Journal of Hydrology 640 (2024) 131590

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Goodness-of-fit, identifiability and extrapolation: Can the Two-Component Extreme Value distribution be used in at-site flood frequency analysis?

Vincenzo Totaro[a,*], George Kuczera[b], Vito Iacobellis[a]

[a] Dipartimento di Ingegneria Civile, Ambientale, del Territorio, Edile e di Chimica, Politecnico di Bari, Bari 70125, Italy
[b] School of Engineering, University of Newcastle, Callaghan, NSW, Australia

A B S T R A C T

When fitting three-parameter flood frequency models to annual maximum (AM) flood series, the lack-of-fit can be mitigated by censoring potentially influential low flows (PILFs). An alternative and less-studied approach is to apply mixture probability models with four or more parameters, which trade off greater flexibility to fit AM series against the need to deal with degeneracy which arises when there is insufficient information to identify mixture components. However, the issue of degeneracy and the lack of a robust inference framework present a significant barrier to adoption. This study investigated the potential of the most parsimonious mixture model, the four-parameter Two-Component Extreme Value (TCEV) model. A Bayesian framework using Markov chain Monte Carlo sampling was developed to robustly characterize parameter uncertainty even in the presence of degeneracy. Two new posterior diagnostics based on the strength of the TCEV components were developed to aid identification of degeneracy. Armed with a robust inference framework, the study evaluated the potential of TCEV using a case study based on 31 catchments in eastern Australia with records exceeding 70 years. The evaluation used short to long records to compare TCEV and Log Pearson III fit and extrapolative uncertainty. The Log Pearson approach (LP3-PILF) censors PILFs following the approach described in Australian Rainfall and Runoff. With respect to goodness-of-fit, we found in most cases that TCEV fitted AM flood peaks well without the need to censor or stratify data, and overall, no clear difference emerged between TCEV and LP3-PILF. However, and contrary to expectation, TCEV produced high flow quantile confidence intervals consistently narrower than LP3-PILF even in the presence of degeneracy. While more case studies in different regions are needed to confirm the potential of TCEV, this study is a reminder that goodness-of-fit is a necessary but not sufficient criterion for selecting the probability model that best represents flood frequency.

1. Introduction

The current flood frequency analysis (FFA) paradigm mainly involves fitting a parametric probability model to annual maximum (AM) discharge records. A broad consensus has developed over the choice of the most appropriate parametric models with the three-parameter Generalized Extreme Value (GEV) and Log Pearson III (LP3) models seeing widespread usage (Stedinger et al., 1993; Vogel and Wilson, 1996). However, there would be few, if any, who would argue that these models represent the true AM flood distribution, at best they represent an approximation. It is, therefore, not uncommon to find these models fail to achieve a good fit over the whole frequency range.

The issue of lack-of-fit has raised concerns about the relevance of the smallest AM floods to the estimation of large flood risk (Cohn et al., 2013). The fundamental concern was expressed by Klemeš (1986, pp. 183S–184S): “For it is by no means hydrologically obvious why the regime of the highest floods should be affected by the regime of flows in years when no floods occur, why the probability of a severe storm hitting this basin should depend on the accumulation of snow in the few driest winters”.

Recently Lamontagne et al. (2016) observed that: “Annual maximum flood records may contain observations dominated by different physical processes, and we should not extrapolate the distribution of the largest floods from the distribution of small floods, when small floods are likely due to different processes”. Turning to the lack-of-fit issue, they identified two broad strategies to address lack-of-fit. The first retains the three-parameter probability model but adopts estimators that put emphasis on the largest flood observations at the expense of the smaller floods. The second uses a more flexible probability model with the intent of producing a better fit and so obviating the need to weight more heavily larger flood observations.

The first strategy has been the focus of a concerted research effort in recent years with one approach adopted in national guidelines. This approach relies on identifying potentially influential low flows (PILFs) and censoring them when estimating the probability model parameters. In the context of the LP3 distribution, Bulletin 17C (England et al., 2019) has adopted the multiple Grubbs-Beck test for identifying PILFs and censoring them using the EMA estimator (Cohn et al., 1997; Griffis et al., 2004). Likewise Australian Rainfall and Runoff (Ball et al., 2019) recommends the multiple Grubbs-Beck test for identifying PILFs which are

* Corresponding author.
E-mail address: vincenzo.totaro@poliba.it (V. Totaro).

https://doi.org/10.1016/j.jhydrol.2024.131590
Received 22 January 2024; Received in revised form 17 May 2024; Accepted 21 June 2024
Available online 1 July 2024

treated as censored observations in a Bayesian framework. Another approach uses LH moments (Wang, 1997) which generalize L-moments using higher-order statistics that emphasize the upper part of the distribution.

The second strategy, the focus of this paper, is motivated by the mounting evidence that different physical mechanisms are responsible for producing AM floods and hence the flood record contains samples from multiple populations (Alila and Mtiraoui, 2002; Barth et al., 2019).

The parametric modelling of multiple populations in flood records has been largely implemented using mixture distributions (recent examples include Barth et al., 2019; Iacobellis et al., 2011; Kjeldsen et al., 2018; Shin et al., 2015). Mixture distributions provide more flexibility in fitting AM series than three-parameter distributions at the expense of a higher number of parameters requiring inference and, in turn, an increase in uncertainty of high quantile estimates. Waylen et al. (1993) note that mixture distributions can be represented as finite additive or compound models. The most common FFA applications of mixture distributions use an additive mixture of two populations with each population represented by a two or three-parameter distribution. Allowing for the mixture weight, the number of parameters can range from five to seven.

The application of mixture distributions in FFA has been limited by the need for long AM records (Alila and Mtiraoui, 2002). When stratification of AM records is possible using independent information, the need for long AM records is reduced. In such cases, two- or three-parameter distributions can be fitted to the stratified records (Barth et al., 2019; Franks and Kuczera, 2002; Kiem et al., 2003; Micevski et al., 2006; Webb and Betancourt, 1992; Yan et al., 2019). It is noted that Bulletin 17C does not discourage use of well-constructed stratified records. Nonetheless, while the ability to stratify samples reduces the need for long AM records, the stratified samples need to be sufficiently long to identify the individual populations.

Despite the issue of overparameterization and identification of populations, there has only been limited assessment of uncertainty in the parameters of mixture distributions (e.g., Evin et al., 2011; Grego and Yates, 2010; Yan et al., 2019). Evin et al. (2011) used a Gibbs sampler to infer the Bayesian posterior distribution. Yan et al. (2019) coupled the parametric bootstrap with maximum likelihood estimators to approximate the sampling distributions and found this procedure was not robust. To our knowledge, there has been no comprehensive assessment of uncertainty in FFA mixture models and the relationship between identifiability and record length.

The current practice of applying mixture distributions only in cases with long AM records or with shorter AM records when independent information permits stratification, leaves open two mutually related questions: how useful are short to moderate AM records and how robust are current estimation procedures when one population is dominant resulting in poor identifiability of the other population, a situation referred to as *degeneracy*? These questions were investigated by Zheng and Frey (2004) who considered a two-population additive mixture model. They examined the influence of the weight, sample size and degree of separation between populations on identifiability. They found in cases of limited separation between populations, the parametric bootstrap using maximum likelihood estimators could fail and recommended use of a single population.

Within the framework of mixture models, compound models have received far less attention than additive ones. They offer a compromise between the flexibility offered by two different populations and the advantage of a more parsimonious distribution. In return for a reduction in flexibility, compound models offer the prospect of reducing or even eliminating the reliance on stratifying AM records into two populations.

In summary, the mixture distribution approach requires adopting probability models with four or more parameters to provide the additional degrees of freedom to improve model fit. If well chosen, such models may avoid the need to identify and manage PILFs. They may also offer the opportunity to explicitly represent the contribution of different physical processes to AM floods and thus provide a stronger foundation for extrapolation. However, the issue of degeneracy and availability of adequate records present a significant barrier to adoption.

Several compound models have been proposed in the literature, among which there is the Two Component Extreme Value (TCEV) distribution, developed by Rossi et al. (1984). Exploiting the theory of exact distribution of extremes (Todorovic, 1970, 1978), TCEV represents the exact distribution of extremes arising from a mixture of exponential exceedances. With four parameters, TCEV is more parsimonious than additive mixture models using two- or three-parameter components.

The purpose of this paper is twofold. The first is to develop an inference scheme for TCEV that robustly deals with degeneracy and its impact on parameter identifiability. The second is to assess the potential of TCEV in at-site applications using short to long AM records with particular focus on goodness of fit and understanding how degeneracy impacts on parameter and high quantile uncertainty.

Section 2 of this paper describes the TCEV model and reviews approaches used to estimate its parameters. It then develops a Markov chain sampling implementation of the Bayesian framework for at-site analysis. It proposes parameterizations to manage and diagnostics to identify degeneracy and illustrates their application. This represents the first comprehensive treatment of identifiability in mixture distributions.

Section 3 explores the potential of Bayesian TCEV using a case study based in eastern Australia focusing on catchments with long (70 or more years) records. It compares implementation of the TCEV and LP3-PILF approaches for these catchments, first using the full records and then shorter records. The comparison focuses on goodness of fit and extrapolative quantile uncertainty. The paper ends with a critical discussion of the findings and future directions.

2. Robust inference of TCEV parameters using annual maximum data

This section presents and illustrates a Bayesian framework for inferring TCEV parameters that robustly handles degeneracy which arises when the parameters of one component are poorly identified. To our knowledge, this is the first rigorous treatment of identifiability in flood frequency mixture models with four or more parameters. It makes possible the application of TCEV to sites with records much shorter than the long records typically used with the additive mixture flood models and the robust quantification of quantile uncertainty for magnitudes beyond the observed data.

2.1. Conceptual foundation

The TCEV model is a compound mixture model. It assumes there are two processes contributing to the annual maximum flood. However, unlike additive mixture models, there is no explicit assignment of weight to the probability models associated with each process. The conceptual framework underpinning TCEV assumes in any year that each flood process i ($i = 1,2$) generates a random number of events whose number is sampled from a Poisson distribution with parameter λ_i (mean number of events in a year) and whose peak flow is sampled from an exponential distribution with parameter θ_i (mean peak flow). Rossi et al. (1984) showed the distribution of annual maximum floods q resulting from these two processes is:

$$F(q|\boldsymbol{\Theta}) = \begin{cases} \exp\left(- \lambda_1 e^{-\frac{q}{\theta_1}} - \lambda_2 e^{-\frac{q}{\theta_2}} \right) & q > 0 \\ \exp(- \lambda_1 - \lambda_2) & q = 0 \\ 0 & q < 0 \end{cases} \qquad [1]$$

where $\boldsymbol{\Theta}^{\mathrm{T}} = (\lambda_1, \theta_1, \lambda_2, \theta_2)$ with $\lambda_1, \lambda_2 \geq 0$ and $\theta_1, \theta_2 > 0$.

Two important tail behaviors of the TCEV need to be highlighted.

First, the right tail of the TCEV converges to a Gumbel distribution,

$$F(q|\mathbf{\Theta}) \underset{q\to\infty}{\to} \exp\left(-\lambda_r e^{-q/\theta_r}\right), \quad r = \underset{i}{\mathrm{argmax}}(\theta_i, i = 1, 2) \qquad [2]$$

Given the primary interest in FFA is estimation of the right tail, the identifiability of the right-tail component parameters λ_r and θ_r and the goodness-of-fit to the right tail are key considerations.

Second, there is a finite probability that neither process generates an event in a year resulting in a zero annual maximum flood. This may be useful in applications involving arid and semi-arid catchments where flows are ephemeral. However, this means $F(q|\mathbf{\Theta})$ is not differentiable at $q = 0$ requiring the probability density function to be expressed as a generalized probability density

$$f(q|\mathbf{\Theta}) = \begin{cases} \left(\frac{\lambda_1}{\theta_1}\exp\left(-\frac{q}{\theta_1}\right) + \frac{\lambda_2}{\theta_2}\exp\left(-\frac{q}{\theta_2}\right)\right)F(q|\mathbf{\Theta}) & q > 0 \\ \exp(-\lambda_1 - \lambda_2) & q = 0 \\ 0 & q < 0 \end{cases} \qquad [3]$$

2.2. Bayesian framework

Estimation of TCEV parameters has been largely confined to regionalization applications using maximum likelihood and L-moments. Mindful of their experience with Italian data where the largest annual maximum floods were represented by an "outlier" component, Rossi et al. (1984) commented "*when the four parameters of the TCEV distribution are estimated from a single AFS, the uncertainty is great, particularly as regard the parameters of the outlying component*". However, the right-tail component does not necessarily have to be an "outlier". Connell and Pearson (2001) applied TCEV to sites in New Zealand and found distinct dog-leg signatures in Gumbel frequency plots, a clear sign that both components are well-identified. Totaro et al. (2024) applied TCEV to long-record sites in eastern Australia where flood regime is linked with decadal-scale climatic phenomena (Erskine and Warner, 1988; Franks and Kuczera, 2002; Micevski et al., 2006). Like Connell and Pearson (2001) they found both TCEV components were well identified. TCEV was also applied to the seasonal components of AM flood series (Strupczewski et al., 2012).

To date, most of TCEV applications have used either maximum likelihood (ML) or L-moments to fit parameters. The ML method developed by Rossi et al. (1984) and modified by Fiorentino et al. (1987) uses the method of successive approximations to solve the nonlinear equations obtained by setting the derivative of the log-likelihood function to zero. A limitation of this ML approach is that AM records must be strictly positive. Arnell and Beran (1988) investigated an alternative approach using L-moments estimators. In synthetic experiments they found the L-moments estimators failed to reach a solution in a significant proportion of cases (41 % for the UK-based parent and 9 % for the Italian-based parent), while ML found a solution in all cases. They concluded that the failure of the L-moments procedure was due to the limited range of L-skew and L-kurtosis covered by the TCEV distribution.

In view of the concern about the robustness of L-moments estimators, the Bayesian framework was chosen to infer TCEV parameters. Marin et al. (2005) summarize the literature on Bayesian approaches to mixture distributions which have identifiability issues similar to TCEV. Their work suggests that with care, Markov chain sampling of the posterior distribution offers a robust tool to evaluate uncertainty about TCEV parameters even in the case of degeneracy.

A key step in Bayesian inference is the evaluation of the posterior distribution (Gelman et al., 1995)

$$p(\mathbf{\Phi}|D) = \frac{p(\boldsymbol{D}|\Phi)p(\mathbf{\Phi})}{p(\boldsymbol{D})} \qquad [4]$$

where $\mathbf{\Phi}$ represents some transformation of the native TCEV parameters $\mathbf{\Theta}$, $\boldsymbol{D}$ represents the data used to update the prior density $p(\mathbf{\Phi})$, $p(\boldsymbol{D}|\mathbf{\Phi})$ represents the likelihood function which assigns a data-based weight to the prior, and $p(\boldsymbol{D})$ is a normalizing constant independent of $\mathbf{\Theta}$.

A problem peculiar to mixture distributions (Evin et al., 2011; Marin et al., 2005) also arises for TCEV. If the component indices *1* and *2* are swapped, the distribution function $F(q|\mathbf{\Theta})$ and density $f(q|\mathbf{\Theta})$ remain unchanged. To avoid computing a posterior distribution having two regions, one a perfect copy of the other, an additional constraint $\theta_1 < \theta_2$ is imposed. This means that component *2* will describe the right tail behavior of the TCEV. While the imposition of the constraint $\theta_1 < \theta_2$ is well founded in the case of TCEV, Marin et al. (2005) note that imposing constraints on location parameters for mixture distributions with more than two populations can be problematic. If the TCEV were generalized to three or more components, a different strategy would be required to deal with the issue of invariance of $F(q|\mathbf{\Theta})$ under permutation of component indices.

Samples from the posterior are drawn using a Metropolis-Hasting sampler (Gelman et al., 1995). The sampling involves the following steps:

1) Find the maximum posterior estimate $\widehat{\mathbf{\Theta}}$ using the Shuffled Complex Evolution optimization algorithm (Duan et al., 1993) to maximize the log-posterior.
2) Estimate the posterior covariance $\widehat{\mathbf{\Sigma}}$ by computing the Hessian of the log-posterior at $\widehat{\mathbf{\Theta}}$ using central differences. If this fails, assign an identity matrix to the covariance.
3) Generate multiple Metropolis-Hastings chains using a multivariate normal proposal with covariance $\mathbf{\Sigma}_{\boldsymbol{P}}$:
   i. Burn-in phase: Starting at $\widehat{\mathbf{\Theta}}$ with $\mathbf{\Sigma}_{\boldsymbol{P}} = \widehat{\mathbf{\Sigma}}$, update $\mathbf{\Sigma}_{\boldsymbol{P}}$ at specified intervals using generated samples and adjust the covariance scaling factor with the goal of achieving an acceptance rate in a target range.
   ii. Production phase: Generate a specified number of samples using the $\mathbf{\Sigma}_{\boldsymbol{P}}$ tuned in the burn-in phase.
4) Compute Gelman-Rubin statistic to assess the mixing of chains.
5) If the maximum Gelman-Rubin statistic< 1.01, stop. Otherwise go to step 3 and rerun with an increased number of samples.

For well-identified posteriors, 60,000 production samples were found more than adequate with run times of the order of 10 s on an Intel I7 processor. For degenerate posteriors, the number of samples was increased by a factor up to 5 to allow the sampler to visit the entire posterior domain.

2.3. Degeneracy

In mixture distributions, degeneracy occurs when one or more weights approach zero making the affected populations unidentifiable. With TCEV, a similar form of degeneracy can arise. The annual maximum is the largest of the peaks generated by two independent component processes. If one component produces all the annual maximum peaks, the other component is unidentifiable.

Degeneracy can occur in two ways:

1) Component degeneracy occurs when the contribution of component *i* to $F(q|\Theta)$ vanishes for all q in the data $\boldsymbol{D}$: this condition arises when $\lambda_i \exp\left(-\frac{q}{\theta_i}\right) \to 0$ for all $q \in \boldsymbol{D}$. Either $\theta_i \to 0$ or $\lambda_i \to 0$ will satisfy this condition.
2) Gumbel degeneracy occurs when both components become identical: This condition arises when $|\lambda_1 - \lambda_2| \to 0$ and $|\theta_1 - \theta_2| \to 0$. The location constraint $\theta_1 < \theta_2$ forces this condition to manifest itself as component *1* vanishing.

2.4. Identifying degeneracy

When degeneracy occurs, the data $\boldsymbol{D}$ does not contain enough information to meaningfully identify the two components of the TCEV. At best, there may be useful information about the dominant component and little (or none) about the degenerate component. It is therefore essential that such a circumstance be clearly identified and managed.

The robustness of the Metropolis-Hasting sampler can be enhanced by selecting a transformation $\boldsymbol{\Theta}\rightarrow\boldsymbol{\Phi}$ which allows the multivariate normal proposal to efficiently sample feasible parameters over the posterior domain.

In the presence of degeneracy, the posterior is expected to have significant density close to zero for one or both of the degenerate-component parameters. This suggests that a log transformation be used, namely $\boldsymbol{\Phi_k} = \log_e(\boldsymbol{\Theta}_k), k = 1, .., 4$. As the degenerate component parameters $\theta\rightarrow 0$, $\lambda\rightarrow 0$, the parameters in the posterior domain $\log_e\theta\rightarrow -\infty, \log_e\lambda\rightarrow -\infty$. Such a domain ideally suits the unbounded multivariate normal proposal used in the Metropolis-Hastings sampler.

To further improve the performance of the Metropolis-Hastings algorithm in cases of degeneracy, a weak prior is specified. For $\log_e\theta_i$ and $\log_e\lambda_i$, the prior is uniform bounded by -5 and 15. This prior has effectively no influence on the posterior for well-identified cases in this study but limits the feasible domain in degenerate cases. Of particular importance is the lower bound of -5. This bound would only be active in cases of degeneracy and prevents the Metropolis-Hastings algorithm from sampling values of θ_i and λ_i infinitesimally close to zero, a domain of no practical relevance.

2.4.1. Posterior degeneracy diagnostics

The posterior has distinctly different signatures when a TCEV component is well identified and when it is degenerate. When both components are well identified, the posterior can be approximated by a multivariate normal distribution, which is the asymptotic form of the posterior. In the case of degeneracy, the multivariate normal signature is absent. It is replaced by large regions of near-uniform density where the degenerate component parameters ensure the component has near-zero probability of contributing to the annual maximum.

In addition to study of the posterior, it is useful to compute diagnostics that alert to the presence of degeneracy. We present two posterior diagnostic measures for identifying degeneracy and a discussion of their use. We note that Strupczewski et al. (2012) assessed the contribution of seasonal components to AM quantile uncertainty in their investigation of a two-season TCEV model.

The first diagnostic is based on the strength of the contribution of component 2 to $F(q|\boldsymbol{\Theta})$

$$\Psi_2(\boldsymbol{\Theta}, q) = \frac{\lambda_2 \exp\left(-\frac{q}{\theta_2}\right)}{\lambda_1 \exp\left(-\frac{q}{\theta_1}\right) + \lambda_2 \exp\left(-\frac{q}{\theta_2}\right)} \quad [5]$$

Frequency plots of $\Psi_2(\boldsymbol{\Theta}, q)$ for a range of q evaluated over the posterior can be used to assess the strength of component 2. When both components are well identified, one expects to see the strength $\Psi_2(\boldsymbol{\Theta}, q)$ vary from 0 to 1 over the range of q. Furthermore, $\Psi_2(\boldsymbol{\Theta}, q)$ helps identify the flow beyond which the right tail of the TCEV is described by a Gumbel distribution.

The second diagnostic is the probability that an annual maximum is produced by component 2. Beran et al. (1986) show that the probability of the annual maximum produced by component 2 is

$$P_2(\boldsymbol{\Theta}) = -\frac{\lambda_*}{\theta_*} \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \lambda_*^j \Gamma\left(\frac{j+1}{\theta_*}\right) \quad [6]$$

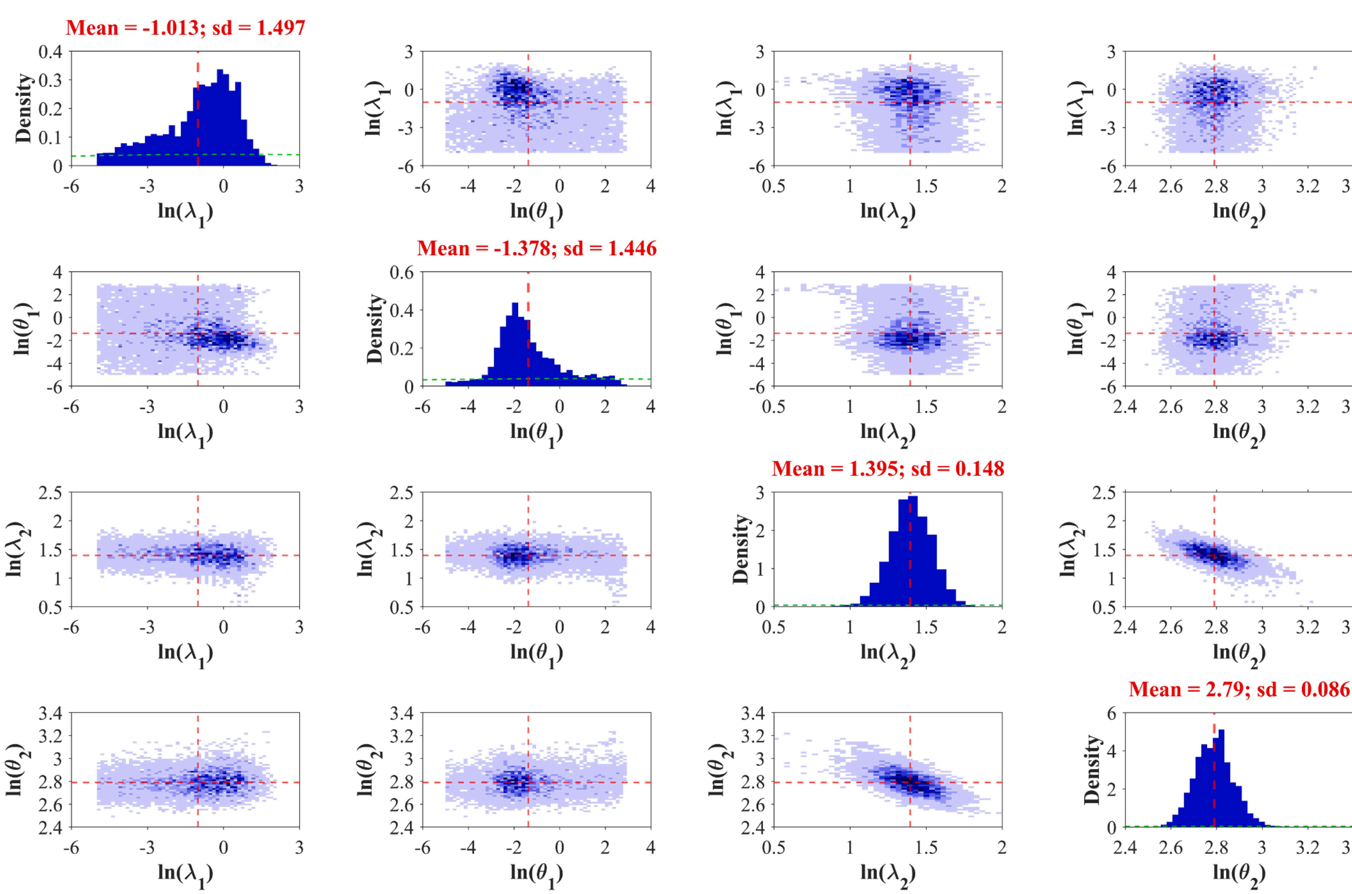

Fig. 1. Posterior TCEV parameter plot for 100 observations from Gumbel distribution with true parameters $log_e\lambda = 1.39$ and.$log_e\theta = 2.71$.

where $\theta_* = \frac{\theta_2}{\theta_1}$ and $\lambda_* = \frac{\lambda_2}{\lambda_1^{1/\theta_*}}$. This diagnostic is useful because it is independent of q and therefore can be used to obtain the expected number of annual maxima in a sample produced by each component.

2.4.2. Degeneracy signatures

Two synthetic cases are presented to illustrate the posterior signatures associated with degeneracy.

Case 1: Gumbel degeneracy

This case involves a sample of 100 generated by a Gumbel distribution with true parameters $\log_e\lambda = 1.39$ and $\log_e\theta = 2.71$. Fig. 1 shows the posterior parameter plot. The plots for $\log_e\lambda_2$ and $\log_e\theta_2$ show the signature of well-identified parameters, the plots exhibit Gaussian-like behavior with bell-shaped densities (which overlap with the true parameters) and elliptically-shaped 2-d frequency plots. In contrast, the plots for $\log_e\lambda_1$ and $\log_e\theta_1$ show a typical degeneracy signature. The 2-d frequency plots for $\log_e\lambda_1$ and $\log_e\theta_1$ show a wide low-density coverage with all points within this domain corresponding to pairs of $\log_e\lambda_1$ and $\log_e\theta_1$ which result in near-zero strength for component *1* for all flows in the record. The densities of $\log_e\lambda_1$ and $\log_e\theta_1$ are skewed and have large standard deviations. Finally, the 2-d frequency plots between components 1 and 2 are indicative of no interaction between the parameters of components 1 and 2. Fig. 2 shows the posterior strength of component 2 for floods with Annual Exceedance Probabilities (AEPs) ranging from 1 in 1.01 to 1 in 100. The dominance of component 2 is unequivocal. This is confirmed by Fig. 3 which shows that posterior probability of the annual maximum produced by component 2 is virtually one.

Case 2: Component degeneracy

In the case of Gumbel degeneracy, the degeneracy will arise regardless of sample size. The more likely form of degeneracy arises when $P_2(\boldsymbol{\Theta})\rightarrow 1$. As the dominant component probability approaches 1 and the sample size decreases, the chance that the record contains annual maxima produced by the dominated component decreases ultimately leading to degeneracy.

Fig. 4 to Fig. 6 show the posterior parameter plot and degeneracy diagnostic plots for a 100-year synthetic record generated by a TCEV distribution with parameters matching those fitted to the Barron River at Picnic Crossing site considered in Section 3. The posterior plot in Fig. 4 shows the signature of well-identified components with Gaussian-like behaviour. Fig. 5 shows that component 2 becomes dominant for flows with AEP less than 1 in 2, while Fig. 6 shows the expected probability of component 2 producing an annual maximum is 62.8% with the probability ranging from 30% to 90%.

The posterior plots are repeated in Fig. 7 to Fig. 9 using the first 25 years of the synthetic record. Given that component 2 has an expected 69.9% chance of producing an annual maximum, one would expect on average in a sample of 25, 17 samples from component 2 and 8 from component 1.

The posterior parameter plot in Fig. 7 shows that component-2 parameters are well-identified and consistent with those in Fig. 4. However, the component 1 plots suggest there is limited information about the parameters. The posterior density of $\log_e\theta_1$ is bimodal with the right mode close to the posterior mean in Fig. 4. The 2-d frequency plot for $\log_e\lambda_1$ and $\log_e\theta_1$ shows a narrow-curved region of high density bordering a large region of low density. Fig. 8 shows that component 2 is largely dominant for flows with AEP less than 1 in 1.10, which accounts for over 90 % of the record. Fig. 9 confirms the overall dominance of component 2. The posterior plots show that, while there is useful inference of component-2 parameters, there is limited and difficult-to-interpret information about component 1.

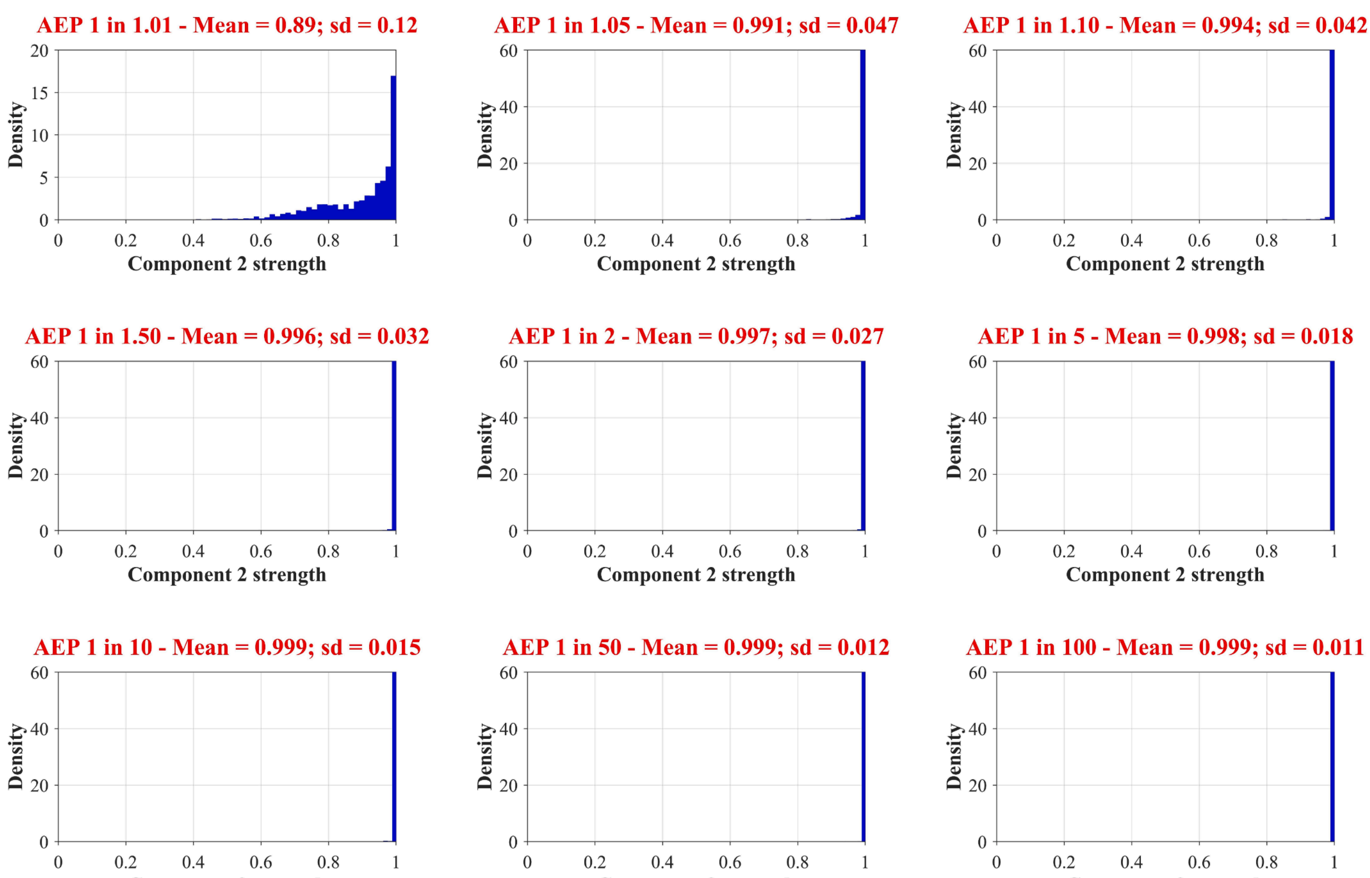

Fig. 2. Posterior component-2 strength plots for a range of AEPs: 100 observations from Gumbel distribution with true parameters $\log_e\lambda = 1.39$ and.$\log_e\theta = 2.71$.

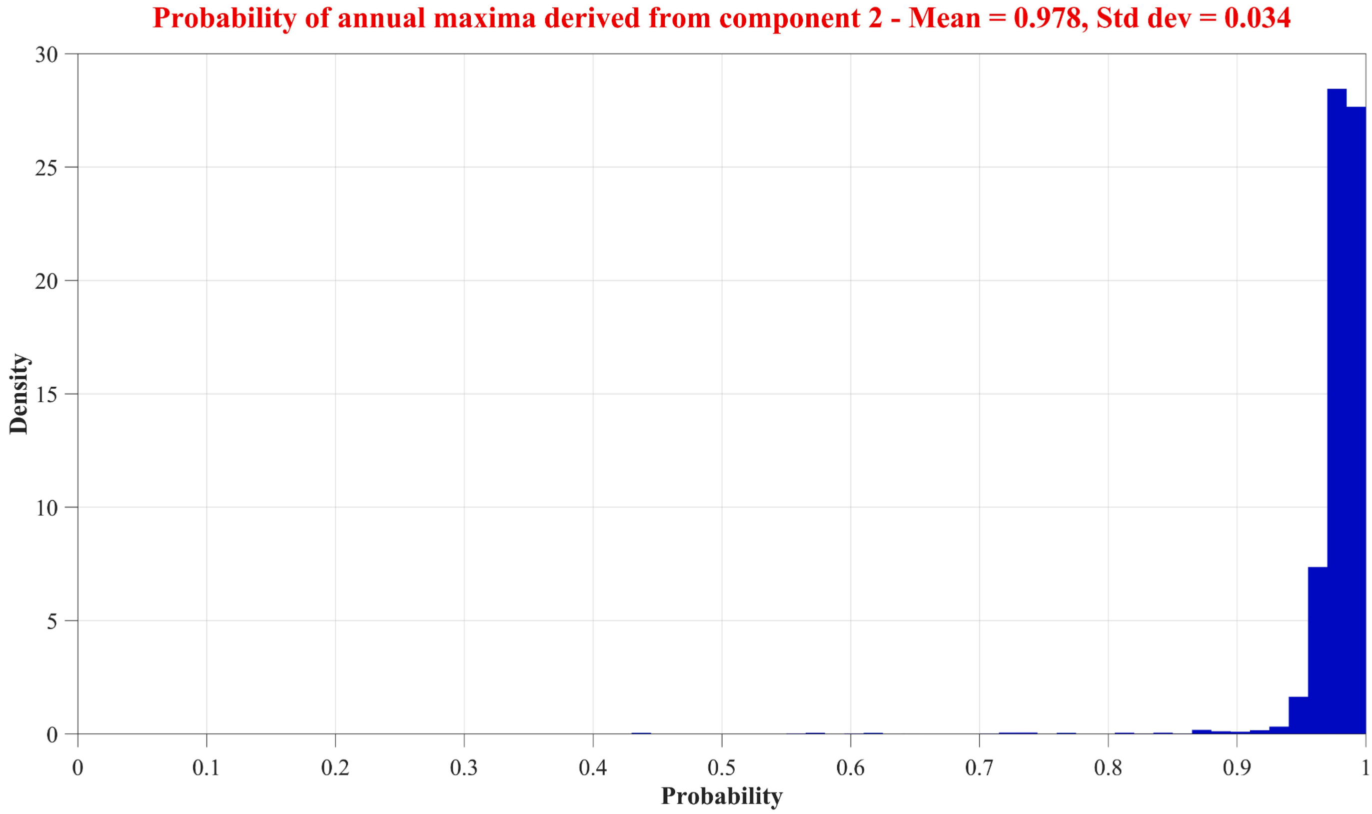

Fig. 3. Posterior plot of probability of the annual maximum produced by component-2: 100 observations from Gumbel distribution with true parameters $log_e\lambda = 1.39$ and.$log_e\theta = 2.71$.

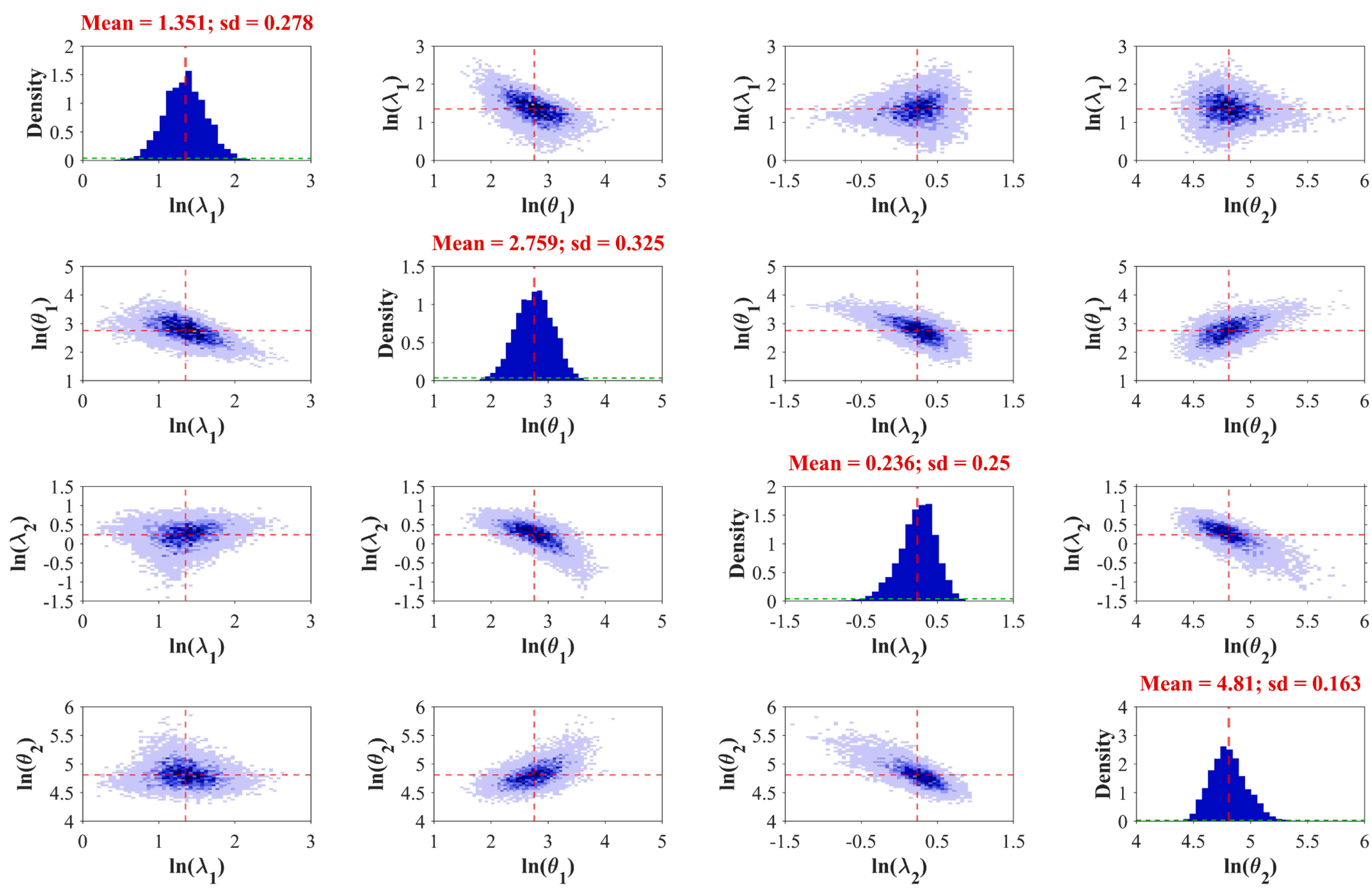

Fig. 4. Posterior TCEV parameter plot for 100 synthetic observations generated from TCEV with parameters estimated for Barron River at Picnic Crossing.

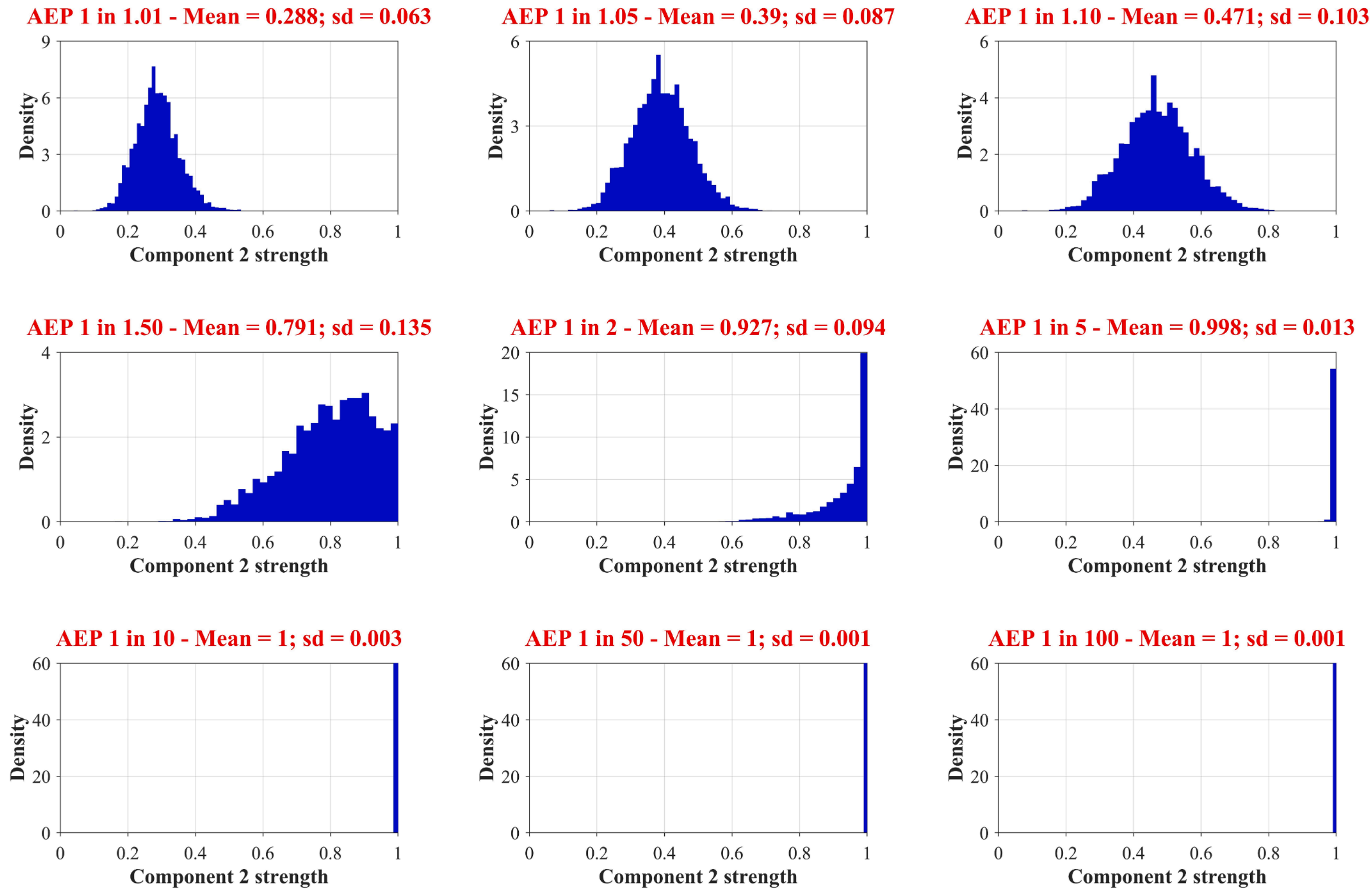

Fig. 5. Posterior component-2 strength plots for a range of AEPs: 100 synthetic observations generated from TCEV with parameters estimated for Barron River at Picnic Crossing.

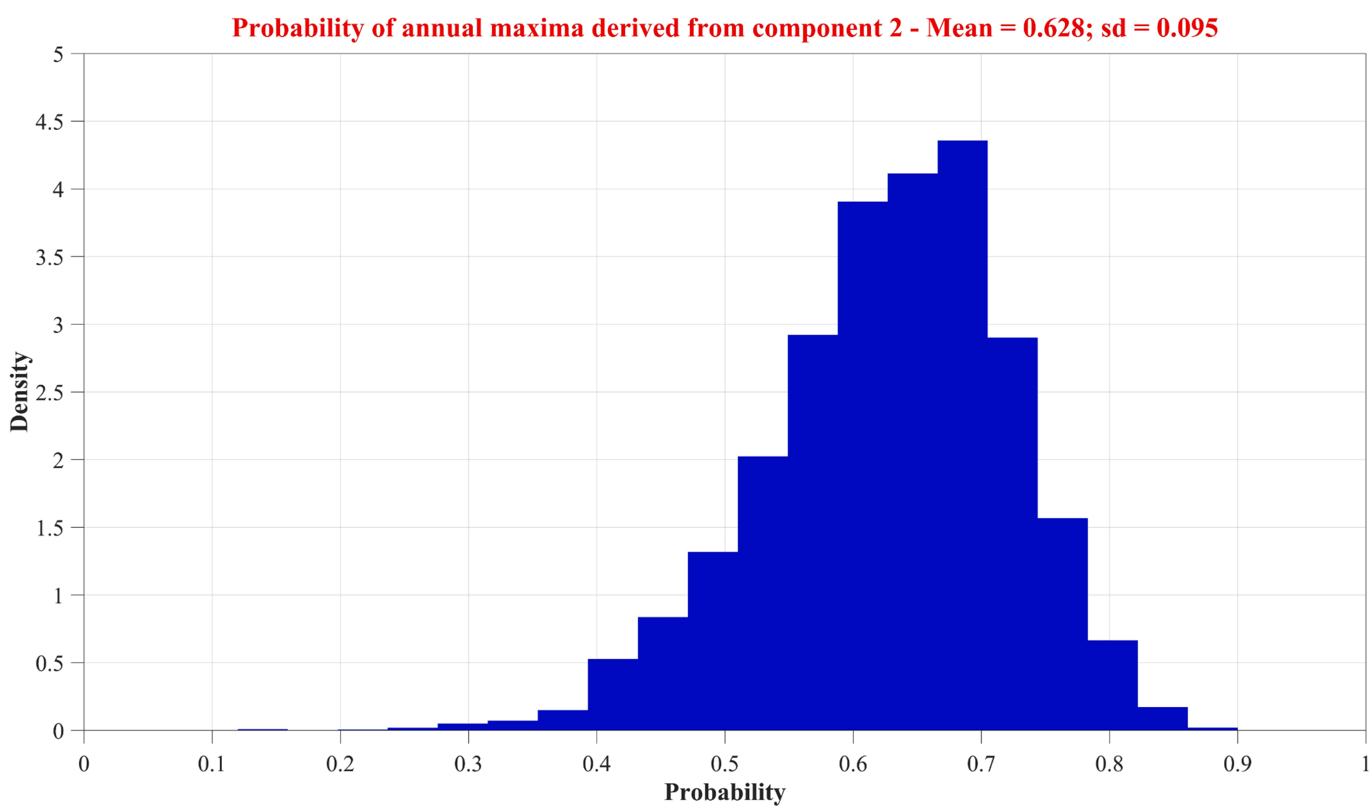

Fig. 6. Posterior plot of probability of the annual maximum produced by component-2: 100 synthetic observations generated from TCEV with parameters estimated for Barron River at Picnic Crossing.

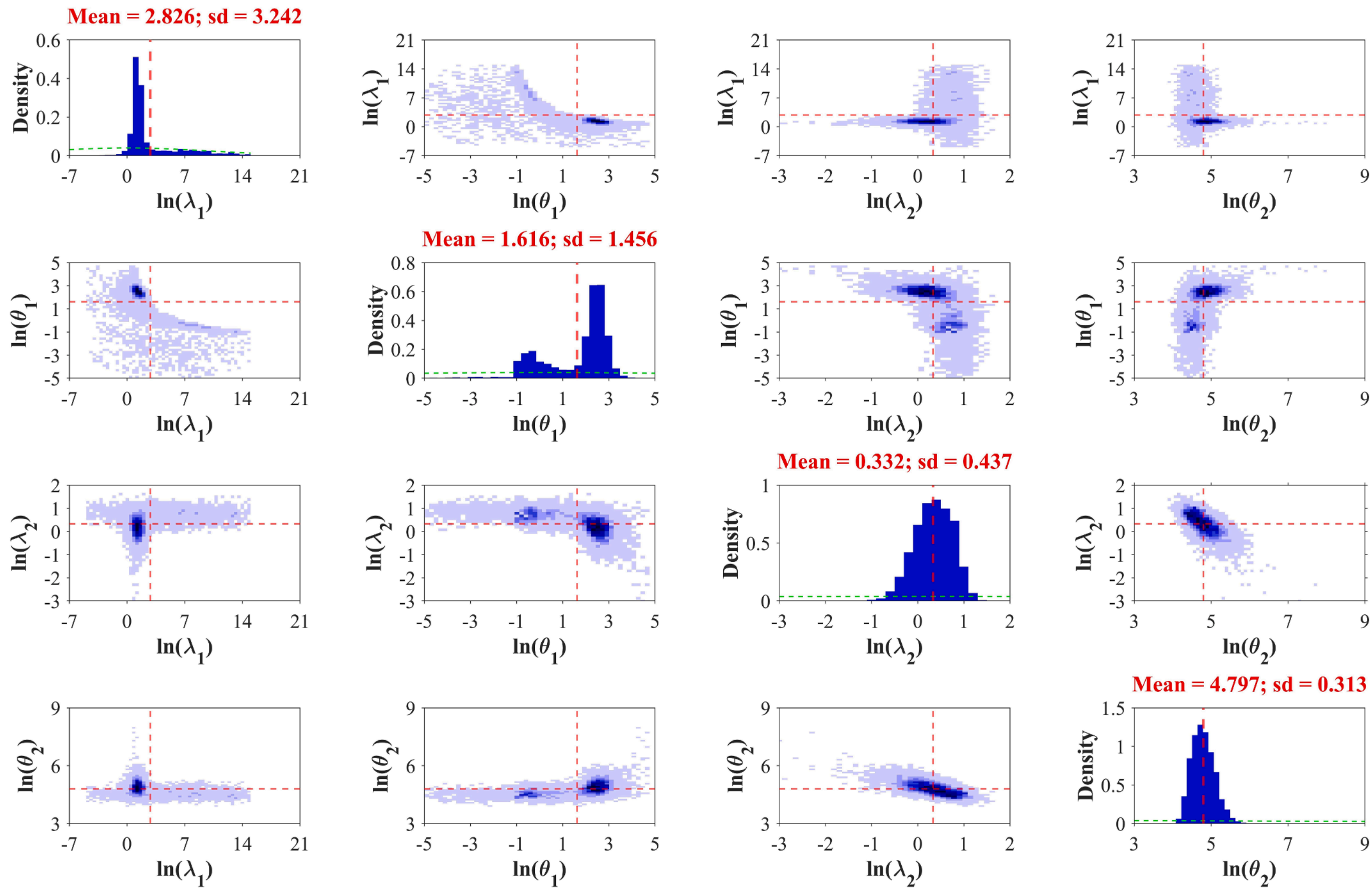

Fig. 7. Posterior TCEV parameter plot for 25 synthetic observations generated from TCEV with parameters estimated for Barron River at Picnic Crossing.

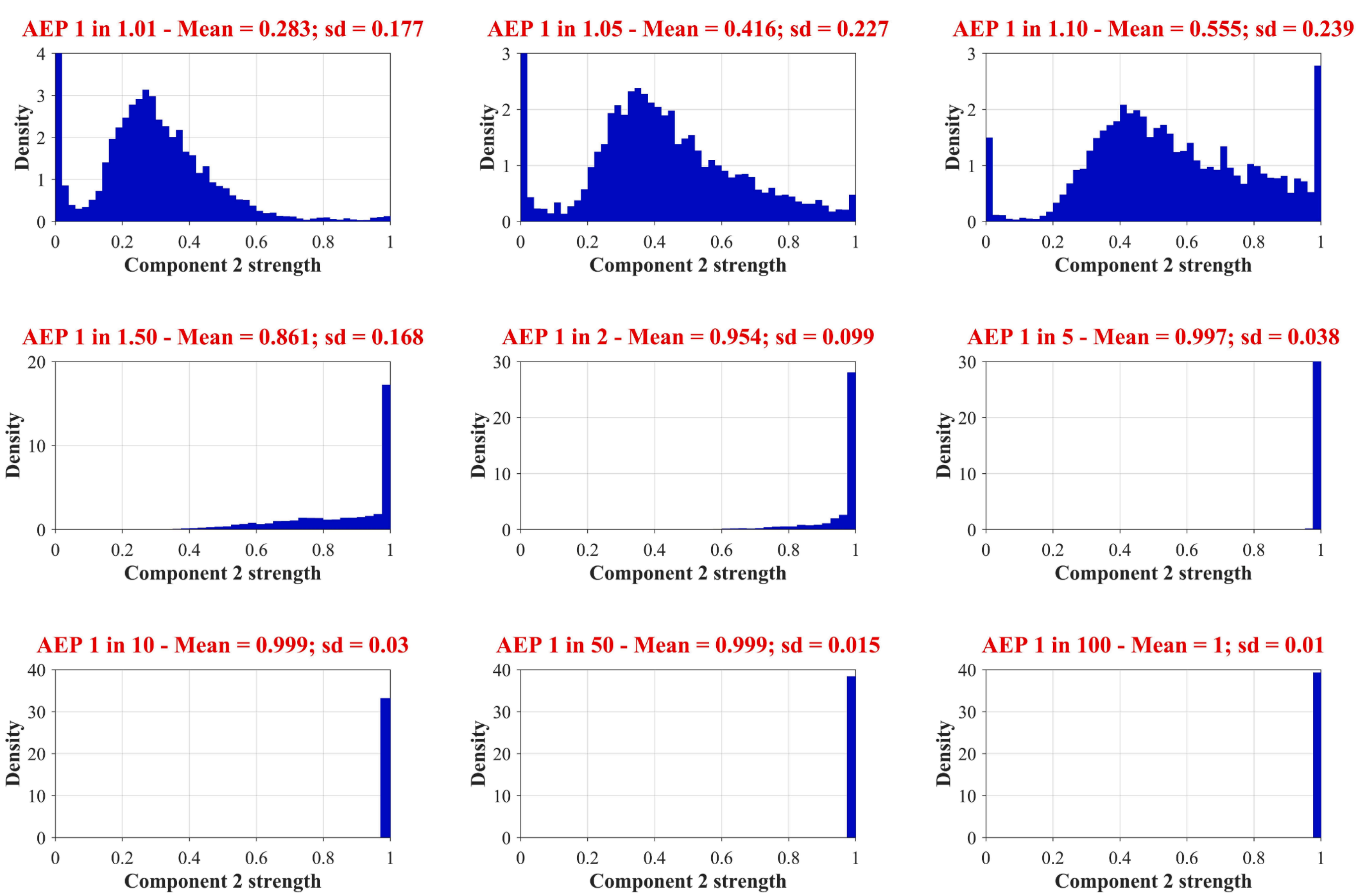

Fig. 8. Posterior component-2 strength plots for a range of AEPs: 25 synthetic observations generated from TCEV with parameters estimated for Barron River at Picnic Crossing.

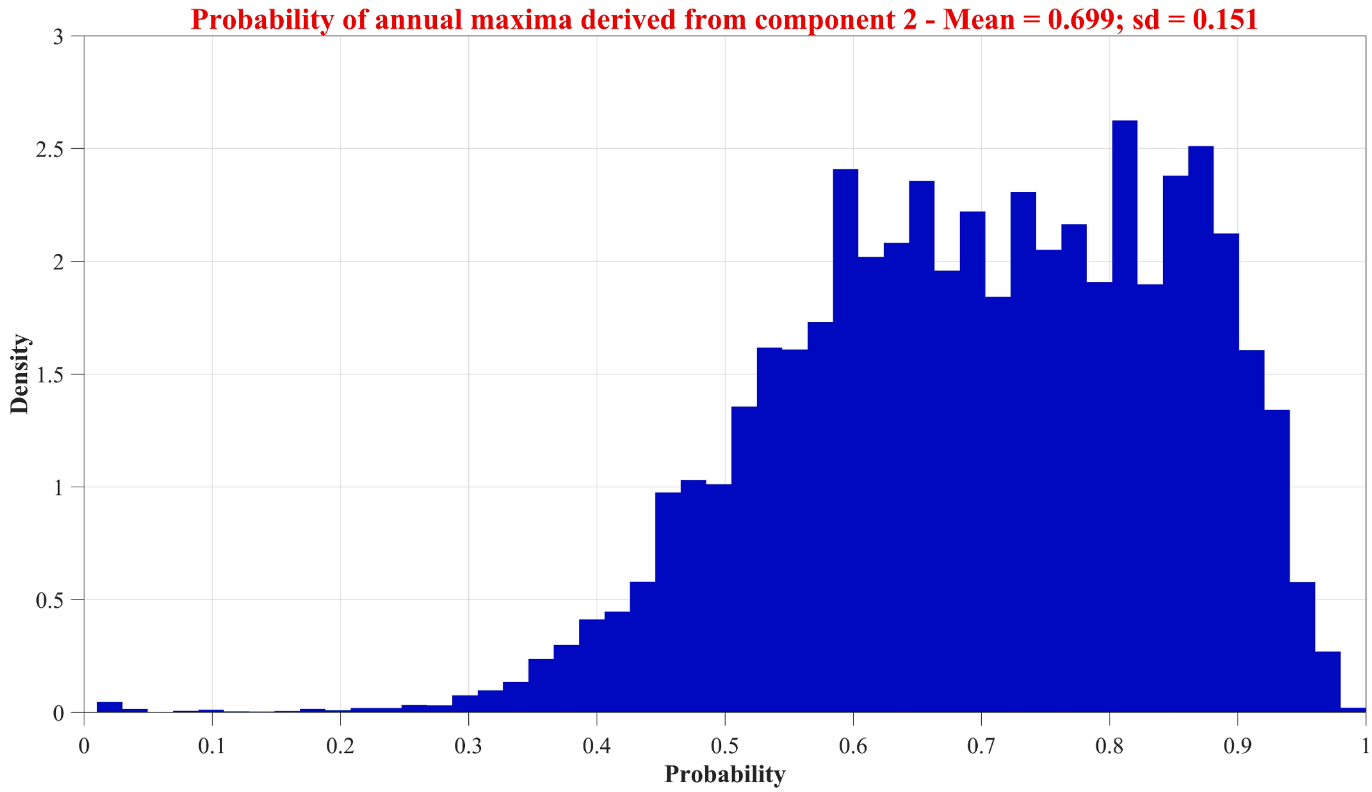

Fig. 9. Posterior plot of probability of the annual maximum produced by component-2: 25 synthetic observations generated from TCEV with parameters estimated for Barron River at Picnic Crossing.

3. Comparative case study evaluating goodness-of-fit and predictive uncertainty

With a robust inference framework in place for the TCEV, the goal of this section is to explore its practical potential in at-site analysis. This will be done by comparing TCEV against LP3 in a case study involving 31 catchments from eastern Australia. The LP3 analysis, referred to as LP3-PILF, uses the procedure recommended in Australian Rainfall and Runoff (Ball et al., 2019) which involves identification of PILFs using the multiple Grubbs-Beck test followed by Bayesian censoring. There are two parts to the case study: the first compares TCEV and LP3 for long records of at least 70 years, the second repeats the comparison using split samples.

3.1. Selection of sites

The sites used in the case study were taken from the P5 database used to develop regional frequency methods in Australian Rainfall and Runoff (Ball et al., 2019). All sites from NSW and Queensland that had 70 or more years were selected. The rationale for this cutoff was to maximize the chance of identifying two TCEV components in the full record. Table 1 summarizes the key characteristics of the 31 sites. Catchment areas range from 16 to 2972 km^2 and record lengths from 70 to 101 years. Fig. 10 shows the location of the catchments which shows the sites cluster in four regions.

3.2. Full record analysis

The section compares the performance of TCEV and LP3-PILF using the full records. Because these record lengths are representative of the longest available, they showcase outcomes least affected by sampling variability. The first part focuses on a qualitative assessment of TCEV and LP3 goodness-of-fit and predictive uncertainty for high quantiles. The second focuses on the identifiability of TCEV parameters.

3.2.1. Comparison of TCEV and LP3-PILF goodness-of-fit and predictive uncertainty

The comparison focusses assessment of goodness-of-fit and right tail quantile uncertainty. We use frequency plots to qualitatively assess goodness-of-fit. Following a scheme similar to that of Barth et al. (2019), we use *normalized confidence intervals*, defined as the ratio of the 95% confidence interval and the expected probability quantile to report predictive uncertainty. Mindful that plotting positions are correlated, we also use the number of outlying observations, defined as observations lying outside the 95% quantile confidence limits, as a guide of misfit. We stress that comparison of TCEV and LP3-PILF fits cannot be done using traditional goodness-of-fit metrics such as Akaike or Bayesian Information Criteria, because the information used in fitting differs (TCEV is fitted to all the data while LP3 is fitted to a subset of the data).

By way of illustration, consider Fig. 11 which shows the LP3 and TCEV frequency plots for site 31. The plots use an AEP Gumbel scale up to the 1 in 200 AEP and have identical vertical scales. For a given AEP, the uncertainty in the quantiles is expressed by 95% confidence limits

Table 1
Summary of case study sites characteristics.

Site	Station	Name	Latitude	Longitude	Area (km^2)	Record length, (yrs)
1	110003	Barron River at Picnic Crossing	−17.26	145.54	220.6	96
2	112002	Fisher Creek at Nerada	−17.57	145.91	15.8	93
3	136203	Barker Creek at Brooklands	−26.74	151.82	246.5	72
4	143303	Stanley R at Peachester	−26.84	152.84	103.5	94
5	145102	Albert River at Bromfleet	−27.91	153.11	543.9	94
6	208001	Barrington River at Bobs Crossing	−32.03	151.47	20.4	71
7	208003	Gloucester River at Doon Ayre	−31.9	152.1	1630.6	76
8	208005	Nowendoc River at Rocks Crossing	−31.77	152.08	1893.8	76
9	208006	Barrington River at U/S Rocky Crossing	−32.04	151.87	606.6	76
10	208008	Gloucester River at Forbesdale	−32.06	151.9	197.8	73
11	208009	Barnard River at Barry	−31.58	151.32	152.7	72
12	210011	Williams River at Tillegra	−32.32	151.69	202.8	88
13	210014	Rouchel Brook at Rouchel Brook	−32.15	151.05	401.8	75
14	210017	Moonan Brook at Moonan Brook	−31.94	151.28	102.7	73
15	210018	Hunter River at Moonam Dam Site	−31.92	151.22	741.2	77
16	210022	Allyn River at Halton	−32.31	151.51	191.8	78
17	214003	Macquarie Rivulet at Albion Park	−34.58	150.71	35.5	72
18	215004	Corang River at Hockeys	−35.15	150.03	163.7	94
19	218001	Tuross River at Tuross Vale	−36.26	149.51	91.4	73
20	219001	Rutherford Creek at Brown Mountain	−36.59	149.44	15.5	79
21	219003	Bemboka River at Morans Crossing	−36.67	149.65	316.5	78
22	219006	Tantawangalo Creek at Tantawangalo Mtn	−36.78	149.54	82	70
23	222004	Little Plains River at Wellesley	−37.00	149.09	612.4	77
24	222007	Wullwye Creek at Woolway	−36.42	148.91	542.6	72
25	410044	Muttama Creek at Coolac	−34.93	148.16	1063.6	83
26	410059	Gilmore Creek at Gilmore	−35.34	148.17	278.4	73
27	410061	Adelong Creek at Batlow Road	−35.33	148.07	148.1	72
28	410062	Numeralla River at Numeralla School	−36.18	149.35	675.5	74
29	412029	Boorowa River at Prossers Crossing	−34.14	148.81	1557	83
30	422202	Dogwood Creek at Gilweir	−26.71	150.18	2972.1	71
31	422306	Swan Creek at Swanfels	−28.16	152.28	82	101

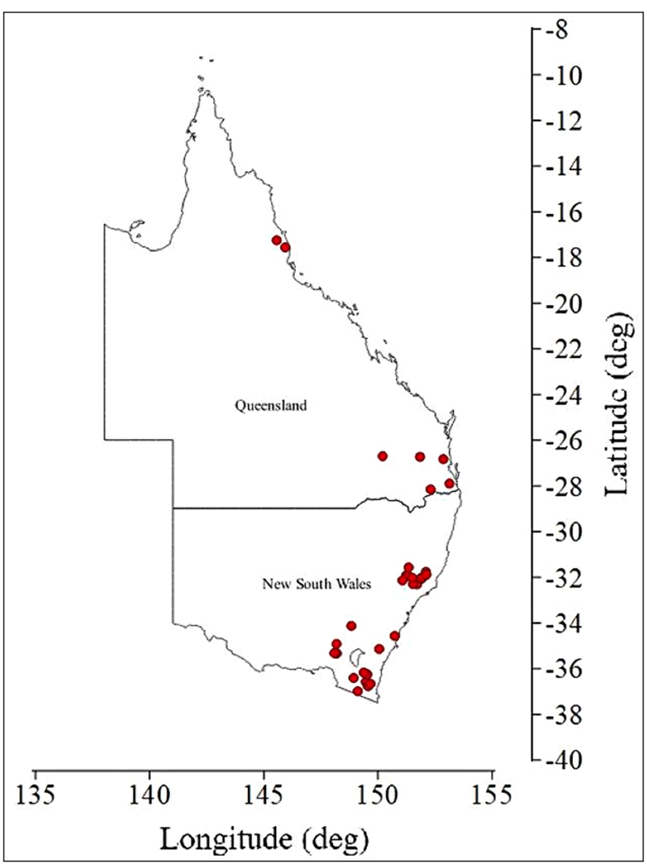

Fig. 10. Location of case study sites in eastern Australia.

(or more accurately, credible limits). Censoring of PILFs is denoted by a square box located at the censoring threshold. For both TCEV and LP3 all observations lie within the 95% confidence limits with the principal differences occurring in the way the distributions fit the right tail - LP3 overshoots the top 4 ranked flows, while TCEV undershoots the rank 2–4 flows but matches the rank 1 flow. Both distributions are judged as fitting the data satisfactorily. The most obvious difference between the plots is the width of the 95% confidence intervals with TCEV exhibiting relatively narrow intervals at high quantiles which contrast with the rapidly widening intervals for LP3.

This result is typical for the large majority of LP3 and TCEV site comparisons. The LP3 was judged to fit all the sites satisfactorily in the sense the 95 % confidence limits largely contained the observations. In contrast for TCEV, it was found at several sites that a larger than expected number of observations were outside the 95% confidence limits. Table 2 reports the number of observations lying outside the 95% confidence limits for each site. For LP3, only six sites had outlying observations with a maximum of 3. In the case of TCEV, sixteen sites had outlying observations with five sites having 4 or more flows lying outside the 95% limits. Table 2 also reports the percentage of flows identified as PILFs and censored in the LP3 fit. By censoring PILFs, LP3 is being preferentially fitted to the larger flows. As shown in Table 2, one or more PILFs were identified in twenty sites (out of 31) with sixteen sites having ≥30% PILFs and nine sites with ≥40% PILFs. As a practical consequence, LP3 has three parameters to fit in many cases to the top 50–75% of the data while TCEV has to fit the whole series. This observation also supports the fairness of the proposed comparison.

An example of a site with significant outlying observations in the TCEV fit is site 29, Boorowa River at Prossers Crossing. Fig. 12 shows the TCEV and LP3 frequency plots for this site. TCEV has six outlying observations - three lie close to the confidence limits, while the other three correspond to the top three ranked flows which lie well above the upper confidence limit. In contrast, LP3 has no observations lying outside the limits, though 29% of the original observations were censored. Both the LP3 and TCEV fits overshoot the flows between 1 in 10 and 1 in 5 AEP.

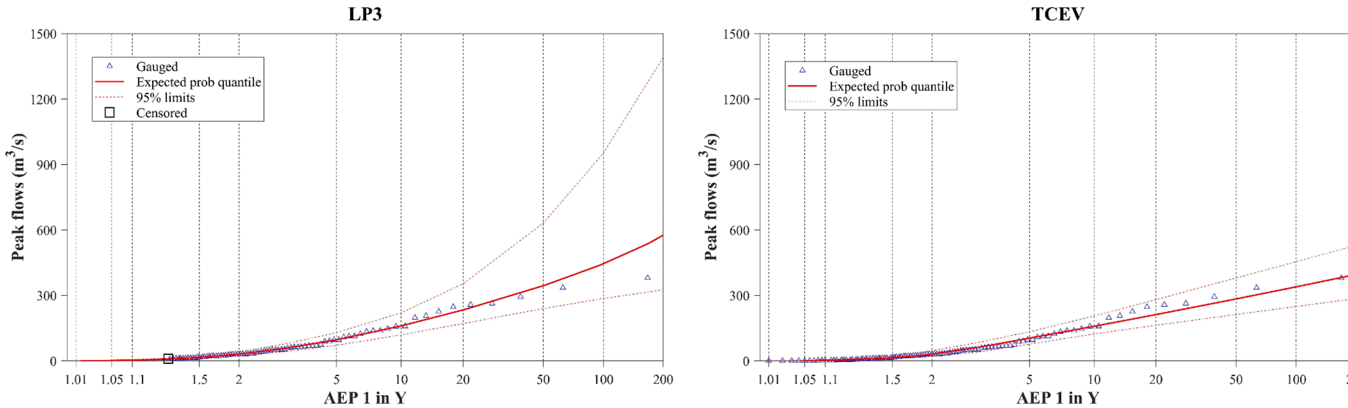

Fig. 11. LP3 and TCEV frequency plots for site 31, Swan Creek at Swanfels.

Table 2
Summary of goodness-of-fit indicators.

Site	Record length (yrs)	Observations lying outside 95 % confidence limits		% of years identified as PILFs and censored in LP3 fit
		LP3	TCEV	
1	96	0	1	0
2	93	2	0	0
3	72	0	2	32
4	94	0	0	43
5	94	0	0	37
6	71	0	0	0
7	76	0	0	38
8	76	0	0	49
9	76	1	0	24
10	73	0	0	47
11	72	0	0	10
12	88	0	4	0
13	75	0	0	0
14	73	1	0	1
15	77	1	1	0
16	78	1	1	32
17	72	0	6	21
18	94	0	1	36
19	73	0	1	49
20	79	0	1	48
21	78	3	3	49
22	70	0	1	44
23	77	0	0	0
24	72	0	5	32
25	83	0	1	37
26	73	0	7	47
27	72	0	1	29
28	74	0	0	43
29	83	0	6	29
30	71	0	0	8
31	101	0	0	22

However, LP3 has more flexibility to follow the path of the top ranked flows than TCEV but this comes at the price of significantly inflated confidence intervals.

In summary, the comparison of LP3 and TCEV fits suggests there is little practical difference for the bulk of the sites. There were, however, about five sites where LP3 better fitted the highest-ranking flows in the sense of observations lying within the 95 % confidence limits. This is in part attributed to the censoring of PILFs which allows more flexibility fitting the right-hand tail but, more likely, is due to the fact the LP3-PILF 95% confidence intervals are much wider than the TCEV intervals for high quantiles. This is clearly demonstrated in Fig. 13 which shows a scatter plot of normalized confidence intervals for 1 in 50 and 1 in 100 AEP quantiles. Using the 1:1 red line as a reference, it can be seen that the LP3 normalized quantile uncertainty is roughly double that of TCEV.

3.2.2. *TCEV parameter identifiability*

The identifiability of the TCEV parameters was assessed studying the posterior parameter and degeneracy diagnostic plots. An example of degeneracy is shown for site 11, Barnard River at Barry. Fig. 14 shows the TCEV frequency plot and the parameter posterior. While the fit is satisfactory and the high quantile confidence intervals much smaller than the LP3 intervals, the posterior exhibits a complex degeneracy signature. The posterior of component 2 is uni-modal but heavily skewed. The posterior of $\ln(\theta_1)$ is bimodal. This bimodality is very evident in the 2-d frequency plots involving $\ln(\theta_1)$ and the component-2 parameters. The 2-d frequency plots involving $\ln(\lambda_1)$ and the component-2 parameters have the characteristic L-shape consistent with degeneracy. The component 1 standard deviations are about 6 times greater than their component-2 counterparts.

The practical focus is identification of the right tail of the TCEV. In this regard, two items of information are particularly relevant, the posterior of the component-2 parameters and the AEP beyond which component 2 dominates the right tail of the TCEV.

Four figures are used to summarize the identifiability issue across the 31 sites. Fig. 15 shows for each site the posterior mean with standard deviation error bars for $\ln(\theta)$, while Fig. 16 does so for $\ln(\lambda)$. Fig. 17 shows for each site the posterior mean with standard deviation error bars of the component-2 strength for AEPs 1 in 2, 1 in 5 and 1 in 10 as well as $1 - \mathrm{P}_2(\boldsymbol{\Theta})$. Fig. 18 offers a different perspective showing box plots of the component-2 strength posterior means for AEPs from 1 in 1.01 to 1 in 100.

These figures offer the following insights:

1) The AM records contain more samples from component 2 than 1. Fig. 18 shows the transition from component 1 to component 2 occurs in the AEP range 1 in 1.01 to 1 in 2. This is consistent with the posterior mean of $P_2(\Theta)$ lying above 0.5 for all but four sites.
2) For all sites, the component-2 posteriors are well identified with Gaussian-like behavior. Fig. 15 and Fig. 16 show that the component-2 parameters have much tighter standard deviations than their component-1 counterparts. This is attributed to the greater number of samples in the AM records arising from component 2.
3) Unlike component 2, the uncertainty in the component-1 parameters exhibits much more variability. The posterior standard deviations

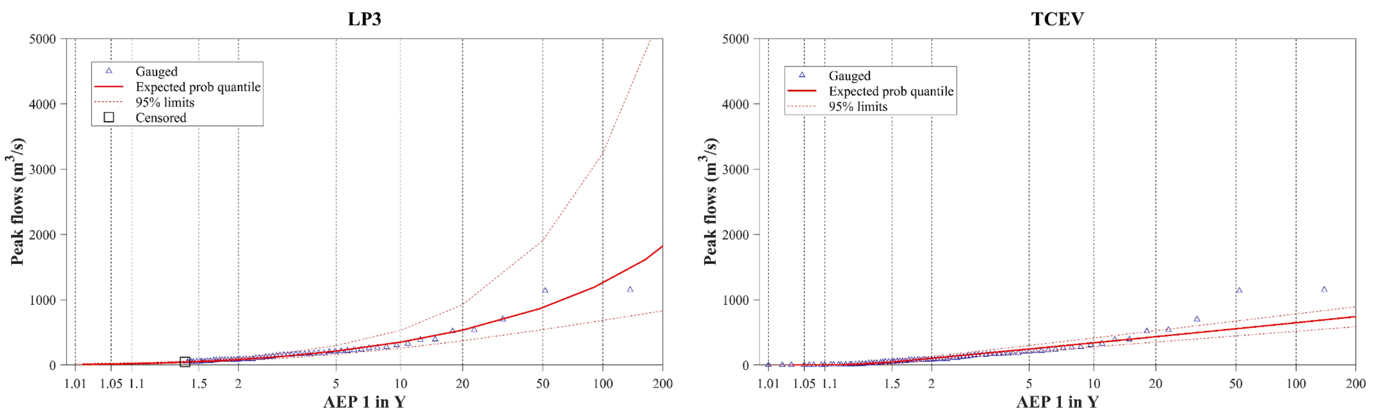

Fig. 12. LP3 and TCEV frequency plots for site 29, Boorowa River at Prossers Crossing.

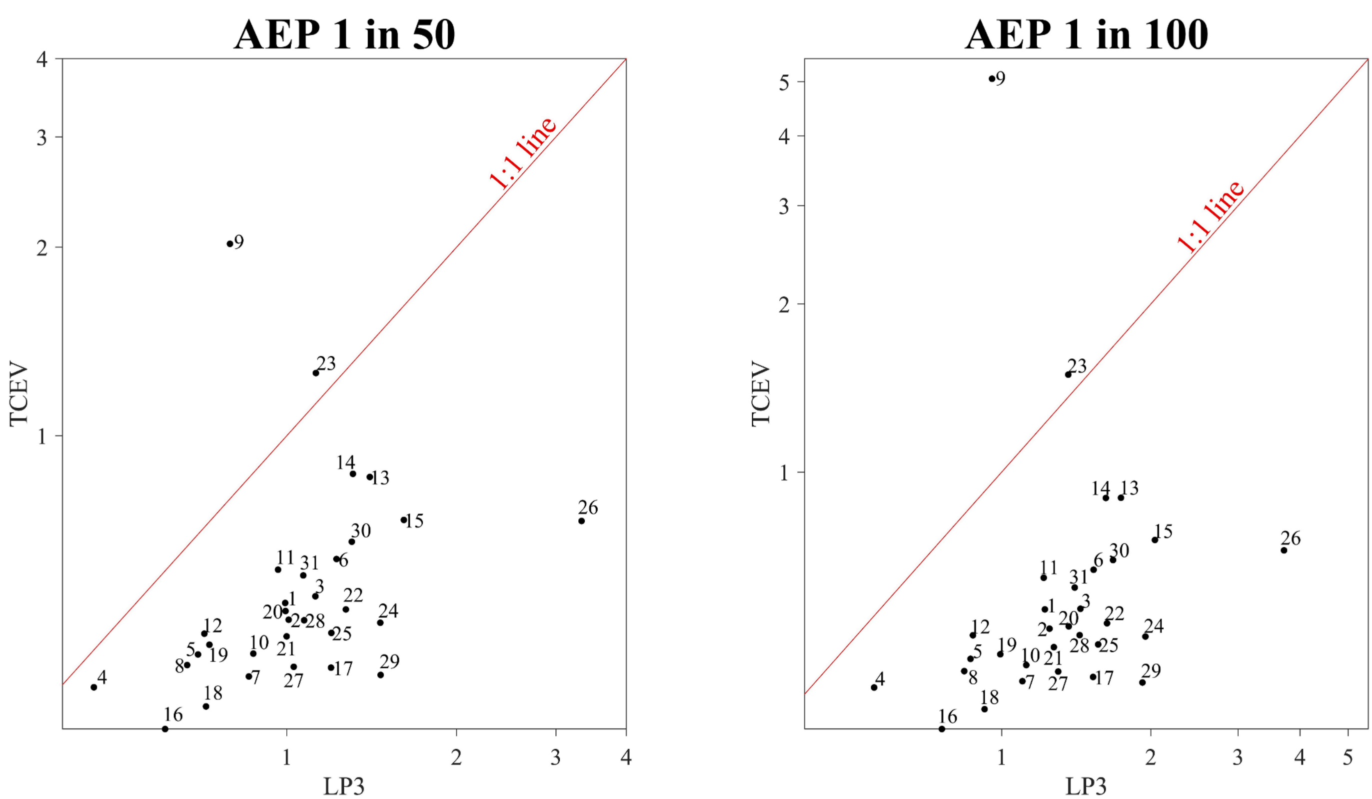

Fig. 13. Scatter plot of normalized uncertainty intervals for LP3 and TCEV for 31 sites.

range from small values comparable to component 2 to large values associated with degeneracy. When degeneracy occurs, the posterior standard deviations of both component-1 parameters are large.

4) Component 2 is dominant for quantiles with AEPs less than 1 in 5 meaning the right tail closely follows a Gumbel distribution for AEPs less than 1 in 5. Fig. 17 shows that for an AEP of 1 in 5, the expected strength of component 2 is greater than 0.9 for all but two sites. For an AEP of 1 in 10, only one site has an expected strength less than 0.95. The tight confidence limits for high quantiles arise from the fact that well over half of the AM record is available to identify the component-2 parameters, which define the right-tail Gumbel distribution.

3.3. Split-sample analysis

The goal of this section is to assess the identifiability of TCEV parameters when fitted to records short to moderate in length. Such records are obtained by splitting site records in two and four nonoverlapping sub-samples. The results are reported here for the same three sites used in the previous section, sites 11, 29 and 31.

Fig. 19 shows plots of posterior mean and two standard deviation bars for the TCEV parameters at three sites for all sample combinations. The leftmost sample is the full record. As expected, uncertainty increases with decreasing sample size. Nonetheless, the component-2 parameters remain well-identified. In contrast, the component 1 parameters are subject to much more uncertainty and in many cases are poorly identified. Their posteriors are often bimodal with complex non-Gaussian

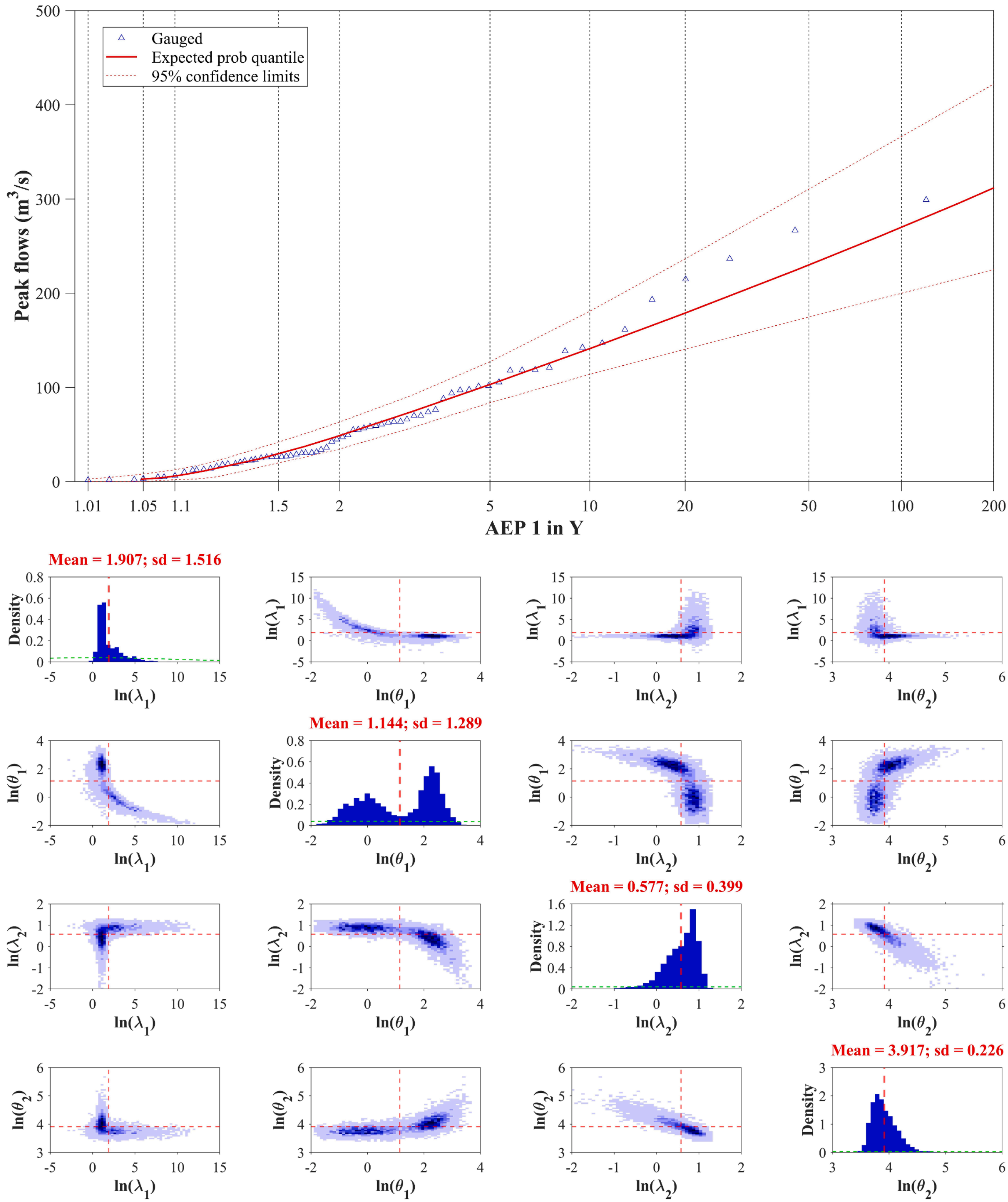

Fig. 14. TCEV frequency plot (a) and posteriors (b) for site 11, Barnard river at Barry.

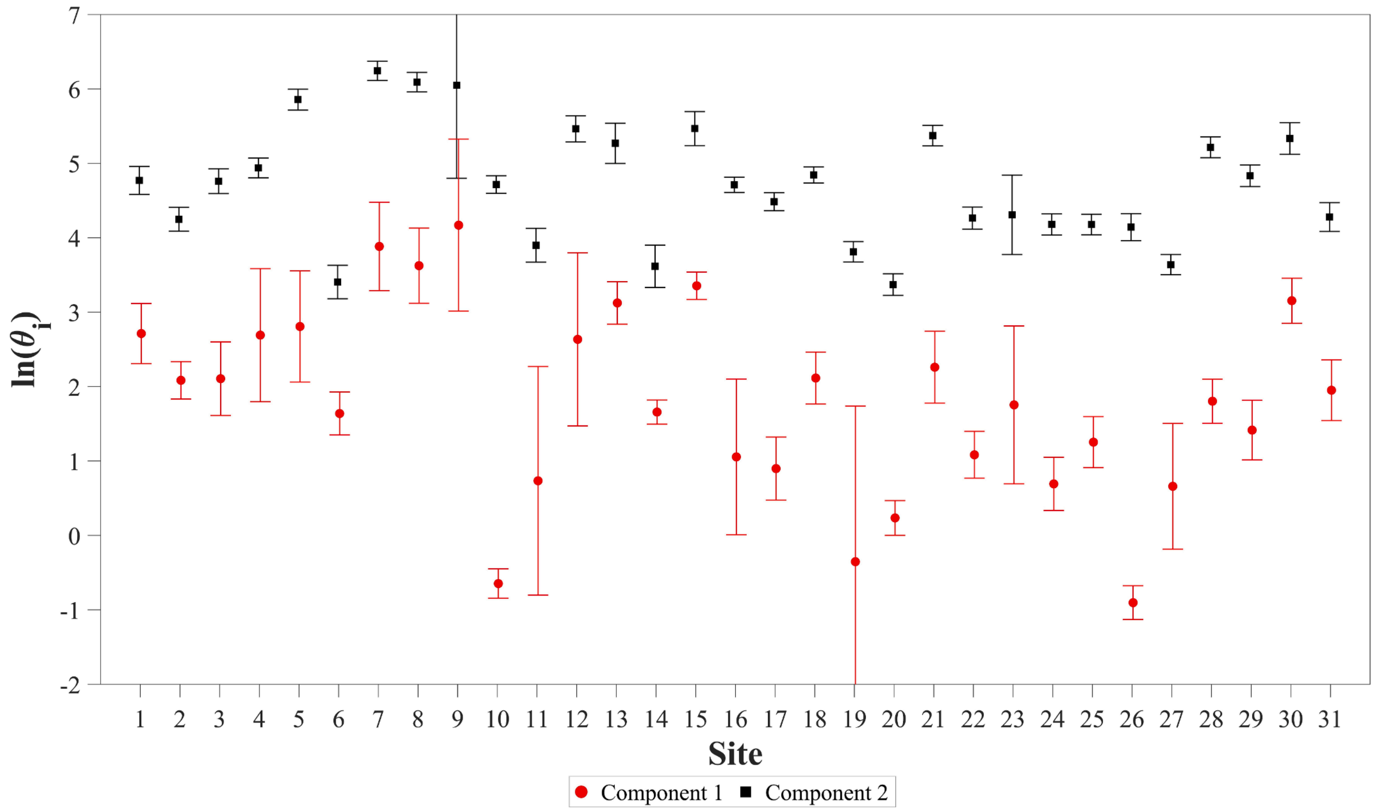

Fig. 15. Component posterior mean of $ln(\theta)$ with standard deviation error bars.

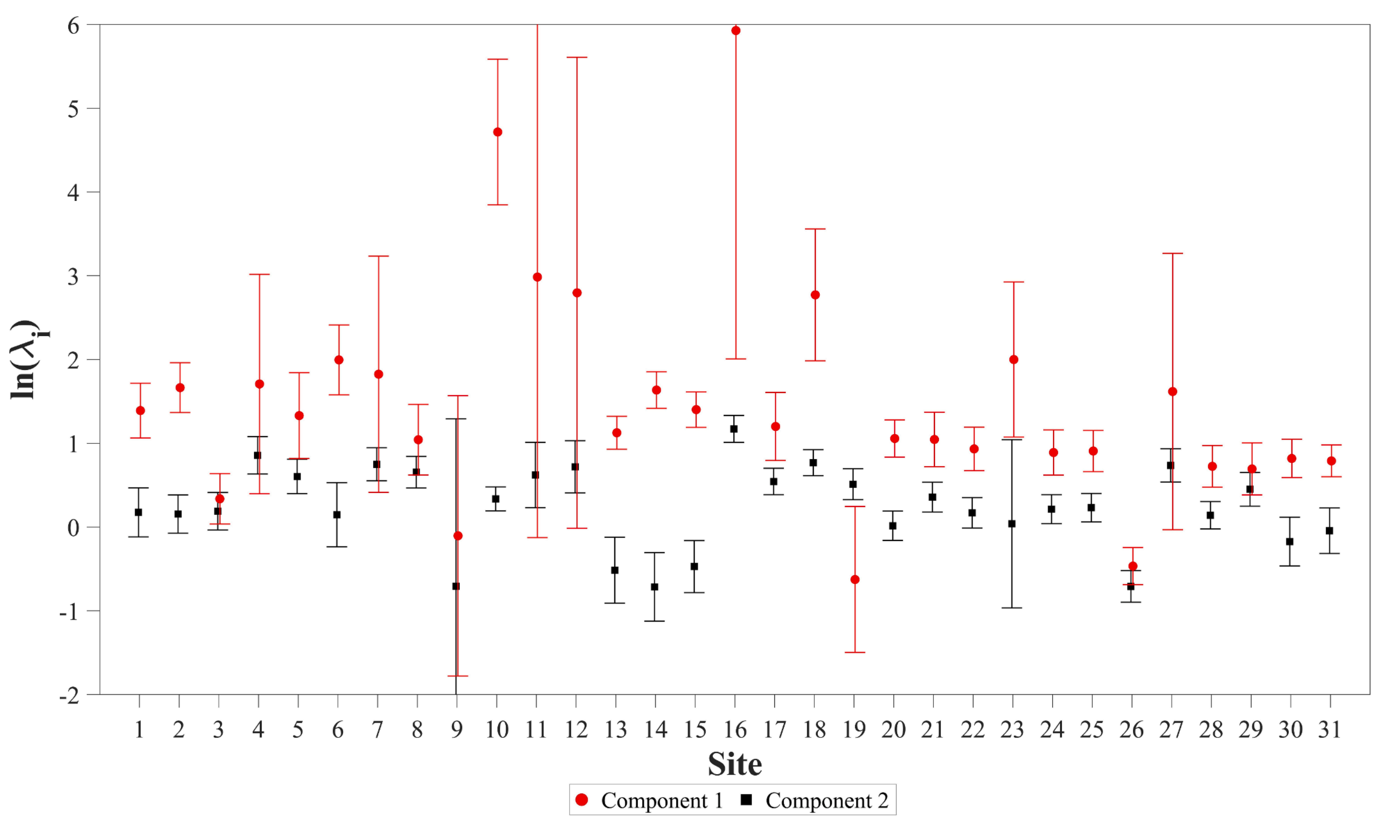

Fig. 16. Component posterior mean of $ln(\lambda)$ with standard deviation error bars.

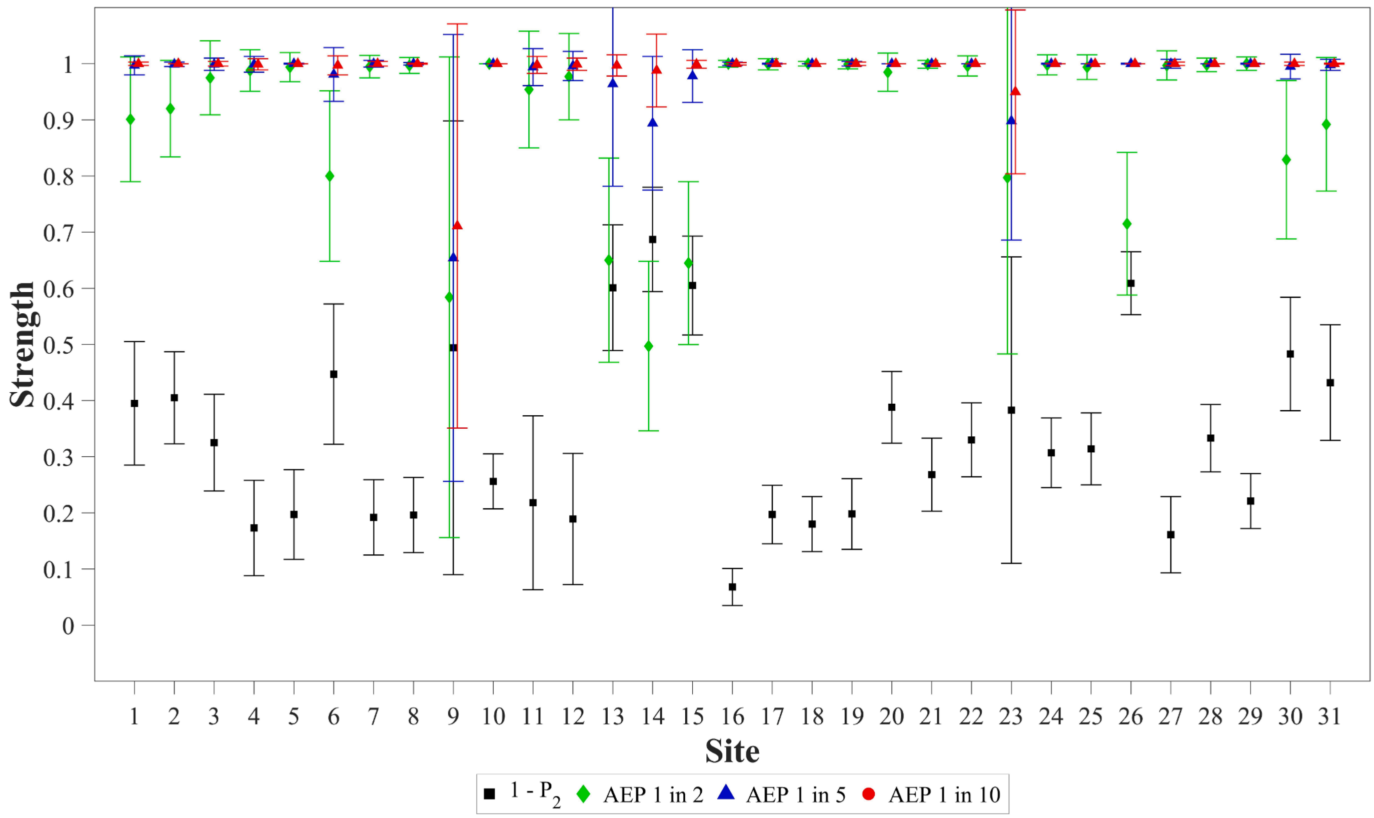

Fig. 17. Posterior mean and standard deviation error bars for component-2 strength for AEPs 1 in 2, 1 in 5 and 1 in 10 and for $1-P_2(\boldsymbol{\Theta})$, probability of component 1 producing annual maximum, for all 31 sites.

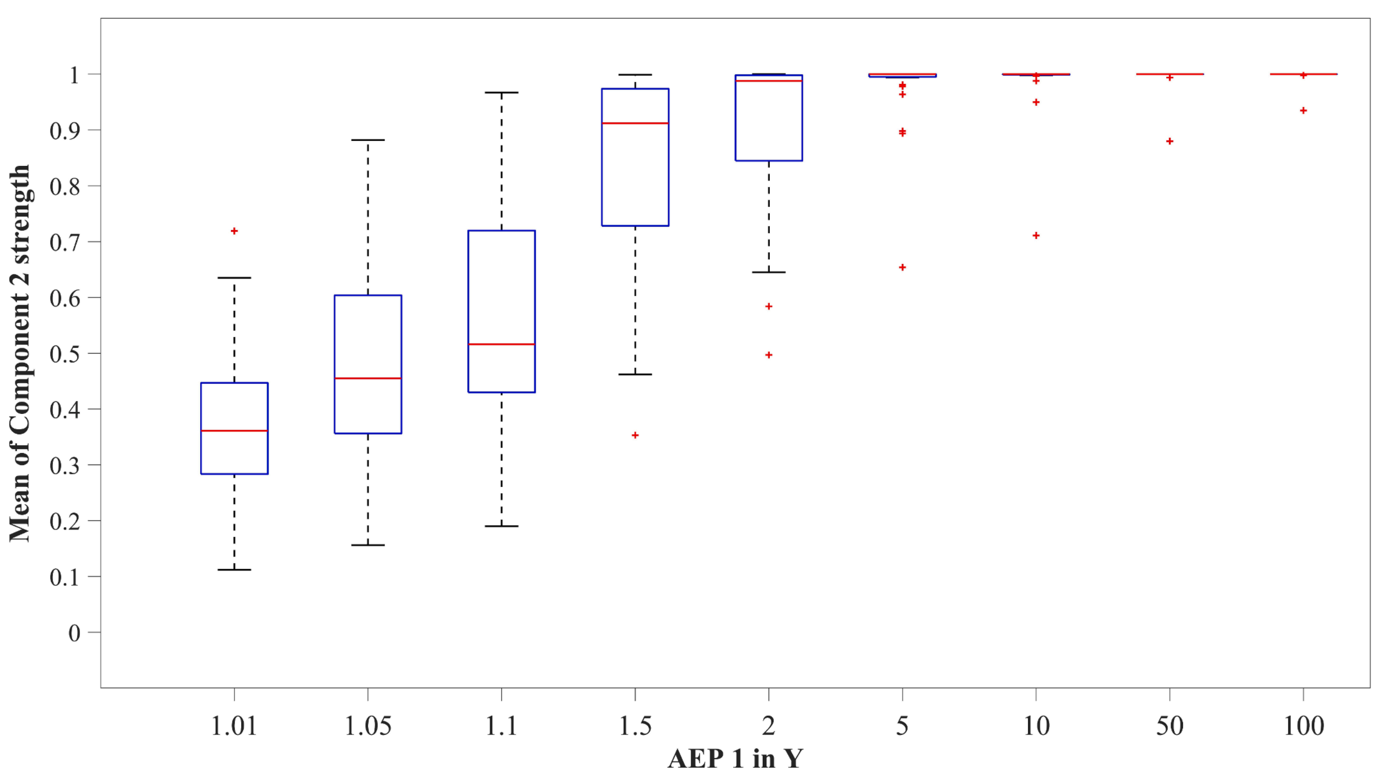

Fig. 18. Boxplots of posterior mean of component-2 strength for AEPs from 1 in 1.01 to 1 in 100.

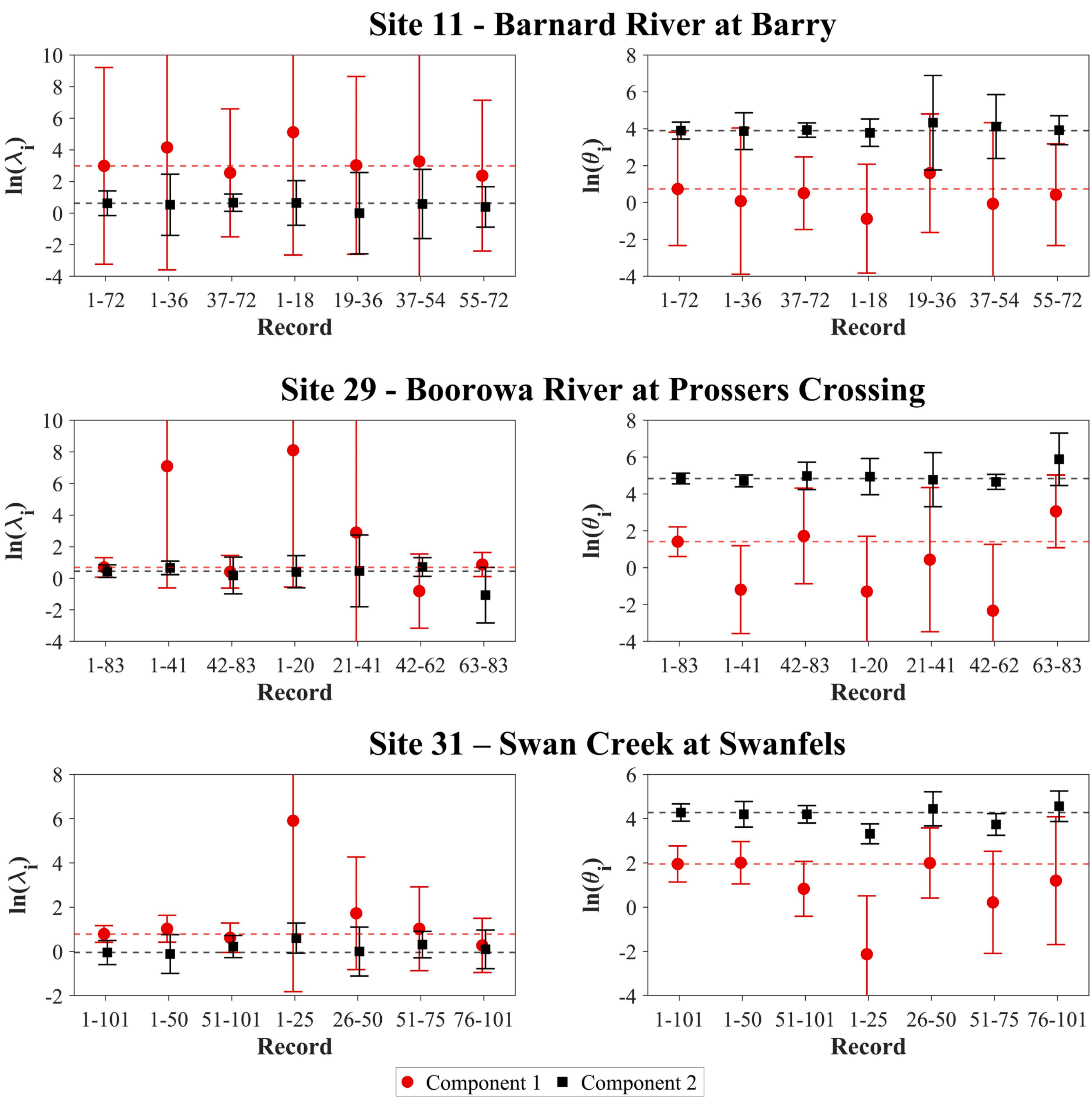

Fig. 19. Posterior mean and two standard deviation error bars for TCEV parameters for sub-samples at sites 11,29,31.

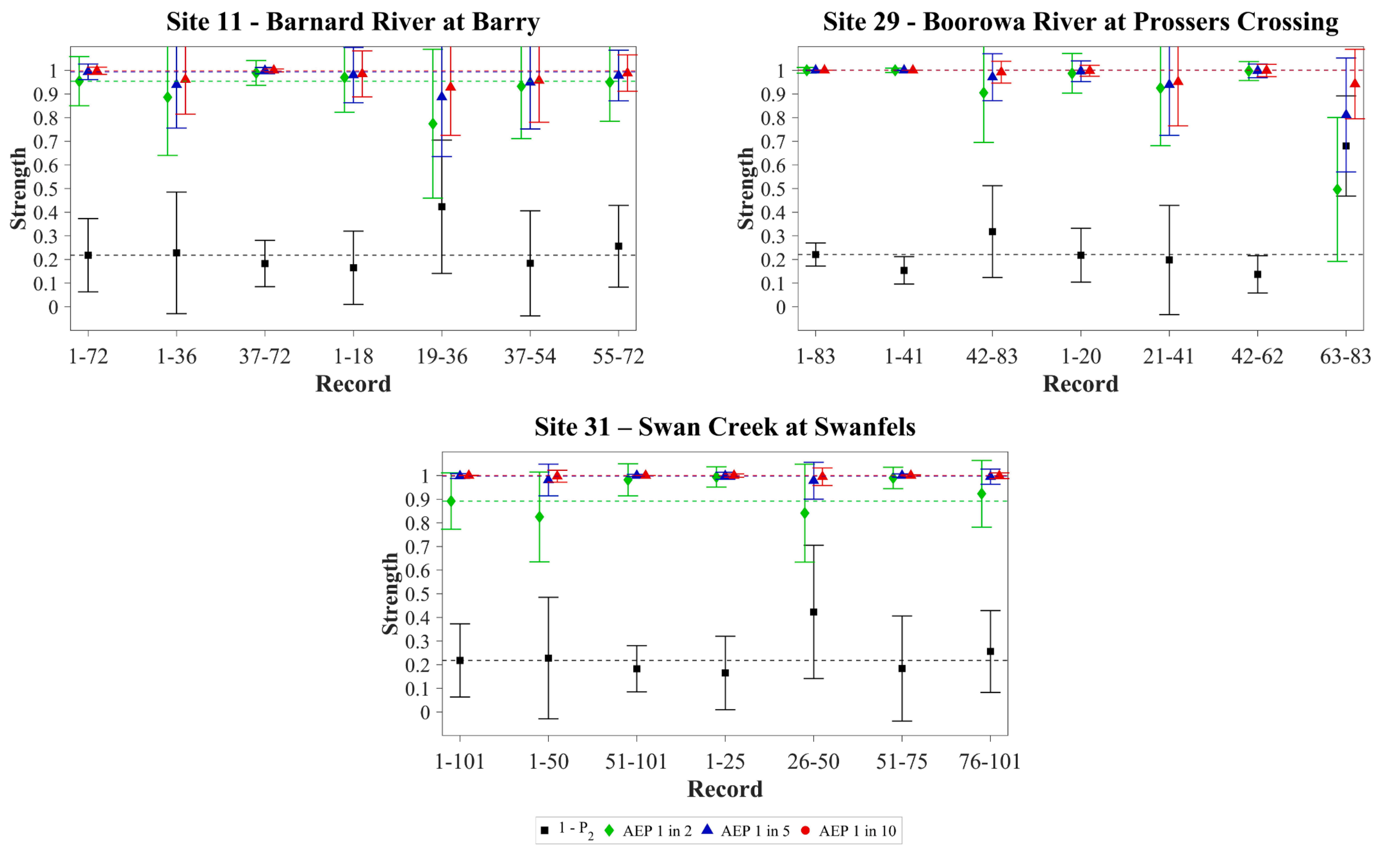

Fig. 20. Posterior mean and standard deviation error bars for component-2 strength for AEPs 1 in 2, 1 in 5 and 1 in 10 and for $1-P_2(\Theta)$, probability of component 1 producing annual maximum, for sub-samples at sites 11,29,31.

signatures. This poor identifiability is consistent with low values of $1-P_2(\Theta)$, shown in Fig. 20, which indicates only a small number of AM come from component 1 in all cases except one. The exception occurs for the fourth split sample of site 29, where component 2 is dominated by component 1. Likewise, the component-2 strength plots show that component 2 is dominant for all AEPs lower than 1 in 2.

As before, the confidence limits for TCEV are much narrower than LP3-PILF due to convergence of the TCEV right tail to a Gumbel distribution. However, due to the increased quantile uncertainty, only a limited number of observations lie outside the TCEV confidence limits. These results are not reported here because they are consistent with those of the previous section.

3.4. Discussion

The purpose of the case study was to empirically assess the practical potential of TCEV in at-site FFA. In assessing its potential, two questions need to be answered.

The first is whether the TCEV inference framework is robust. Clearly, without a reliable inference framework, there is little prospect of practical application. The Bayesian framework presented in Section 2 proved to be robust in the case study in the sense that the Metropolis-Hastings algorithm produced chains that adequately mixed in all cases including a significant number involving degeneracy. The degenerate posteriors were found to have quite complex signatures (including bi-modality and non-linear interaction) which are very different to the Gaussian signature expected in large samples. While there is no guarantee the Bayesian framework will be robust in application to other catchments, the widespread usage and success of Markov chain sampling in Bayesian computation suggests it is reasonable to expect similar performance elsewhere.

The second question is whether TCEV produces FFA outcomes that are unambiguously better than existing methods. Unlike the first question, this one is more difficult to answer with confidence. The ultimate practical goal of FFA is to estimate high flow quantiles as accurately as possible - ideally the quantile estimates should be unbiased and have narrow confidence intervals.

The case study compared the performance of TCEV against the LP3-PILF approach. The first part of the case study focused on long records as they provide the best opportunity to assess goodness-of-fit. Overall, there is little practical difference in the LP3-PILF and TCEV fits. If one uses the number of observations lying outside 95 % confidence limits as a measure of misfit, then LP3-PILF fits would be judged to be better - TCEV has five sites with 4 or more observations lying outside 95 % confidence intervals, while LP3-PILF has six sites with outlying observations with a maximum of 3. While this result may be due to censoring of PILFs which allow LP3 to better fit the right tail, it is more likely a consequence of LP3-PILF having much wider confidence intervals than TCEV for high quantiles.

The main difference to emerge is the consistently narrower confidence intervals produced by TCEV. This result runs counter to the expectation that flood probability models with more parameters are subject to greater extrapolative uncertainty. However, as the case study shows this may not always be the case. In the case study, component 2 produced more annual maximum samples than component 1. This resulted in component-2 parameters being well-identified in all cases including records of the order of 20 years. Because the right tail of the TCEV approaches a Gumbel distribution as component 2 becomes dominant, the confidence intervals for high quantiles will be tighter compared to LP3-PILF whose right-tail behavior is quite sensitive to uncertainty in the skew parameter.

Which high-quantile confidence intervals, those produced by TCEV or LP3-PILF, better represent the true uncertainty? While it is beyond the scope of this study to seek an answer to this question, it is worth noting some of the awkward issues that need to be confronted.

Given that goodness-of-fit assessment is limited by historical record lengths, the credibility of extrapolation becomes central to this question. To our knowledge, there is no known physical justification for LP3 as a model of flood frequency. The skew parameter can be thought of as a fitting device to allow curvature in the log-normal probability domain. On the other hand, TCEV has a physical foundation in so far that annual maxima are sampled from flow events generated by two independent processes. Totaro et al. (2024) provided a theoretical framework supported by physical evidence for justifying the application of TCEV distribution in eastern Australia. Extrapolation of the TCEV rests on the assumption that its right tail is sampled from events whose peaks are exponentially distributed and whose number is Poisson distributed. In principle, these are testable hypotheses using data other than AM. The fact that both TCEV components were identifiable in the majority of the long records in the case study (see Figs. 15 and 16) suggests that AM flows with AEP greater than 1 in 2 may be the product of processes different to those producing the rarer flows. Again, in principle, this is a testable hypothesis using data other than AM.

Two probabilistic approaches have been proffered as providing a plausible foundation for extrapolation. One is based on the theory of asymptotic distributions of extremes, the other on the theory of exact distributions. Each approach makes assumptions which must be validated if the approach is to provide a credible basis for extrapolation. If the number of independent events in a year is insufficient to ensure convergence, the asymptotic distributions approach is incapable of providing a credible basis for extrapolation. In a similar vein, the lack of knowledge about the true underlying distribution of event exceedances (Koutsoyiannis, 2004; Lombardo et al., 2019) undermines the extrapolative credibility of the exact distributions approach. However, the exact distributions approach is open to refinement. The challenge is to make the best use of our accumulated knowledge on the physics of floods and the extensive data on floods to identify with more confidence the underlying event exceedance distribution. We are of the view that the growing evidence of multiple flood populations and the derived distribution framework (Eagleson, 1972) will contribute to the refinement.

The Bayesian scheme for inferring TCEV parameters can exploit additional sources of information (e.g., historical or pre-gauged information) which can lead to significant reductions in quantile uncertainty. Viglione et al. (2013) highlight the versatility of Bayesian inference in using spatial, temporal and causal information in flood frequency hydrology. The use of historical information has received considerable attention in the literature (see, Stedinger and Cohn, 1986; Reis and Stedinger, 2005; Kjeldsen et al., 2014), to the point that its use has been incorporated in national flood frequency guidelines (e.g., Bulletin 17C, England et al., 2019; Australian Rainfall and Runoff, Ball et al., 2019). Furthermore, the physical basis underlying the derivation of the TCEV distribution supported by physical arguments for the case study of Eastern Australia (Totaro et al., 2024), offers the prospect of developing regional relationships that may lead to further reduction in quantile uncertainty.

While this study has presented empirical evidence to suggest that the TCEV has practical potential in at-site FFA, it is important to stress the tentative nature of this conclusion. The case study is representative of four clusters of catchments in eastern Australia with areas up to about 1000 km^2. The success of TCEV rests on the identifiability of the right-tail component. In this case study, more than half of the AM records were produced by the right-tail component so its two parameters were identifiable even using short records. In other regions, the right-tail component may be dominated by the left-tail component. How TCEV compares with LP3-PILF or other approaches in such regions needs to be investigated.

4. Conclusions

The issue of lack-of-fit in AM flood series can be tackled by censoring PILFs when fitting three-parameter probability models. An alternative and less-studied approach is to apply mixture probability models which offer greater flexibility to fit AM series. Mixture models may be additive or compound with compound models being more parsimonious. This study investigated the potential of using the four-parameter TCEV compound model in at-site flood frequency analysis. The conceptual foundation of TCEV relies on two independent processes contributing to AM floods. If one process dominates, degeneracy arises with the parameters of the other process not being identifiable. For this reason, a robust inference of TCEV parameters is required.

This study produced two significant outcomes.

The first addressed the issue of robust inference of TCEV parameters. Given L-moment estimation has been found to be questionable for TCEV, a Bayesian framework using Markov chain Monte Carlo sampling was developed. It produced new insight into the nature of degeneracy associated with TCEV. In particular, the posterior signatures for well-posed and degenerate cases were described. It was found that the signatures in degenerate cases were quite complex. To aid assessment of such degeneracy, two new posterior diagnostics based on the strength of the TCEV components were developed. This work represents the first known comprehensive assessment of identifiability in FFA using mixture models with four or more parameters. In the still wide-open field of model selection, the Bayesian framework and tools developed in this study offer interesting perspectives for further methodological developments in the comparison between different types of heavy tailed distributions, not excluding mixture distributions.

Armed with a robust inference framework, the study evaluated the potential of TCEV using a case study based on 31 catchments in eastern Australia with record lengths 70 or more years. The evaluation involved comparing TCEV fit and high quantile uncertainty against the LP3-PILF approach recommended in Australian Rainfall and Runoff (Ball et al., 2019).

The second significant outcome showed that, at least for the case study, the TCEV distribution is a suitable choice for at-site flood frequency analysis. TCEV was shown to deal with the dog-leg signatures encountered in the flood frequency plots. Overall, there appears to be no clear difference in goodness-of-fit between TCEV and LP3-PILF. Unlike LP3-PILF, TCEV was sufficiently flexible to fit all AM data without the need to censor low flows. However, and contrary to the expectation that more parameters lead to greater high quantile uncertainty, TCEV produced consistently narrower high quantile confidence intervals than LP3-PILF. It was found that TCEV extrapolative uncertainty depends on the number of AM floods in the record being generated by the right-tail component process. While more case studies in different regions using the TCEV Bayesian framework are needed, this study has demonstrated that TCEV can produce fits comparable to LP3-PILF with significantly narrower high quantile confidence intervals.

This study does raise a significant and difficult-to-answer question, which high-quantile confidence intervals, those produced by TCEV or LP3-PILF, better represent the true uncertainty? While it is unlikely this is a genuinely binary choice, the question deserves serious consideration. Goodness-of-fit is a necessary but not sufficient criterion for selecting the probability model that best represents flood frequency. The path to sufficiency will likely require probability models to better reflect the physical basis of AM floods.

CRediT authorship contribution statement

Vincenzo Totaro: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **George Kuczera:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Vito Iacobellis:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Annual maxima of peak flows are provided by the Bureau of Meteorology (BoM) of Australian Government and are freely available at the website http://www.bom.gov.au/waterdata/.

Acknowledgements

The authors thank Mark Babister and Monique Retallick for their advice and access to LP3 results for a range of Australian catchments.

Data statement.

Annual maxima of peak flows are provided by the Bureau of Meteorology (BoM) of Australian Government and are freely available at the website http://www.bom.gov.au/waterdata/.

References

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

Arnell, N., Beran, M., 1988. Probability-weighted moments estimators for TCEV parameters. Wallingford.

Ball, J., Babister, M., Nathan, R., Weeks, W., Weinmann, E., Retallick, M., Testoni, I., 2019. A guide to Australian Rainfall and Runoff. Retrieved from http://hdl.handle.net/11343/119609.

Barth, N.A., Villarini, G., White, K., 2019. Accounting for mixed populations in flood frequency analysis: Bulletin 17C PERSPECTIVE. J. Hydrol. Eng. 24 https://doi.org/10.1061/(asce)he.1943-5584.0001762.

Beran, M., Hosking, J.R.M., Arnell, N., 1986. Comment on "Two-component extreme value distribution for flood frequency analysis" by fabio rossi, mauro fiorentino, and pasquale versace. Water Resour. Res. 22, 263–266. https://doi.org/10.1029/WR022i002p00263.

Cohn, T.A., Lane, W.L., Baier, W.G., 1997. An algorithm for computing moments-based flood quantile estimates when historical flood information is available. Water Resour. Res. 33, 2089–2096. https://doi.org/10.1029/97WR01640.

Cohn, T.A., England, J.F., Berenbrock, C.E., Mason, R.R., Stedinger, J.R., Lamontagne, J. R., 2013. A generalized grubbs-beck test statistic for detecting multiple potentially influential low outliers in flood series. Water Resour. Res. https://doi.org/10.1002/wrcr.20392.

Connell, R.J., Pearson, C.P., 2001. Two-component extreme value distribution applied to Canterbury annual maximum flood peaks. J. Hydrol. (new Zealand) 40, 105–127.

Duan, Q.Y., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and efficient global minimization. J. Optim. Theory Appl. 76, 501–521. https://doi.org/10.1007/BF00939380.

Eagleson, P.S., 1972. Dynamics of flood frequency. Water Resour. Res 8, 878–898. https://doi.org/10.1029/WR008i004p00878.

England, J.F., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Veilleux, A.G., et al., 2019. Guidelines for determining flood flow frequency-Bulletin 17C. Techniques and Methods, Book 4, chap. B5.

Erskine, W.D., Warner, R.F., 1988. Geomorphic effects of alternating flood- and drought-dominated regimes on NSW coastal rivers. In: Warner, R.F. (Ed.), Fluvial Geomorphology of Australia. Academic Press, Marrickville, NSW, pp. 223–244 (Chapter 11).

Evin, G., Merleau, J., Perreault, L., 2011. Two-component mixtures of normal, gamma, and Gumbel distributions for hydrological applications. Water Resour. Res. 47, 1–21. https://doi.org/10.1029/2010WR010266.

Fiorentino, M., Gabriele, S., Rossi, F., Versace, P., 1987. Hierarchical approach for regional flood frequency analysis. In: Singh, V.P. (Ed.), Regional Flood Frequency Analysis. D. Reidel, Norwell, Mass, pp. 35–49.

Franks, S.W., Kuczera, G., 2002. Flood frequency analysis: Evidence and implications of secular climate variability, New South Wales. Water Resour. Res. 38 (5) https://doi.org/10.1029/2001wr000232.

Gelman, A., Carlin J., Stern H.S., Rubin D.B., 1995. Bayesian Data Analysis, Chapman and Hall, New York, 526 pp.

Grego, J.M., Yates, P.A., 2010. Point and standard error estimation for quantiles of mixed flood distributions. J. Hydrol. 391, 289–301. https://doi.org/10.1016/j.jhydrol.2010.07.027.

Griffis, V.W., Stedinger, J.R., Cohn, T.A., 2004. Log pearson type 3 quantile estimators with regional skew information and low outlier adjustments. Water Resour. Res. 40 https://doi.org/10.1029/2003WR002697.

Iacobellis, V., Gioia, A., Manfreda, S., Fiorentino, M., 2011. Flood quantiles estimation based on theoretically derived distributions: Regional analysis in southern Italy. Nat. Hazards Earth Syst. Sci 11 (3), 673–695. https://doi.org/10.5194/nhess-11-673-2011.

Kiem, A.S., Franks, S.W., Kuczera, G., 2003. Multi-decadal variability of flood risk. Geophys. Res. Lett. 30 (2) https://doi.org/10.1029/2002GL015992.

Kjeldsen, T.R., Macdonald, N., Lang, M., Mediero, L., Al-buquerque, T., Bogdanowicz, E., Brázdil, R., Castellarin, A., David, V., Fleig, A., Gül, G.O., Kriauciuniene, J., Kohnová, S., Merz, B., Nicholson, O., Roald, L.A., Salinas, J.L., Sarauskiene, D., Šraj, M., Strupczewski, W., Szolgay, J., Toumazis, A., Vanneuville, W., Veijalainen, N., Wilson, D., 2014. Documentary evidence of past floods in Europe and heir utility in flood frequency estimation. J. Hydrol. 517, 963–973. https://doi.org/10.1016/j. hydrol.2014.06.038.

Kjeldsen, T.R., Ahn, H., Prosdocimi, I., Heo, J.H., 2018. Mixture Gumbel models for extreme series including infrequent phenomena. Hydrol. Sci. J. https://doi.org/10.1080/02626667.2018.1546956.

Klemeš, V., 1986. Dilettantism in hydrology: Transition or destiny? Water Resour. Res. 22, 177S–188S. https://doi.org/10.1029/WR022i09Sp0177S.

Koutsoyiannis, D., 2004. Statistics of extremes and estimation of extreme rainfall: I. Theoretical Investigation. Hydrol. Sci. J. 49 (4), 575–590. https://doi.org/10.1623/hysj.49.4.575.54430.

Lamontagne, J.R., Stedinger, J.R., Xin, Y., Whealton, C.A., Xu, Z., 2016. Robust flood frequency analysis: Performance of EMA with multiple Grubbs-Beck outlier tests. Water Resour. Res. 52, 3068–3084. https://doi.org/10.1002/2015WR018093.

Lombardo, F., Napolitano, F., Russo, F., Koutsoyiannis, D., 2019. On the exact distribution of correlated extremes in hydrology. Water Resour. Res. 55, 10405–10423. https://doi.org/10.1029/2019WR025547.

Marin, J.M., Mengersen, K., Robert, C.P., 2005. Bayesian modelling and inference on mixtures of distributions. Handb. Stat. 25, 459–507. https://doi.org/10.1016/S0169-7161(05)25016-2.

Micevski, T., Franks, S.W., Kuczera, G., 2006. Multidecadal variability in coastal eastern Australian flood data. J. Hydrol. 327 (1–2), 219–225. https://doi.org/10.1016/j.jhydrol.2005.11.017.

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

Rossi, F., Fiorentino, M., Versace, P., 1984. Two-component extreme value distribution for flood frequency analysis. Water Resour. Res. https://doi.org/10.1029/WR020i007p00847.

Shin, J.Y., Lee, T., Ouarda, T.B.M.J., 2015. Heterogeneous mixture distributions for modeling multisource extreme rainfalls. J. Hydrometeorol. 16, 2639–2657. https://doi.org/10.1175/JHM-D-14-0130.1.

Stedinger, J.R., Cohn, T.A., 1986. Flood frequency analysis with historical and paleoflood information. Water Resour. Res. 22 (5), 785–793.

Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events. In: Maidment, D. (Ed.), Handbook of Hydrology. McGraw-Hill, N. Y (chap. 18).

Strupczewski, W.G., Kochanek, K., Bogdanowicz, E., Markiewicz, I., 2012. On seasonal approach to flood frequency modelling. Part I: Two-component distribution revisited. Hydrol. Process. 26 (5), 705–716. https://doi.org/10.1002/hyp.8179.

Todorovic, P., 1970. On some problems involving random number of random variables. Ann. Math. Stat. 41, 1059–1063.

Todorovic, P., 1978. Stochastic models of floods. Water Resour. Res. 14, 345–356.

Totaro, V., Gioia, A., Kuczera, G., Iacobellis, V., 2024. Modelling multidecadal variability in flood frequency using the two-component extreme value distribution. Stoch Environ. Res. Risk Assess 1–18. https://doi.org/10.1007/s00477-024-02673-8.

Viglione, A., Merz, R., Salinas, J.L., Bloschl, G., 2013. Flood frequency hydrology: 3. A Bayesian Analysis. Water Resour. Res. 49 (2), 675–692.

Vogel, R., Wilson, I., 1996. The probability distribution of annual maximum, minimum and average streamflow in the United States. J. Hydrol. Eng. 1, 69–76. https://doi.org/10.1061/(ASCE)1084-0699(1996)1:2(69).

Wang, Q.J., 1997. LH moments for statistical analysis of extreme events. Water Resour. Res. 30, 2841–2848. https://doi.org/10.1029/97WR02134.

Waylen, P.R., Caviedes, C.N., Juricic, C., 1993. El nino-southern oscillation and the surface hydrology of Chile: A window on the future? Can. Water Resour. J. 18, 425–441. https://doi.org/10.4296/cwrj1804425.

Webb, R.H., Betancourt, J.L., 1992. Climatic variability and flood frequency of the Santa Cruz River, Pima County. US Geol. Surv. Water-Supply Pap. 2379.

Yan, L., Xiong, L., Ruan, G., Xu, C.Y., Yan, P., Liu, P., 2019. Reducing uncertainty of design floods of two-component mixture distributions by utilizing flood timescale to classify flood types in seasonally snow covered region. J. Hydrol. 574, 588–608. https://doi.org/10.1016/j.jhydrol.2019.04.056.

Zheng, J., Frey, H.C., 2004. Quantification of variability and uncertainty using mixture distributions: Evaluations of sample size, mixing weights, and separation between components. Risk Anal. 24, 553–571. https://doi.org/10.1111/j.0272-4332.2004.00459.x.