* 2009). To investigate the importance of fitting the correct mixture model containing different types of events, a bespoke two-step Monte Carlo procedure is developed for generating an annual maximum series with the correct mixture of Type 1 and 2 events, as well as maintaining the constraint that a Type 2 event is only included if its magnitude exceeds the magnitude of the Type 1 event for the same year. This procedure is subsequently used to compare the performance of commonly used distributions for describing annual maximum series originating as a result of events from two different populations.

3.1 A new Monte Carlo simulation procedure for infrequent events

To generate an annual maximum series of length n containing a mixture of Type 1 and Type 2 events according to the weight parameter ω, the following steps were implemented:

Step 1 Using the Gumbel distribution specified for Type 1 events, a Type 1 AM event is generated for each of the $t = 1, \ldots, n$ years.

Step 2 For each of the $t = 1, \ldots, n$ years in turn the occurrence of a Type 2 event is generated with probability $P\{F_2\}$. In practice this is implemented as n independent Bernoulli distributions with parameter $P\{F_2\}$.

Having specified the mixture model according to Equation (2), this probability is defined per Equation (7) as:

$$P\{F_2\} = \frac{\omega}{P\{X_2 - X_1 > 0\}} \tag{10}$$

where the denominator is derived as described in Section 2 based on the two specified Gumbel distributions. If a Type 2 event occurs, then the associated magnitude is generated based on the Gumbel distribution of the Type 2 events. Finally, the actual annual maximum event for this particular year is defined as the largest of either the Type 1 from Step 1, or the Type 2 event generated in Step 2.

Examples of generated annual maximum series are shown in Figure 1 as flood frequency curves for a range of parameter values of the underlying mixture distribution. The parameter values are representative of the two populations identified in the South Korea rainfall data presented in Section 4 of this paper. As expected, the more different the two populations become in terms of the location and scale parameters of the two Gumbel distributions, the more separated the two samples become. For example, when the location parameter of the Type 2 process is twice as large as for the Type 1 process $(\mu_2/\mu_1 = 2)$, then the largest events are almost exclusively Type 2 events.

Figure 1 also shows the quantile function of the fitted Gumbel mixture distribution, clearly demonstrating the flexibility of the distribution to represent different sample compositions.

The estimator introduced in Equation (10) was validated using the Monte Carlo simulation procedure and compared to the naïve estimator in Equation (6). Annual maximum series of length between 25 and 1000 years were generated from a Gumbel mixture model with fixed parameters $(\mu_2/\mu_1 = 1.5, \alpha_2/\alpha_1 = 1.5, \omega = 0.20)$. For each record length a total of $M = 500$ individual annual maximum series were generated, with the information recorded on whether the annual maximum is the

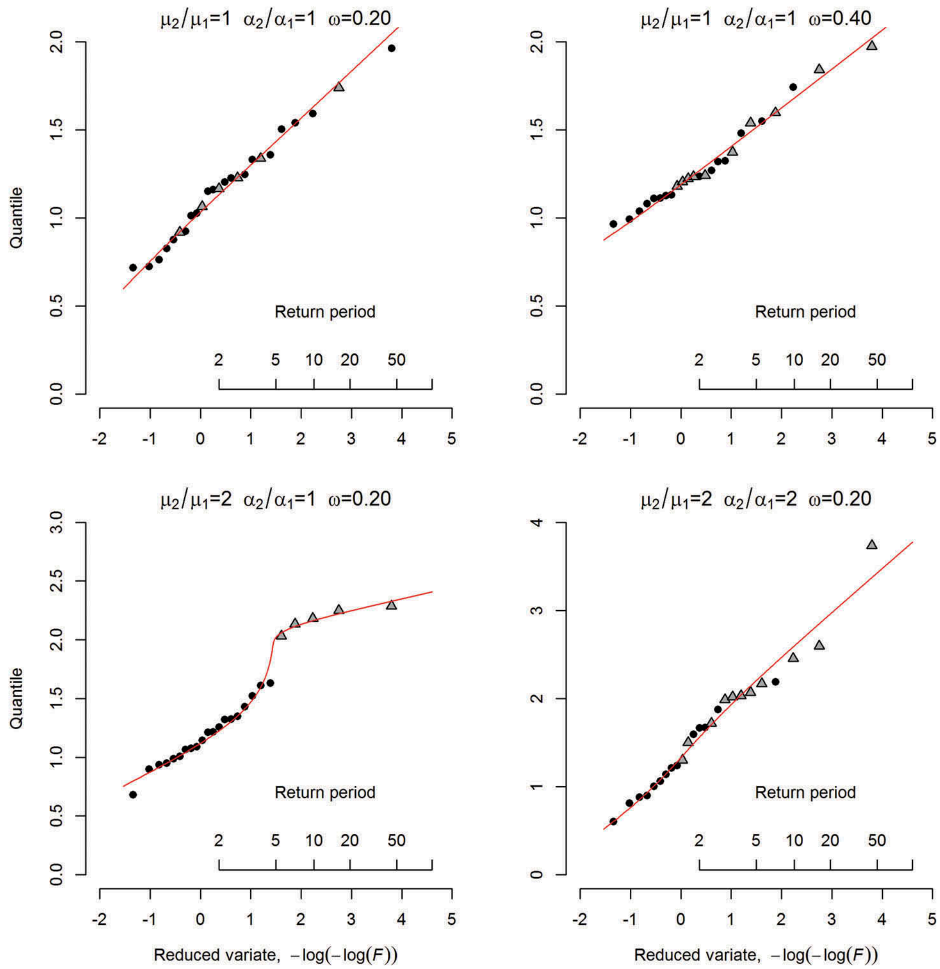

Figure 1. Examples of annual maximum series covering $n = 25$ years derived from a mixture of Type 1 and Type 2 events, including flood frequency curves fitted to the sample data using the Gumbel mixture model combined with the method of L-moments. Black dots represent Type 1 events and triangles are Type 2 events.

result of a Type 1 or a Type 2 event, For each series the five parameters of the Gumbel mixture model (including the weight parameter ω) were estimated using L-moments for the parameter of each individual Gumbel distribution as per Equations (5a) and (5b), and the estimator in Equation (6) for the weight parameter ω. Estimates of the adjusted weight derived using Equation (10) were obtained with sample estimates of $(\mu_1, \mu_2)$and(α_1, α_2) used to calculate $P\{X_2 - X_1 > 0\}$. Figure 2 shows the median of the $M = 500$ series generated for each record length for both the weight parameter ω estimated directly from the generated samples using Equation (6), and the adjusted weight estimated using Equation (10) plotted against record length.

The results given in Figure 2 show that the adjusted estimate (Equation (10)) ensures that the estimated weight parameter represents more closely the specified population value of the mixture distribution ($\omega = 0.20$). In contrast, the naïve estimator consistently underestimates the true model parameters as it ignores the requirement for the infrequent phenomena simultaneously to both occur (F_2) and generate an event larger than the Type 1 events ($X_2 - X_1 > 0$) in order for the annual maximum to be a Type 2 event. Note also that the relatively large sampling uncertainty of the weight parameter (the 95% confidence interval shown by the shaded area in Fig. 2) is a manifestation of the difficulty of obtaining robust results from a statistical analysis of infrequent events.

3.2 Choice of distribution

Monte Carlo experiments were conducted to compare the robustness (as measured by the bias and root mean square error) of the five-parameter Gumbel mixture model (GMM) with that of other, commonly used, candidate distributions, specifically the two-parameter Gumbel distribution (GUM) and the three-parameter GEV

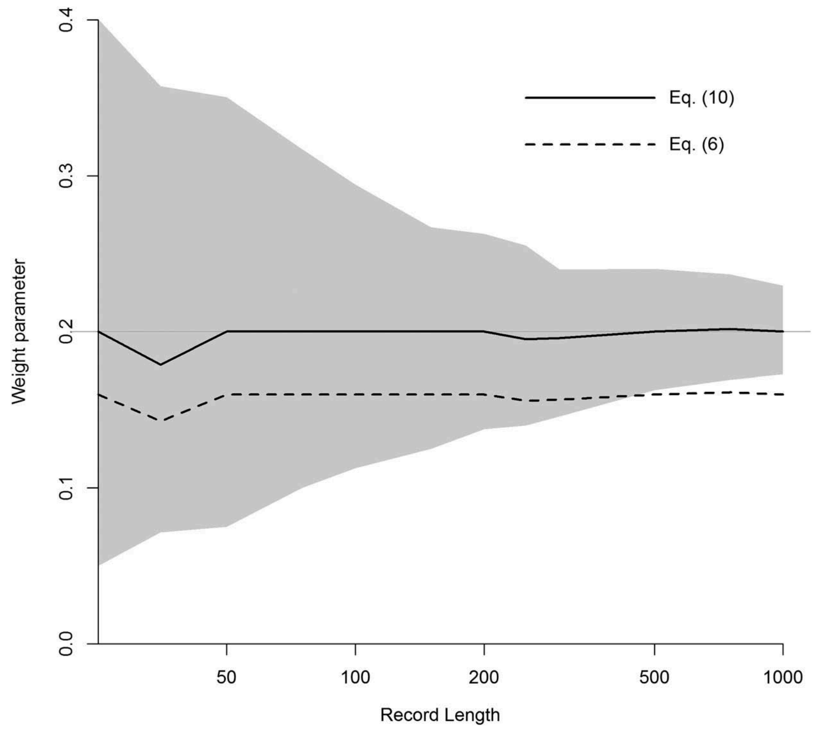

Figure 2. Median of the weight parameter ω estimated using the naïve estimator (dashed line) and the adjusted estimator (solid line). The shaded area represents the 95% confidence interval around the adjusted estimator.

distribution. The model parameters for each of the candidate distributions were estimated using the method of L-moments, as described by Hosking and Wallis (1997).

The performance of each distribution was quantified using the bias and root mean square error (RMSE), as recommended by Stedinger *et al.* (1993). The quantity of interest in this study is the estimated design event for a 100-year event. Based on $M = 10\ 000$ Monte Carlo generated annual maximum series of record length n, Bias and RMSE are calculated as:

$$\text{Bias} = \frac{1}{M}\sum_{m=1}^{M}(\hat{x}_m - x) \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{x}_m - x)^2} \tag{12}$$

where $\hat{x}_m$ is the estimated design event for the mth Monte Carlo generated annual maximum series, and x is the true design event as specified by the mixture model. No explicit analytical expression for the quantile of the Gumbel mixture model exists, and instead the design event with a return period T is derived numerically using bisection to solve the equation below for the design quantile x_T:

$$\begin{aligned} F(x_T) &= \exp\left(-\exp\left[\frac{x_T - \mu_1}{\alpha_1}\right]\right)(1-\omega) \\ &\quad + \exp\left(-\exp\left[\frac{x_T - \mu_2}{\alpha_2}\right]\right)\omega \\ &= 1 - \frac{1}{T} \end{aligned} \tag{13}$$

The performance (Bias and RMSE) of the GMM was compared to results obtained by fitting two-parameter Gumbel and three-parameter GEV distributions to the same annual maximum series using the method of L-moments (Hosking and Wallis 1997). The comparisons were conducted by assuming the underlying distributions are Gumbel mixture distributions classified according to the ratios between location parameters, $\mu_2/\mu_1 = 1.0, 1.5, 2.0$ (assuming the location parameter of Type 2 events is equal to or larger than that for Type 1 events), and the ratios between scale parameters, $\alpha_2/\alpha_1 = 1.0, 1.5, 2.0$, assuming the standard deviation for Type 2 events is equal to or larger than that for Type 1 events. Finally, the weight parameter ω varies between 0.01 and 0.50, thus assuming that Type 2 events are less frequent than Type 1 events. These parameter ratios were selected based on analysis of annual maximum series of 24-hour rainfall events extracted from 64 raingauges located in South Korea, recording extreme rainfall associated with both typhoon events and non-typhoon events.

Figure 3 shows the Bias of the five-parameter (GMM), two-parameter Gumbel and three-parameter GEV distributions plotted as a function of weight parameter for all nine combinations of μ_2/μ_1 and α_2/α_1. The results are based on a record length of $n = 50$ years. When the distributions of Type 1 and Type 2 events are very similar, i.e. both μ_2/μ_1 and α_2/α_1 are equal to 1 (upper left corner), then the five-parameter Gumbel mixture model exhibits a high level of bias compared to more parameter parsimonious Gumbel or GEV models. Maintaining the ratio of the location parameters at 1, but increasing the ratio of the scale parameters to 1.5, both the Gumbel and GEV distributions are generally less biased than the mixture model, except for weight parameters close to zero. For all other combinations, the mixture model is generally less biased than the other candidate distributions, especially for weight parameters exceeding 0.10. For all considered cases, these results show that, if distinctly different types of events can be identified in an annual maximum series, and the statistical properties (mean and standard deviation) of the events generated by the two different processes are sufficiently different, then there is merit, if minimizing Bias is of interest, in choosing a five-parameter mixture model over more parameter parsimonious and popular alternatives such as Gumbel or GEV distributions. The results in Figure 3 are based on a record length of 50 years; further Monte Carlo simulations (reported in the Supplementary Material) show that the patterns seen in Figure 3 maintain their general form for a range of record lengths typically reflected in observed records and for different design events of different rarity.

Next, the performance of the mixture model was assessed based on the RMSE as defined in Equation (12). Unlike Bias, RMSE can only take positive values. Therefore, the results from the Monte Carlo experiment are reported in Figure 4 as the ratio between RMSE for any distribution and the RMSE obtained by using the mixture distribution, e.g. $\mathrm{RMSE}_{\mathrm{GUM}}/\mathrm{RMSE}_{\mathrm{GMM}}$. Values of the ratio of less than 1 suggest that the candidate distribution has a lower RMSE than the GMM, even if the true underlying population consists of two distinct flood populations.

Except for isolated instances where $\mu_2/\mu_1 > 1$ and $\alpha_2/\alpha_1 = 1$, the RMSE is consistently lowest for the Gumbel distribution and highest for the GEV distribution, with the Gumbel mixture model located between the two. These results suggest that the added uncertainty introduced by having to estimate the relatively

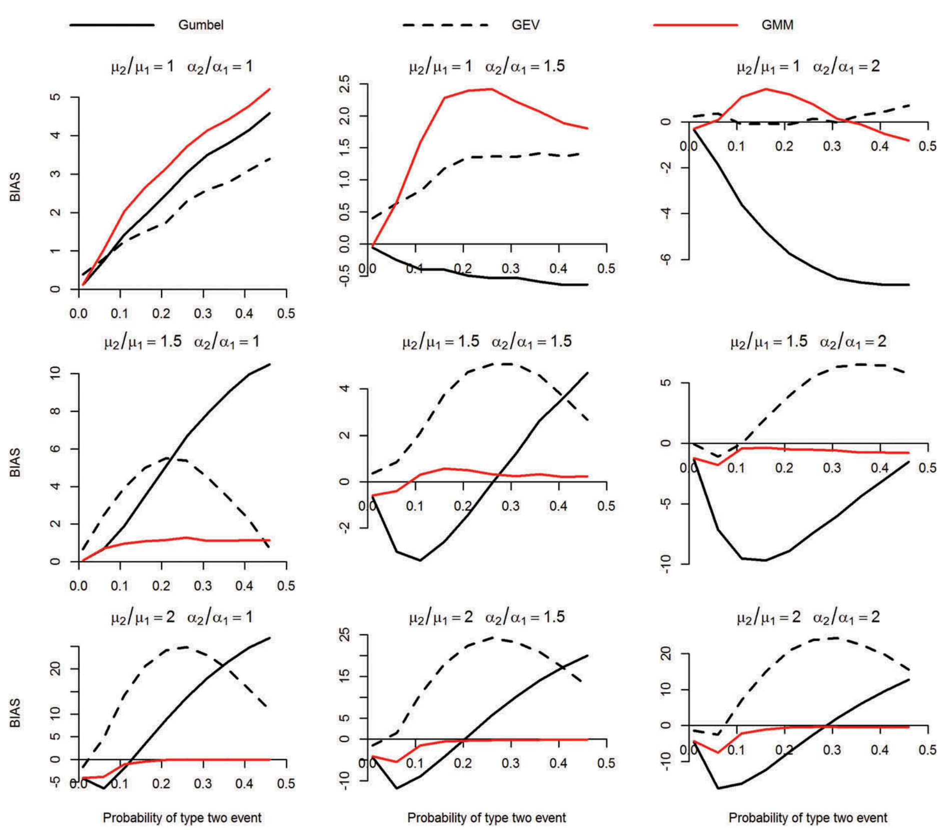

Figure 3. Bias for the 100-year event of Gumbel mixture, two-parameter Gumbel and three-parameter GEV distributions plotted as a function of weight parameter for all nine combinations of μ_2/μ_1 and α_2/α_1.

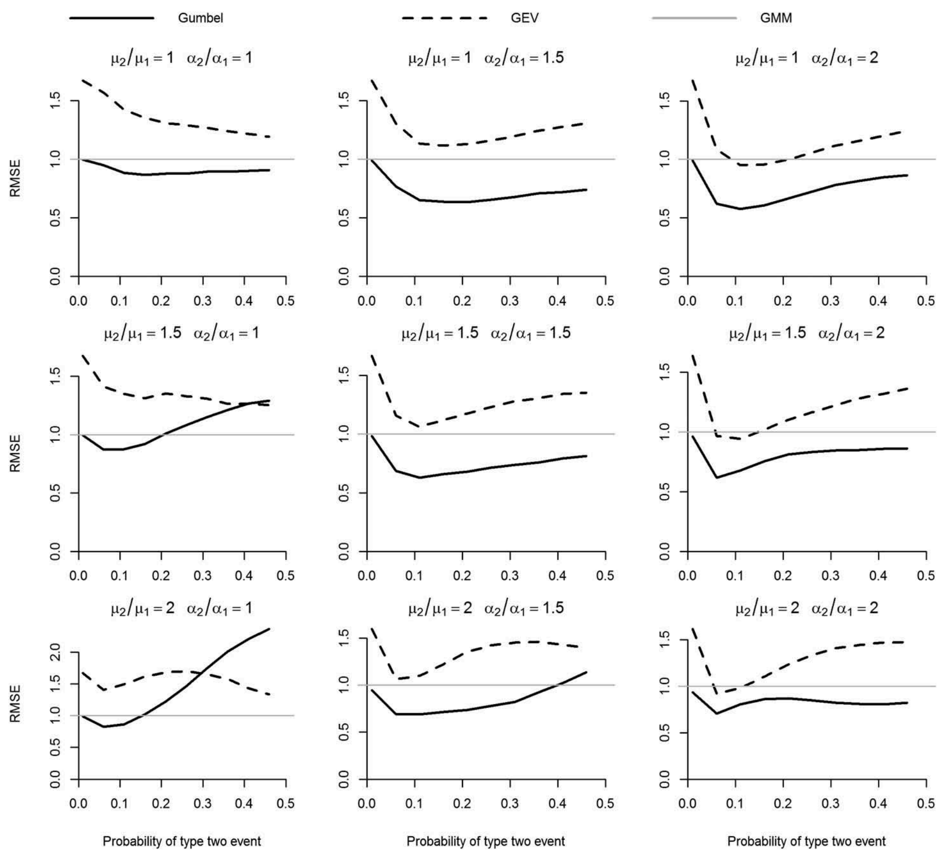

Figure 4. Ratios between RMSE for the 100-year event for candidate distributions (Gumbel: solid line, GEV: dashed line) and the Gumbel mixture model as a function of weight parameter for all nine combinations of μ_2/μ_1 and α_2/α_1.

high number of model parameters of the mixture distribution (five in total) is offsetting the advantages obtained by specifying the correct model structure. As for Bias, the effect on RMSE of using different record lengths and target return periods was found not to impact on the general conclusions derived from Figure 4 with regard to the relative performance of the distributions. These results support conclusions from other studies (e.g. Lu and Stedinger 1992) that parameter parsimonious models are often characterized by lower RMSE than more complex models.

4 Case study: Korean rainfall

The practical implications of choosing a mixture model were investigated through analysis of extreme rainfall data from South Korea. High-quality data from 64 raingauges with long-term records (more than 20 years) were acquired from the Korea Meteorological Administration (KMA). The raingauges are evenly distributed over inland and mountainous, and coastal areas in South Korea (Fig. 5). For each raingauge the annual maximum series of 24-h rainfall events were extracted, resulting in series with durations of between 30 and 56 years.

Rainfall in South Korea is characterized by strong seasonality, with typically 80% of the total annual rainfall occurring in the summer rainy season between May

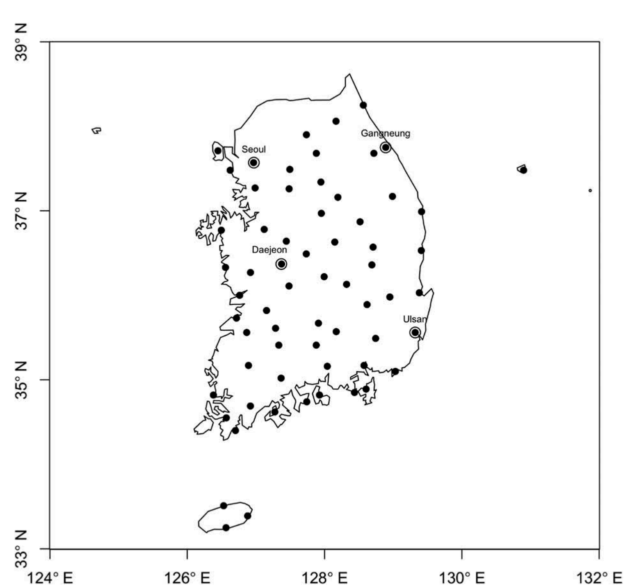

Figure 5. Map showing the location across South Korea of the 64 raingauges used in this study.

and October. Another important aspect of rainfall in the region is that rainfall events are generally the result of either non-typhoon rainfall or recorded during the passing of a typhoon. Usually, the monsoon season begins in late June and lasts until late July, whereas typhoons normally occur between August and September. The largest rainfall event on record is 870 mm in 24 h measured during Typhoon Rusa in August 2002 on the northeast coast (Park and Lee 2007). For a more detailed discussion of the rainfall types in South Korea see Yoon *et al.* (2013).

Information is available from the Korean Meteorological Administration (KMA, 2011) as to which raingauges were impacted by typhoon events. This information was used to classify each of the annual maximum events as either a typhoon event (Type 2) or a non-typhoon event (Type 1). An example of how the information from KMA (2011) was used is shown in Figure 6. The example in Figure 6 represents an entire year that has been subdivided into three periods where rainfall is attributed to typhoon events (black) and four periods where rainfall is attributed to non-typhoon events. If the annual maximum event occurs within one of the typhoon periods, it is classified as a Type 2 event. Conversely, if the annual maximum is recorded in a non-typhoon period, then it is classified as a Type 1 event. The number of "black" typhoon periods is based on the actual occurrence of typhoons in any particular year. Figure 7 shows two examples where the annual maximum event for a - particular year is classified as either a typhoon event (Fig. 7(a)) or a non-typhoon event (Fig. 7(b)).

A total of 154 typhoon events were recorded during the period covered by the observed rainfall records (1961–2010), about three per year. Note that the spatio-temporal characteristics of each typhoon are unique. Therefore, the impact of typhoon rainfall is not uniform across raingauges and years. Using this information, the annual maximum series extracted at each of the 64 raingauges were divided into two sub-samples representing events generated by typhoon and non-typhoon rainfall.

The Monte Carlo simulations reported in Section 3 showed that the Gumbel mixture model is generally only preferable for at-site analysis if the two populations are sufficiently different. Using the observed annual maximum series from the 64 raingauges, the implications of assuming a Gumbel mixture distribution for design rainfall estimation rather than a Gumbel or GEV distribution were investigated. The ratio between the 100-year 24-h design rainfall was estimated for each raingauge assuming a Gumbel mixture distribution (GMM), a two-parameter Gumbel distribution and a three-parameter GEV distribution, i.e. $\text{ratio} = Q_{100}^{i}/Q_{100}^{\text{GMM}} \times 100\%, i = \text{GUM}, \text{GEV}$. The ratios of design rainfall estimates were plotted against estimated parameters of the GMM to investigate what aspects of the annual maximum samples most influence the magnitude of the design events. Figure 8 shows the ratios for Gumbel *vs* GMM (top row) and GEV *vs* GMM (bottom) plotted against the estimated weight parameter (column one), the ratios between estimated location parameters (column two), and the ratio between estimated scale parameters (column three).

The results in Figure 8 show that, in general, fitting a GMM adopting a Gumbel distribution will in most cases result in a lower design event, and for GEV in a higher design event. Interestingly, there is a strong relationship between the ratio of scale parameters α_2/α_1 and the ratio between the Gumbel and the Gumbel mixture model, such that increasing the difference between the scale parameters results in a consistent decrease of the ratio between 100-year design event magnitudes. There is a similar, though less striking, relationship observed when the Gumbel ratio is plotted against the ratio of the location parameters μ_2/μ_1. Similar patterns are not observed when considering the ratios of design events derived using the GEV and GMM distributions. This suggests that adopting a Gumbel distribution will potentially result in underestimation of design events, especially where Type 2 events are consistently larger than Type 1 events.

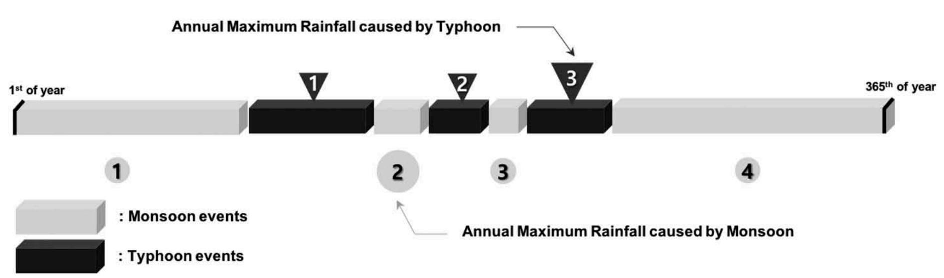

Figure 6. Example of how the year is split into rainfall types using information from KMA (2011).

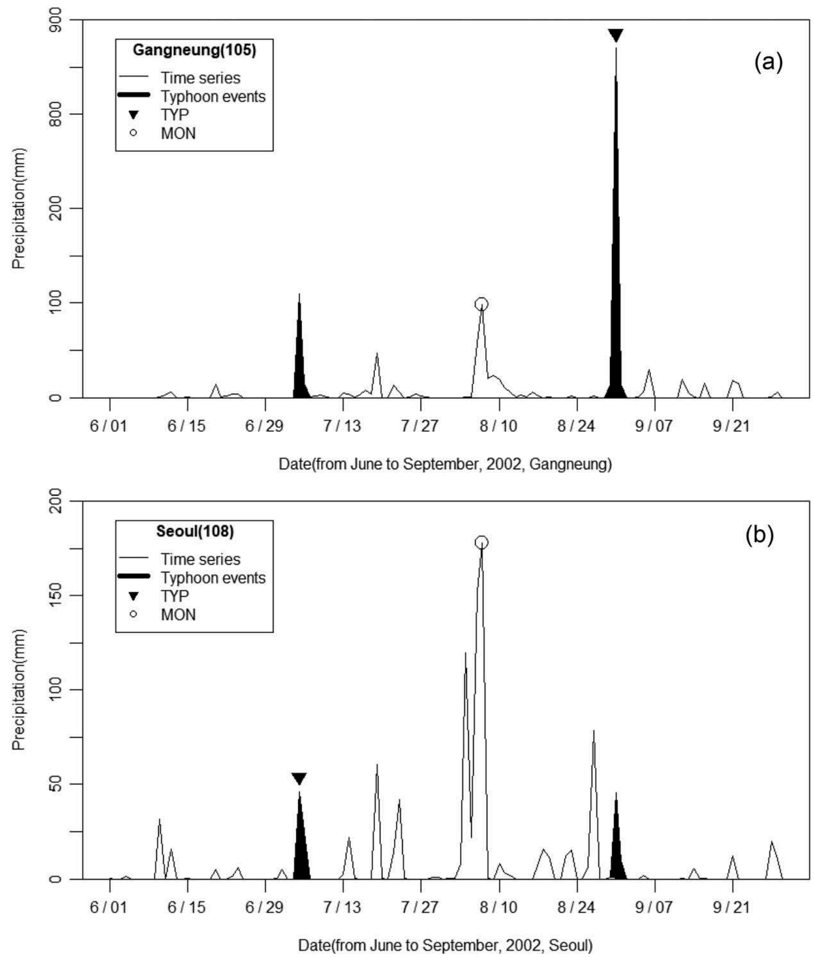

Figure 7. (a) Annual maximum from a typhoon event: Gangneung raingauge (June–September 2002), and (b) annual maximum from a non-typhoon event: Seoul raingauge (June–September 2002).

To illustrate the impact of fitting different distributions, Figure 9 shows the annual maximum 24-h rainfall (AMAX) events plotted against return period, using the Gringorten plotting position, for four different raingauges. Each data point was classified, as described above, as being a result of either a typhoon event (triangle) or a non-typhoon rainfall event (dot). The rainfall frequency curves in Figure 9 show that, in all four cases, the choice of a GEV distribution results in a larger 100-year design event, especially at higher return periods.

5 Discussion

Combining the outcome of the Monte Carlo simulations and the analysis of the Korean rainfall data, the results show that the use of mixture models appears most useful when the two flood populations are distinctly different. In contrast, if the location and shape parameters of the two populations do not differ much, a simple Gumbel model provides a more robust flood frequency model. In contrast, design events derived from the GEV model generally have a higher degree of uncertainty (RMSE), thus highlighting the need for a more detailed understanding of the impacts and trade-offs between model complexity and parameter uncertainty driven by the relatively short data series. This conclusion is based on the use of the flood frequency model as a tool for predicting design flood events of moderate return period (100 years in this study). However, there might be several reasons why a mixture model might be preferred over a single population model, such as studies aimed at defining flood topologies (Merz and Blöschl 2008). Another reason could be in considering the behaviour of the flood frequency curve when extrapolating to very high return periods. For example, De Niel *et al.* (2017) developed an 18-parameter mixed model based on classification of events according to weather type, and used historical events to validate the extrapolation of this mixture model. Also, if a particular type of events is consistently more severe than the other populations, it could be argued that the risk analysis should be focused exclusively on these events only as, for example, infrastructure failure will

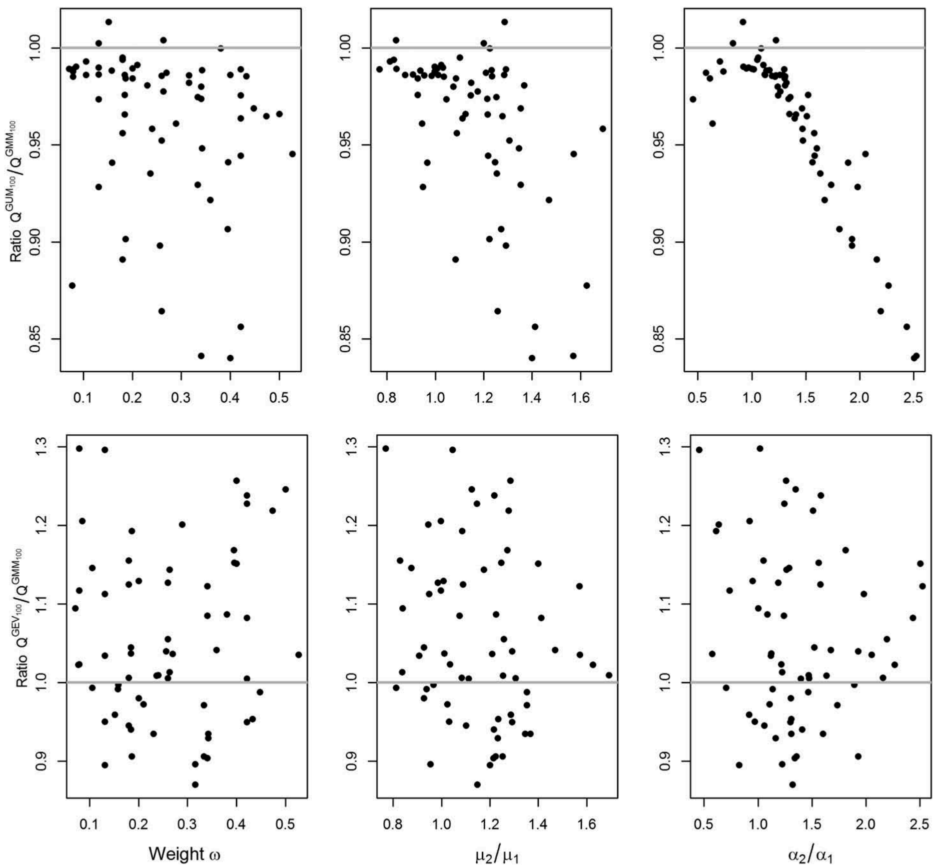

Figure 8. Ratio between 100-year 24-h design events estimated using Gumbel and GMM (top row) and GEV and GMM (bottom row) for 64 raingauges.

inevitably occur as a result of these events rather than the consistently smaller events from a different population.

Several decisions were made to constrain the extent of the Monte Carlo simulation study. This was necessary as the full combination of possible values of the five-parameter mixture model, record length and target return periods would have led to an intangible number of results. Consequently, the results presented were limited by considering only design floods with a return period of 100 years as estimated from a record length of 50 years. The sensitivity of the results with regard to these decisions was investigated and found to be so minor as not to fundamentally change the overall patterns presented (see Supplementary Material).

A new two-step Monte Carlo procedure is presented for use in generating annual maximum series from a two-component mixture model with specified weight parameter. It was necessary to introduce this additional complication in the Monte Carlo simulations to ensure that the generated annual maximum series reflected the correct value of the weight assigned to each of the conditional distributions. While this step resulted in a somewhat more burdensome set of calculations, it is argued here that a more formal definition of the weight parameter and the introduction of a separate probability of a Type 2 event occurring is an original and useful contribution to further studies of rare or infrequent phenomena causing large events.

The case study investigated the impacts of choosing a Gumbel mixture model over a single-population Gumbel or GEV distribution. The results highlight the danger of adopting a single-population Gumbel distribution, as this resulted in consistently lower estimates of design floods than both the GEV and Gumbel mixture model. The choice between GEV and Gumbel mixture model is less clear, with a slight tendency for the GEV distribution to produce higher design flood estimates. A key difficulty was to develop an operational and robust method for classifying events as typhoon or non-typhoon events. Further research is needed to investigate the use of additional data sources to allow for a more objective classification. The advantages brought by mixture distributions in terms of their ability to faithfully describe the extreme generating process must be weighed against the practical complexity of separating the different types of events before suggesting their widespread use.

Finally, it should be noted that the results presented in this study pertain exclusively to the use of mixture

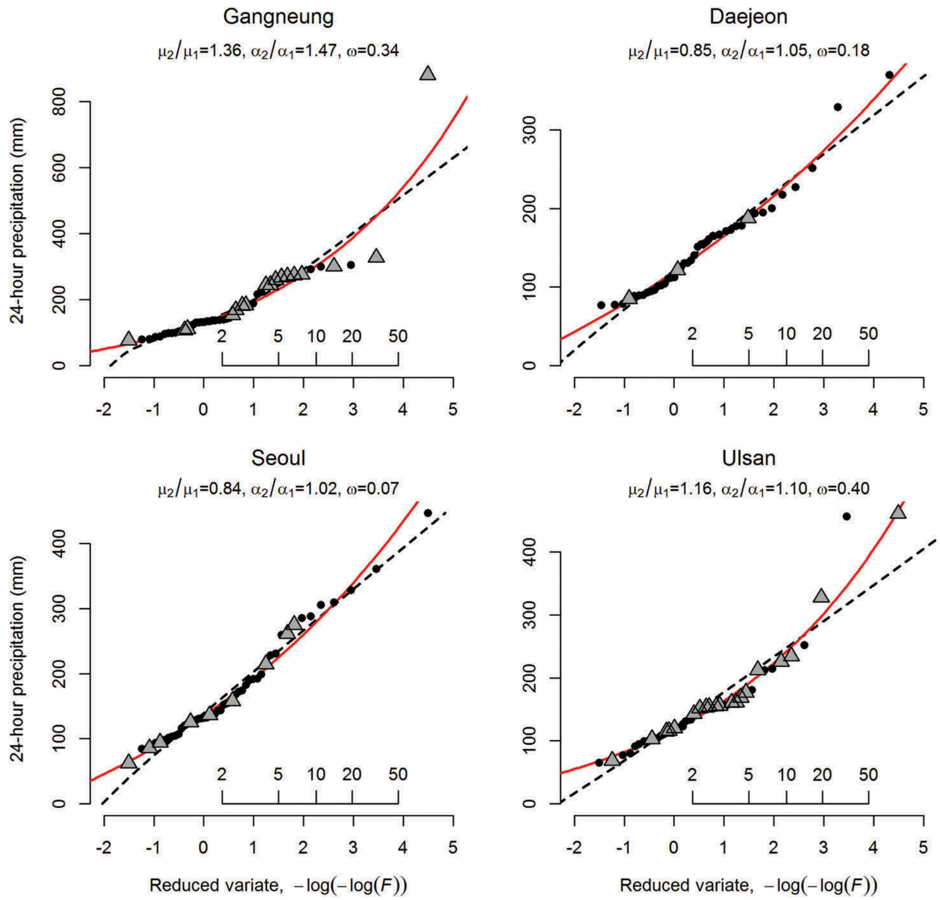

Figure 9. Example of frequency curves for 24-h annual maximum rainfall from four raingauges in South Korea impacted by monsoon and typhoon rainfall events. Triangles: AMAX from Typhoon, dots: AMAX from monsoon rainfall. Solid (red) line: GEV fitted to the entire series, dashed line: mixture Gumbel model fitted to non-typhoon and typhoon events. Location of raingauges is indicated in Figure 5.

models for at-site analysis. Further challenges and developments are needed in order to apply these models in a regional frequency analysis.

6 Conclusions

The aim of this study was to investigate the usefulness of mixture models to describe annual maximum series generated by the existence of two different populations. The conclusions of the study are:

- A mixture model becomes increasingly desirable as the difference between the two populations increases.
- When the two populations are similar or almost similar, a Gumbel distribution is preferable over both a mixture and a GEV model as it tends to give unbiased estimates for the design events of interest.
- A new Monte Carlo simulation procedure was proposed for correctly specifying the occurrence of infrequently occurring extreme events.
- Future research should consider robust procedures enabling objective and practical tools for identification of event types so mixture models can be moved into the regional flood frequency domain.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors acknowledge funding from the University of Bath Global Mobility Grant. Ilaria Prosdocimi is supported by an NPIF Innovation Fellowship from the Natural Environment Research Council (NERC), under the project NE/R013152/1.

References

Alila, Y. and Mtiraoui, A., 2002. Implications of heterogeneous flood-frequency distributions on traditional stream-discharge prediction techniques. *Hydrological Processes*, 16 (5), 1065–1084. doi:10.1002/(ISSN)1099-1085

Barth, N.A., *et al.*, 2017. Mixed populations and annual flood frequency estimates in the western United States: the role of atmospheric rivers. *Water Resources Research*, 53 (1), 257–269. doi:10.1002/2016WR019064

Calenda, G., Mancini, C.P., and Volpi, E., 2009. Selection of the probabilistic model of extreme floods: the case of the River Tiber in Rome. *Journal of Hydrology*, 371 (1), 1–11. doi:10.1016/j.jhydrol.2009.03.010

Cunnane, C., 1985. Factors affecting choice of distribution for flood series. *Hydrological Sciences Journal*, 30 (1), 25–36. doi:10.1080/02626668509490969

De Niel, J., Demarée, G., and Willems, P., 2017. Weather typing-based flood frequency analysis verified for exceptional historical events of past 500 years along the Meuse river. *Water Resources Research*, 53 (10), 8459–8474. doi:10.1002/wrcr.v53.10

Hosking, J.R.M. and Wallis, J.R., 1997. *Regional frequency analysis: an approach based on L-moments*. Cambridge, UK: Cambridge University Press.

Kjeldsen, T.R., Ahn, H., and Prosdocimi, I., 2017. On the use of a four-parameter kappa distribution in regional frequency analysis. *Hydrological Sciences Journal*, 62 (9), 1354–1363. doi:10.1080/02626667.2017.1335400

KMA (Korean Meteorological Administration), 2011. *Typhoon White Book*. Seoul: Korea Meteorological Administration, 358.

Laio, F., Di Baldassarre, G., and Montanari, A., 2009. Model selection techniques for the frequency analysis of hydrological extremes. *Water Resources Research*, 45 (7), W07416. doi:10.1029/2007WR006666

Lavers, D.A., *et al.*, 2011. Winter floods in Britain are connected to atmospheric rivers. *Geophysical Research Letters*, 38, 23. doi:10.1029/2011GL049783

Lindenschmidt, K.E., *et al.*, 2018. Lessons learned from past ice-jam floods concerning the challenges of flood mapping. *International Journal of River Basin Management*, 1–48. just-accepted.

Lu, L.H. and Stedinger, J.R., 1992. Variance of two-and three-parameter GEV/PWM quantile estimators: formulae, confidence intervals, and a comparison. *Journal of Hydrology*, 138 (1–2), 247–267. doi:10.1016/0022-1694(92)90167-T

Macdonald, N., *et al.*, 2006. Historical and pooled flood frequency analysis for the River Tay at Perth, Scotland. *Area*, 38 (1), 34–46. doi:10.1111/area.2006.38.issue-1

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

Murphy, P.J., 2001. Evaluation of mixed-population flood-frequency analysis. *Journal of Hydrologic Engineering*, 6 (1), 62–70. doi:10.1061/(ASCE)1084-0699(2001)6:1(62)

Nadarajah, S. and Kotz, S., 2005. A generalized logistic distribution. *International Journal of Mathematics and Mathematical Sciences*, 2005 (19), 3169–3174. doi:10.1155/IJMMS.2005.3169

Natural Environment Research Council, 1975. Flood Studies Report (5 Volumes). London, UK.

Park, J.S., *et al.*, 2001. Modelling summer extreme rainfall over the Korean peninsula using Wakeby distribution. *International Journal of Climatology*, 21 (11), 1371–1384. doi:10.1002/joc.701

Park, S.K. and Lee, E., 2007. Synoptic features of orographically enhanced heavy rainfall on the east coast of Korea associated with Typhoon Rusa (2002). *Geophysical Research Letters*, 34, 2. doi:10.1029/2006GL028592

Rossi, F., Fiorentino, M., and Versace, P., 1984. Two-component extreme value distribution for flood frequency analysis. *Water Resources Research*, 20 (7), 847–856. doi:10.1029/WR020i007p00847

Salinas, J.L., *et al.*, 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, 4381–4389. doi:10.5194/hess-18-4381-2014

Serago, J.M. and Vogel, R.M., 2018. Parsimonious nonstationary flood frequency analysis. *Advances in Water Resources*, 112, 1–16. doi:10.1016/j.advwatres.2017.11.026

Shin, J.Y., Lee, T., and Ouarda, T.B., 2015. Heterogeneous mixture distributions for modeling multisource extreme rainfalls. *Journal of Hydrometeorology*, 16 (6), 2639–2657. doi:10.1175/JHM-D-14-0130.1

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events. Chapter 18. *In*: D.R. Maidment, ed. *Handbook of hydrology*. New York, NY: McGraw-Hill.

Villarini, G. and Smith, J.A., 2010. Flood peak distributions for the eastern United States. *Water Resources Research*, 46, 6. doi:10.1029/2009WR008395

Waylen, P. and Woo, M.K., 1982. Prediction of annual floods generated by mixed processes. *Water Resources Research*, 18 (4), 1283–1286. doi:10.1029/WR018i004p01283

Yoon, P., Kim, T.W., and Yoo, C., 2013. Rainfall frequency analysis using a mixed GEV distribution: a case study for annual maximum rainfalls in South Korea. *Stochastic Environmental Research and Risk Assessment*, 27 (5), 1143–1153. doi:10.1007/s00477-012-0650-5

Appendix

The probability density function (pdf) of the difference between two independent random variables $Z = X_2 - X_1$ is denoted $f_Z(z)$ and can be derived as:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X1}(z) f_{X2}(x - z) \mathrm{d}x \tag{A1}$$

where f_{X1} and f_{X2} are the pdfs of X_1 and X_2, respectively. In this study the difference between two Gumbel distributed variables is considered as $Z = X_2 - X_1$, each with a pdf given by:

$$f_{X1}(x) = \alpha_1^{-1} \exp\left(-\frac{x-\mu_1}{\alpha_1}\right) \exp\left(-\exp\left[-\frac{x-\mu_1}{\alpha_1}\right]\right) \tag{A2}$$

$$f_{X2}(x) = \alpha_2^{-1} \exp\left(-\frac{x-\mu_2}{\alpha_2}\right) \exp\left(-\exp\left[-\frac{x-\mu_2}{\alpha_2}\right]\right) \tag{A3}$$

where μ_1, μ_2 and α_1, α_2 are the location and scale parameters, respectively, of the Gumbel distributions describing the Type 1 and Type 2 events. Both pdfs exist on the interval $-\infty < x < \infty$. Substituting Equations (A2) and (A3) into (A1) and making the substitution $u = \exp(-x/\alpha_2)$ gives the following pdf for the difference:

$$f_Z(z) = \alpha_1^{-1} \exp\left(\frac{\mu_2}{\alpha_2} + \frac{z+\mu_1}{\alpha_1}\right) \Omega(\mu_1, \alpha_1, \mu_2, \alpha_2, z) \tag{A4}$$

where

$$\Omega(\mu_1, \alpha_1, \mu_2, \alpha_2, z) = \int_0^{\infty} u^{\frac{\alpha_2}{\alpha_1}} \exp\left(-\exp\left(\frac{\mu_2}{\alpha_2}\right) u - \exp\left(\frac{\mu_1 + z}{\alpha_1}\right) u^{\frac{\alpha_2}{\alpha_1}}\right) \mathrm{d}u \tag{9}$$

A similar result was reported by Nadarajah and Kotz (2005), defining the distribution in Equation (A4) as a generalized logistic distribution.