

Rethinking FFA?

A tale about a tail

George Kuczera¹, Vincenzo Totaro², Vito Iacobellis²,
Mark Babister³, Monique Retallick³

¹The University of Newcastle, Australia

²Politecnico di Bari, Italy

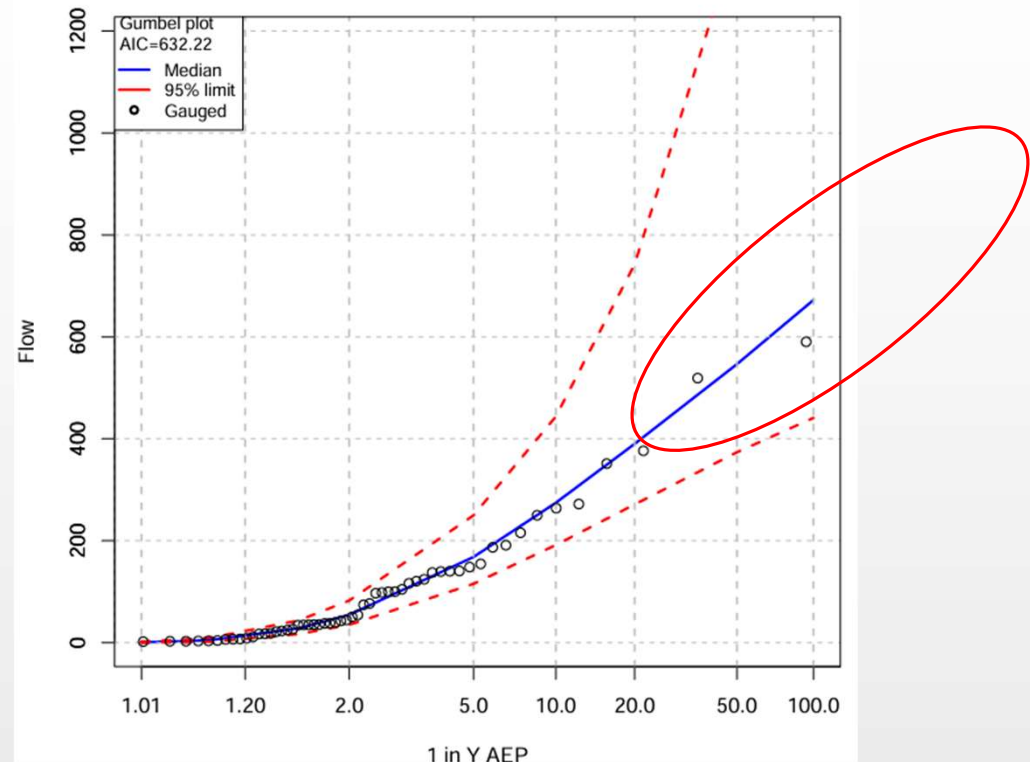
³WMAwater, Australia

Fit-for-purpose FFA

Practical aim of FFA is to provide estimates of flood frequency for large floods whose magnitudes are likely to be outside the observed range.

A fit-for-purpose FFA model should:

- ❑ Fit observed data satisfactorily
- ❑ Be well behaved (accurate and unbiased) when extrapolated

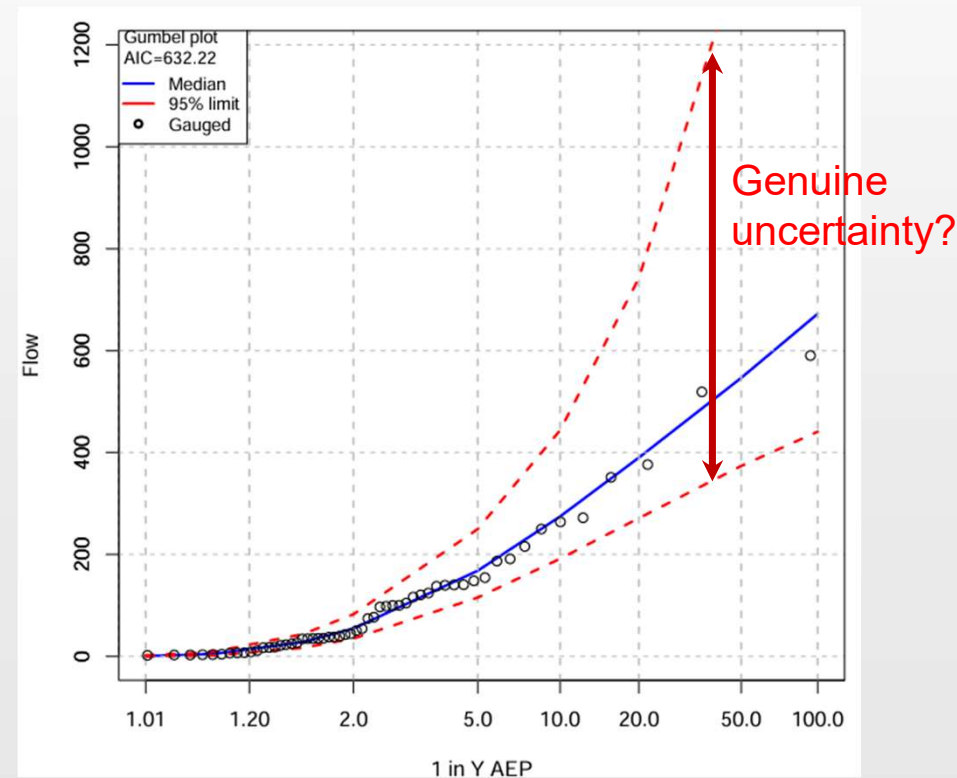


Key question?

How do we choose between models that have similar goodness-of-fit but different right tail behaviour?

All is not well with the right tail!

- ❑ FFA is quite mature but extrapolation has rocky foundations:
 - We are quite clever at obtaining good fits to observed AM
 - Good fit is necessary but **not sufficient** for well-behaved extrapolation
- ❑ Long-held concerns:
 - Extrapolated quantiles can differ significantly between good-fit models
 - Some models can produce explosively large quantile uncertainty
 - Not all annual maxima (AM) are floods
 - Small and large AM flows can be driven by different processes



Not all annual maxima are floods?

Some sage insights

- ❑ “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” *Klemes (1986)*
- ❑ “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” *Lamontagne et al. (2016)*



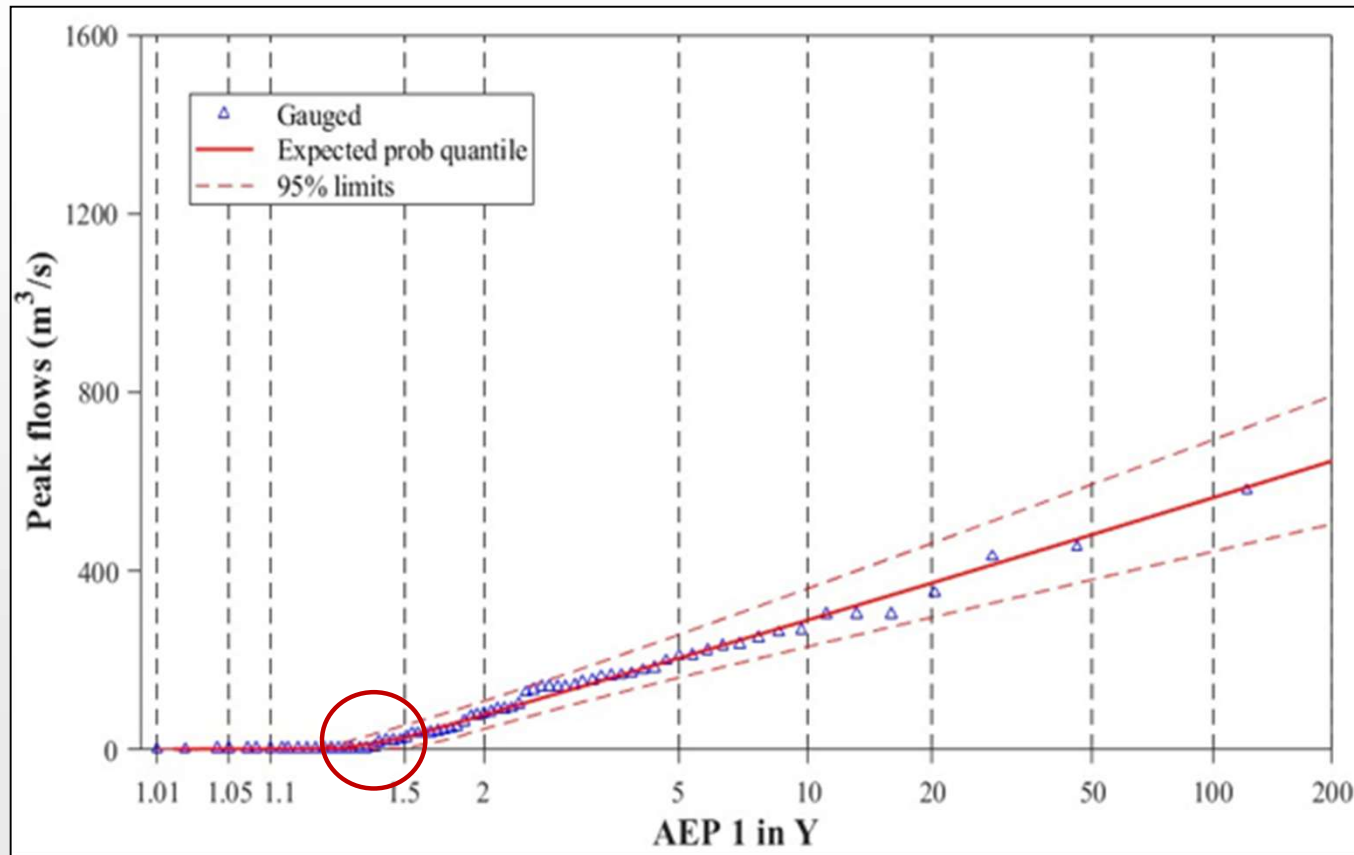
Many annual maxima are within-bank flows



Some annual maxima are genuine floods

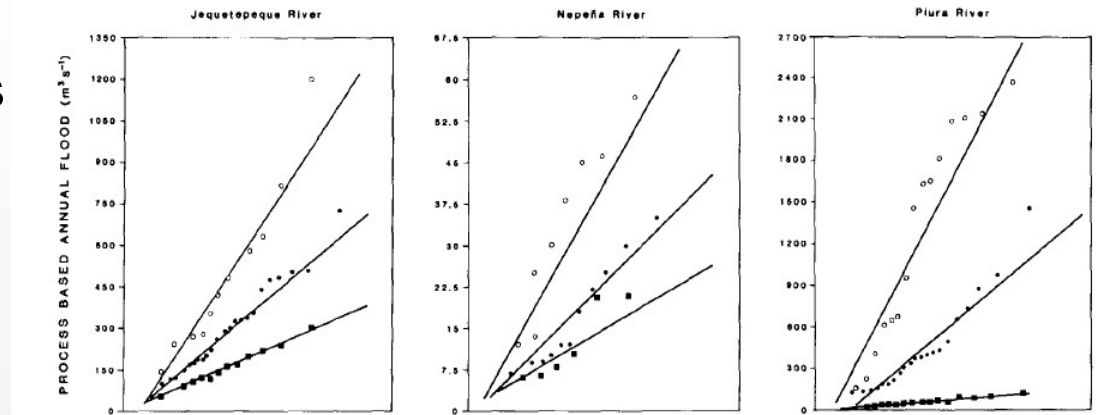
Evidence for different AM processes

“Dog leg” signature → transition from one to another process?



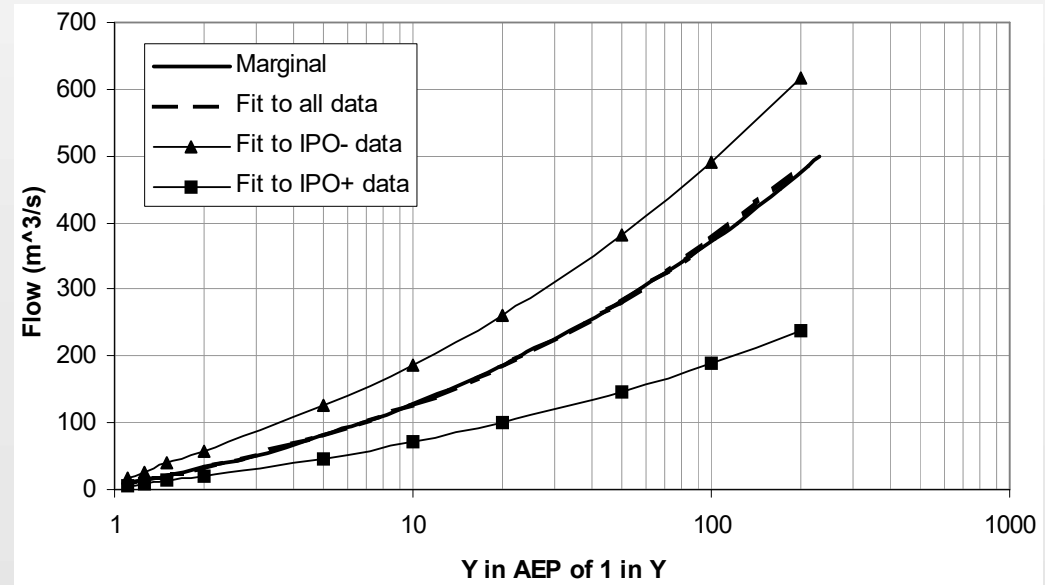
Evidence for different AM processes

- Different FF curves when we stratify El-Nino and La-Nina years



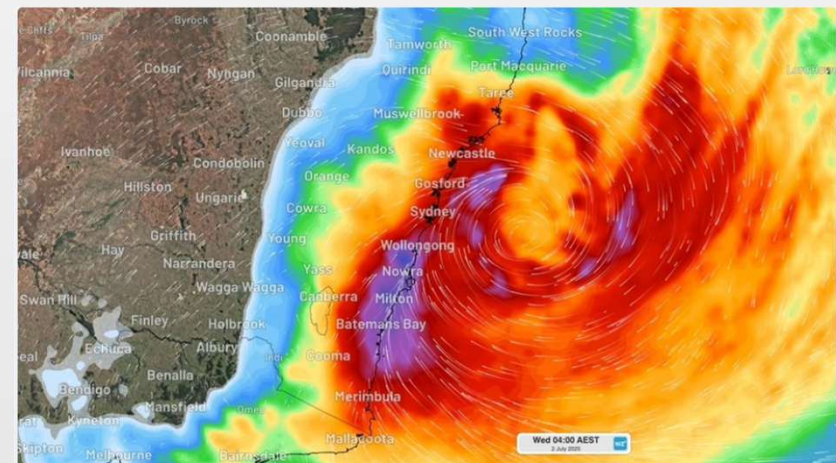
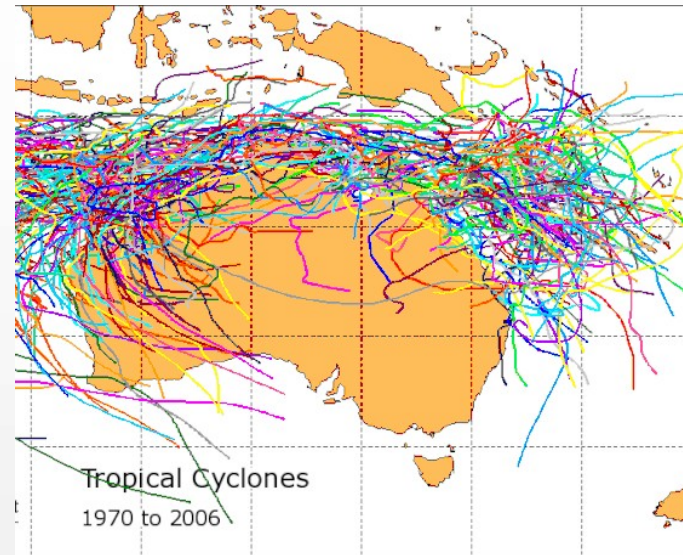
- Different FF curves for Inter Pacific Oscillation (IPO) wet and dry epochs

- IPO wet epochs have much higher frequency of La Nina



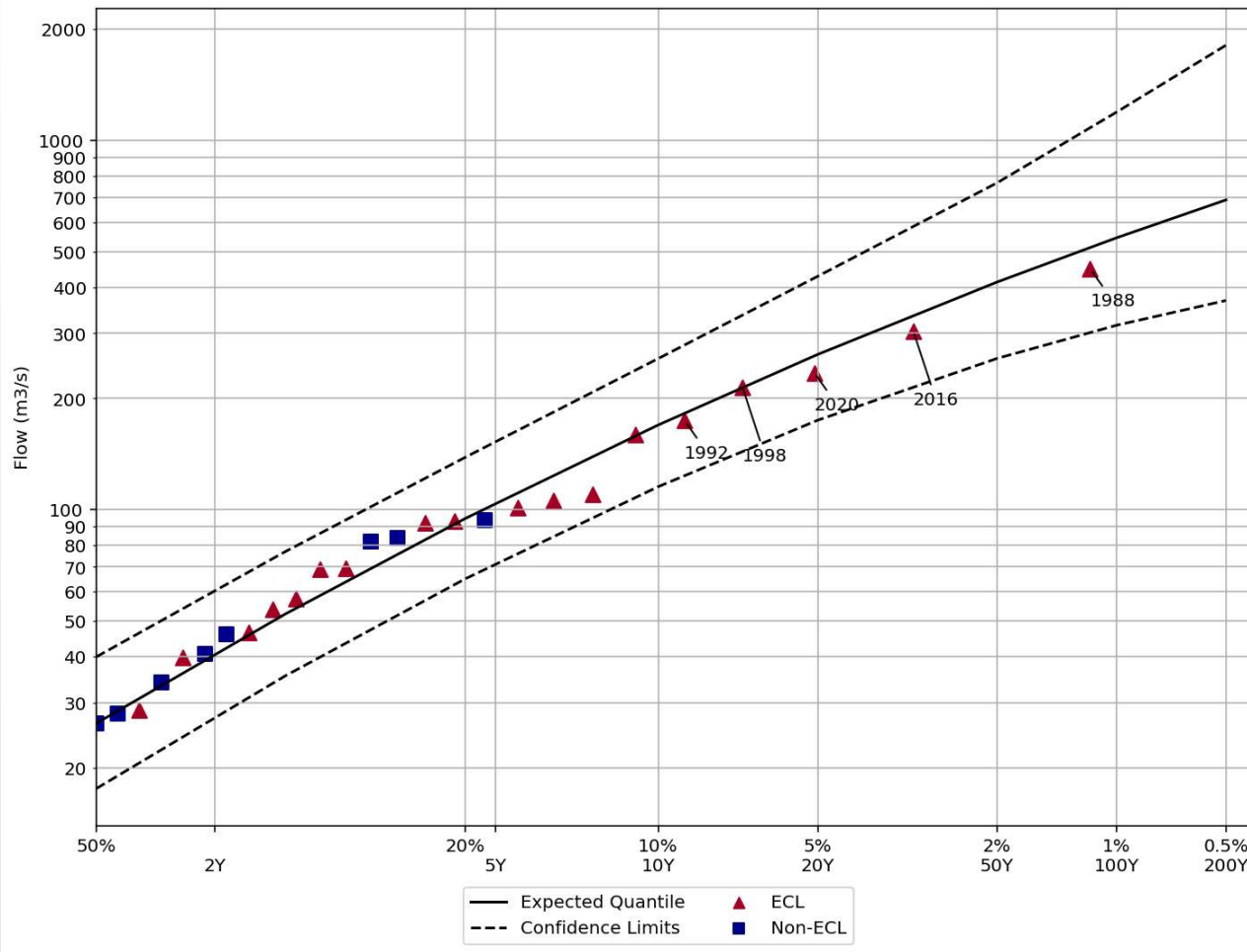
Evidence for different AM processes

- ❑ Cyclones in Qld and WA drive large floods
- ❑ East coast lows in NSW and Qld drive large floods



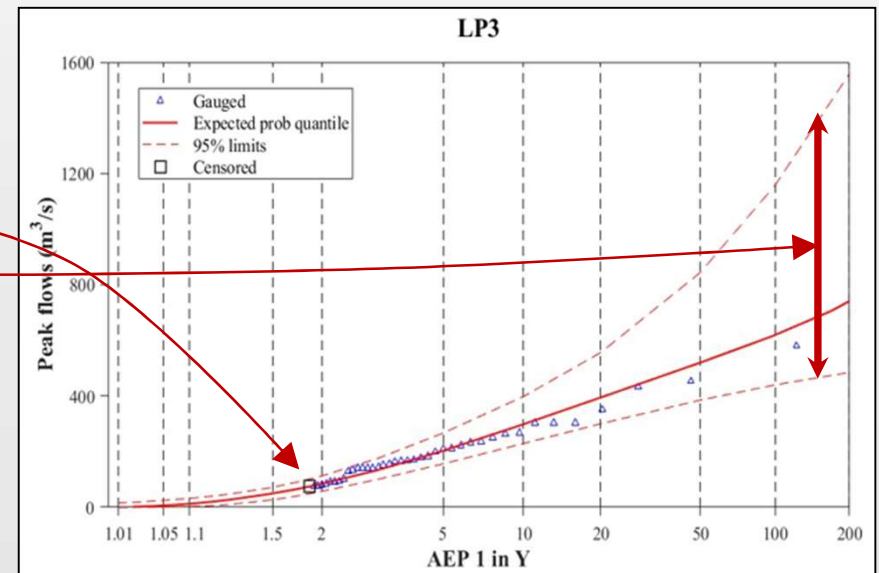
Evidence for different AM processes

Storm typing of AM events



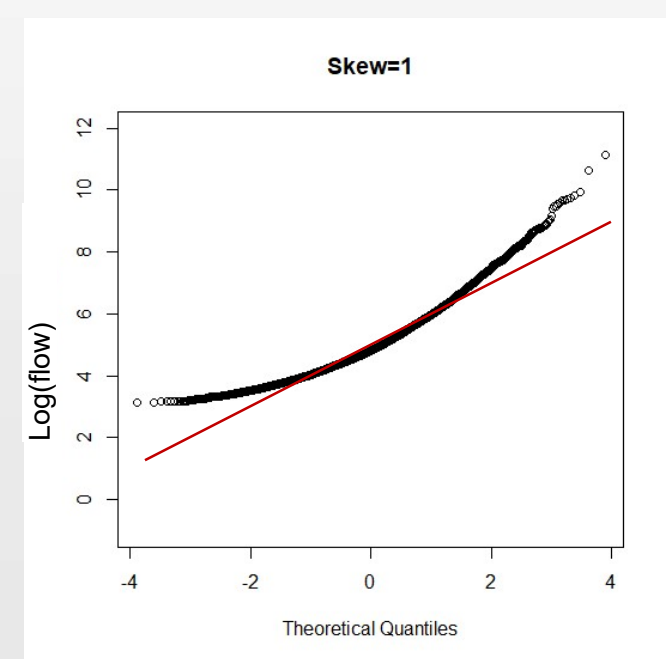
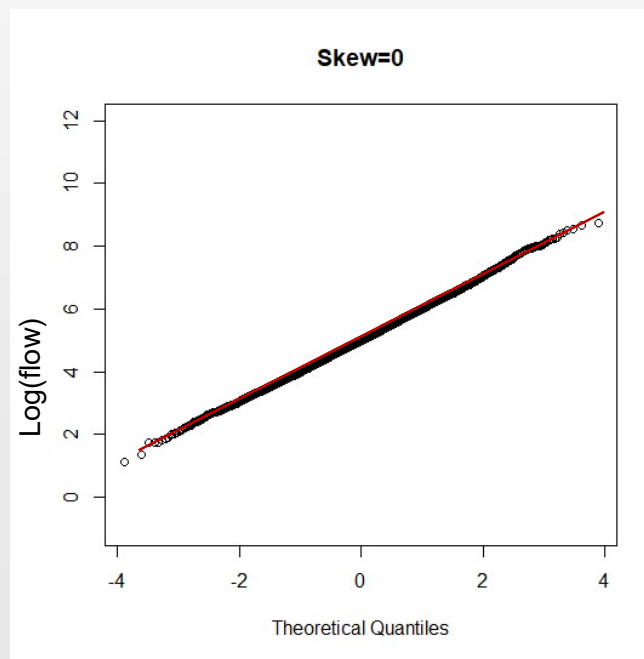
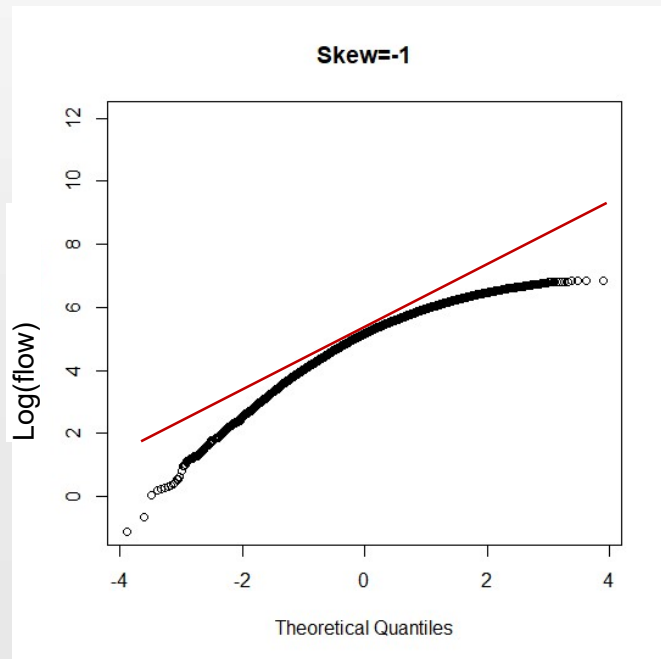
Why do we fit to small AM flows?

- Most common distributions used in practice
 - Log Pearson III (LP3) → 3 parameters (mean, std dev, skew or shape)
 - Generalized Extreme Value (GEV) → 3 parameter (location scale, shape)
- We fit LP3 and GEV to all AM data to minimize uncertainty in the shape parameter and in quantile extrapolation
- **Except** when potentially influential low flows (PILFs) are detected:
 - Censor PILFs to favour right tail fit
 - Can lead to larger quantile uncertainty



LP3 tail behaviour

- ❑ LP3 has NO physical basis
- ❑ The skew parameter of the LP3 distribution is purely geometric
- ❑ It allows “quadratic” curvature on a **log-normal probability plot** to improve fit:
 - Skew = 0 plots a straight line
 - Skew > 0 plots with upward curvature
 - Skew < 0 plots with downward curvature



The sting in the tail

LP3 1-in-Y AEP¹ quantile can be written as

$$w_Y = \frac{\log q - \mu}{\sigma} \approx Z_Y + \frac{k}{6} (Z_Y^2 - 1) \text{ for } |k| < 1$$

where μ is mean of $\log q$,

σ is std dev of $\log q$,

k is log skew, and

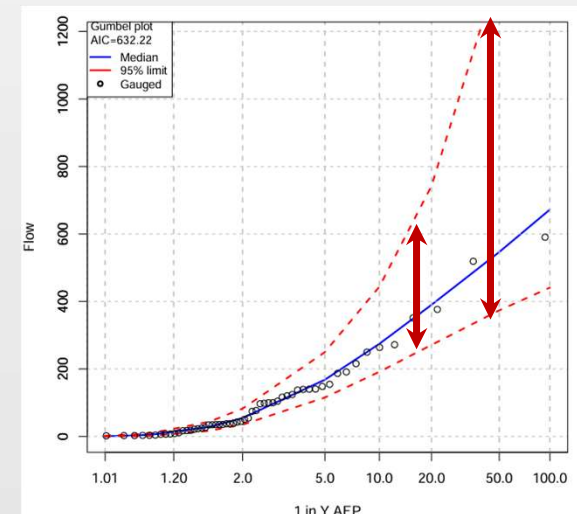
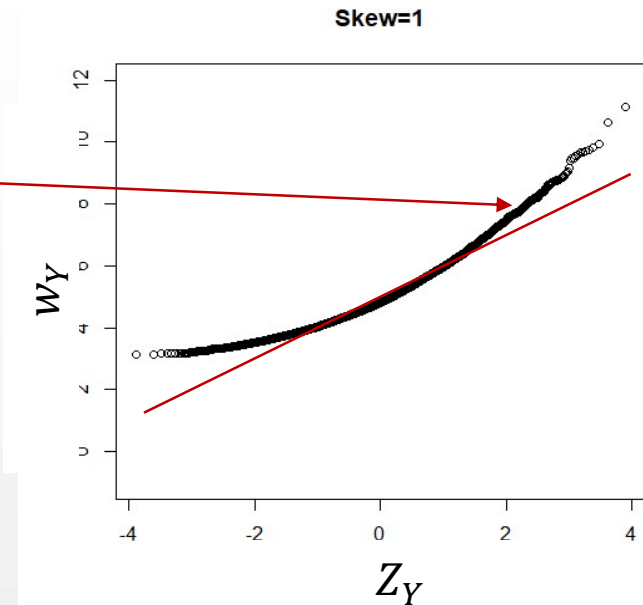
Z_Y is normal 1-in-Y AEP quantile.

Standard deviation of quantile uncertainty is

$$\sigma(w_Y) \approx \left| \frac{1}{6} (Z_Y^2 - 1) \right| \sigma(k)$$

Quantile uncertainty grows quadratically with Z_Y

¹AEP annual exceedance probability



Alternatives to LP3 and GEV?

- ❑ If different processes are responsible for small and large AM flows
 - Small AM flow data convey little useful information about the large flow process
 - Little sense using shape parameter to quadratically link small and large AM flows
- Want to decouple small and large AM flows

- ❑ Two-component mixture models can "decouple" small and large AM flows

- ❑ Application is problematic

- ❑ Most mixture models have five or more parameters → unstable estimation
 - ❑ Need to stratify long records by component process

- ❑ Most promising mixture model is the four-parameter Two-Component Extreme Value (TCEV) (Rossi et al., 1984)

Two-Component Extreme Value (TCEV)

- Two independent processes (or components) generate peak flows during year

Component 1 (small flow events)

- Sample number of events in a year from a Poisson distribution with mean λ_1
- Sample peak flow for each event from an exponential distribution with mean θ_1

Component 2 (large flow events)

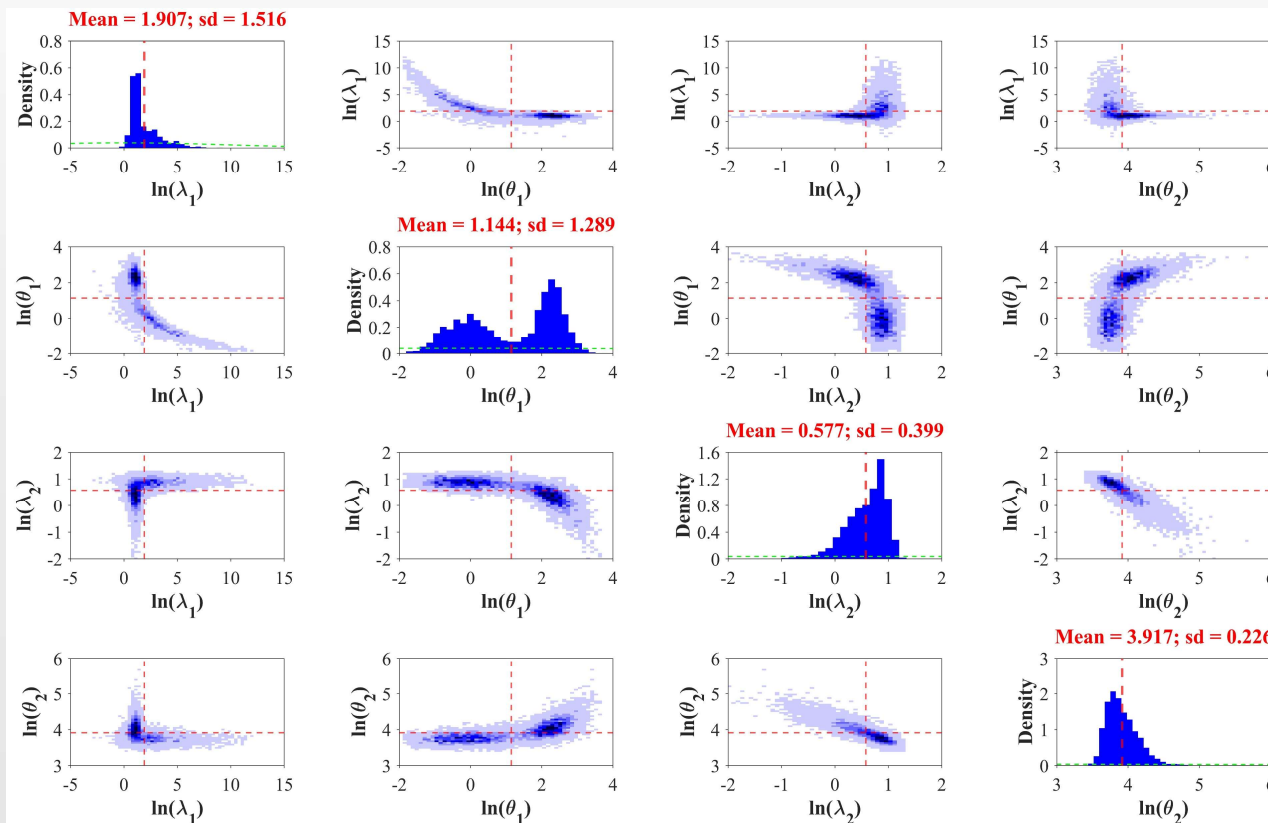
- Sample number of events in a year from a Poisson distribution with mean λ_2
- Sample peak flow for each event from an exponential distribution with mean θ_2

- Rossi et al. (1984) used total probability theorem to get distribution of annual maximum flow

$$F(q) = P(Q \leq q) = \exp(-\lambda_1 \exp(-q/\theta_1)) \exp(-\lambda_2 \exp(-q/\theta_2)), \quad \theta_1 \leq \theta_2$$

TCEV inference

- Inference of TCEV parameters is much less straightforward than for LP3 or GEV
 - One component may not be identifiable if the other component dominates the sample
 - L moments don't work
- Totaro et al. (2024) show Bayesian inference provides robust inference & diagnostics



Barnard river at Barry
Parameter posterior illustrating
poorly identified component 1

TCEV has a parsimonious right tail

- A very important property of the TCEV is its right tail behaviour

$$F(q) \rightarrow \exp(-\lambda_2 \exp(-q/\theta_2)) \text{ as } q \rightarrow \infty$$

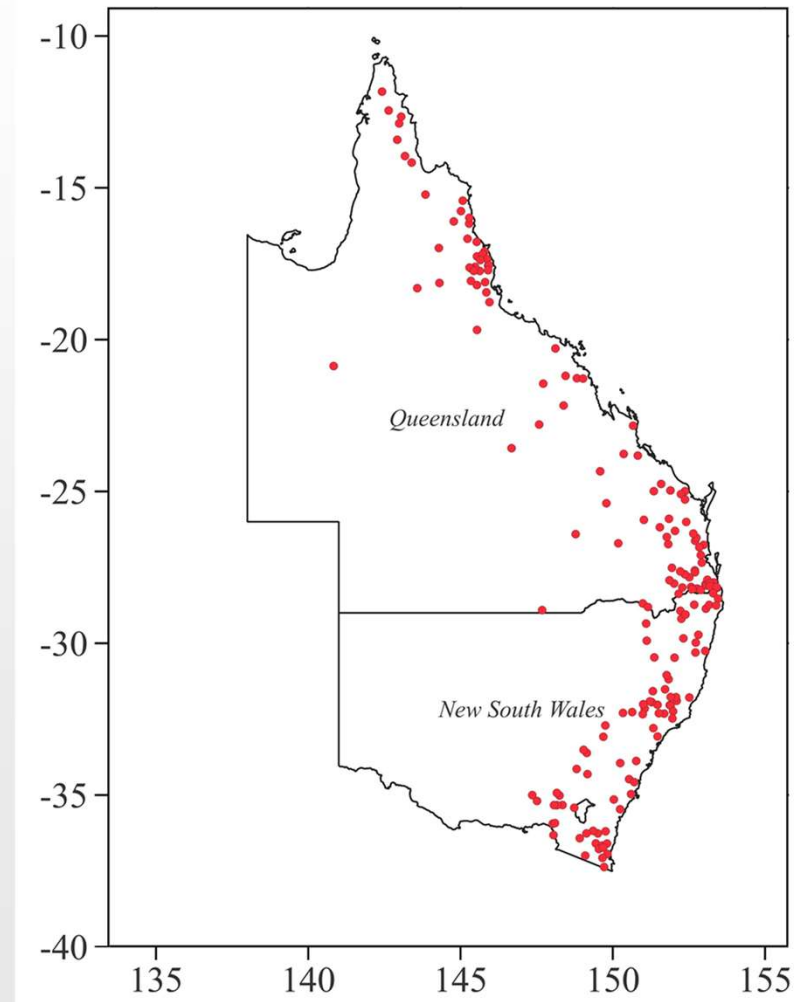
For large q , TCEV right tail is a 2-parameter Gumbel distribution

- If the two parameters of the big flood component are well identified by the data, quantile confidence limits are unlikely to explode when TCEV is extrapolated

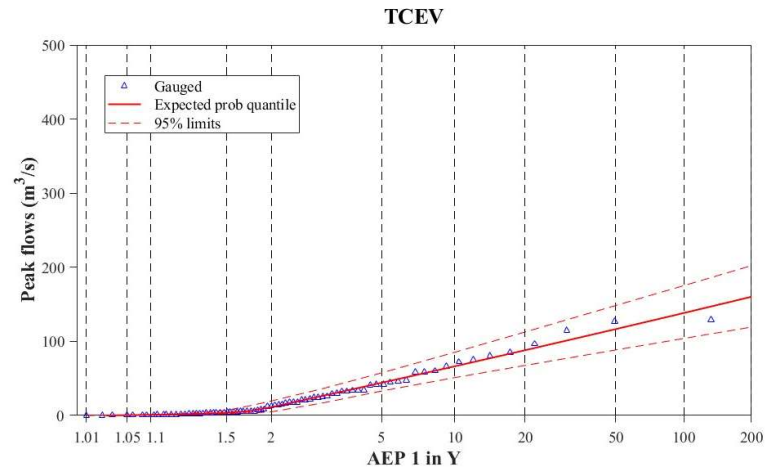
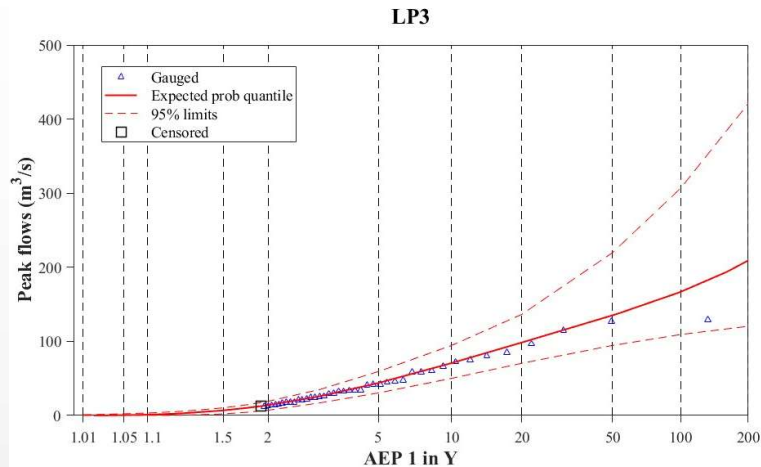
Case study: Eastern Australia

- ❑ 178 stations in NSW and Qld
 - ARR flood data base
 - Records ≥ 50 years
 - Areas range from 7 to 4930 km²
 - Average area 642 km²

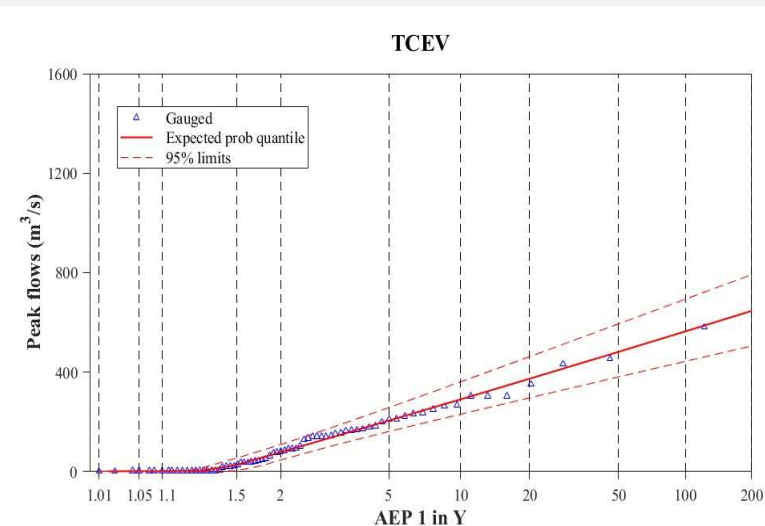
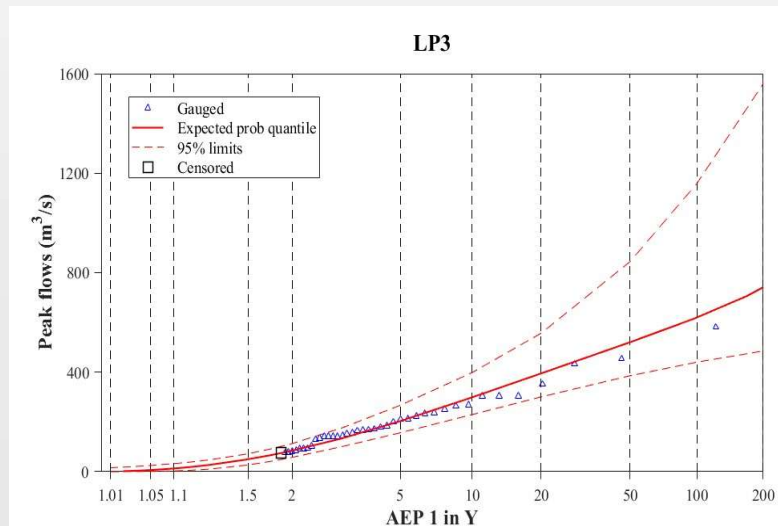
- ❑ Compare LP3, GEV and TCEV
 - no censoring
 - no priors



LP3 vs TCEV fits



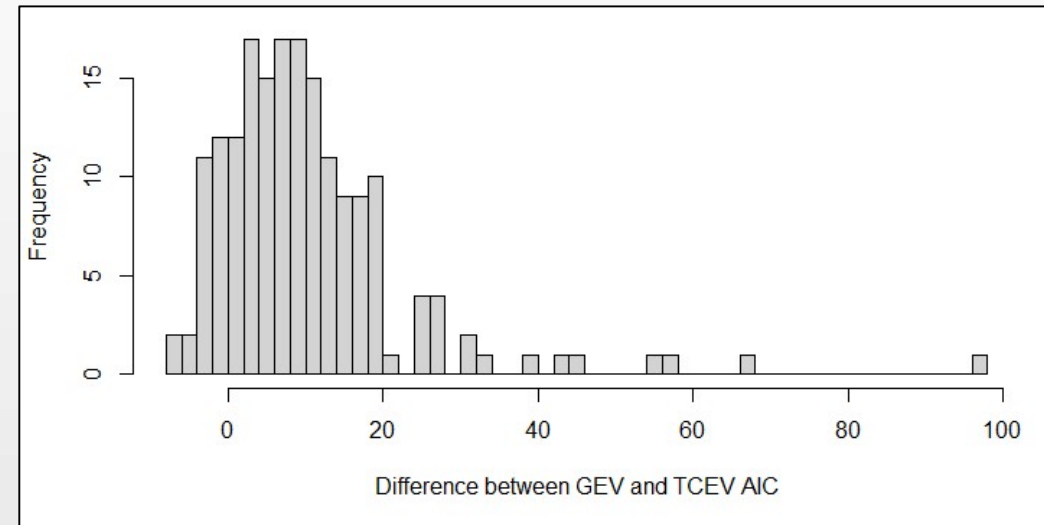
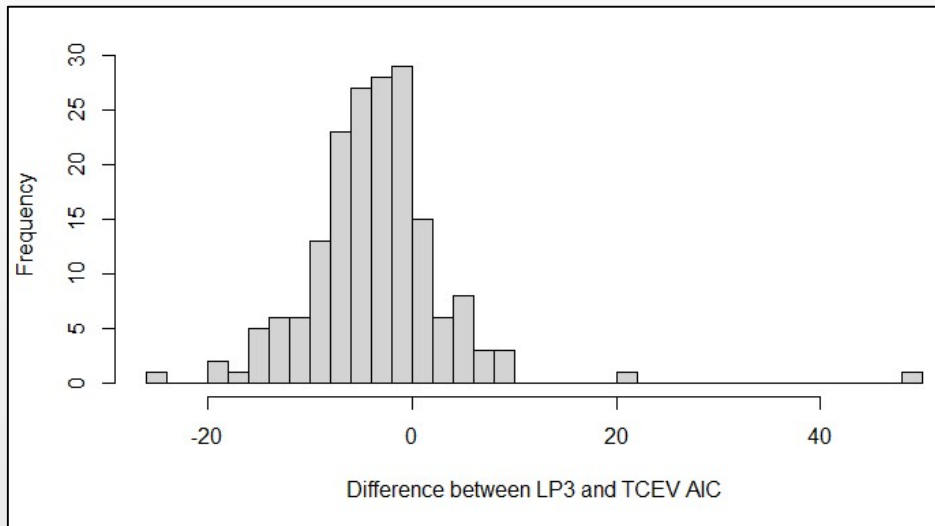
219001 - Rutherford Creek at Brown Mountain



208008 - Gloucester River at Forbesdale

Case study: Goodness-of-Fit

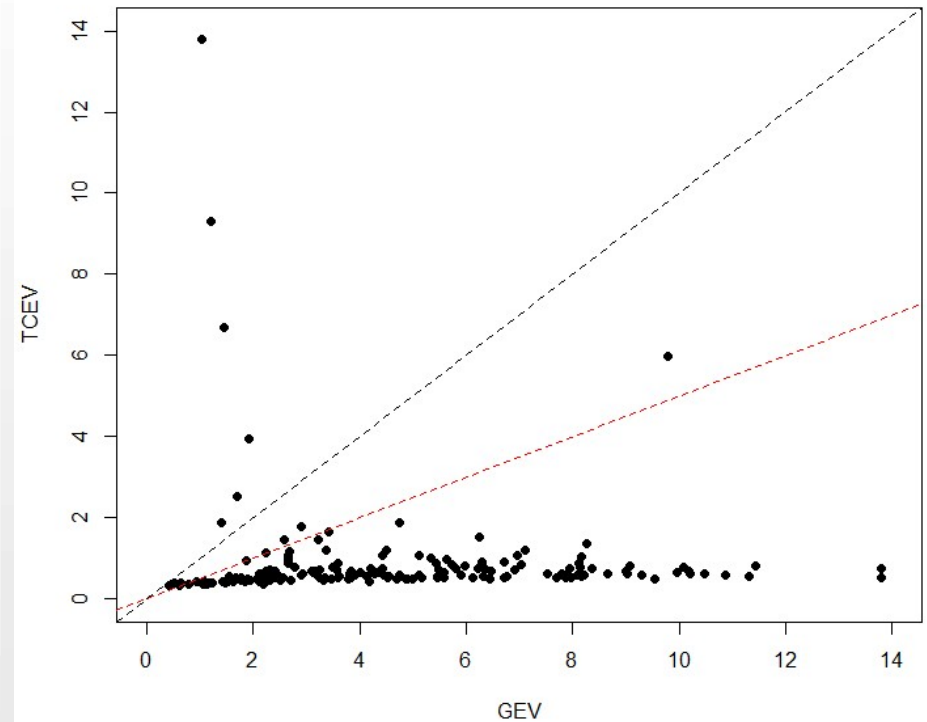
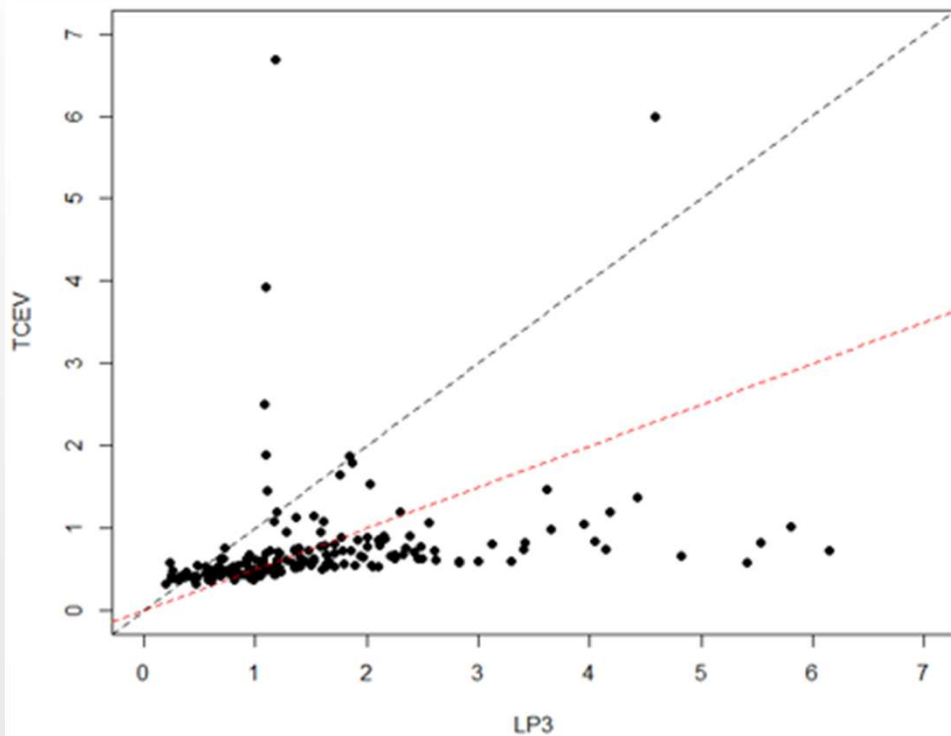
- ❑ Compare LP3, GEV and TCEV with no censoring or priors
- ❑ Use Akaike information criterion (lower is better)



- ❑ LP3 marginally better than TCEV
- ❑ GEV clearly inferior to LP3 and TCEV

Case study: Quantile uncertainty

- Compare 1-in-100 AEP normalized uncertainty = $\frac{95\% \text{ predictive interval}}{\text{Estimated quantile}}$



Dilemma

- ❑ LP3 and TCEV have competitive goodness of fit
- ❑ LP3 and TCEV have different right tail behaviour when extrapolated
- ❑ LP3 large quantile uncertainty > TCEV quantile uncertainty
- ❑ TCEV conceptually more appealing than LP3

We do NOT KNOW which model better describes large quantiles beyond the observed data!

One solution is to institutionalize model(s)

A better solution is to devise a data-based test that can discriminate between different right tail behaviour

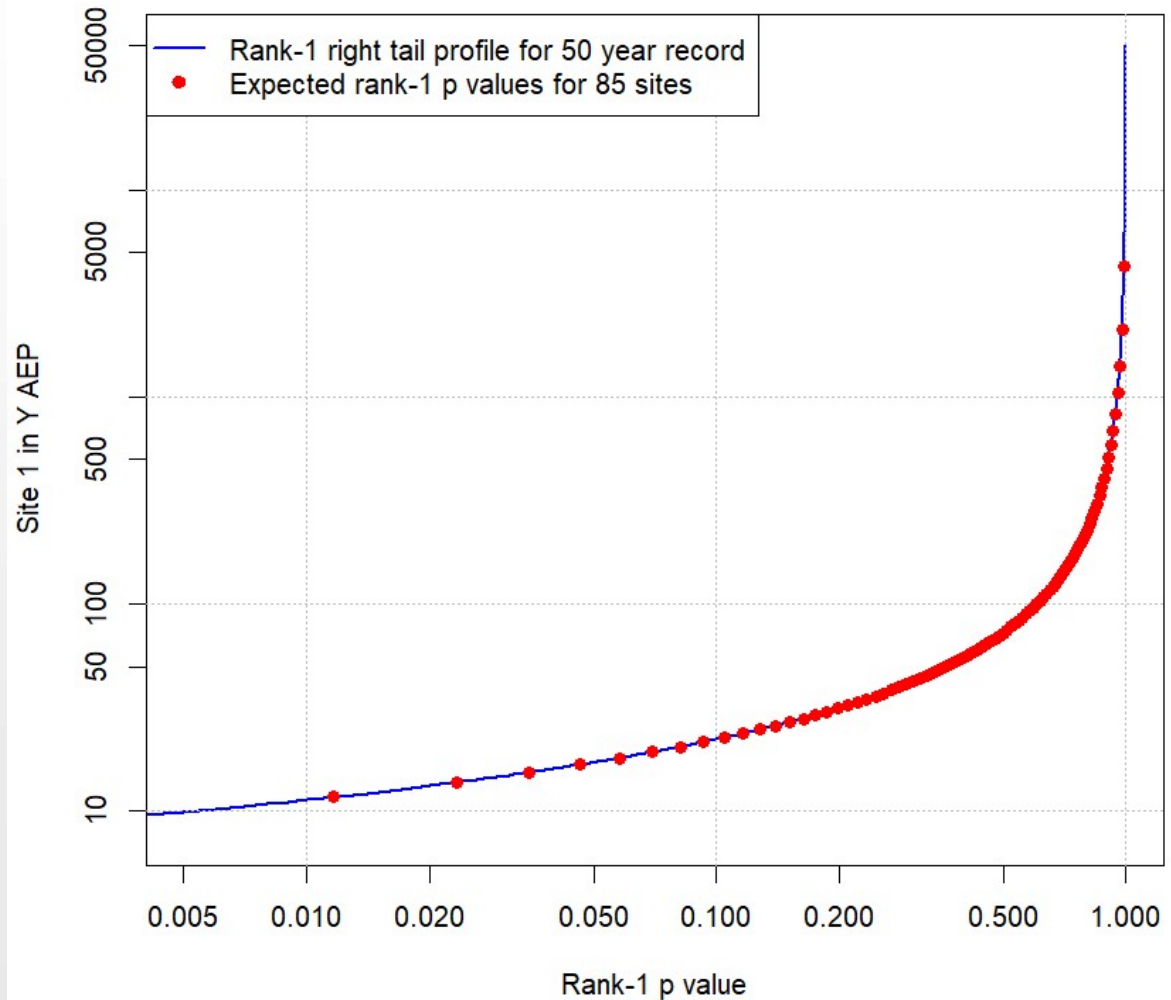
- ❑ Longer gauged records X
- ❑ Paleo records?
- ❑ Trade space for time → use regional data

Motivation for data-based test

What are the AEPs of rank-1 (or maximum) AM flows in a region with 85 independent 50-year records?

- ❑ 50% of rank-1 flows will have AEP values less than 1 in 73
- ❑ 25% will AEP values less than 1 in 170.
- ❑ Largest rank-1 flow has an expected AEP of 1 in 4275

Rank-1 AM flows in a region have a lot of information about right tail



Rank-1 order statistics

Plotting position of rank-1 flow $q^{(1)}$

$$F(q^{(1)}|\mathcal{M}, \theta) = P(Q \leq q^{(1)}|\mathcal{M}, \theta)$$

where $\mathcal{M}$ is the model, and

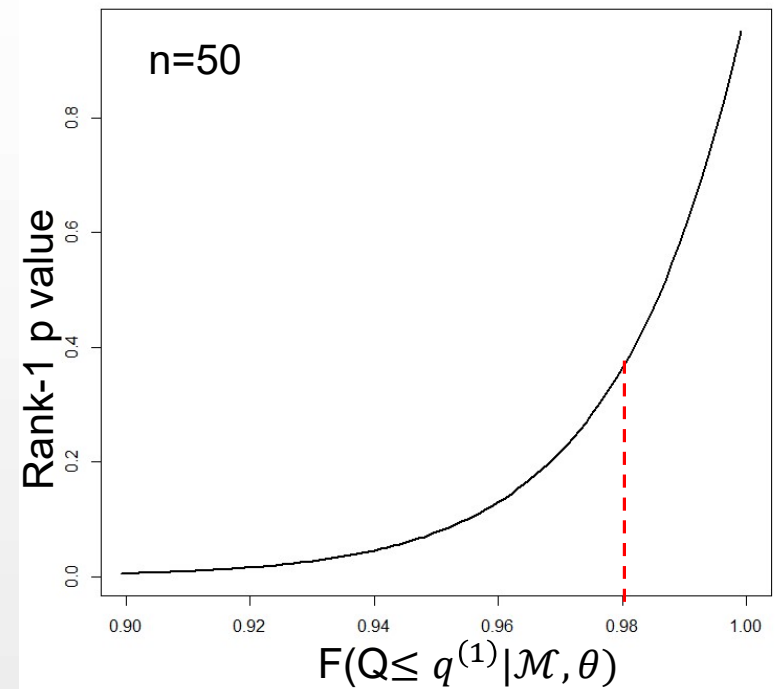
θ is the model parameter.

The distribution function for the rank-1 flow is

$$\mathcal{P}^{1,\mathcal{M}} = P(Q^{(1)} \leq q^{(1)}|\mathcal{M}, \theta) = F(q^{(1)}|\mathcal{M}, \theta)^n$$

where n is the record length.

Joint
probability of
all observed
 $AM \leq q^{(1)}$



Suppose we have a set of observed rank-1 flows

Rank-1 p value $\mathcal{P}^{1,\mathcal{M}}$ will follow a **uniform** distribution provided model $\mathcal{M}$ and its parameter θ are true

Independent of record length n

Understanding the rank-1 p value

Use R script to simulate distribution of rank-1 p value:

Repeat

- Draw 50 samples from a normal distribution with mean 10 and std dev 3

- Get rank-1 (or largest) value from sample

- Compute p value

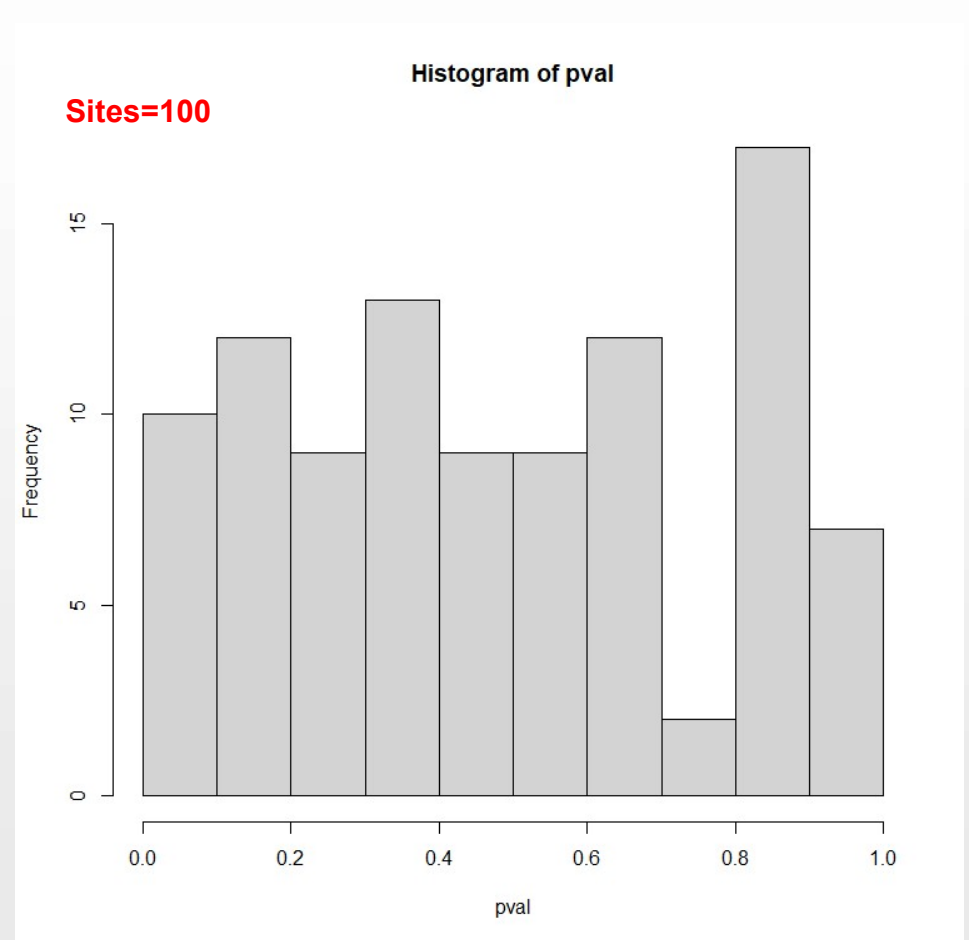
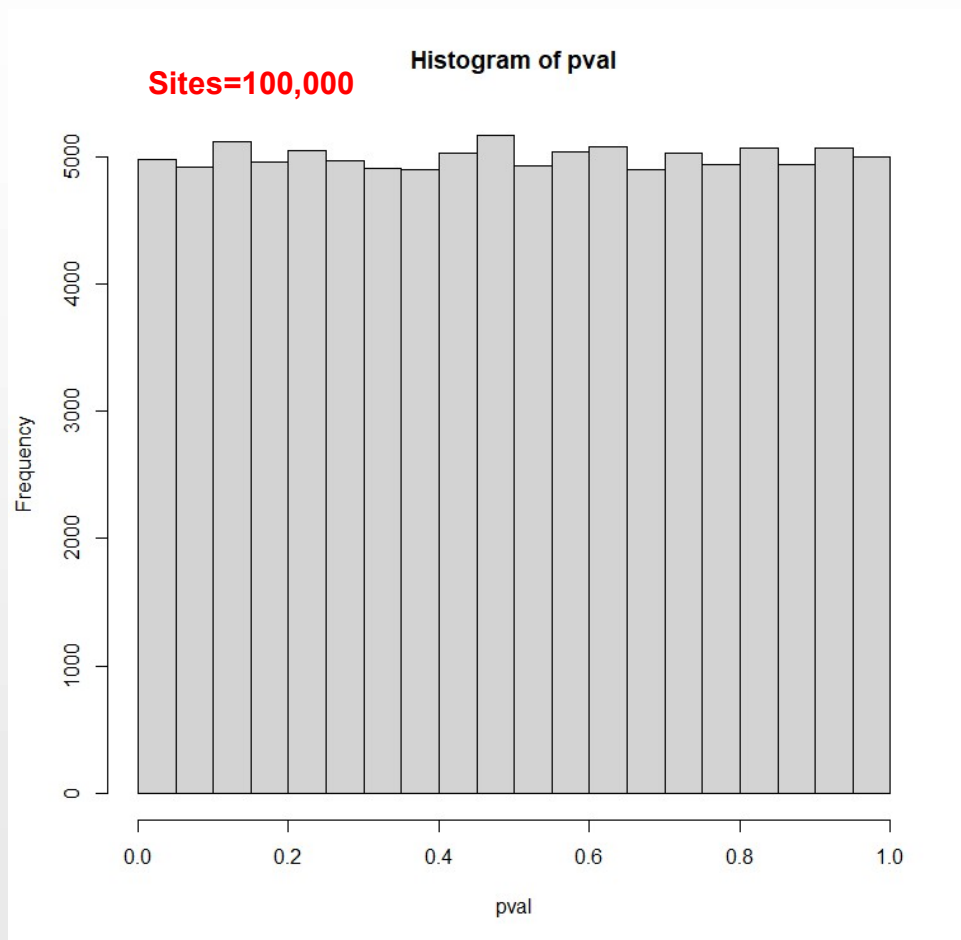
End repeat

```
n=50          # record length
sites=100     # number of replicates (independent sites)

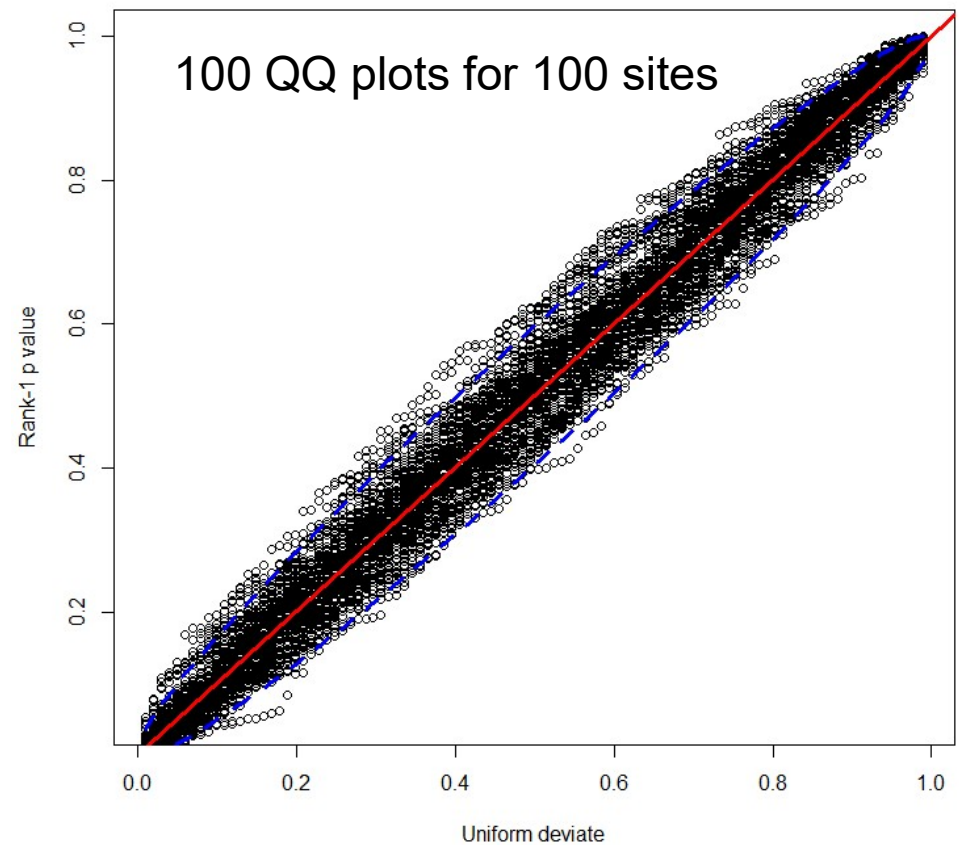
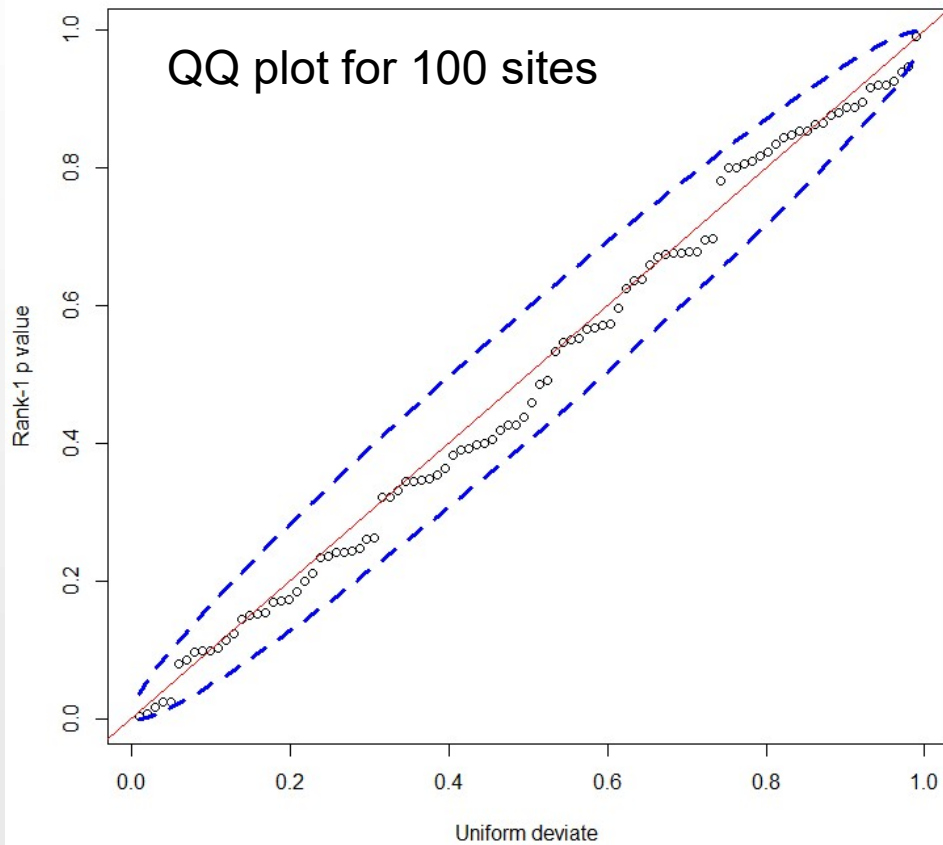
mn = 10       # mean of normal distribution
sig = 3       # std dev of normal distribution

for (r in 1:sites) {
  x[] = sort(rnorm(n,mn,sig))          # Sort n samples drawn from normal distribution
  pval[r] = pnorm(x[n],mean=mn,sd=sig)^n # Compute rank-1 p value
}
```

Understanding the rank-1 p value



Rank-1 p value signature



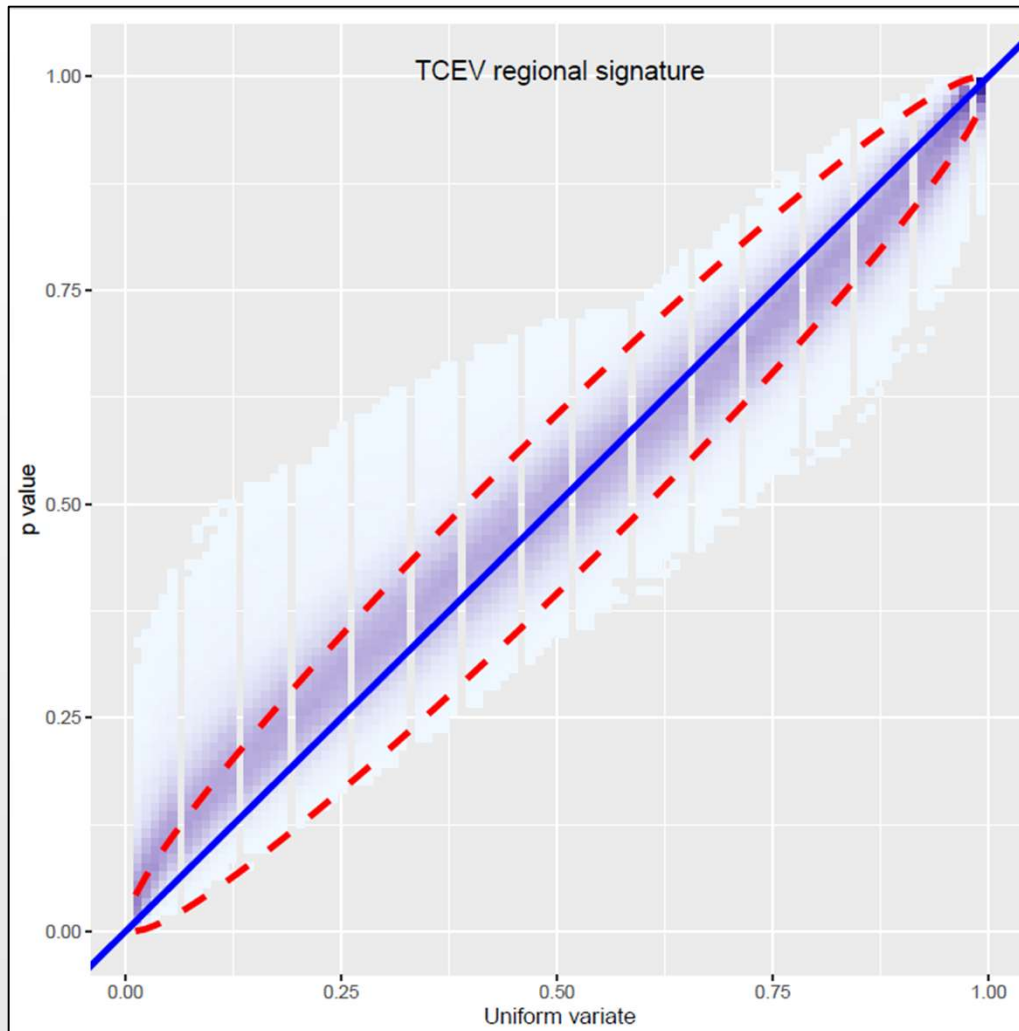
The fine print: It is not that simple!

- Rank-1 p value $\mathcal{P}^{1,\mathcal{M}}$ is uniformly distributed **only** if we know true parameter θ
- In practice we replace θ by its most likely value $\hat{\theta}$ to compute rank-1 p value using the observed rank-1 flow

$$\hat{\mathcal{P}}^{1,\mathcal{M}} = F(q^{(1)} | \mathcal{M}, \hat{\theta})^n$$

- Statistical theory shows that as record length n increases:
 - Most likely value of θ converges to true θ
 - Posterior of θ converges to a normal distribution with standard deviation $\propto 1/\sqrt{n}$
 - Distribution of $\hat{\mathcal{P}}^{1,\mathcal{M}}$ converges a uniform distribution
- Do 50-year AM records produce a sufficiently accurate estimate of θ ?
- Use Monte Carlo simulation to determine whether the posterior is sufficiently accurate

TCEV regional signature



For each of 178 case study sites:

Fit model TCEV to observed site data $\rightarrow \theta$

Generate 50 rank-1 p value QQ plots

Repeat 85 times:

Generate 50-year record using TCEV and θ

Fit model to generated record $\rightarrow \hat{\theta}$

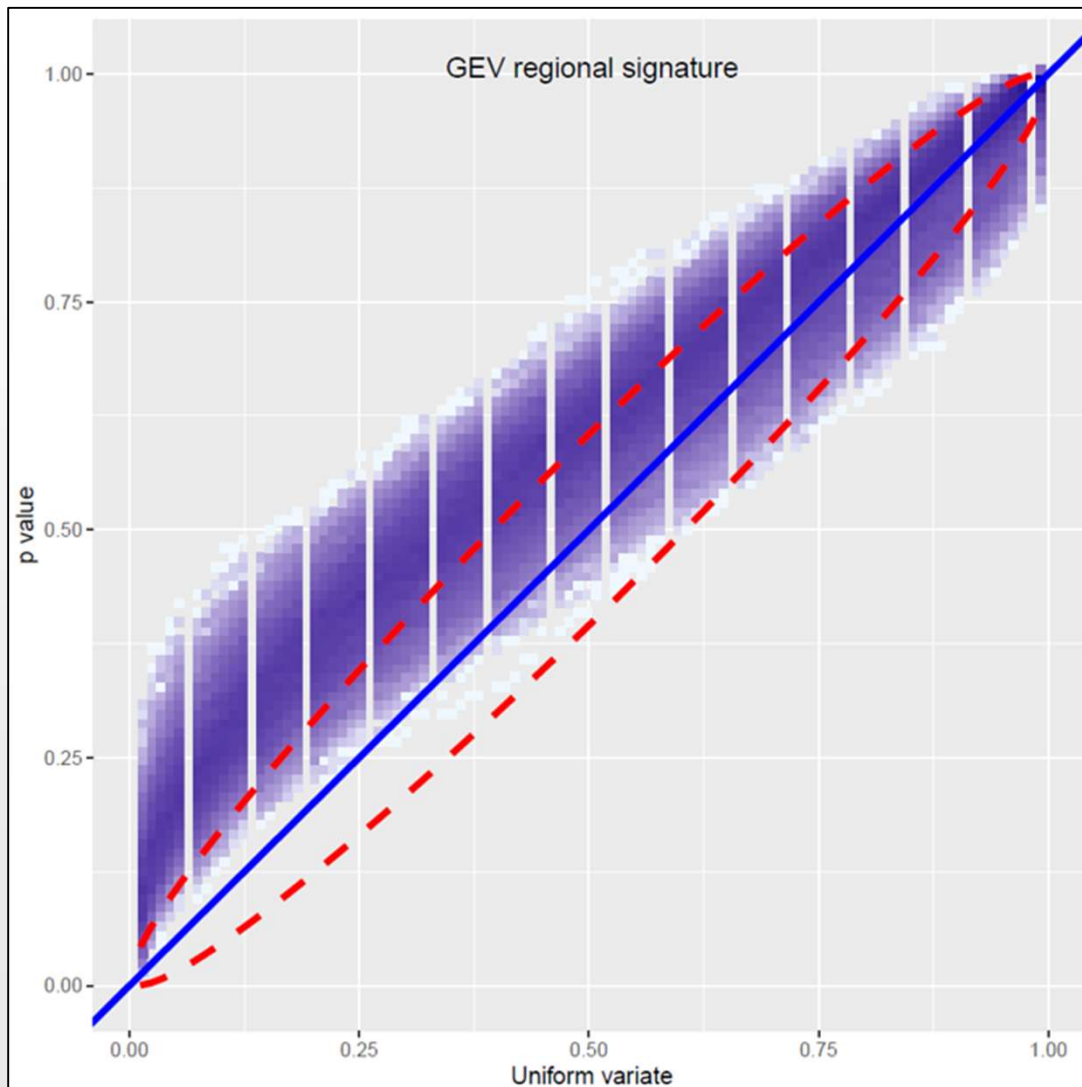
Compute rank-1 p value using $\hat{\theta}$

Draw p value QQ plot

Regional signature is the density plot of the 178 x 50 rank-1 p value QQ plots

TCEV posterior is sufficiently accurate for $\hat{\mathcal{P}}^{1,\mathcal{M}}$ to be close to uniform in distribution and independent of site parameters

GEV regional signature



For each of 178 case study sites:

Fit model GEV to observed site data $\rightarrow \theta$

Generate 50 rank-1 p value QQ plots

Repeat 85 times:

Generate 50-year record using GEV and θ

Fit model to generated record $\rightarrow \hat{\theta}$

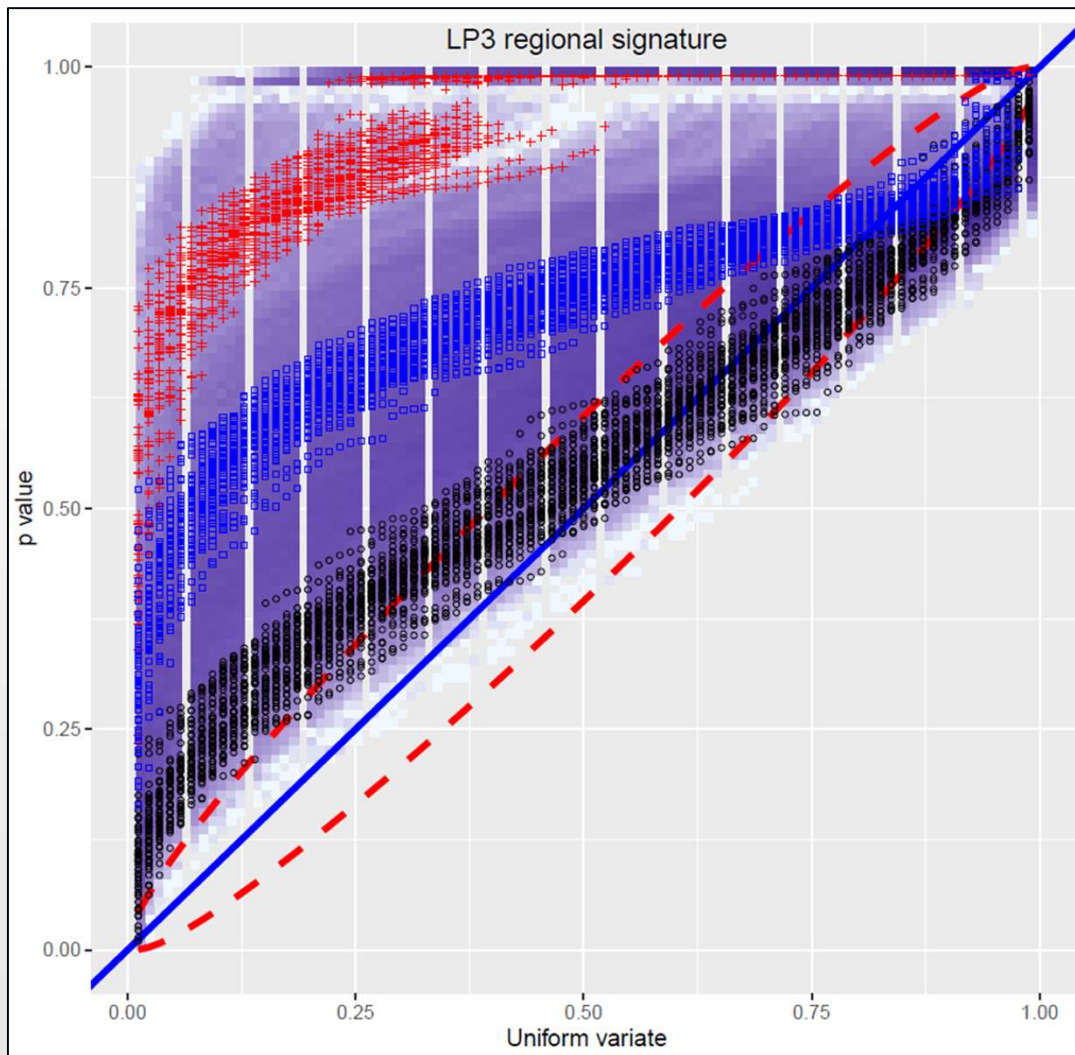
Compute rank-1 p value using $\hat{\theta}$

Draw p value QQ plot

Regional signature is the density plot of the 178 x 50 rank-1 p value QQ plots

GEV posterior is sufficiently accurate for $\hat{\mathcal{P}}^{1,\mathcal{M}}$ to be largely independent of site parameters but not uniform in distribution

LP3 regional signature



For each of 178 case study sites:

Fit model LP3 to observed site data $\rightarrow \theta$

Generate 50 rank-1 p value QQ plots

Repeat 85 times:

Generate 50-year record using LP3 and θ

Fit model to generated record $\rightarrow \hat{\theta}$

Compute rank-1 p value using $\hat{\theta}$

Draw p value QQ plot

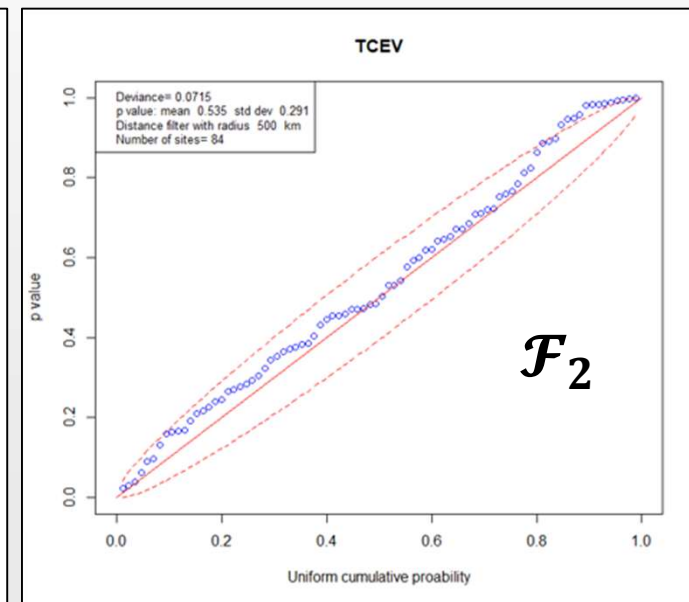
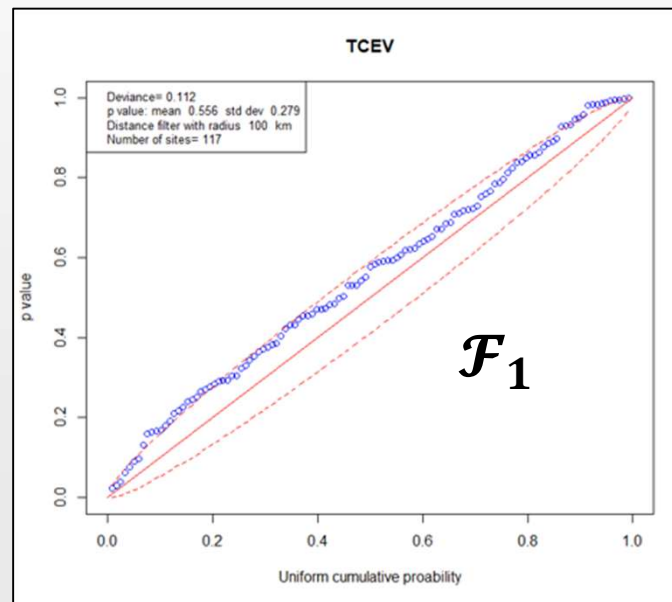
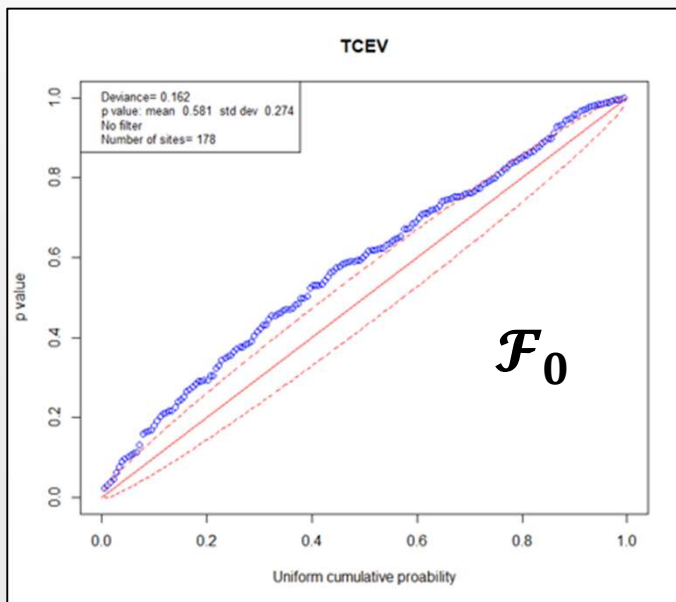
Regional signature is the density plot of the 178 x 50 rank-1 p value QQ plots

LP3 posterior of θ is not sufficiently accurate for $\hat{\mathcal{P}}^{1,\mathcal{M}}$ to be independent of site parameters

TCEV rank-1 p value plots

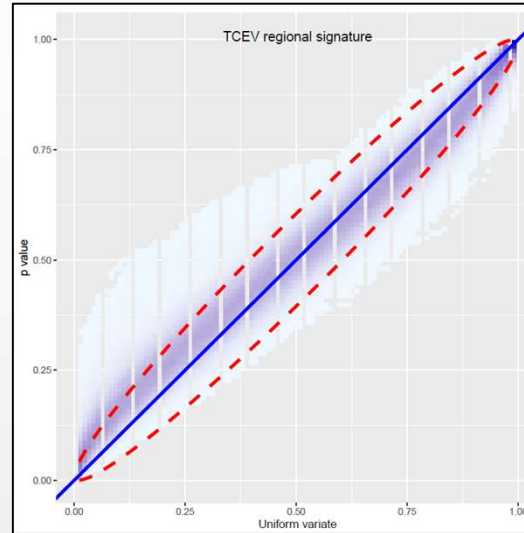
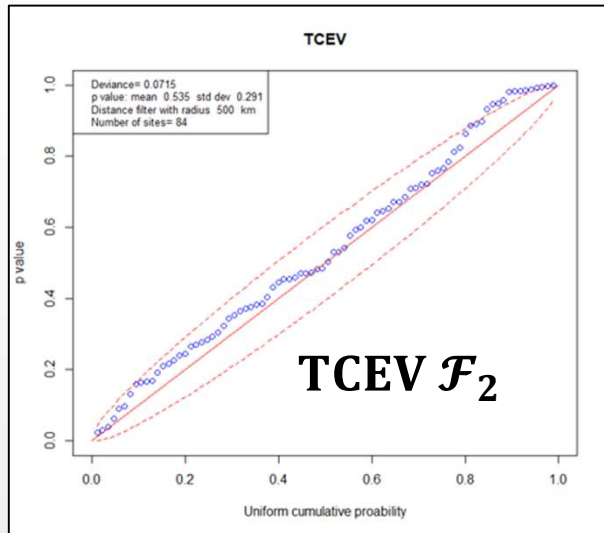
Filter case study sites to reduce site dependence:

- ❑ $\mathcal{F}_0$ - 178 sites
- ❑ $\mathcal{F}_1$ - 117 sites (no pairs of sites within 100 km with same rank-1 year)
- ❑ $\mathcal{F}_2$ - 84 sites (no pairs of sites within 500 km with same rank-1 year)

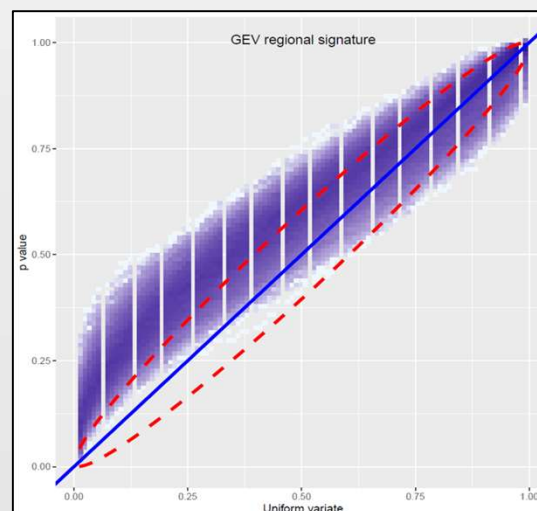
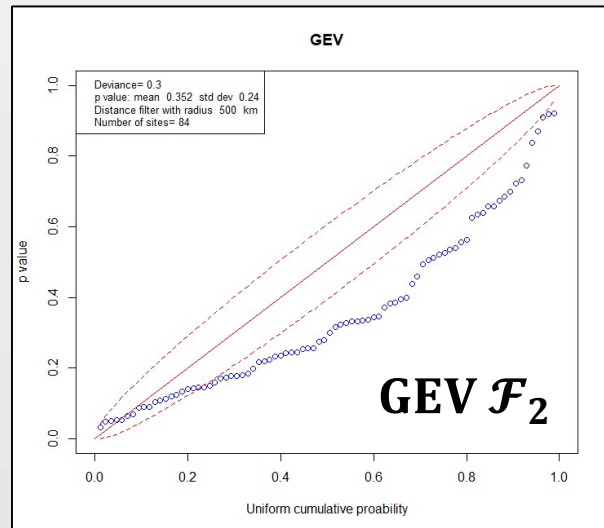


Rank-1 p value conclusions

$\mathcal{F}_2$ - 84 sites (no pairs of sites within 500 km with same rank-1 year)



TCEV rank-1 p values consistent with uniform distribution and regional signature



TCEV cannot mimic LP3 right tail
→ conclude LP3 is **inconsistent** with regional rank-1 data

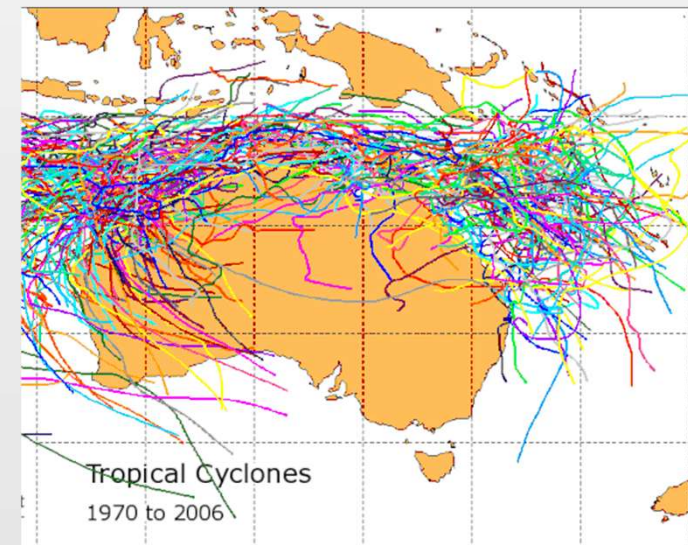
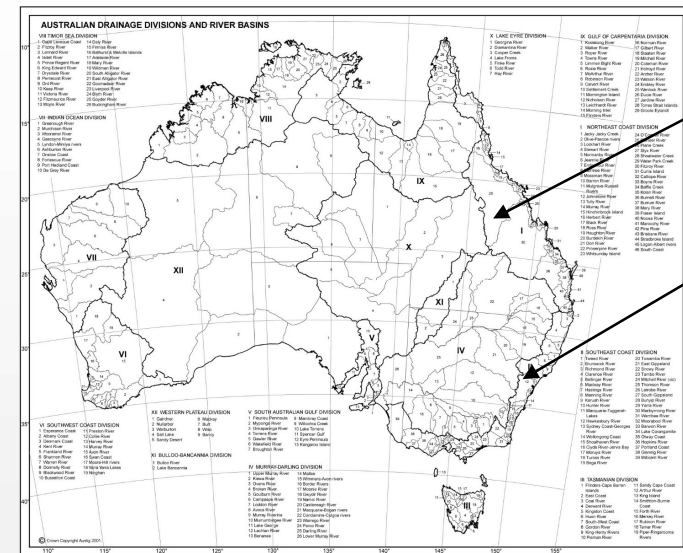
GEV rank-1 p values **inconsistent** with regional signature

TCEV regionalization

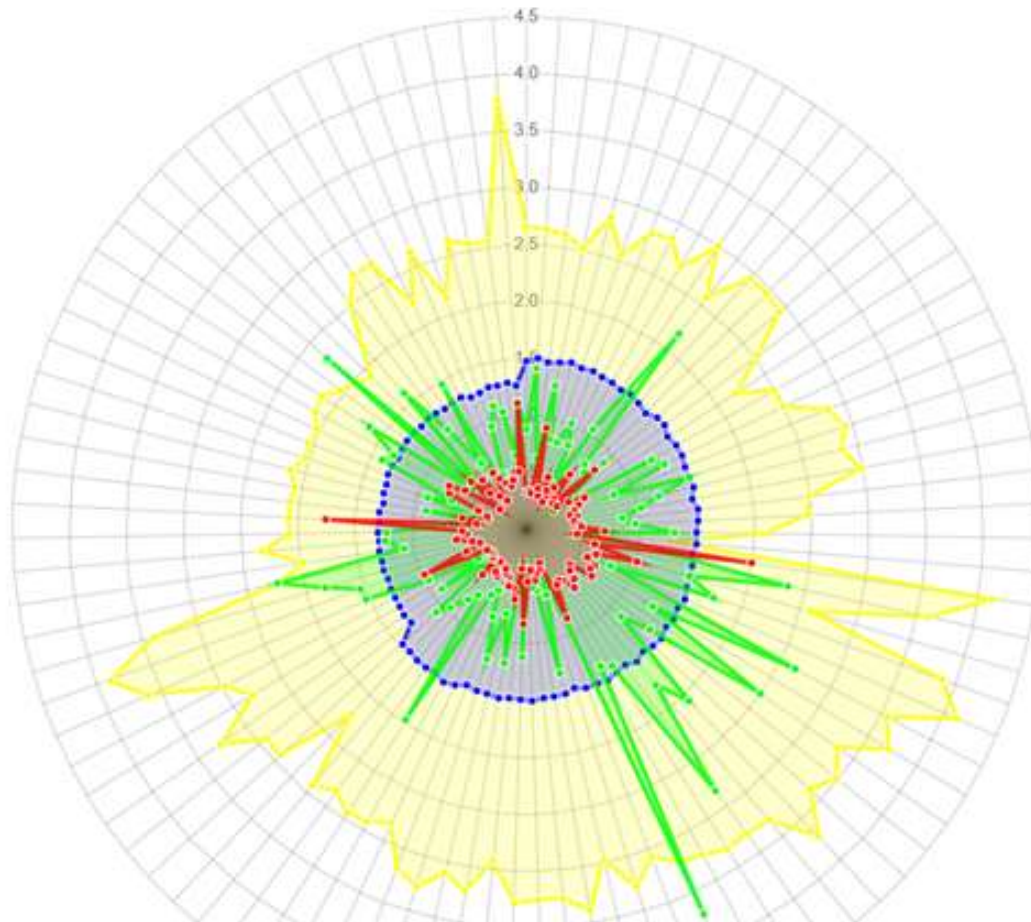
- Regionalize right-tail TCEV parameters $\{\lambda_2, \theta_2\}$ for drainage divisions 1 and 2
- Choice of predictor variables motivated by observation that tropical cyclones and East coast lows weaken and decrease in frequency as they move inland
- Mean number of component-2 events per year

$$\lambda_2 = f(\text{distance from coast})$$
- Mean component-2 peak flow

$$\theta_2 = f(\text{distance from coast, catchment area})$$



Spider plot: 1 in 100 AEP



— At-site TCEV
— At-site LP3
— Regional TCEV
— Regional LP3

- ❑ Regional TCEV much more accurate ARR LP3 regional method
- ❑ At-site LP3 is significantly less accurate than regional TCEV at many sites

The uncertainty of each method is normalized by dividing its 1 in 100 AEP 95% confidence interval by the 1 in 100 AEP regional TCEV quantile estimate

Conclusions

- ❑ How do we choose between models that have similar goodness-of-fit but different right tail behaviour?
- ❑ Regional rank-1 test probes consistency of right tail behaviour
 - Exploits information about observed rare floods over a large region
 - Sites must be statistically independent and parameter estimates sufficiently precise
- ❑ Concerns about LP3 and GEV
 - Not all annual maxima (AM) are floods, different processes drive small and large AM flows
 - (Mis)behaviour of shape parameter
- ❑ TCEV offers an alternative to LP3/GEV, decoupling low and high AM flows
- ❑ Eastern Australia case study with 178 50+ year sites:
 - LP3 fit marginally better than TCEV but different right tail behaviour, GEV not competitive
 - Regional rank-1 test accepts TCEV, rejects LP3 and GEV
- ❑ Prima facie case for FFA rethink with focus on right tail
 - More work to refine and apply rank-1 test to other regions
 - Better-behaved TCEV right tail → opportunities, other mixture models
 - Build more trust in FFA