

Simulating flood event sets using extremal principal components

Alexandre Wan, Benjamin Lapostolle

December 2023

- Paper: *Simulating flood event sets using extremal principal components* by **Christian Rohrbeck** and **Daniel Cooley**.
- Objective: simulate extreme flooding events in the UK.

Context

PCA and Implementation

Data processing

Generative framework and sampling algorithm:

Conclusion

Context

- Source: UK's National River Flow Archive.
- January 1980 - September 2018: weekly maximum of **45** gauges recorded in southern Scotland and northern England.
- After processing: 780 weeks of data point.
- Hypothesis: Stationarity of the data.

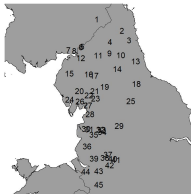


Figure: Locations of the gauges

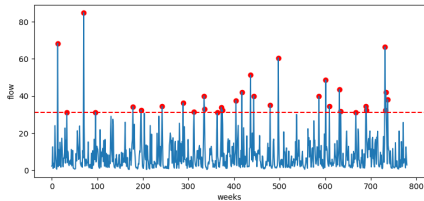


Figure: Distribution of gauge 1

Extremal PCA

- Let $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_K)$ such that $\mathbb{P}(\tilde{X}_k \leq x) = \exp(-x^{-2})$ for $x > 0, k = 1, \dots, K$.
- \tilde{X} regularly varying with index $\alpha = 2$: with $H_{\tilde{X}}$ the angular measure

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

Definition

We define the tail pairwise dependence matrix $K \times K$ (TPDM) Σ of \tilde{X} , which is positive semi-definite under the previous hypothesis, as:

$$\forall i, j \leq K, \Sigma_{i,j} = \int_{S_+^{K-1}} \omega_i \omega_j dH_{\tilde{X}}(\omega)$$

Estimator of the TPDM

Given $(\tilde{X}_t)_{t \in [0, T]}$, $r_t = \|\tilde{X}_t\|$, $r_0 \in \mathbb{R}$, $T^* = \{t \in [0, T], r_t > r_0\}$ and $\omega_{t,i} = \frac{X_t^i}{r_t}$, we estimate Σ using the formula:

$$\forall i, j \leq K, \Sigma_{i,j} = \frac{K}{|T^*|} \sum_{t \in T^*} \omega_{t,i} \omega_{t,j}$$

Then, obtain V , the representation of \tilde{X} in the eigen space with $\tau^{-1}(\cdot) = \log[\exp(\cdot) - 1]$:

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

V is regularly varying with $\alpha = 2$.

From X to \tilde{X}

- Normalize the data.
- $\forall k \leq K, \forall t \in T \quad \tilde{X}_{t,k} = \left[-\log \hat{F}_k(X_{t,k}) \right]^{-1/2}.$
- Estimator of \hat{F}_k : set to empirical cdf below the 96% quantile, set to a GPD(σ_k, ξ_k) above.

Fitting the GPD to the marginals

- High estimation errors: too few data points to fit a Generalized Pareto Distribution (GPD) on every marginal.
- Strategy: regroup the gauges into clusters.
- Creation of clusters: hierarchical clustering using the Wasserstein distance.
- Original paper: Clustering taking into consideration the spatial locations.

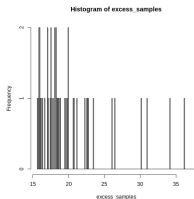


Figure: Histogram of the exceedances for one gauge

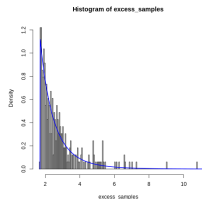


Figure: Fit of a GPD distribution of the exceedances for one cluster

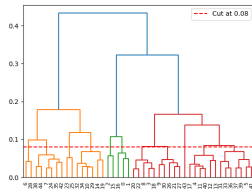


Figure: Dendrogram using the Weierstrass distance

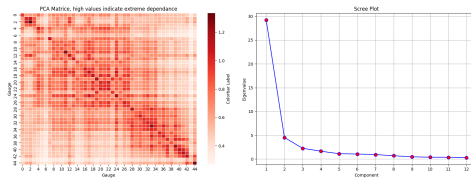
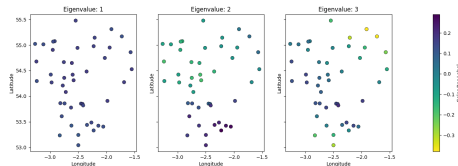
Figure: Heatmap of Σ and scree plot

Figure: Visual representation of the eigenvectors in space

Sampling extreme values for X

- Get X by sampling \tilde{X} and then inverting:

$$\tilde{X}_k = \left[-\log \hat{F}_k(X_k) \right]^{-1/2} \quad (k = 1, \dots, K),$$

- Get \tilde{X} by sampling V and then inverting:

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

- Sample V by sampling its radius r^* and direction W independently.
- Sample r^* from a Fréchet distribution with $\mathbb{P}(\|\mathbf{V}\|_2 \leq r) = \exp \left[-(r/K)^{-2} \right]$.
- We are left to sample W .

Sampling values for W

- The idea: reducing the problem to a dimension $m + 1$ instead of K
- Sample a random variable Z on the $(m + 1)$ -dimensional unit sphere \mathbb{S}^m .
- The first m components of Z represent the information contained in $W_{1:m}$.
- Z_{m+1} summarizes some aspects of the random vector $W_{(m+1):K}$.

Sampling values for Z

- Let $Z = (Z_1, \dots, Z_{m+1})$
- Sample a random variable Z on the $(m+1)$ -dimensional unit sphere \mathbb{S}^m .
- The first m components of Z represent the information contained in $W_{1:m}$. Z_{m+1} summarizes some aspects of the random vector $W_{(m+1):K}$.
- Let $Z = (Z_1, \dots, Z_{m+1})$ with $Z_j = W_j$ ($j = 1, \dots, m$) and

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

- Given observations $\{\mathbf{z}_i : i = 1, \dots, n\}$, density kernel estimate (von Mises Fisher):

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

with:

$$h(z; \mu, \kappa) = c_0(\kappa) \exp(\kappa z^T \mu)$$

Constructing w^* from a sample z^*

- We set $w_j^* = z_j^*$ ($j = 1, \dots, m$)
- For ($j = m + 1, \dots, K$), nearest-neighbour approach:

$$q = \operatorname{argmax}_{i=1, \dots, n} z_i^\top z^*.$$

Then:

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

Selecting a value for m

Different approaches:

- Elbow rule from PCA scree plot.
- Select m by leave-one-out cross validation. Remove the i -th extreme event $\tilde{\mathbf{x}}^{(i)}$ and generate 2,000 samples $\tilde{\mathbf{x}}_1^{(-i)}, \dots, \tilde{\mathbf{x}}_{2000}^{(-i)}$. Compute:

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

and select m minimizing: $\bar{D}^m = n^{-1} (D_1^m + \dots + D_n^m)$

Selecting a value for m

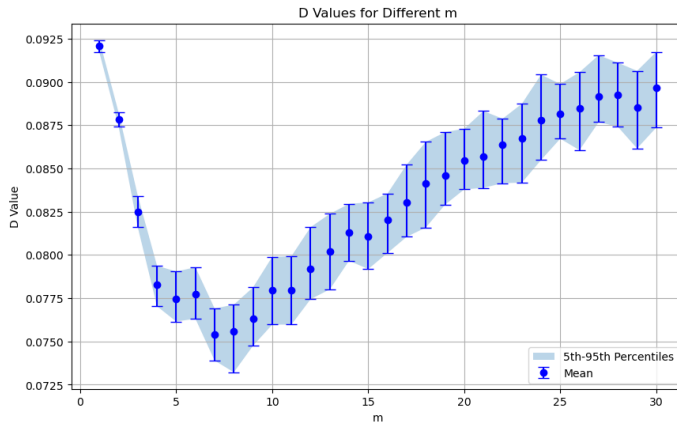


Figure: Graph showing the distance metric D for different values of m

Sampling values for W

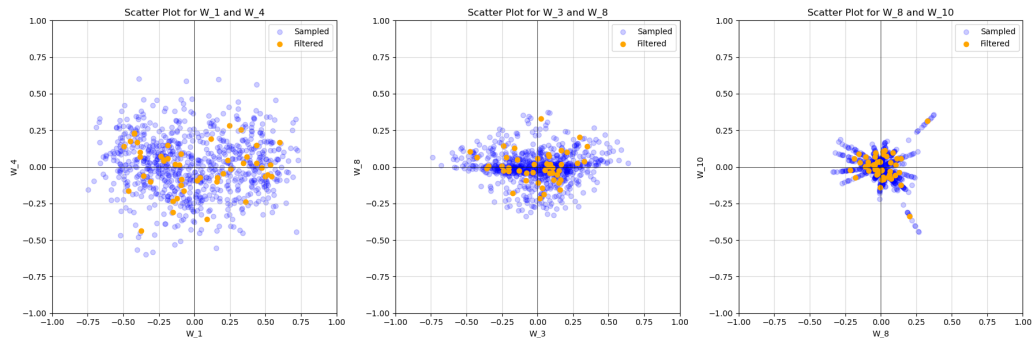


Figure: Pairwise plots of generated (blue) and observed (orange) values for (W_1, W_4) (left), (W_3, W_8) (middle) and (W_8, W_{10}) (right) for the UK river flow data.

Sampling values for \tilde{X}

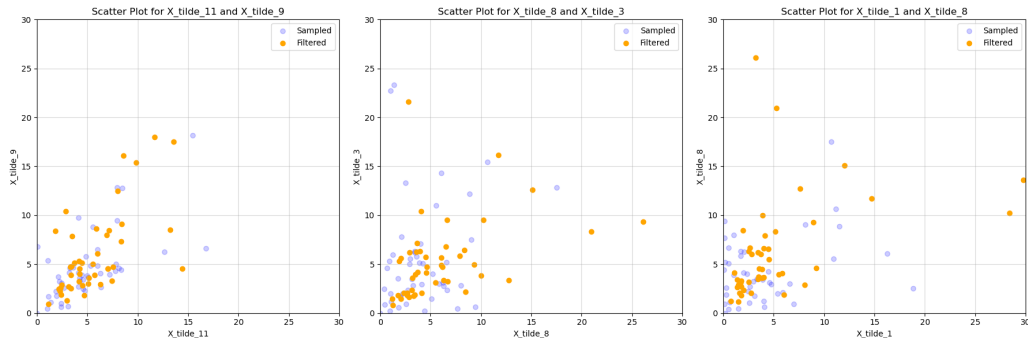


Figure: Pairwise plots of generated (blue) and observed (orange) values for $(\tilde{X}_9, \tilde{X}_{11})$ (left), $(\tilde{X}_3, \tilde{X}_8)$ (middle) and $(\tilde{X}_8, \tilde{X}_1)$ (right) for the UK river flow data.

Sampling values for X

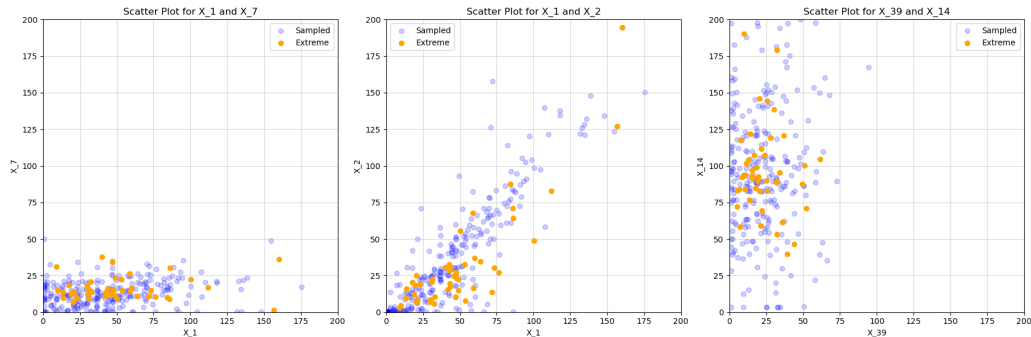


Figure: Pairwise plots of generated (blue) and observed (orange) values for (X_1, X_7) (left), (X_1, X_2) (middle) and (X_{39}, X_{14}) (right) for the UK river flow data.

Q-Q plots values for X

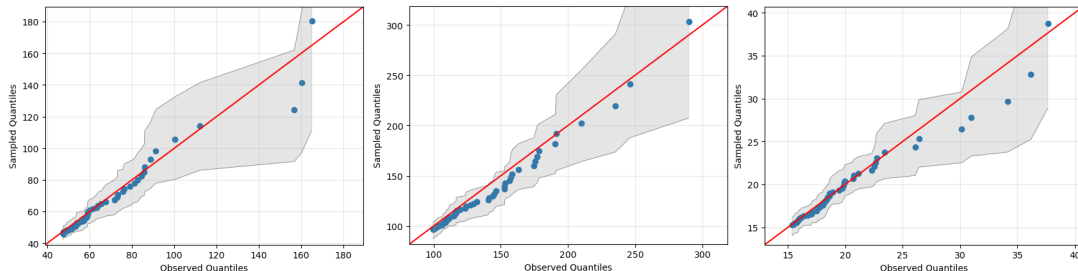


Figure: Quantile-quantile plots for the fifty largest observations of gauge 2 (left), 5 (middle) and 8 (right).

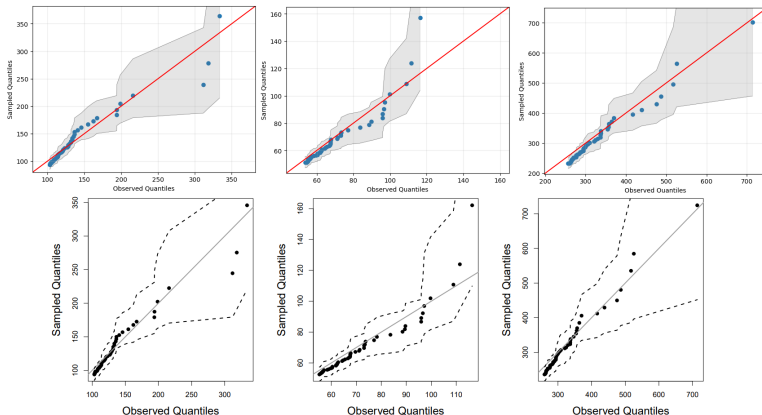







Figure: Quantile-quantile plots for the fifty largest observations of $\|X_G\|_2$ for the groups of gauges $G_1 = 3, 4, 16, 20, 24$ (left), $G_2 = 1, 27, 36, 41, 43$ (middle) and $G_3 = 6, 15, 21, 25, 33$ (right) sampled by us (top) and sampled in the paper (bottom).

- We were able to produce clusters to fit more accurately a GPD on the marginals.
- The extremal PCA allows us to work on a smaller space to sample extreme values.
- The extremes values sampled keep the dependence between the variables.

-  D. Cooley and E. Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019.
-  P. Hall, G. S. Watson, and J. Cabrera. Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762, 1987.
-  C. Rohrbeck and D. Cooley. Simulating flood event sets using extremal principal components. *Journal Name*, 2023. In press.
-  C. Rohrbeck and J. A. Tawn. Bayesian spatial clustering of extremal behavior for hydrological variables. *Journal of Computational and Graphical Statistics*, 30(1):91–105, 2021.
-  S. F. Sweere, I. Valtchanov, M. Lieu, A. Vojtekova, E. Verdugo, M. Santos-Lleo, F. Pacaud, A. Briassouli, and D. Cámpora Pérez. Deep learning-based super-resolution and de-noising for xmm-newton images. *Monthly Notices of the Royal Astronomical Society*, 517(3): 4054–4069, 2022.

-  A. Vojtekova, M. Lieu, I. Valtchanov, B. Altieri, L. Old, Q. Chen, and F. Hroch. Learning to denoise astronomical images with u-nets. *Monthly Notices of the Royal Astronomical Society*, 503(3):3204–3215, 2021.