# A study of: Simulating flood event sets using extremal principal components

Benjamin Lapostolle[1], Alexandre Wan[1]

**Abstract**

The aim of this project is to explore the research paper titled "Simulating Flood Event Sets using Extremal Principal Components" authored by Christian Rohrbeck and Daniel Cooley. In their study, they present a unique approach for generating extreme values based on the distribution of river flows observed in Northern England and Southern Scotland. Their methodology involves conducting extremal principal components analysis and working with the resulting principal components. This technique enables the manipulation of data in lower dimensions and demonstrates good results in comparison with existing statistical methods.

## 1. Introduction

The analysis of substantial values in a stochastic process holds significant importance across various domains, including insurance and finance. Effectively capturing the extremal dependence of marginal distributions poses a challenging endeavor. In the context of this project, our aim is to generate extreme events from a given distribution of interest. As outlined in the reviewed paper of Rohrbeck and Cooley (2023), methods for sampling from the joint distribution of the vector are constrained, particularly in higher dimensions. In response to this limitation, the proposed approach employs extremal principal component analysis on the data. The essence of this technique involves retaining only the most crucial variables and sampling from these variables to generate extreme events. In an extremal principal component analysis, the principal components are dependant. A kernel density estimate suitable for spherical data is used to generate new samples of principal components, which are then used to produce new samples of extremes of the original distribution.

## 2. Methods

### 2.1. Data

Our data is sourced from the UK's National River Flow Archive, which provides data from multiple gauges. The measurements span from January 1980 to September 2018, and we focus on the maximum value per week for each gauge. It's crucial to highlight that measurements vary significantly among gauges, influenced by the specific water sources in which the river flow is measured, see figure 1. Hence, it will be necessary to standardize the data. To replicate the methodology of the paper, we process our database in the following way:

- From the available gauges, we specifically chose the same 45 gauges as the authors did for the study, located in southern Scotland and northern England.

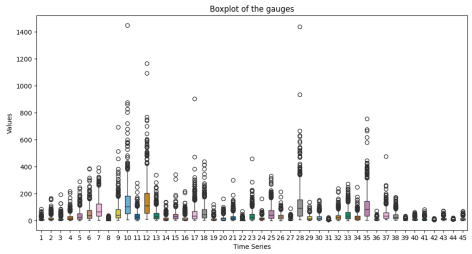


Figure 1: Distribution of the gauges

- Given that maximum floods are typically observed during winter, we restrict our analysis to data recorded between November and March, resulting in a total of 848 data points.

- To address missing measurements, we exclude weeks with such occurrences, resulting in a dataset of 780 data points.

- In this study, we treat the distribution of floods as stationary and refrain from modeling any form of temporal dependency. Given the potential influence of climate change on the distribution of floods over time, we assess this hypothesis using the Augmented Dickey-Fuller (ADF) Test. The results suggest that this hypothesis appears to be valid.

### 2.2. Extremal principal component analysis

Our goal is to generate extreme events according to the distribution of the extremes of our data. To do so, we will apply the method from the paper of Cooley and Thibaud (2019). In the following, we denote our data as $f\tilde{X}_t = (\tilde{X}_{t,1}, \tilde{X}_{t,2}, ..., \tilde{X}_{t,K})$, with $K$ being the number of gauges and $t$ the week. Let's assume that the marginal distributions of our vector follow a Fréchet distribution of parameter $\alpha = 2$, i.e., $\forall t$, $P(\tilde{X}_{t,k} \leq x) = \exp(-x^{-2})$, $k = 1, \ldots, K$. We also assume that $\tilde{\mathbf{X}}$ is regularly varying with index 2, which means that, for any Borel set $B \in S_+^{K-1} = \{\omega \in \mathbb{R}_+^K : \|\omega\|_2 = 1\}$ with $H_{\tilde{X}}$ the angular measure:

$$\lim_{r\to\infty} \mathbb{P}\left(\|\tilde{X}\|_2 > rz, \frac{\tilde{X}}{\|\tilde{X}\|_2} \in B \middle| \|\tilde{X}\|_2 > r\right) = z^{-2} H_{\tilde{X}}(B) \quad (1)$$

Given these assumptions, the tail dependence of $\tilde{X}$ can be examined through the tail pairwise dependence matrix (TPDM) $\Sigma$, representing the second-order properties of $H_X$. This matrix, denoted as a $K \times K$ matrix, is defined as follows:

$$\forall i, j \leqslant K, \Sigma_{i,j} = \int_{S_+^{K-1}} \omega_i \omega_j \, dH_X(\omega) \quad (2)$$

Because we have assumed that $\mathbf{X}$ is regularly varying with an index of 2, we can demonstrate that $\Sigma$ is positive semidefinite. Elevated values of $\Sigma_{i,j}$ signify significant extremal dependence between the marginals $X_i$ and $X_j$, whereas low values indicate weak dependence. Similar to a traditional principal component analysis, we conduct an eigenvalue decomposition of $\Sigma$ to obtain eigenvalues and eigenvectors. Subsequently, we calculate the principal components of $\tilde{X}$ by applying the following transformation:

$$V = U^T \tau^{-1}(\tilde{X}) \quad (3)$$

with $\tau^{-1}(.) = \log(\exp(.) - 1)$ applied component-wise and $U$ defined as $\Sigma = UDU^T$ in the eigendecomposition of $\Sigma$. The results from Cooley and Thibaud (2019) indicate that $V$ is regularly varying with $\alpha = 2$ and can assume any real value. Hence, the angular measure $H_V$ will operate on the entire unit sphere $S^{K-1}$, allowing for the sampling of directions across the entire space.

To practically estimate the TPDM, we will select weeks whose norms are above a threshold set to the empirical 96% quantile. Let $T^*$ be the set of selected weeks, and define new variables $W_{t,k} = \frac{X_{t,k}}{r_t}$, with $r_t = \|\mathbf{X}_t\|$. We will compute $\Sigma$ as follows:

$$\Sigma_{i,j} = \frac{K}{|T^*|} \sum_{t \in T^*} \omega_{t,i} \omega_{t,j} \quad (4)$$

### *2.3. Marginales*

To apply this method, we need to have data that satisfies the given assumptions. To achieve this, we utilize the following transformation:

$$\forall k \leqslant K, \ \forall t \in T \quad X_{t,k} = \left[-\log \hat{F}_k(\tilde{X_{t,k}})\right]^{-1/2} \quad (5)$$

where $T$ is the set of weeks, and $\tilde{X}_{t,k}$ represents our data standardized such that each marginal has a mean of 0 and a standard deviation of 1. This transformation ensures that the new marginals follow the Fréchet distribution, enabling the application of our PCA. To compute an estimation of the cumulative distribution function (CDF), we will use the empirical CDF for values below the threshold $u_k$, set at the 96% quantile. Additionally, we will employ a Generalized Pareto Distribution $GPD(\sigma_k, \xi_k)$. According to the Extreme Value Theory, for a suitably high threshold $u$, the exceedances of a real random variable $Y$ are well modeled by:

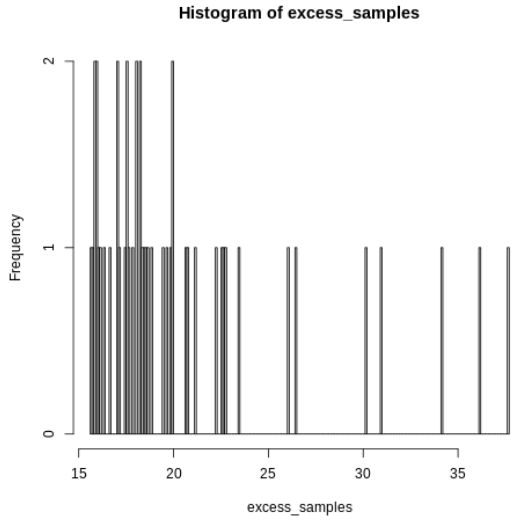


Figure 2: Distribution of the extremes of the gauge 7

$$P(Y > x + u \,|\, Y > u) = \left(1 + \frac{\xi_k x}{\sigma_k}\right)_+^{-\frac{1}{\xi_k}}, \ z_+ = \max(0, z) \quad (6)$$

Employing a Generalized Pareto Distribution to characterize the tail distribution of marginals is crucial for two principal reasons. Firstly, the scarcity of observations in the tail distribution, resulting from the low probabilities associated with this region, makes the empirical cumulative distribution function (CDF) less effective as an estimator. Given that our dataset comprises around 800 points, the occurrence of extreme events is minimal. Secondly, and more importantly, our primary focus is on replicating extreme events. As we will utilize the inverse CDF to generate new data, as explained in subsequent sections, relying on the empirical CDF as the inverse would limit us to generating extremes that mirror the values in our existing data. To overcome this limitation, we opt for a continuous distribution instead.

When modeling a Generalized Pareto Distribution (GPD) on the extremes distribution of the gauges, a significant challenge arises: the number of extremes is too low to obtain relevant $\sigma$ and $\xi$ parameters, as depicted in Figure 2. To augment the number of samples for fitting the GPD, a potential solution is to fit a GPD simultaneously on several gauges. It is observed that when using a textbook example distribution such as the Fréchet distribution and setting the threshold as the 96% quantile, the estimation of parameters becomes meaningful when the number of samples exceeds 3000. Consequently, we will create clusters with a minimum size of 5 and model the same GPD for gauges within the same cluster.

### *2.4. Clusters*

Clusters can be constructed in various ways. The authors refer to Rohrbeck and Tawn (2021), which studies the same data points to build its clusters. In constructing our clusters, we

chose to implement a hierarchical clustering method based on the distance between the distributions of the gauges. Using the Weierstrass distance, we compute the distributions that appear close to each other and form clusters. While this method is not as refined as the one employed in the paper, it allowed us to generate parameters with a relatively good estimation error compared to the numbers provided in the original paper. To potentially enhance our results, we could also consider working with the positions of the gauges and utilizing a combination of the Weierstrass distance and the Euclidean distance to build the clusters. However, given the satisfactory performance of our current method, we did not pursue further improvements.

### 2.5. Generating hazard event sets

The proposed framework is based on sampling values for the extremal principal components $V$, and then applying the inverse transformation given by Eq. (3) to obtain $\tilde{X}$. Subsequently, the inverse transformation (5) is applied again to obtain the reconstructed dataset $X$. As $\mathbf{V}$ is regularly-varying with index $\alpha = 2$ and $\mathbb{P}(\|\mathbf{V}\|_2 \leq r) = \exp\left[-(r/K)^{-2}\right]$ $(r > 0)$, we can sample the radius $r$ with a Frechet distribution. We will then have to estimate $H_V(\cdot)$ which describes the behavior of $\mathbf{W} = \frac{\mathbf{V}}{\|\mathbf{V}\|_2} \,\Big|\, \|\mathbf{V}\|_2 > r_V$ as $r_V \to \infty$. We will make the assumption that the limits holds above a sufficiently high $r_V$, which we set to the empirical 94% quantile of $\|\mathbf{V}\|_2$.

The crux of the matter is then to efficiently sample $\mathbf{W}$. We denote $\{\mathbf{w}_i \in \mathbb{S}^{K-1} : i = 1, \ldots, n\}$ as the observations for $\mathbf{W}$. The proposed generative sample propose to sample the first $m$ componets of $W$ and then the last $K - m$.

*Modelling the first m components.* Given the observations $\{\mathbf{w}_{i,1:m} : i = 1, \ldots, n\}$ for $\mathbf{W}_{1,m}$, we model the distribution of $\mathbf{W}_{1:m} / \|\mathbf{W}_{1:m}\|_2$ with the von Mises-Fisher distribution as our chosen kernel (as recommended by Hall et al. (1987)) characterized by the density $h(\mathbf{z}; \mu, \kappa) = c_0(\kappa) \exp\left(\kappa \mathbf{z}^{\mathrm{T}} \mu\right)$ where $\mathbf{z} \in \mathbb{S}^{m-1}$, $\mu \in \mathbb{S}^{m-1}$, and $\kappa > 0$ represent the sample point, mean direction, and concentration parameter, respectively. Here, $c_0(\kappa)$ denotes the normalizing constant. The kernel density estimate for $\mathbf{W}_{1:m} / \|\mathbf{W}_{1:m}\|_2$ is given by

$$\hat{h}_V(\mathbf{z}; \kappa) = \frac{1}{n} \sum_{i=1}^{n} h\left(\mathbf{z}; \frac{\mathbf{w}_{i,1:m}}{\|\mathbf{w}_{i,1:m}\|_2}, \kappa\right) \qquad (\mathbf{z} \in \mathbb{S}^{m-1}), \quad (7)$$

where $\mathbf{w}_{i,1:m}$ represents the $i$-th observation vector, and $n$ is the total number of observations.

*Modelling the last K − m components.* The idea is to map $\mathbf{W}$ into a random variable $\mathbf{Z}$ on the $(m+1)$-dimensional unit sphere $\mathbb{S}^m$. The intuition behind that is that the first $m$ components of $\mathbf{Z}$ represent the information contained in $\mathbf{W}_{1:m}$, while the final component $Z_{m+1}$ summarizes the $K-m$ components of $W$ which are of less importance (PCA). Therefore, let $\mathbf{Z} = (Z_1, \ldots, Z_{m+1})$ with $Z_j = W_j$ $(j = 1, \ldots, m)$ and

$$Z_{m+1} = \begin{cases} \sqrt{1 - \sum_{j=1}^{m} W_j^2} & \text{if } W_{m+1} \geq 0, \\ -\sqrt{1 - \sum_{j=1}^{m} W_j^2} & \text{if } W_{m+1} < 0. \end{cases} \quad (8)$$

That way, we have indeed constructed $Z$ in the $(m+1)$-dimensional unit sphere $\mathbb{S}^m$. Given observations $\{\mathbf{z}_i : i = 1, \ldots, n\}$, we obtain the kernel density estimate

$$\hat{h}_Z(\mathbf{z}, \kappa_Z) = \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{z}; \mathbf{z}_i, \kappa_Z) \qquad (\mathbf{z} \in \mathbb{S}^m). \quad (9)$$

This approach captures the dependence between capture the dependence in $W_1, \ldots, W_m$ but also some dependence between $\mathbf{W}_{1:m}$ and $\mathbf{W}_{(m+1):K}$.

Being able to sample $Z$, we construct $W$ with it. Based on the defined connection between $\mathbf{W}$ and $\mathbf{Z}$, the first $m$ components of the sample $\mathbf{w}^*$ are set to those of the sample $\mathbf{z}^*$, i.e., $w_j^* = z_j^*$ $(j = 1, \ldots, m)$. We are left to determine how to transform the single value $z_{m+1}^*$ into the last $K - m$ values for $\mathbf{w}^*$ i.e $\mathbf{w}_{(m+1):K}^*$

We utilize observations from $\mathbf{Z}$ that are most similar to the sampled vector $\mathbf{z}^*$ to identify potential patterns for $w_1^*, \ldots, w_K^*$, conditioned on $w_j^* = z_j^*$ $(j = 1, \ldots, m)$. Our proposed method employs a nearest-neighbor approach, where the index $q$ of the observation closest to the sample is determined by

$$q = \underset{i=1,\ldots,n}{\operatorname{argmax}} \, \mathbf{z}_i^{\mathsf{T}} \mathbf{z}^*. \quad (10)$$

Subsequently, the sample $\mathbf{w}_{(m+1):K}^*$ is set to a scaled version of $\mathbf{w}_{q,(m+1):K}$, resulting in

$$\mathbf{w}^* = \left(z_1^*, \ldots, z_m^*, \left|\frac{z_{m+1}^*}{z_{q,m+1}}\right| w_{q,m+1}, \ldots, \left|\frac{z_{m+1}^*}{z_{q,m+1}}\right| w_{q,K}\right). \quad (11)$$

**Algorithm 1** Steps to sample $X$

- Obtain a sample $\mathbf{z}^* \in \mathbb{S}^m$ from the distribution with the estimated density (9).
- Extract the index $q$ in (10) and derive the angular component $\mathbf{w}^* \in \mathbb{S}^{K-1}$ using (11).
- Sample $r^*$ from a Fréchet distribution with $\mathbb{P}(\|\mathbf{V}\|_2 \leq r) = \exp\left[-(r/K)^{-2}\right]$.
- Calculate the generated sample $\mathbf{v}^*$ for $\mathbf{V}$, $\mathbf{v}^* = (v_1^*, \ldots, v_K^*) = (r^* w_1^*, \ldots, r^* w_K^*) \in \mathbb{R}^K$.
- Apply the inverse of (3) to obtain a sample $\tilde{\mathbf{x}}^* \in \mathbb{R}_+^K$ for $\tilde{\mathbf{X}}$,

$$\tilde{\mathbf{x}}^* = \tau\left\{\sum_{j=1}^{K} v_j^* \mathbf{U}_{\cdot,j}\right\},$$

where $\tau(\cdot) = \log[\exp(\cdot) + 1]$ is applied component-wise, and $\mathbf{U}_{\cdot,j}$ is the $j$-th eigenvector of the estimated TPDM. Note, $\tilde{\mathbf{x}}^* \in \mathbb{R}_+^K$.

- Apply the inverse of (5) to obtain a sample $\mathbf{x}^*$.

*Finding the optimal m.* In order to sample $X$ with the proposed algorithm, a value for $m$ need to be chosen. For a small value of $m$, the distribution of the random vector $\mathbf{Z}$ can be estimated to a reasonable degree of accuracy, but there is a trade-off: we may underestimate the variance of $\mathbf{W}$ due to the restrictive sampling

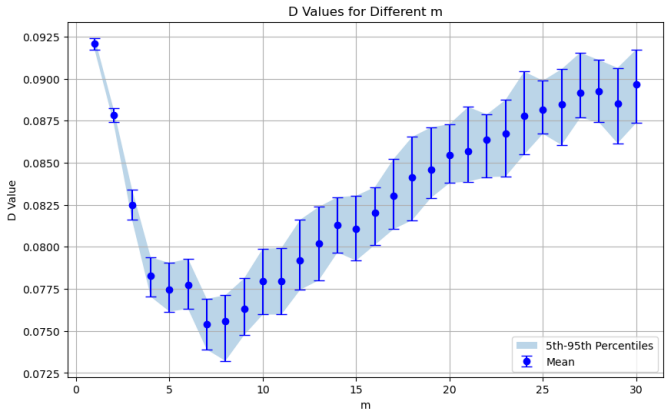


Figure 3: Graph showing the distance metric $D$ for different values of $m$

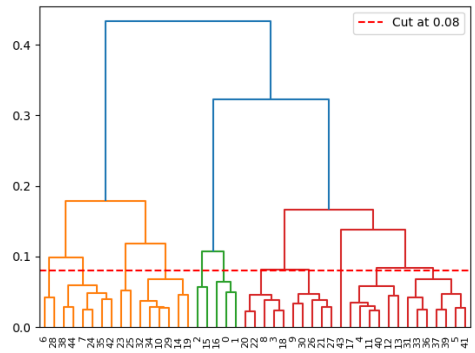


Figure 4: Dendrogram of the gauges using the Weierstrass distance

approach for $\mathbf{w}^*_{(m+1):k}$. A large $m$ provides greater model flexibility, but the approach suffers from the curse of dimensionality: if $m$ increases, the relatively few extreme observations of $\mathbf{Z}$ become more isolated in $\mathbb{S}^m$.//

We suggest choosing the value of $m$ using leave-one-out cross-validation. This method checks how well the model, focusing on the leading $m$ eigenvalues, captures dependence and considers how effectively the nearest neighbor approach deals with any leftover patterns.

To implement this, for each sample $i$ (where $i$ ranges from 1 to $n$), we exclude the $i$-th extreme event $\tilde{\mathbf{x}}^{(i)}$ from the data for $\tilde{\mathbf{X}}$ and estimate $\hat{\Sigma}^{(-i)}$ using the remaining data, with $r_V$ fixed at the 94% empirical quantile across all samples. Thus, each estimated TPDM $\hat{\Sigma}^{(-i)}$ comes from $n - 1 = 46$ extreme events in our river flow application.

Next, we generate 2,000 samples $\tilde{\mathbf{x}}_1^{(-i)}, \ldots, \tilde{\mathbf{x}}_{2000}^{(-i)}$ for $\hat{\Sigma}^{(-i)}$ using the generative framework, considering various possible values of $m$. We evaluate performance based on the proximity of $\tilde{\mathbf{x}}_1^{(-i)}, \ldots, \tilde{\mathbf{x}}_{2000}^{(-i)}$ and the $i$-th removed extreme event $\tilde{\mathbf{x}}^{(i)}$ using the distance metric:

$$D_i = 1 - \max_{j=1,\ldots,2000} \left( \frac{\tilde{\mathbf{x}}^{(i)}}{||\tilde{\mathbf{x}}^{(i)}||} \cdot \frac{\tilde{\mathbf{x}}_j^{(-i)}}{||\tilde{\mathbf{x}}_j^{(-i)}||} \right),$$

where $\cdot$ represents the dot product. A smaller $D_i$ indicates closer proximity between the samples and the removed extreme event.

The optimal $m$ is the one that minimizes the average error $\bar{D} = n^{-1}(D_1 + \cdots + D_n)$. This process provides a practical way to choose an appropriate $m$ for our analysis.

The graph in Figure 3 illustrates the distance metric $D$ for various values of $m$. It shows that the optimal value for $m$ falls within the range of 6 to 8, where the average error $\bar{D}$ is minimized. We will take $m = 7$.

## 3. Results

### 3.1. Transformation the data

Before fitting the Generalized Pareto Distribution (GPD) on the marginals, we computed clusters using the hierarchical clustering method with the Weierstrass distance. It is important to note that, given the differences in the distribution of the gauges, we standardized our data to ensure that each gauge is centered and has a scale of 1. The dendrogram in Figure 4 illustrates our clustering results.

We observe that, for most gauges, we can form clusters of consistent size. However, some gauges, such as gauge number 43, appear isolated. Since clusters of size 1 or 2 are not suitable for obtaining parameters, we opted to assign these isolated gauges to the nearest cluster based on the Weierstrass distance.

As mentioned earlier, there is room for improvement in our clustering method, particularly in handling isolated gauges. Our intuition suggests that, given that some gauges are situated in the same river or close proximity, the method should easily identify clusters due to the closeness of gauge distributions. Consequently, we did not attempt to implement a more refined clustering method, given our time constraints.

Using the obtained clusters, we can compute the $\sigma$ and $\xi$ parameters. The fitting of one of the clusters is illustrated in Figure 5. It is observed that the estimation error for $\sigma$ is acceptable. However, the $\xi$ parameter appears to be more uncertain. The values of the parameters used are provided in the appendix.

### 3.2. PCA

In this section, we present the results obtained from our Principal Component Analysis (PCA). The scree plot and the heatmap of $\Sigma$ are displayed in Figure 6. It is evident that the information contained in $\Sigma$ is well approximated by the first eigenvalues. This justifies our approach of focusing on the first principal components to generate new samples. Additionally, we observe regions with low values in the top-right and bottom-left corners of the matrix. As the gauges in these regions are not located close to each other, the coefficients of $\Sigma$ indicate that the extremum values of these gauges seem to be less dependent.

We also can interpret the eigenvectors we obtain. Figure 7 displays eigenvectors $U_1, U_2$ and $U_3$ based on the spatial locations of each gauge. The first eigenvector exhibits solely positive values, capturing the overall magnitude of extreme river

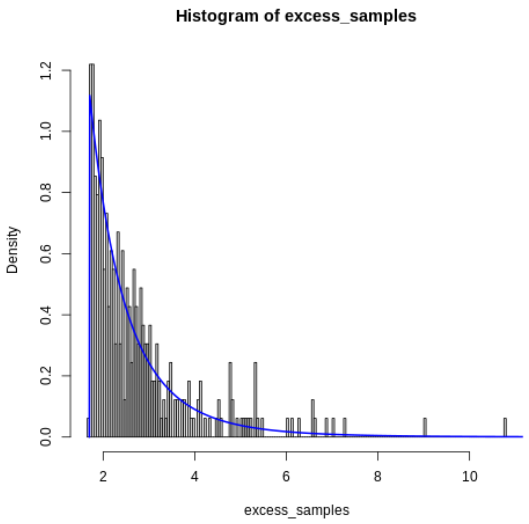


Figure 5: Fit of a GPD distribution of the exceedances for one cluster

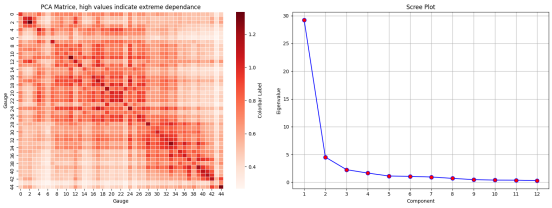


Figure 6: Heatmap of $\Sigma$ and scree plot

flow events, with peak values in the central study area. The second eigenvector reveals a north-south division, indicating a tendency for extreme river flow events to disproportionately impact either the northern or southern half of the study area. Additionally, $U_3$ demonstrates a west-to-east linear trend, excluding the southernmost gauges, which is attributed to their exposure to weather fronts from a westerly or south-westerly direction. Notably, the north-eastern gauges benefit from protection by the Pennines, while the most southern gauges reside in the rain shadow of Snowdonia in northern Wales.

### *3.3. Sampling W*

The results of our sampling of $W$ is shown in Figure 8. The number of observed values is $n = 47$, and $n = 848$ weeks samples were generated using the procedure to reproduce the same time frame as in the data collected. In the left plot, the

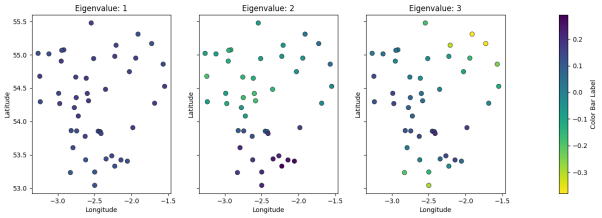


Figure 7: Visual representation of the eigenvectors in space

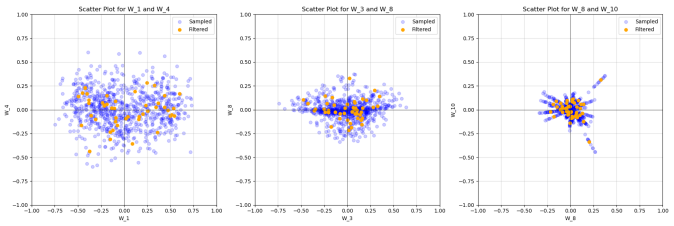


Figure 8: Pairwise plots of generated (blue) and observed (orange) values for $(W_1, W_4)$ (left), $(W_3, W_8)$ (middle) and $(W_8, W_{10})$ (right) for the UK river flow data.

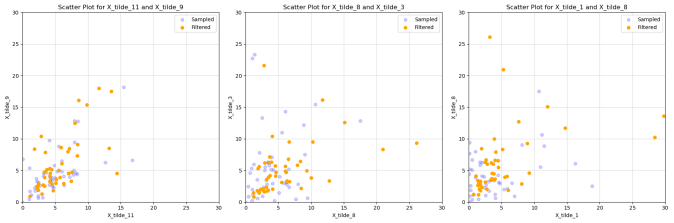


Figure 9: Pairwise plots of generated (blue) and observed (orange) values for $(\tilde{X}_9, \tilde{X}_1 1)$ (left), $(\tilde{X}_3, \tilde{X}_8)$ (middle) and $(\tilde{X}_8, \tilde{X}_1)$ (right) for the UK river flow data.

components were sampled using the von Mises Fisher Kernel, resulting in a "spherical" shape for the sampled data focused on previous extreme values. Conversely, the right graph displays components greater than $m = 7$, forming a "star" shape as these components essentially represent shifted points from the observations due to the nearest-neighbor approach outlined in the framework.

### *3.4. Sampling $\tilde{X}$*

The results of our sampling of $\tilde{X}$ is shown in Figure 9. We sample random variable from a Frechet distribution for the radius and we used the inverse transformation of Equation 3 to sample $\tilde{X}$ from the sampled $W$ from the precedent section. We can indeed see that the extreme $\tilde{X}$ generated are similar from the observed $\tilde{X}$ in terms of angles and norm.

### *3.5. Sampling X*

The results of our sampling of $X$ is shown in Figure 10. We sample by inverting the transformation of Equation 5. We can indeed see that the extreme $X$ generated are similar from the observed $X$ in terms of angles and norm.

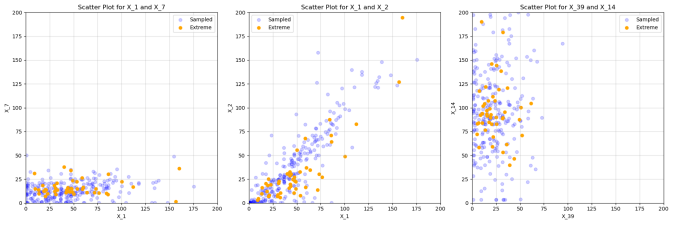


Figure 10: Pairwise plots of generated (blue) and observed (orange) values for $(X_1, X_7)$ (left), $(X_1, X_2)$ (middle) and $(X_{39}, X_{14})$ (right) for the UK river flow data.

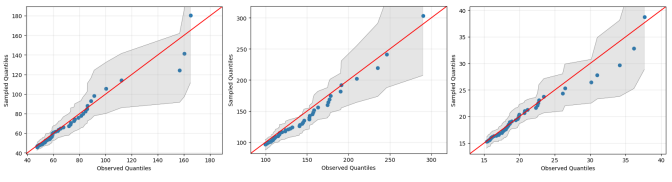

Figure 11: Quantile-quantile plots for the fifty largest observations of gauge 2 (left), 5 (middle) and 8 (right).

### *3.6. Assessing our results for X*

To validate our described approach, we simulate 100 event sets of the same length as the original data, each comprising $T = 848$ samples. In this context, we compare our generated values to the original river flow measurements, specifically analyzing the performance of the random vector $\mathbf{X}$. Initially, we assess the samples at a gauge level by contrasting the fifty largest order statistics between the observed and simulated event sets. Our findings indicate that the observed order statistics generally fall within their corresponding central 95% sampling intervals. Additionally, the marginal distributions of the sampled and observed values of $\mathbf{X}$ exhibit agreement as shown in Figure 11.

## 4. Summary and conclusions

This project aimed to replicate the work of Rohrbeck and Cooley (2023). It introduces an innovative generative framework, utilizing principal component analysis on the most severe events, to derive sets of extreme river flow hazard events across 45 gauges in northern England and southern Scotland. The methodology models the marginals of the distributions by employing a hierarchical clustering approach. Extremal dependence is then analyzed using the principal component decomposition method of Cooley and Thibaud (2019). The generative framework is based on a statistical model for extremal principal components, reducing dimensionality by modeling leading components with a kernel density estimate and employing an empirical estimate for the remaining components. Cross-validation is used to determine the number of principal components to model. In conclusion, we observe strong agreement between generated and observed extreme events. Therefore, the implementation of this method could be used in other scenarios to produce realistic extreme events.

## Appendix A. Appendix title 1

### *Appendix A.1. $\sigma$ and $\xi$ parameters of the GPD*

In the table A.1, the loc parameter is the threshold above which the data is considered as an extremum in a cluster. It was taken as the minimum over all the 96% quantile of the gauges in the cluster.

Table A.1: Parameters of the GPD

| custers | scale | shape | loc |
|---|---|---|---|
| 1 | 0.71843852 | 0.09368397 | 1.78708799865882 |
| 2 | 0.8681187 | 0.1547781 | 1.54888431354673 |
| 3 | 1.0377688 | 0.1689711 | 1.56718121474168 |
| 4 | 0.9124137 | 0.221533 | 1.62293110548231 |
| 5 | 0.9467621 | 0.2005266 | 1.5108382151862 |
| 6 | 0.9368907 | 0.1048575 | 1.58764319080148 |
| 7 | 0.8891009 | 0.1572848 | 1.64360110261808 |

Table A.2: Estimation errors for the parameters.

| clusters | std_scale | std_shape |
|---|---|---|
| 1 | 0.05398283 | 0.05622811 |
| 2 | 0.06272289 | 0.05360964 |
| 3 | 0.12481034 | 0.09031149 |
| 4 | 0.09021372 | 0.07698098 |
| 5 | 0.09301581 | 0.07711365 |
| 6 | 0.07252329 | 0.05900736 |
| 7 | 0.07505217 | 0.06562725 |

## Appendix B. Appendix title 2

## References


Cooley, D., Thibaud, E., 2019. Decompositions of dependence for high-dimensional extremes. Biometrika 106, 587–604.

Hall, P., Watson, G.S., Cabrera, J., 1987. Kernel density estimation with spherical data. Biometrika 74, 751–762.

Rohrbeck, C., Cooley, D., 2023. Simulating flood event sets using extremal principal components. Journal Name In press.

Rohrbeck, C., Tawn, J.A., 2021. Bayesian spatial clustering of extremal behavior for hydrological variables. Journal of Computational and Graphical Statistics 30, 91–105.