Modelling joint extreme events with Stan

MEDiate Workpackage 2

Talfan barnie

August 22, 2023

1. The statistical model

We desire to model the co-occurence of extreme values of tidal surge, s, and river flow, f, to assess the multihazard of compound flood events for the city of Oslo. The extreme values of surge and flow are selected using the Peaks Over Threshold (POT) technique, where values above a threshold μ, specified as a specific quantile of the dataset, are chosen. By the Pickands-Balkema-De Haan theorem, we expect these data to be well approximated by the Generalised Pareto Distribution with location, scale and shape parameters μ, σ and ξ, where the location parameter is equal to the threshold used in the POT procedure. Thus the marginal probability density distributions $p_m(s)$ and $p_m(f)$ over s and f are given by

$$p_m(s) = gpd(s|\mu_s, \sigma_s, \theta_s) \tag{1.1}$$
$$p_m(f) = gpd(f|\mu_f, \sigma_f, \theta_f) \tag{1.2}$$

Where we indicate a density distrbution by lower case p and a marginal distrbution by subscript m, and gpd is the probability density function for the Generalised Pareto Distribution. However, we want the joint density distribution over s and f, not just the marginals, because it is the joint distribution which allows us to describe the co-occurrence of extreme surge and flow. By Sklar's theorem, any joint distribution can be described in terms of it's marginals, which we already have, and a function known as a copula that accounts for the correlation. In our case, we choose a Gumbel extreme value copula, which gives the following equation for our joint probability density function p_j:

$$\begin{aligned} p_j(s,f) = gumbel\left(GPD(s|\mu_s, \mu_f s, \sigma_s), GPD(f|\mu_f, \mu_f, \sigma_f)|\theta\right) \\ *gpd(s|\mu_s, \mu_f s, \sigma_s) \\ *gpd(f|\mu_f, \mu_f, \sigma_f) \end{aligned} \tag{1.3}$$

Where the subscript j indicates a joint distribution, $gumbel$ is the pdf of the Gumbel copula, which has one parameter θ, and we indicate the cumulative distribution function of the Generalised Pareto Distribution by capitalisation, GPD. We can then say that our paired extreme observations (s_i, f_i) are distributed as follows:

$$(s_i, f_i) \sim p_j(\mu_s, \mu_f, \sigma_s, \sigma_f, \xi_s, \xi_f, \theta) \tag{1.4}$$

This is the likelihood of our model, to which we add uniform priors (remember the GPD location parameters μ_s and μ_f are fixed values set by the POTs procedure).

$$\sigma_s \sim uniform(0, \infty) \tag{1.5}$$
$$\sigma_f \sim uniform(0, \infty) \tag{1.6}$$
$$\xi_s \sim uniform(-\sigma_s/(s_{max} - \mu_s), \infty) \tag{1.7}$$
$$\xi_f \sim uniform(-\sigma_f/(f_{max} - \mu_f), \infty) \tag{1.8}$$
$$\theta \sim uniform(1, \infty) \tag{1.9}$$
$$\tag{1.10}$$

The likelihood and the priors define a posterior over our five dimensional parameters space of $(\sigma_s, \sigma_f, \xi_s, \xi_f, \theta)$.

2. Implementation in Stan

We implemented this model in the Bayesian probabilistic programming language Stan, which comes with Hamiltonian Monte Carlo with the No U-Turn Sampler (NUTS) built in. Briefly, Hamiltonian Monte Carlo is a Markov Chain Monte Carlo method that seeks to approximate a probability density distribution. This is achieved by drawing random samples from the domain of the distribution in such a way that the density of those samples will tend towards the probability density as the number of samples tends to infinity. In our case the probability density distribution is the posterior distribution specified by our model. We used the Stan language implementations of the Genralised Pareto Distribution of Aki Vehtari [1] except for the inverse cumulative distribution function / quantile function / ppf which we implemented based on [2]. For the Gumbel copula pdf we used Ben Goodrich's Stan implementation [3], while we coded up the cdf based on [4]. For random draws from the Gumbel copula for posterior predictive checks we implemented in Stan the procedure outlined in [5].

3. Model diagnostics

The Hamiltonian Monte Carlo procedure draws successive samples from the posterior, in sequences we call chains. After a warm up period, the chain can be considered to have 'converged' on the density distribution and to be adequately sampling it, if certain criteria are met. One way to check for convergence is to run multiple randomly initialised chains and see if they converge on the same result. The $\hat{R}$ statistic is a measure of this - it is considered good practice to ensure values are below 1.1, the closer to 1.0 the better. Chains should also be checked for autocorrelation, which reduces the accuracy of the parameter estimates for a given chain length. This is measured by the Effective Sample Size. Finally we also check for divergences, which occurr when the posterior distribution is too tightly curved for Hamiltonian Monte Carlo to sample properly. This can be an indication of an improperly specified model, or simply that the model as formulated is too difficult to sample from and needs to be reformulated. We ran four chains for 2000 samples each, as shown in Figure 1. There were no divergences, and by inspection the chains look well mixed with little autocorrelation. The effective sample size statistics (ess_{bulk} and ess_{tall}), as well as $\hat{R}$ are shown in Figure 2, and confirm the model is adequately sampling the posterior.

[1] https://mc-stan.org/users/documentation/case-studies/gpareto_functions.html

[2] https://real-statistics.com/other-key-distributions/pareto-distribution/generalized-pareto-distribution/

[3] https://spinkney.github.io/helpful_stan_functions/group__copula.html

[4] https://en.wikipedia.org/wiki/Copula_(probability_theory)

[5] https://www.caee.utexas.edu/prof/bhat/abstracts/supp_material.pdf

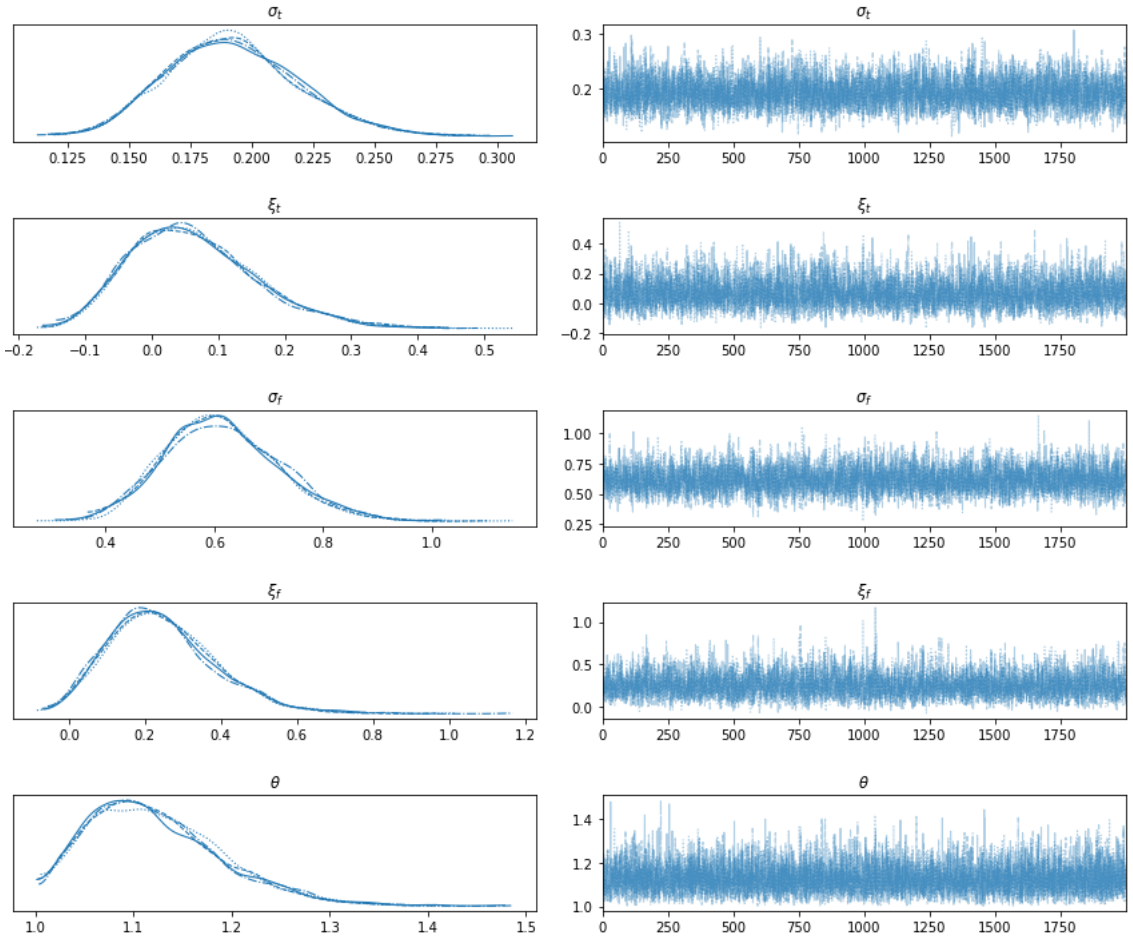

Figure 1: Trace plot for the five parameters of the joint distribution function.

```
In [14]: df_summary.loc[['xit','xif','sigmat','sigmaf','theta']]
```

Out[14]:

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
xit	0.063	0.096	-0.099	0.250	0.001	0.001	4965.0	4719.0	1.0
xif	0.245	0.142	0.006	0.514	0.002	0.001	4732.0	4552.0	1.0
sigmat	0.192	0.027	0.145	0.244	0.000	0.000	5819.0	5262.0	1.0
sigmaf	0.614	0.104	0.420	0.809	0.001	0.001	5407.0	5406.0	1.0
theta	1.123	0.069	1.012	1.249	0.001	0.001	6871.0	4154.0	1.0

Figure 2: Model diagnostics for the five parameters.

4. Model assessment

Having found that we are able to adequately sample from the posterior distribution defined by our model, we need to then assess whether the model is a good fit to the data. The former is a technical question about how our model is implemented, whereas this is a scientific question about how well our model describes our data. We can assess the fit of our marginal Generalised Pareto Distributions using Bayesian Q-Q plots. Every sample in the chains gives us a single value of the scale and shape parameters for both distributions, which can be used to calculate quantiles. Thus we get one Q-Q curve for each sample, and our posterior distribution defines a distribution over Q-Q curves, as shown in Figure 3. As with a regular Q-Q plot, a perfect fit would be indicated by the curves lying along the 1:1 relationship between empirical quantiles and those calculated from the distribution. Here we can see the model fits the smaller extreme values well, but not the four or five largest ones.

We can also test our whole model using Posterior Predictive Checks (PPCs). For each sample in each chain, we make a random draw from the p_j defined by that samples values of $\sigma_s, \sigma_f, \xi_s, \xi_f, \theta$. We can think of these as draws as predictions, or simulated observations. We then plot the distribution of those simulated observations, compare then with the original observations, to see if the latter could plausibly be drawn from the former. As shown in Figure 4, they look pretty similar.

The posterior distribution over the model parameters is shown as a corner plot in Figure 5. Here we can see some correlation between the scale and shape parameters for the Generalised Pareto Distributions but nothing pathological. The values for the θ parameters of the Gumbel copula are close to 1, showing that the correlation between the surge and flow data is weak.

5. Calculating return periods

Having established that we can adequately sample from the posterior distribution and that it describes our data reasonably well, we can use the posterior distribution over the model parameters to calculate Return Periods (RPs). As we saw with the Q-Q curves, each sample in each chain is set of model parameters that define a single marginal RP curve for surge, another for flow, and a JRP curve for both combined. Thus we wind up with a distributions over these curves, just as we did with the Q-Q curves, and for a given pair of (surge, flow) values, there is a distribution of marginal RPs and joint RPs, $p(RP|s), p(RP|f), p(JRP|s,f)$. These RPs are calculated as follows:

$$RP(s) = \frac{\lambda}{Pex_m(s)} \tag{5.11}$$

$$RP(f) = \frac{\lambda}{Pex_m(f)} \tag{5.12}$$

where $RP(s)$ is the return periods for surges greater than s, $RP(f)$ is the return period for flows greater than f, λ is mean interval between successive events, and P_{ex} is the

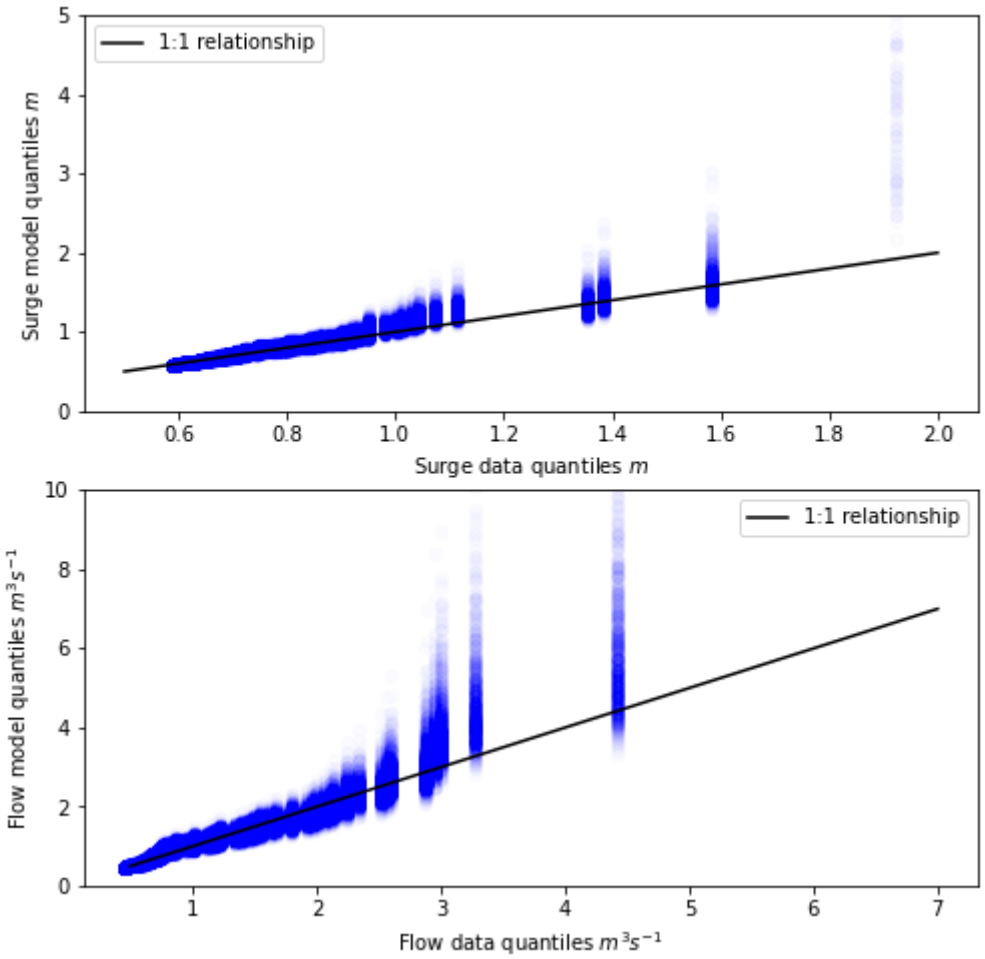

Figure 3: Bayesian Q-Q plot for the two marginal Generalised Pareto Distribution

exceedence probability which is just one minus the cdf:

$$Pex_m(s) = 1 - P_m(s) \tag{5.13}$$

$$Pex_m(f) = 1 - P_m(f) \tag{5.14}$$

Joint Return Periods (JRPs) are calculated in a similar way:

$$JRP(s,f) = \frac{\lambda}{Pex_j(s,f)} \tag{5.15}$$

where Pex_j is the joint excedence, the probability of both the value of s and f being exceeded. Pex_j is given by:

$$Pex_j(s,f) = 1 - P_m(s) - P_m(f) + P_j(s,f) \tag{5.16}$$

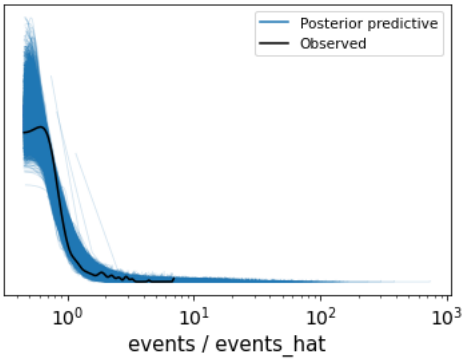

Figure 4: Posterior predictive checks.

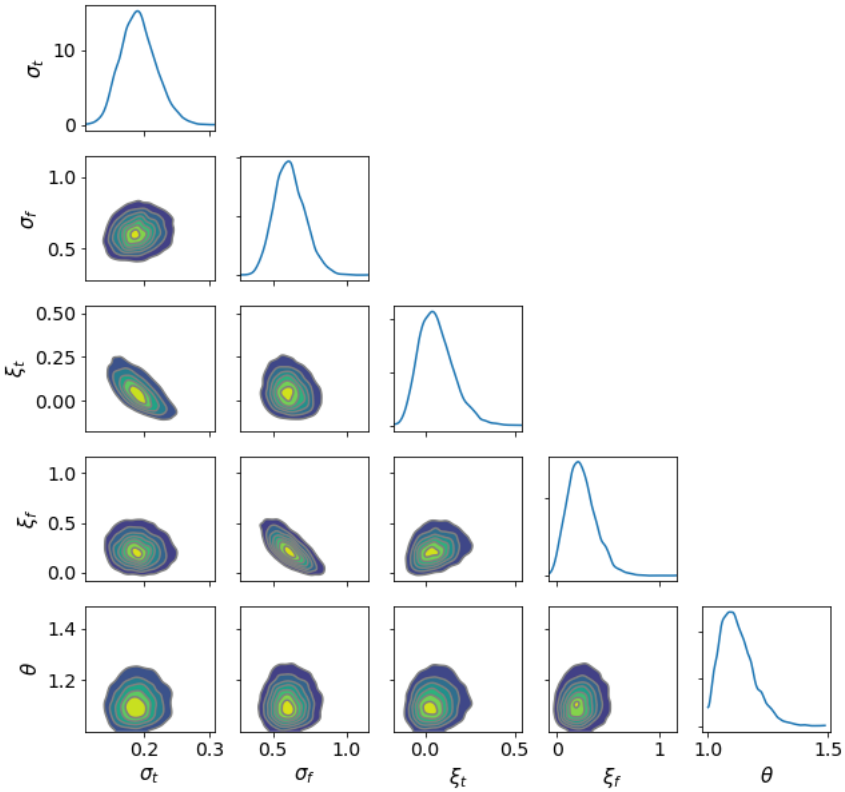

Figure 5: Corner plot.

Where we indicate the cumulative distribution function by capitalization again, and $P_j(s,f)$, $P_m(s)$ and $P_m(f)$ are given by:

$$P_j(s,f) = GUMBEL\left(GPD(s|\mu_s,\mu_f s,\sigma_s), GPD(f|\mu_f,\mu_f,\sigma_f)|\theta)\right) \tag{5.17}$$

$$P_m(s) = GPD(s|\mu_s,\sigma_s,\xi_s) \tag{5.18}$$

$$P_m(f) = GPD(f|\mu_f,\sigma_f,\xi_f) \tag{5.19}$$

6. Results

The JRP for surges and flows is shown in Figure 6. Here we can see that the JRPs for smaller events are quite well constrained, whereas for the largest event, October 1987 the JRP is barely constrained at all. This is to be expected given that this observation is a significant outlier. One solution to narrow the distribution of JRPs for this event, would be to add extra pairs of tide and river gauges, repeating the procedure for each, but assuming that each of the $\sigma_s, \sigma_f, \xi_s, \xi_f, \theta$ parameters are themselves draw from some unkown distribution, itself also with parameters we try to estimate. These higher order parameters (known as hyperparameters) describe how alike these tide/river gauge pairs are, and the extent information is shared between them in making statistical inferences about them.

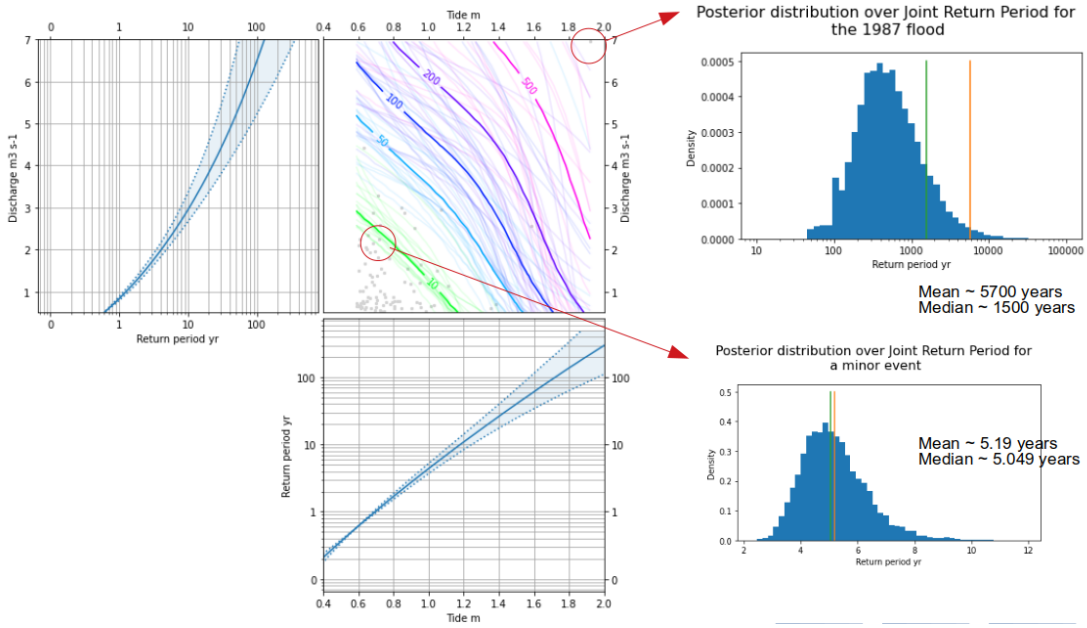

Figure 6: Left: Plots showing the 10, 50 and 90$^{\text{th}}$ posterior marginal probabilities over return period for surge and flow ($RP(s)$ and $RP(f)$), as well as the distribution of Joint Return Periods (JRPs) over values of surge and flow. The median JRP contour for values of 10, 50, 100, 200 and 500 years are shown, as well as random draws from the distribution to give a sense of the variability. Right: Two examples of distributions over JRP, one for a small event, and one for the 1987 flood. Note the log scale on the x axis for the 1987 flood - here most of the probability mass is in the tails.

We can also calculate distributions over JRP conditional on marginal RPs, i.e. $p(JRP|RP_s, RP_f)$, examples are shown in Figure 7. It is important to note that $p(JRP|RP_s, RP_f)$ and $p(JRP|s, f)$ cannot be used interchangeably, as there is a stochastic relationship between marginal return periods and physical values of surge and flow.

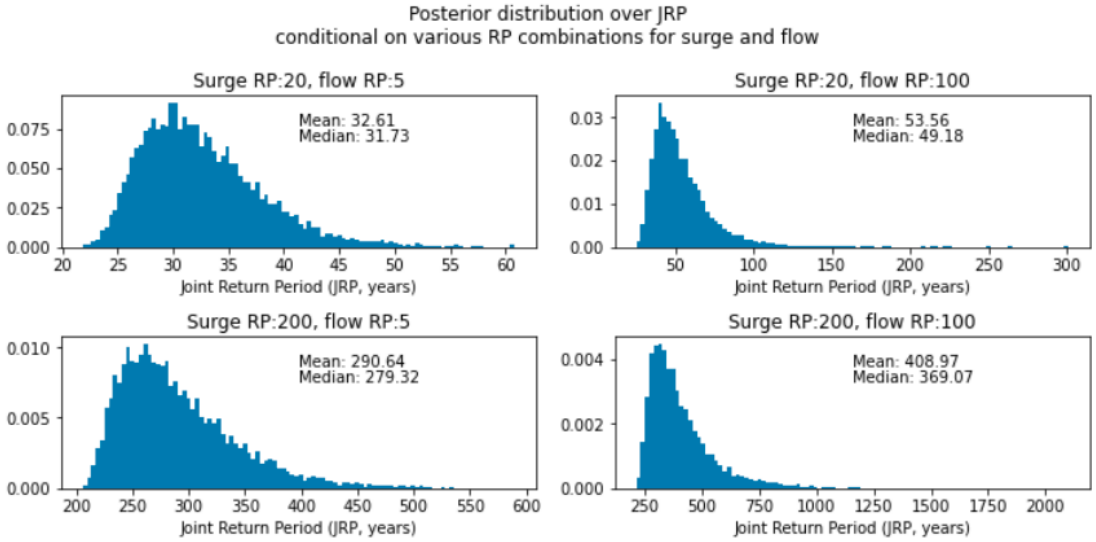

Figure 7: Distributions over JRP conditioned on RP(i.e. $p(JRP|RP_s,RP_f)$)