Discrete Multivariate Generalized Pareto Distribution with application to dry spells

Samira Aka[1,2], Marie Kratz[2], Philippe Naveau[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement (LSCE)
CEA-CNRS-UVSQ

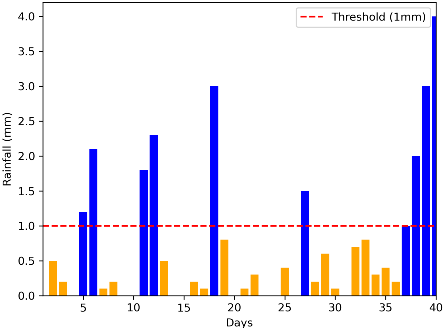
[2]Center for Research in Econo-Finance and Actuarial Sciences on Risk (CREAR)
ESSEC Business School

Data science pour les Risques Hydro-Climatiques et Côtiers 2025,
Roscoff, March 31, 2025

Defining a dry spell

Dry spell=number of consecutive dry days (below a precipitation amount threshold) ([Raymond et al., 2016, Lana et al., 2006])

Example of dry spell



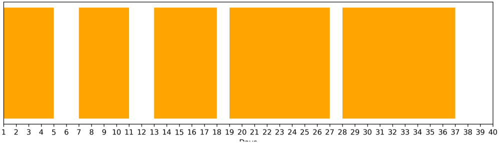
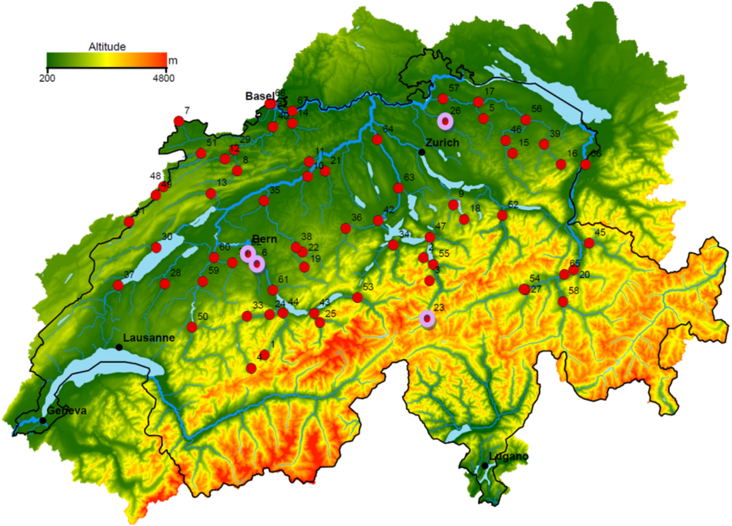
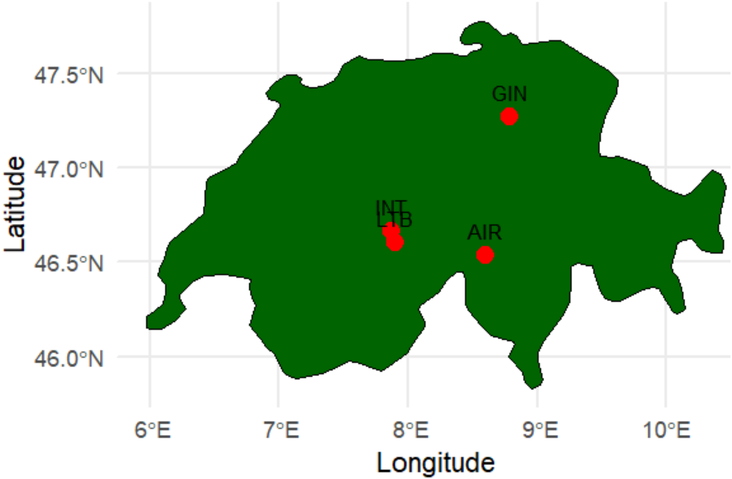


Figure: Precipitation amount with a threshold (blue for precipitation over the threshold and orange otherwise) and corresponding dry spells (orange)

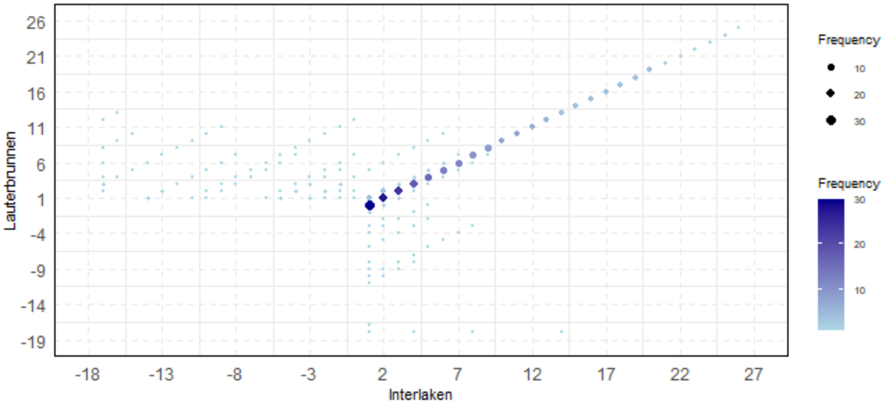
Precipitation stations in Switzerland



Selected Stations



Dry spells exceedances at two locations in Switzerland



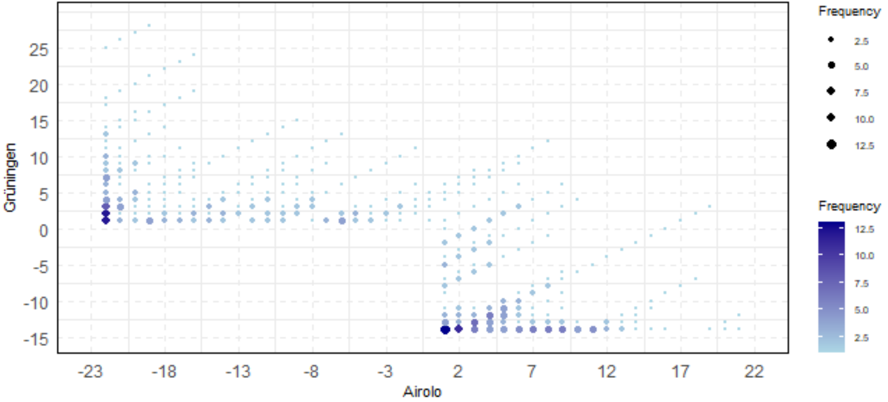
Dry spells exceeding quantile 99% : 17 days for Interlaken and 18 days for Lauterbrunnen (7 km between the stations)

What is the joint distribution ?

Motivation

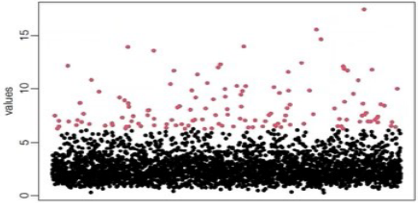
- **Main interest:** Modeling dry spells—successions of days with little or no precipitation—across various stations.
- **Broader question:** How can we model extremal dependence in multivariate discrete vectors?
- **Real-life applications:** Insurance claims, wildfire occurrences, and more.

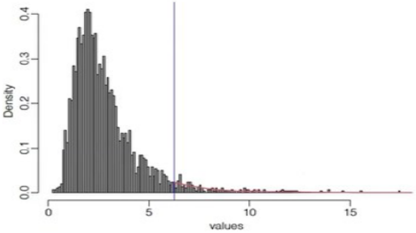
Dry spells exceedances at two locations in Switzerland



Dry spells exceeding quantile 99% : 22 days for Airolo and 14 days for Grüningen (80 km between the stations)

Exceedances in the continuous case





Generalized Pareto Distribution (GPD)

The distribution of exceedances above a large threshold u can be approximated as (Pickands-Balkema-de Haan) :

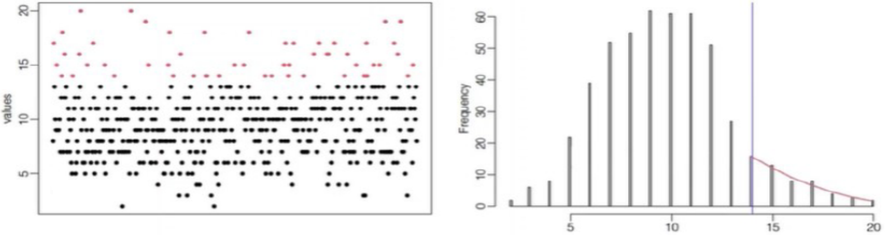
$$\mathbb{P}(Y - u > y \mid Y \geq u) \approx \overline{\mathrm{GPD}}(y; \sigma_u, \xi)$$

where σ_u depends on u, and

$$\overline{\mathrm{GPD}}(y; \sigma_u, \xi) = \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-\frac{1}{\xi}}, \text{ with } \sigma_u > 0,$$

is the survival function of the GPD.

Discrete GPD



Discrete generalized Pareto distribution (D-GPD) (see [Hitz et al., 2024], [Ahmad et al., 2022], [Daouia et al., 2023]).

Discrete GPD definition

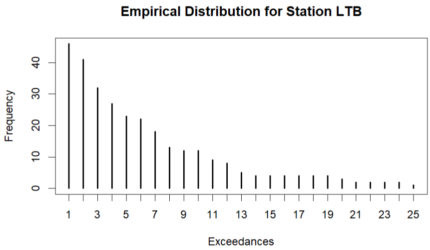
From [Hitz et al., 2024], the probability mass function p_{DGPD} of the discrete GPD is defined as, for $k \in \mathbb{N}$,

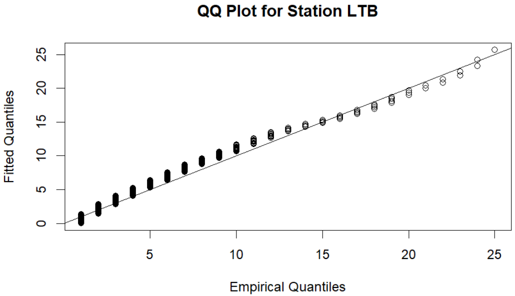
$$p_{\mathrm{DGPD}}(k; \sigma, \xi) = \overline{\mathrm{GPD}}(k; \sigma, \xi) - \overline{\mathrm{GPD}}(k+1; \sigma, \xi).$$

The fit of the DGPD can be done using the code attached to [Hitz et al., 2024].

Fitting the marginal exceedances on a DGPD

$\xi = -0.27$ and $\sigma = 8.78$.





Standard Multivariate GPD

Definition ([Rootzén et al., 2018], Theorem 7)

$\boldsymbol{Z} \in \mathbb{R}^d$ follows a $MGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ if:

- $\max(\boldsymbol{Z})$ unit exponential distribution,
- $\boldsymbol{S} = \boldsymbol{Z} - \max(\boldsymbol{Z})$ with $\boldsymbol{S}$ independent of $\max(\boldsymbol{Z})$.

Non-Standard Multivariate GPD

Definition ([Rootzén et al., 2018])

$\boldsymbol{Z} \in \mathbb{R}^d$ follows a $MGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ then

$$\boldsymbol{X} = \boldsymbol{\sigma} \frac{e^{\gamma \boldsymbol{Z}} - 1}{\gamma},$$

follows a $MGPD(\boldsymbol{\sigma}, \gamma, \boldsymbol{S})$.

State-of-the-Art for Generalized Pareto Distribution (GPD)

What is well-known about the GPD:

- Univariate GPD for exceedances in continuous univariate data,

State-of-the-Art for Generalized Pareto Distribution (GPD)

What is well-known about the GPD:

- Univariate GPD for exceedances in continuous univariate data,
- Univariate discrete GPD for exceedances in discrete univariate data,

State-of-the-Art for Generalized Pareto Distribution (GPD)

What is well-known about the GPD:

- Univariate GPD for exceedances in continuous univariate data,
- Univariate discrete GPD for exceedances in discrete univariate data,
- Multivariate GPD for exceedances in continuous multivariate data.

State-of-the-Art for Generalized Pareto Distribution (GPD)

What is well-known about the GPD:

- Univariate GPD for exceedances in continuous univariate data,
- Univariate discrete GPD for exceedances in discrete univariate data,
- Multivariate GPD for exceedances in continuous multivariate data.

$\rightarrow$ **Aim: Construction of a GPD distribution for discrete multivariate data.**

Standard MDGPD

Definition ([Aka et al., 2024])

$\boldsymbol{N} \in \mathbb{Z}^d$ follows a multivariate discrete Generalized Pareto Distribution $MDGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ if :

- $\max(\boldsymbol{N})$ geometric distribution with parameter $1 - e^{-1}$

$$\mathbb{P}(\max(\boldsymbol{N}) \leq k) = 1 - e^{-k}, k \in \mathbb{N}^*$$

- $\boldsymbol{S} = \boldsymbol{N} - \max(\boldsymbol{N})$ with $\boldsymbol{S}$ independent of $\max(\boldsymbol{N})$.

Threshold stability

Proposition ([Aka et al., 2024])

$\boldsymbol{N} \sim MDGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ *and* $\boldsymbol{m} \in \mathbb{N}^n$ *:*

$$\mathcal{L}(\boldsymbol{N} - \boldsymbol{m} | \boldsymbol{N} \nleq \boldsymbol{m}) = MDGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S_d})$$

Linear Combination

Proposition ([Aka et al., 2024])

$\boldsymbol{N} \sim MDGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$, $\boldsymbol{A} = (a_{ij})$ a matrix $\in \mathbb{N}^{n \times d}$ such that $\mathbb{P}(\sum_{j=1}^{d} a_{ij} N_j > 0) > 0, \forall i = 1, \dots, n$, and $\boldsymbol{m} \in \mathbb{N}^n$, then

$$\mathcal{L}(\boldsymbol{AN} - \boldsymbol{m} | \boldsymbol{AN} \not\leq \boldsymbol{m}) = MDGPD(\mathbf{A1}, \mathbf{0}, \boldsymbol{S_n}).$$

Non-standard *MDGPD*

$\boldsymbol{N} \in \mathbb{Z}^d$ follows a $MDGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ then

$$\boldsymbol{\sigma} \frac{e^{\gamma \boldsymbol{N}} - 1}{\gamma},$$

follows a $MDGPD(\boldsymbol{\sigma}, \gamma, \boldsymbol{S})$.

Non-standard *MDGPD*

$\boldsymbol{N} \in \mathbb{Z}^d$ follows a $MDGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ then

$$\boldsymbol{\sigma} \frac{e^{\gamma \boldsymbol{N}} - 1}{\gamma},$$

follows a $MDGPD(\boldsymbol{\sigma}, \gamma, \boldsymbol{S})$.

NO!

Non-standard *MDGPD*

$\boldsymbol{N} \in \mathbb{Z}^d$ follows a $MDGPD(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ then

$$\boldsymbol{\sigma} \frac{e^{\gamma \boldsymbol{N}} - 1}{\gamma},$$

follows a $MDGPD(\boldsymbol{\sigma}, \gamma, \boldsymbol{S})$.

NO!

Non-standard *MDGPD*

Definition ([Aka et al., 2024])

$\boldsymbol{K} \in \mathbb{Z}^d$ follows a non-standard multivariate discrete Generalized Pareto Distribution *MDGPD* $\left(\boldsymbol{\sigma} = \frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}}, \gamma = \frac{\mathbf{1}}{\boldsymbol{\alpha}}, \boldsymbol{S}\right)$ if

$$\mathbb{P}\left(\boldsymbol{K} \leq \boldsymbol{k}\right) = 1 - \mathbb{E}\left(1 \wedge e^{\max\left(\boldsymbol{S} - \boldsymbol{\alpha} \log\left(\frac{\boldsymbol{k}+\mathbf{1}}{\boldsymbol{\beta}} + \mathbf{1}\right)\right)}\right).$$

Generalization of the *MDGPD* by ratio on a Gamma

Proposition ([Aka et al., 2024])

$\boldsymbol{Z}$ MGPD$(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$ and Λ a Gamma(α, β) random variable independent from $\boldsymbol{Z}$. Then,

$$\left\lfloor \frac{\boldsymbol{Z}}{\Lambda} \right\rfloor \text{ follows a } MDGPD\left(\frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}}, \frac{\mathbf{1}}{\boldsymbol{\alpha}}, \boldsymbol{S}\right).$$

Generalized threshold stability

Proposition ([Aka et al., 2024])

Let $\boldsymbol{M} = \frac{\boldsymbol{Z}}{\Lambda}$, *where* $\boldsymbol{Z}$ *and* Λ *are defined as in the previous proposition,* $\boldsymbol{A} = (a_{ij})$ *be a matrix* $\in \mathbb{N}^{n\times d}$ *such that* $\mathbb{P}(\sum_{j=1}^{d} a_{ij} N_j > 0) > 0, \forall i = 1, \ldots, n$, *and* $\boldsymbol{m} \in \mathbb{N}^n$, *then* $\mathcal{L}(\lfloor \boldsymbol{AM} \rfloor - \boldsymbol{m} | \lfloor \boldsymbol{AM} \rfloor \nleq \boldsymbol{m}) = MDGPD(\frac{\beta}{\alpha}\mathbf{A1}, \frac{\mathbf{1}}{\boldsymbol{\alpha}}, \boldsymbol{S_n})$.

Building $\boldsymbol{N}$ from the increments of a distribution

If $\boldsymbol{N} \sim \boldsymbol{MDGPD}(\mathbf{1}, \mathbf{0}, \boldsymbol{S})$, then:

$$\boldsymbol{N} = \underbrace{\boldsymbol{T} - \max(\boldsymbol{T})}_{\boldsymbol{S} \perp G} + \underbrace{G}_{\max(\boldsymbol{N})},$$

with $\boldsymbol{T}$ any discrete random vector called the generator.

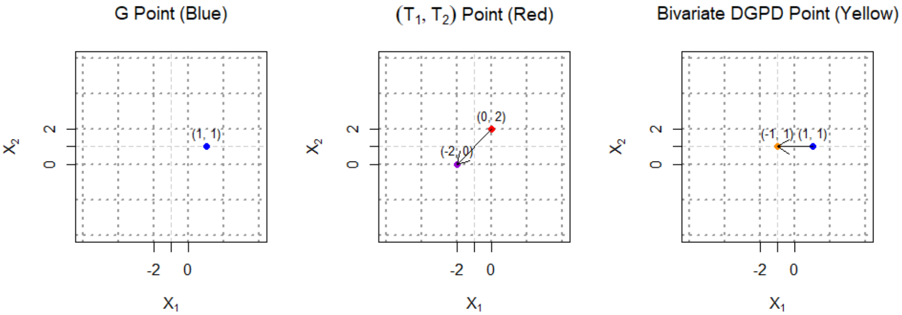
Other definition of N in the bivariate case

In the bivariate case,

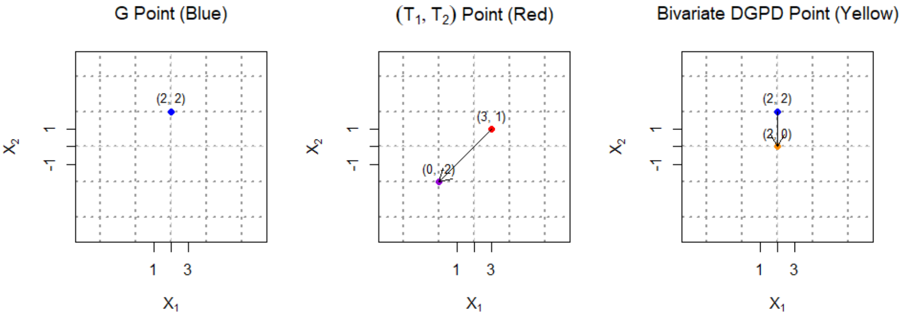
$$N_1 = G + (T_1 - T_2)\mathbb{1}_{((T_1 - T_2)<0)},$$
$$N_2 = G - (T_1 - T_2)\mathbb{1}_{((T_1 - T_2)\geq 0)}.$$

Note that $N_1 - N_2 = T_1 - T_2$.

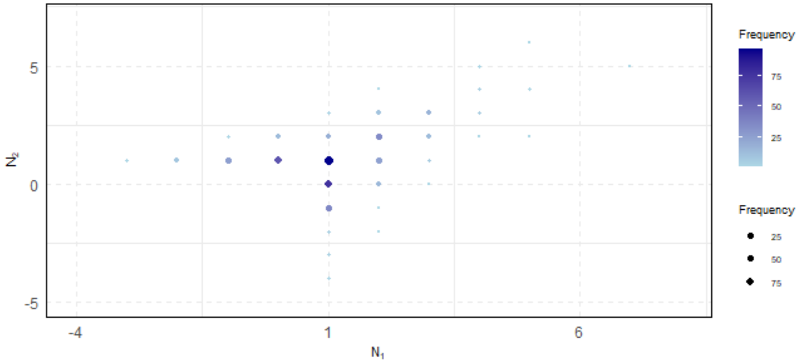
Creating a Bivariate DGPD (1)



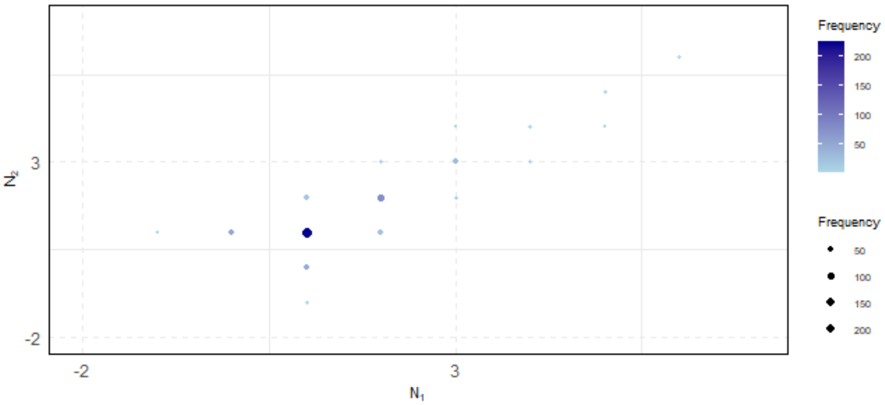
Creating a Bivariate DGPD (2)



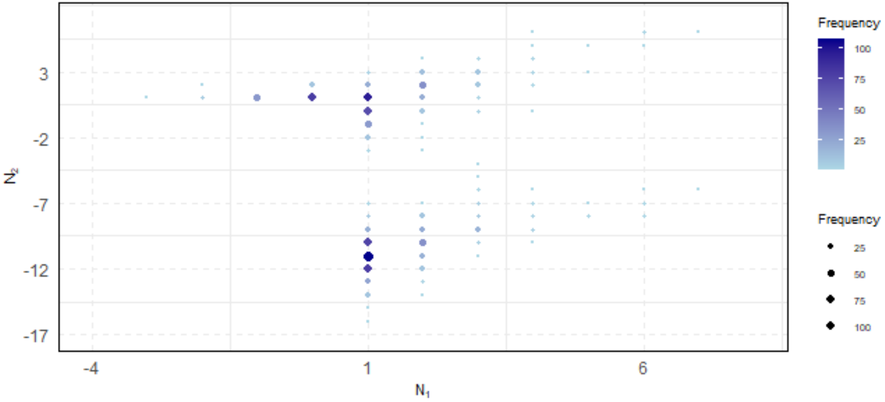
Bivariate DGPD: T_1 and T_2 Independent Poisson Variables



Bivariate DGPD: T_1 and T_2 Dependent Poisson Variables



Bivariate DGPD: T_1 and T_2 Bimodal Poisson Variables



Bootstrap algorithm to simulate a bivariate *DGPD* from unknown (T_1, T_2)

From Bootstrap simulations algorithm in [Legrand et al., 2021].

Algorithm Bootstrap *MDGPD* simulation

Require: A sample of $(N_{1,i}, N_{2,i})_{1\leq i\leq n} \sim MDGPD$

Ensure: A discrete simulated sample $(N^*_{1,k}, N^*_{2,k})_{1\leq k\leq m}$ to choose

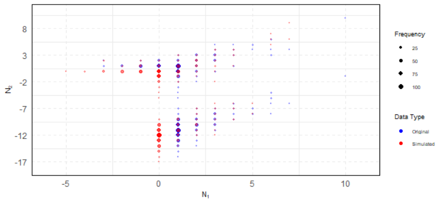
1: Define $\Delta_i = N_{1,i} - N_{2,i}, 1 \leq i \leq n$,

2: Generate m generalizations $G_k \sim Geom(1-e^{-1}), 1 \leq k \leq m$, independently from Δ_i,

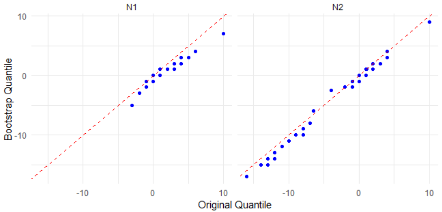
3: Bootstrap m realization Δ^*_k from $(\Delta_1, ..., \Delta_n)$

4: **Return** $N_{1,k} := G_k + \Delta^*_k \mathbb{1}_{(\Delta^*_k<0)}$ and $N_{2,k} := G_k - \Delta^*_k \mathbb{1}_{(\Delta^*_k\geq 0)}$, for $1 \leq k \leq m$.

Bivariate DGPD $N = (N_1, N_2)$ from unknown (T_1, T_2)



(a) Scatter plot of simulated data with the parametric model of sample size $n = 500$ (blue dots) and sampled data from one simulation with sample size $m = 500$ (red dots).



(b) Q-Q plot for marginals N_1 and N_2.
x-axis : Original sample
y-axis : Bootstrap sample

Figure: Bootstrap simulations of (N_1, N_2) using bimodal Poisson (T_1, T_2) and Geometric variable G.

How to Perform Inference on the *MDGPD*?

Discrete Data

$\downarrow$ **Challenges**

$$\text{Marginals} \neq \textit{Geometric}(1 - e^{-1}).$$

How to transform discrete distributions into another?

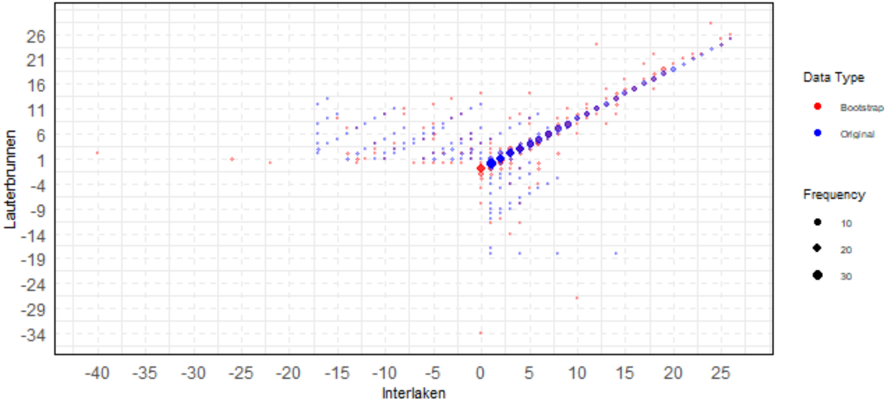
- **If continuous:** 🙂 Easy
- **If discrete:** 🙁 Not so easy

A solution: Simulation-based inference with Neural Networks

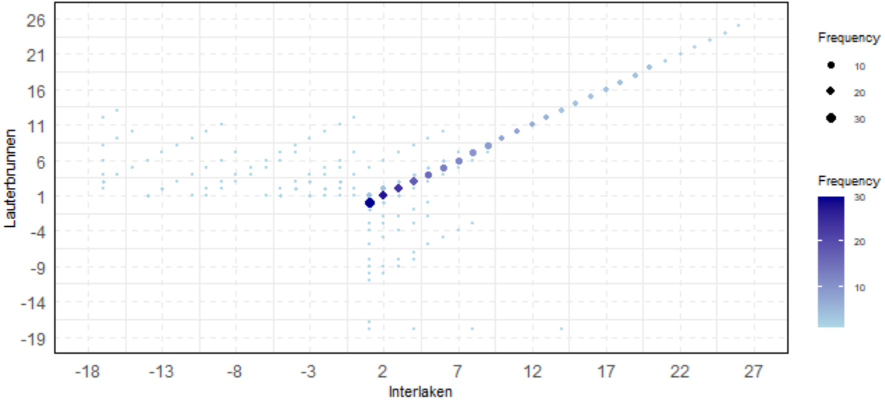
Deep Neural Networks for Inference on the *MDGPD*

- Neural networks (NN) are used for inference in intractable models, such as those in [Pacchiardi et al., 2021] and [Lenzi et al., 2023], which focus on max-stable models.
- In this work, we apply **Neural Bayes Estimators** as introduced by [Sainsbury-Dale et al., 2024].

Bootstrap simulations at Two Close Stations in Switzerland

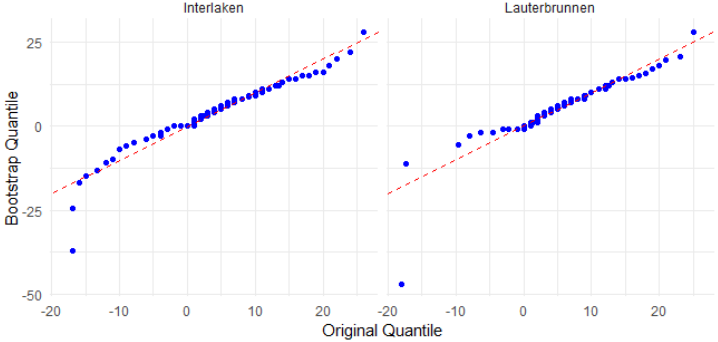


Dry spells exceedances at two locations in Switzerland



Dry spells exceeding quantile 99% : 17 days for Interlaken and 18 days for Lauterbrunnen (7 km between the stations)

Quantile-Quantile plot of marginals (close pair)



Future Directions

- Develop a dedicated package for the MDGPD, facilitating easier application and wider use.
- Investigate extreme value regression using covariates such as geographical features, temperature, and soil moisture, within the MDGPD framework.

References I

Ahmad, T., Gaetan, C., and Naveau, P. (2022).

Modelling of discrete extremes through extended versions of discrete generalized pareto distribution.

Aka, S., Kratz, M., and Naveau, P. (2024).

Discrete multivariate generalized Pareto distribution with application to dry spells.

Balkema, A. A. and de Haan, L. (1974).

Residual life time at great age.

The Annals of Probability, 2.

References II

Daouia, A., Stupfler, G., and Usseglio-Carleve, A. (2023).

Extreme value modelling of sars-cov-2 community transmission using discrete generalized pareto distributions.

Royal Society Open Science, 10(3):220977.

Hitz, A. S., Davis, R. A., and Samorodnitsky, G. (2024).

Discrete extremes.

Journal of Data Science, pages 1–13.

References III

Lana, X., Martínez, M. D., Burgueño, A., Serra, C., Martín-Vide, J., and Gómez, L. (2006).

Distributions of long dry spells in the iberian peninsula, years 1951–1990.

International Journal of Climatology, 26:1999–2021.

Legrand, J., Naveau, P., and Oesting, M. (2021).

Evaluation of binary classifiers for asymptotically dependent and independent extremes.

Lenzi, A., Bessac, J., Rudi, J., and Stein, M. L. (2023).

Neural networks for parameter estimation in intractable models.

Computational Statistics Data Analysis, 185:107762.

References IV

Pacchiardi, L., Adewoyin, R., Dueben, P., and Dutta, R. (2021).

Probabilistic forecasting with generative networks via scoring rule minimization.

Journal of Machine Learning Research, 25:45:1–45:64.

Pickands, J. (1975).

Statistical inference using extreme order statistics.

The Annals of Statistics, 3.

References V

Raymond, F., Ullmann, A., Camberlin, P., Drobinski, P., and Smith, C. C. (2016).

Extreme dry spell detection and climatology over the mediterranean basin during the wet season.

Geophysical Research Letters, 43:7196–7204.

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018).

Multivariate generalized pareto distributions: Parametrizations, representations, and properties.

Journal of Multivariate Analysis, 165:117–131.

References VI

Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2024).

Likelihood-free parameter estimation with neural bayes estimators.

The American Statistician, 78:1–14.