Improved regional frequency analysis of rainfall data

Philomène Le Gall, Anne-Catherine Favre, Philippe Naveau, Clémentine Prieur

Improved Regional Frequency Analysis of rainfall data

Philomène Le Gall

Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, F-38000 Grenoble, France

Anne-Catherine Favre

Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, F-38000 Grenoble, France

Philippe Naveau

Laboratoire des Sciences du Climat et de l'Environnement, ESTIMR, CNRS-CEA-UVSQ, Gif-sur-Yvette, France

Clémentine Prieur

Univ. Grenoble Alpes, CNRS, INRIA, LJK, F-38000 Grenoble, France

Corresponding author address:

E-mail: philomene.le-gall@univ-grenoble-alpes.fr

ABSTRACT

Rainfall are subject to local orography features and their intensities can be highly variable. In this context, identifying climatically coherent regions can greatly help interpreting rainfall patterns and improve the inference of return levels. In practice, partitioning a region of interest into *homogeneous* sub-regions is a delicate statistical task, especially in regards to modeling heavy rainfall features.

In this work, our main goal is to propose and study a fast and efficient clustering algorithm. Compared to classical regional frequency analysis techniques, a key aspect is that our algorithm does not rely on the *a priori* choice of covariates. The proposed numerical scheme is only based on the precipitation dataset at hand, including low, moderate and heavy rainfall. In terms of inference, our approach builds on the easy-to-compute and reliable probability weighted moments commonly used in hydrology. While being in compliance with extreme value theory, we do not impose a parametric form on rainfall distributions and, neither thresholding nor block maxima steps are required in our proposed approach. By construction, our clustering method preserves the scale invariance principle of any classical regional frequency analysis.

The performance of our clustering algorithm is assessed on a detailed experimental design based on the extended Generalized Pareto distribution.

Sensitivity to the number of clusters is carefully analyzed.

We apply our clustering algorithm on Switzerland daily precipitation measured at 191 sites. The found homogeneous regions are consistent with local orography and our approach outperforms the classical regional frequency analysis based on normalized elevation and coordinates as covariates. To complete our analysis of Swiss rainfall, we propose three models based on our clustering outputs. A comparison between our local, semi-regional and regional models indicates that a relatively simple model with two clusters and a spatially varying scale parameter can compete very well against complex models.

Key words : spatial clustering ; precipitation extremes ; probability weighted moments; extended generalized Pareto distribution

1. Introduction

It is well known that heavy rainfall can be responsible for critical floods (e.g., see Gottardi et al. 2012). Physically, the spatial distribution of daily precipitation depends on multivariate and complex factors, in particular on local orography features and small scale and large climatic phenomena (e.g., see Boucefiane and Meddi 2019). As an example, the annual mean precipitation in Switzerland is significantly higher than the European one, 1456 mm *vs* 790 mm (e.g., see Hilker et al. 2009). The specific Helvetic orography greatly influences rainfall probability distributions, see Figure 1 for the elevation map of Switzerland. The annual mean precipitation in Valais, a canton in southern Switzerland, is twice lower than the national one. A great part of these precipitation stems from Atlantic air flows. The regions on the leeward side are drier than the windward regions (e.g., see Zryd 2008). Concerning extremes, 3-day heavy rainfall over Switzerland induced numerous fatalities and huge loss (e.g., see Hilker et al. 2009; Barton et al. 2016). Again, low, moderate and extreme precipitation intensities can be highly variable in space.

In this context, delimiting coherent regions is essential to efficiently capture the distributional climatic features of rainfall at the correct spatial scale. This also makes sense from a statistical point of view. It is almost impossible to detect trends in a single rainfall time series because of its high variability. Combining a few stations together can improve the signal/noise ratio and allows hydrologists and climatologists to detect significant signals, like the impact of anthropogenic forcing on rainfall data. Defining so-called homogeneous regions has been a recurrent theme in hydrology and, in this article, we anchor our work to the following probability based definition (see, e.g. Hosking and Wallis 2005). Given a region of interest, say $\mathcal{R}$, a homogeneous cluster, say $\mathcal{C}$, is defined as a sub-region where all spatial points, say s, have the same marginal distribution up to a normalizing factor, i.e.

$$\mathcal{C} = \Big\{ s \in \mathcal{R} : \;\; Y(s) \stackrel{d}{=} \sigma(s) \times Z(s) \Big\}, \tag{1}$$

where $\stackrel{d}{=}$ means equality in distribution, the positive scalars $\sigma(s)$ are allowed to vary in space, and $Z(s)$ represents a positively-valued stationary process in time and space. In particular, the stationarity of $Z(s)$ implies that its marginal probability density function (pdf) does not depend on s. Eq.(1) is closely linked to the Regional Frequency Analysis (RFA) methodology introduced by Hosking et al. (1985), see Figure 2 and applied by various authors (e.g., see Onibon et al. 2004; Ouarda et al. 2008).

There exists a wide variety of methods that attempt to find homogeneous regions that satisfy Eq.(1). Most approaches rely on auxiliary geographical features and/or climatic information (e.g., see Asadi et al. 2018; Fawad et al. 2018, for recent works on this approach). In nutshell, explanatory covariates characterizing station locations and/or weather patterns are carefully selected to spatially explain rainfall features (e.g., see Burn 1990; St-Hilaire et al. 2003; Evin et al. 2016). For example, Carreau et al. (2017) regressed

non-parametrically $\sigma(s)$ as a function of weather stations latitude and longitude coordinates. This crucial variable selection step requires subjectivity, data availability and may be complex to transfer over regions with different climatic drivers. For example, the chosen covariates tailored to the Valais region could be different for stations a few hundred of kilometers away.

To assess the quality of a given partitioning, RFA homogeneity tests were also proposed to check the validity of model expressed in Eq.(1) (see Hosking and Wallis 2005). A major ingredient is the computation of specific moments and related ratios to measure variability, sknewness and kurtosis for positive random variables. To implement goodness-of-fit tests, another key component to obtain significance levels was the parametric assumption that rainfall follow kappa-like distributions. Such parametric assumptions may be too stringent in practice. In addition, and according to Viglione et al. (2007), RFA homogeneity tests can suffer from a lack of power.

The first objective of this work is to bypass the delicate step of explanatory variables (covariates) selection that produces *a priori* clusters. Instead, our strategy is to cluster from the raw data, i.e. precipitation themselves. A second goal is to avoid imposing a parametric family like the kappa distribution present in the RFA homogeneity tests.

Concerning our first objective and since the work of Hosking and his colleagues, it is known that a few well chosen moments can characterize important features of precipitation intensities. For example, probability weighted moments (PWM) can adequately capture the main features of heavy rainfall distribution (e.g., see Carreau et al. 2017). The theoretical justification of this claim resides in extreme value theory (EVT) (e.g., see Coles et al. 2001; Fougères 2004; Davison and Huser 2015). In particular, mathematical arguments can be used to justify that the upper tail behavior of renormalized rainfall excesses over a high threshold can be well approximated by a generalized Pareto (GP) survival function. If we now look at GP distributed extremes with respect to Eq.(1), then the ratio of survival functions from two stations, say s and s', in $\mathcal{C}$ satisfies

$$\lim_{y\to\infty} \frac{\mathbb{P}[Y(s) > y]}{\mathbb{P}[Y(s') > y]} = \lim_{y\to\infty} \frac{\overline{F}(\sigma(s)y)}{\overline{F}(\sigma(s')y)} = c \in (0, \infty) \tag{2}$$

where $\mathbb{P}[Z(s') > z) = \overline{F}(z)$ and, according to EVT, the finite positive constant c depends on the ratio of $\sigma(s)$ over $\sigma(s')$. In contrast, if the two stations, s and s', belong to two different clusters with non-equal GP shape parameters, then this ratio of two survival functions will either go to zero or infinity. It follows that a homogeneous cluster, by construction, means to be tail invariant, i.e. the ratio goes to the same positive constant c within a cluster. Note that condition (1) implies condition (2), but the converse is not true. So, our simulation study explores a setup based on Eq. (2), a broader family, and our Swiss rainfall will be analyzed with models satisfying both conditions.

Clustering stations in homogeneous regions is a simply way, by gathering stations with an equivalent tail

behavior, to improve the analysis of heavy rainfall (e.g., see Zhang et al. 2012). Traditional at-site method consists in fitting a EVT distribution to each site (e.g., see Li et al. 2019). Then, stations could be gathered according to their parameter estimates. At-site estimators of shape parameter though are known to be poorly robust and require long sequences to be reliable (e.g., see Zhang et al. 2012; Malekinezhad and Zare-Garizi 2014; Jalbert et al. 2017). At the other end of complexity and away from analysis, complicated models based on a hierarchy of layers can also capture spatial dependence by imposing random smooth fields over some EVT parameters (Cooley et al. 2007; de Fondeville and Davison 2018). The main drawback of these techniques is the complexity of implementation in the sense that a deep knowledge of Bayesian hierarchical modeling, in particular of Monte-Carlo sampling techniques, is required to fit and understand such models.

As a computationally simple inference alternative, many authors have studied the links between PWMs and EVT parameters (e.g., see Kojadinovic and Naveau 2017). In particular, explicit expressions of PWMs parameters for EVT distributions have been investigated theoretically (e.g., see Ferreira and de Haan 2015) and used in practical setups (e.g., see Ribereau et al. 2008; Carreau et al. 2017). Besides characterizing EVT distributions, PWMs are also distribution-free moments that can be quickly estimated non-parametrically for any data set. Our idea is to use them as direct inputs of classical clustering algorithms. The only requirement is that the PWM based input remains scale invariant within Eq.(1). Although simple, this strategy can lead to coherent regions with the advantage of avoiding the arbitrary choice of explanatory covariates. In addition, no strong parametric assumptions are needed to implement our approach, see Section 2 for details.

To assess our method with a simulation study, we need a class of distribution that is capable of modeling the whole spectrum of rainfall intensities. In recent years, various approaches (see, e.g. Carreau and Bengio 2009; Naveau et al. 2016; Tencaliec et al. 2020; Stein 2020) have been proposed to combine a Pareto pdf, for modeling the upper tail, with different types of transfer functions to allow the fit of the distribution bulk and its lower tail. Daily rainfall over Switzerland were well captured by the so-called Extended GPD (EGPD) (Evin et al. 2018) studied by Naveau et al. (2016). In addition, PWMs were also explicitly derived for special cases of the EGPD family. Hence, this class offers a well understood benchmark for our clustering approach.

Section 3 details our simulation study and highlights the advantages and limitations of our approach. In Section 4, we apply this algorithm to Swiss precipitation data. Our clustering is compared to the classical RFA approach based on geographical covariates and described in Hosking and Wallis (2005). RFA homogeneity test-based algorithm are applied to assess the two approaches. In addition, we propose a semi-regional fit to improve flexibility within each cluster, see Subsection 4c.

2. Methods

a. PWMs and metrics

To compare distributional features of different rainfall time series, simple and fast summaries are needed. In this context, computing metrics offers mathematically sound tools. For symmetrical distributions with finite variances, $\mathcal{L}^2$-norms of the type $\mathbb{E}\left|Z_1 - Z_2\right|^2$ are convenient to capture relevant information contained in Z_1 and Z_2. The archetypical example is Gaussian random variables that are entirely characterized by their mean and variance. As rainfall are positive, skewed and can be heavy-tailed, other distances need to be proposed and studied. In this work, we focus on $\mathcal{L}^1$–distance of the type $\mathbb{E}\left|Z_1 - Z_2\right|$ because they are closely linked to PWMs, see Appendices A and B for more details. We denote $\alpha_j(Z)$, or merely α_j when the link with the variable Z is self evident, the PWM of order j.

Under the RFA constraint (scale invariance), see Eq. (1), we can write that for any $s \in \mathcal{C}$,

$$\frac{1}{2}\mathbb{E}\left|Y_1(s) - Y_2(s)\right| = \sigma(s) \times \frac{1}{2}\mathbb{E}\left|Z_1(s) - Z_2(s)\right|$$

To build homogeneous regions, one needs to remove the effect of the scaling factor $\sigma(s)$. This can be done by introducing the following ratio

$$\omega = \frac{\mathbb{E}\left|\max\left(Z_1, Z_2\right) - \max\left(Z_1, Z_3\right)\right|}{\mathbb{E}\left|Z_1 - Z_2\right|}.$$

It is possible to show $0 \leq \omega \leq 1$, see Appendix C for a proof. Concerning our RFA problem, one can notice that the ratio ω satisfies

$$\omega(a + b\mathbf{Z}) = \omega(\mathbf{Z}) \text{ for any } a \text{ and } b > 0.$$

The ratio ω can also be rewritten as a ratio of $\mathcal{L}^1$-distance between the triplet (pair) maximum and its mean, see Appendix D. A similar ratio was highlighted in Kojadinovic and Naveau (2017) who studied change point detection in block maximum time series.

By construction, the ratio ω has also a clear connection with PWMs, see Appendix D, and in the special independent and identically distributed (i.i.d.) case, it simply becomes

$$\omega = \frac{3\alpha_2 - \alpha_0}{2\alpha_1 - \alpha_0} - 1. \tag{3}$$

This expression can be rewritten as

$$\omega = \frac{1}{2} - \frac{1}{2}\frac{\lambda_3}{\lambda_2},$$

where λ_2 and λ_3 represent the second and third L-moments studied by Hosking and Wallis (2005), see Appendix A for their definition.

Explicit expressions of ω can be found when the practitioner is ready to impose a parametric family. For example, if Z has a survival Generalized Pareto (GP) function, i.e. $\mathbb{P}(Z > z) = \overline{H}_{\sigma,\xi}(z) = (1 + \xi z/\sigma)_+^{-1/\xi}$ where σ represents the scale parameter and ξ is called the shape parameter and drives the upper tail behavior (see, e.g. Coles et al. 2001). For $\xi = 0$, $\overline{H}_{\sigma,0}(z) = \exp(-z/\sigma)$ for $z > 0$. For $\xi \geq 1$, the mean of Z is not finite anymore and the interpretation with PWMs is lost. So, we always assume that $\xi < 1$ in this work. Applying Eq. (3) for the GP case gives

$$\omega = \frac{5 - \xi}{3 - \xi} - 1.$$

If Z follows a Generalized Extreme Value (GEV) distribution, i.e. $\mathbb{P}(Z \leq z) = \exp(-\overline{H}_{\sigma,\xi}(z))$, then

$$\omega = \frac{3^\xi - 1}{2^\xi - 1} - 1.$$

By construction, the location and scale parameters of the GEV or the GP distributions do not appear in these explicit expressions. The gray dotted and solid black lines in Figure 3 correspond to the GEV and GP cases, respectively. In particular, these convex and increasing functions indicate that ω provides a "standardized" proxy of the upper tail behavior for EVT distributions. Standardized in the sense that ω is always between zero and one, and it is fully decoupled from any scale and location parameters. The GEV and GP examples have been tailored to model extremes, but our goal is to capture information from the full rainfall range. To model the full intensity of precipitation, Naveau et al. (2016) and Tencaliec et al. (2020) proposed and studied a simple scheme to construct a flexible distribution by writing

$$F(z) = G\left(H_{\sigma,\xi}(z)\right)$$

where $G(.)$ can be any cdf such that the limits of $\dfrac{G(u)}{u^\kappa}$ and $\dfrac{1 - G(1-u)}{u}$ have to be finite as u goes to zero and for some positive κ. These two constraints imply that the cdf F is in compliance with EVT for small and heavy rainfall. In practice, the choice of $G(u) = u^\kappa$ appears to be flexible enough to model most observed daily rainfall distributions while keeping parsimony at hand (see, e.g. Evin et al. 2018). It is also possible to define G nonparametrically by using a Bernstein polynomial basis (see Tencaliec et al. 2020). In this case, the ratio ω has also an explicit form expressed as

$$\omega = \frac{3B(3\kappa, 1-\xi) - B(\kappa, 1-\xi)}{2B(2\kappa, 1-\xi) - B(\kappa, 1-\xi)} - 1, \tag{4}$$

where $B(\cdot, \cdot)$ denotes the beta function. The right panel of Figure 3 focuses on values of $\xi \in (0, 1)$, the classical range for hourly and daily rainfall. The effect of κ, see the different colors, appears to be minor for positive ξ and the increase in ω with respect to ξ remains for any $\kappa \in \{.5, .9, 1, 1.3, 1.6\}$. This indicates that a choice based on ω, even if the distribution is not an exact GP distribution, will be robust with respect to ξ, the main driver of the upper tail behavior.

b. Inference of the ratio ω

In the i.i.d. case, Eq. (3) tells us how to get ω from the three PWMs α_0, α_1 and α_2. A natural estimator of ω follows

$$\hat{\omega} = \frac{3\hat{\alpha}_2 - \hat{\alpha}_0}{2\hat{\alpha}_1 - \hat{\alpha}_0} - 1, \tag{5}$$

where the PWM α_j is estimated by a linear combination of order statistics (see e.g. David and Nagaraja 2004) defined as

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i}{n}\right)^j Z_{(i)} \tag{6}$$

with $Z_{(i)}$ the i-th smallest value of the sample $(Z_1, \ldots, Z_n)^t$. Asymptotic properties of these linear combination of order statistics ($\mathcal{L}$-statistics) were studied, among others, by Guillou et al. (2009), see their Theorem 1. The i.i.d. assumption can be replaced by the weaker hypothesis defined by Eq. (12). The asymptotic convergence of our estimator towards a Gaussian distribution will be still valid, but the confidence intervals will differ. Classical delta-method (e.g., see Oehlert 1992) arguments provide convergence results on the ω estimator, see Appendix E for details. Simulations about the inference of ω for different sample sizes are available upon request. In a nutshell, they indicate that accuracy increases with sample size and, for simulations mimicking our Swiss rainfall dataset, the estimation of ω appears reasonable. It does not depend much on κ, the shape parameter ξ being the most relevant one in terms of mean square errors. A value of ξ close to .5 touches the limit of our inference scheme. These inferential conclusions are classical and in compliance with other studies of PWM's estimation (see, e.g. Hosking and Wallis 2005; Carreau and Bengio 2009; Naveau et al. 2016; Tencaliec et al. 2020).

c. Clustering method

Several clustering methods are available in the statistical literature (e.g., see MacQueen et al. 1967; Kaufman and Rousseeuw 1990; Jain et al. 1999; Saxena et al. 2017) with two major types: partitioning and hierarchical. All clustering algorithms need a common ingredient, a dissimilarity measure (e.g., see Saunders 2018). In Subsection 2a, we highlighted that ω can be interpreted as a standardized ratio of two $\mathcal{L}^1$-distances. As such, comparing two values of ω estimated at two different sites provides a simple dissimilarity measure. If the difference between two values of ω is close to zero, it means that the two locations have similar (up to a rescaling constant) marginal distributions, especially in the upper tail, see Figure 3. To keep working with $\mathcal{L}^1$-metrics, the Manhattan distance, i.e. $|\omega_i - \omega_j|$, was used to obtain our dissimilarity matrix (see also Bernard et al. 2013; Bador et al. 2015). The choice of building our clustering dissimilarity on ω is in compliance with our two original desiderata: the RFA constraint and the avoidance of selecting any

covariates such as geographical coordinates, altitudes, distance to sea and so on.

In this work, we focus on a partitioning technique called `PAM` for Partitioning Around Medoid, also called `k-medoids`, that was proposed by Kaufman and Rousseeuw (1990). Its goal is similar to the well-known `k-means` algorithm that returns a partition of the data-set into k clusters. The `PAM` algorithm can be favored over `k-means` with respect to the following aspects: robustness to outliers, determinism, computational cost and ease-of-interpretation. As for `k-means`, the user has to provide the number of clusters, k, beforehand. Centers called medoids, are just a subset of the original data points, so they are easy to interpret. Without the step of recomputing "averaged" centers at each iteration, the only input to run `PAM` is the pairwise dissimilarity matrix that has to be computed only once.

To determine the optimal number of clusters, Rousseeuw (1987) introduced the *silhouette coefficient* that measures tightness of clusters and dissociation between the clusters. The partition with the highest silhouette criterion is the best partition. As k increases, the number of points in each cluster decreases and consequently the uncertainty within each class increases. This leads to a classical trade-off between bias and variance. Another criterion to determine the optimal number of clusters is based on cluster inertia. Best partitions are the ones with lowest relative intracluster inertia. For more details on these existing clustering tools, one can refer to Appendix F.

To assess the efficiency of silhouette coefficient and inertia criterion for our application, we need to define a simulation study mimicking Swiss rainfall intensities.

3. Simulation study

a. Numerical design setup

Eq. (4) provides an explicit expression of ω for a specific EGDP with $G(u) = u^\kappa$ that appears to well capture low, moderate and heavy rainfall intensities (see, e.g. Evin et al. 2018). It is also easy to simulate random samples from this distribution. For these reasons, our simulation study is based on the EGPD(κ, σ, ξ) with the following design.

In Figure 4, the 10 colors represent the different values of ω for 10 pairs of (κ, ξ)'s from $\xi \in \{.0, .1, .2, .3, .4\}$ and $\kappa \in \{.5, .9, 1.3, 1.6\}$. We consider these 10 rectangles as 10 homogeneous regions, the size of the rectangles being proportional to the number of points in each region; 40 in large rectangles and 20 for small ones. Each region is associated with its own 99% return level (to simplify interpretation, the return levels are calculated after renormalizing by the mean, see Appendix H for explicit form), see values between brackets in the color legend. By construction, theoretical values ω do not change with σ. Therefore, ω highlights variations of

both κ and ξ. For a parameter setup where only ξ varies, the homogeneous regions (condition (1)) would match with the regions based on the tail-homogeneity (condition (2)). There would be only 5 homogeneous clusters. However, as mentioned in the introduction, homogeneity of the full distribution (condition (1)) *implies* tail-homogeneity but the converse is not true. In our experimental design, κ varies independently of ξ. Therefore, the experimental setup of Figure 4 has 10 homogeneous regions, i.e. 10 regions satisfying condition (1).

Hence, one question that we want to explore with our design setup is to determine, for finite samples, the robustness of our clustering approach according to the number of clusters.

To mimic the setup of hourly rainfall studies over Europe, we consider 128 wet days in a given year and the number of years varies from 30 to 150. This leads to samples of sizes $128 \times \{30, 50, 80, 150\}$. This step is replicated 100 times. For each of these 100 experiments, the estimate $\hat{\omega}$ is calculated at each location in Figure 4. Then, we can apply the clustering algorithm presented in Section 2c.

b. Sensitivity to the number of clusters

We examine the ability of our clustering algorithm to detect the homogeneous regions of Figure 4. For this simulation setup parametrization, the silhouette coefficient tends to underestimate the number of clusters, we therefore consider this case. In Figure 5, the top and bottom panels compare the `PAM` clustering misspecification rates between two arbitrary pre-determined number of clusters, $k = 10$ and $k = 8$ for the top and bottom panels, respectively. As the true number of clusters is $k = 10$ in our experimental design, see Figure 4, one may be puzzled as misspecification rates in panel (b) of Figure 5 with $k = 8$ appear to be inferior to the rates in panel (a). This result can be explained if we notice that the value of ω in Figure 4 for $\kappa = 1.6$ and $\xi = .2$ (dark green) is .69 and very close to the one obtained with $\kappa = .9$ and $\xi = .1$ (pink), precisely $\omega = .70$. The same can be said between the two setups of $\kappa = 1.3$ and $\xi = .3$ ($\omega = .73$) and $\kappa = .9$ and $\xi = .2$ ($\omega = .72$). To identify small differences in ω like .01, the number of years has to be large. If it is not the case, then the `PAM` algorithm provides stable clusters by joining similar clusters that are undistinguishable with respect to moderate sample sizes. By appropriately joining clusters, misspecification rates then get smaller. Note also that this phenomenon is linked to the rectangle sizes, i.e. the number of locations. The difference of .01 between $\omega = .66$ (dark blue) and $\omega = .67$ (light green) in Figure 4 is of second order because the dark blue and light green rectangles have 40 points each, while the pink, red and light orange rectangles have only 20 locations each. Overall, the `PAM` algorithm appears conservative in terms of the number of clusters. So, the risk of creating artificial clusters is low and a second step may be needed to fine-tune the distribution within each cluster that may combine two clusters, see c. in the next

section. In addition, panel (b) indicates that the misspecification rate is below 5% for 80 years long time series. This temporal length corresponds to our application setup.

4. Regional analysis of Swiss daily precipitation

MeteoSwiss network includes 666 rainfall stations providing daily values from 1930 to 2014. In our study we only consider stations with less of 10 % missing data leading to 191 data series. At each site, we focus on strictly positive precipitation and remove dry events. This leads to years with, in average, around 128 wet days.

a. Number of clusters

After computing the ratio ω using Eq. (5) at each location, a number of clusters has to be chosen when applying PAM. To take into account rainfall variability, we randomly shuffle our whole dataset in space and time. This step should remove all spatial clustering and, hence under the null hypothesis of the absence of clusters, a base level for silhouette coefficients, see Eq. (16), can be obtained and gives us a yardstick. In Figure 6, solid black points indicate the difference between the base level and the level obtained without reshuffling for different cluster numbers, say $k = 1, \ldots, 20$. We suggest to consider the number of clusters for which the discrepancy between partition of white noise and partition of real data is the most significant. A partition with two clusters clearly appears as the optimal choice. To double-check this optimization, we apply the same procedure with the inertia ratio (with the exception that we consider the highest value of the difference between inertia on shuffled data and inertia on observed data), see Eq. (17). The gray diamonds confirm the value of $k = 2$ clusters. Both criteria lead to an optimal value of clusters equal to two and we can visually check if the two clusters provide spatially rainfall patterns. Panel (a) in Figure 7 displays the `PAM` clustered sites with $k = 2$. The first cluster are represented by diamond shapes, while the second cluster corresponds to circles. The size of the points are proportional to their silhouette coefficient: largest points are the best classified ones. Although the geographical covariates have not been used, this division of Switzerland reveals a spatially coherent structure. Our so-called low altitude area cluster (diamonds) mainly covers the plain, the Jura mountains and a few points at the south tip of the Ticino canton. Our so-called alpine cluster (circles) appears to gather sites in the south. Climatologically, the dichotomy appears reasonable with heavier rainfall in the alpine class than in the low altitude area one.

b. Comparison with the RFA approach

A natural question is to wonder if the classical RFA analysis (Hosking and Wallis 1993) would have given the same type of clustering. To apply the RFA approach, covariates have to be given *a priori*. Since orography can largely influence precipitation (e.g., see Gottardi et al. 2012), we consider normalized elevation and coordinates as covariates. They are then clustered by `PAM` algorithm, see Figure 10 in Appendix I. To compare the RFA outputs with our approach, we compute the relative intra-cluster inertia, silhouette criterion and homogeneity tests.

Table 1 shows silhouette criteria and relative inertia for the classical RFA and our clustering approach based on $\hat{\omega}$, see Eq. (5). Clearly, the classical RFA that has a very low mean silhouette coefficient and high inertia, two undesirable features, is outperformed by our method.

To assess homogeneity within each of the two clusters, we compute the three RFA homogeneity tests (see Hosking and Wallis 2005, for the mathematical definition of these tests). We recall that, asymptotically, these three RFA tests reject homogeneity when they are far from a zero-mean Gaussian distribution (the variance depends on the sample size). They are also based on the assumption that rainfall follow a kappa distribution. However, there is no guarantee that a kappa distribution correctly fits the dataset at hand. This may explain why the three tests strongly reject the homogeneity hypothesis when applied to the two RFA clusters. Table 2 tells us that, even small sub-regions with 10% of locations have difficulty to be considered as homogeneous. For example, the second row indicates that the three RFA tests are far from zero, pointing out that the RFA northern cluster is strongly heterogeneous. In contrast, Table 3 shows that our clustering approach provides larger and more homogeneous regions, at least 30% of locations can be kept. Still, this leads us to revisiting the strict definition of homogeneity via Eq. (1). In our Introduction section, the tail condition defined by Eq. (2) offers a less stringent way to define homogeneity. More generally, it is of interest to compare different parametric models within a given region under the tail constraint (2). The next section compares three EGPD-based models for our Swiss data. Each one can be viewed as a different level of flexibly within a regional frequency analysis, in particular within each of our alpine and low altitude area clusters.

c. Local, semi-regional and regional EGPD models

Given a cluster set $\mathcal{C}$, our so-called "local" model is the most flexible and allows variability in each of the EGPD parameters

$$Y(s) \sim \text{EGPD}(\kappa(s), \sigma(s), \xi(s)), \quad s \in \mathcal{C}. \tag{7}$$

The "regional" model is the most stringent one and it is defined by

$$Y(s) \sim \text{EGPD}(\kappa_C, \sigma(s), \xi_C), \quad s \in \mathcal{C}. \tag{8}$$

Between these two cases, the "semi-regional" model consists of regionalizing only the shape parameter ξ and letting the scale σ and flexibility κ parameters vary

$$Y(s) \sim \text{EGPD}(\kappa(s), \sigma(s), \xi_{\mathcal{C}}), \quad s \in \mathcal{C}. \tag{9}$$

Models (8) and (9), but not Model (7), satisfy Eq. (2), and only (8) satisfies Eq. (1). The key aspect is to avoid both overfitting, say with Model (7), and oversimplified models that may not well capture local extremes.

In terms of inference, the fitting of Model (7) is obtained by using the *mev* package in R. Concerning the parameters of Model (9), they are estimated with the following PWM based algorithm. The PWM of order one for EGPD in Appendix B in Naveau et al. (2016) provides the key ingredient of our algorithm, see steps 8 and 10 below. Note that Algorithm 1 can be adapted to the regional version defined by Eq. 8. This inferential procedure performs well on simulated data, results available upon request.

Concerning the fit of our regional model to Swiss rainfall, the .99 return levels of panel (a) in Figure 7 reproduce the expected climatological Helvetic features where the Ticino canton presents the highest return values. Note that within a same cluster non-normalized return levels are allowed to highly vary in space. In a homogeneous cluster, only normalized return-levels (i.e. return levels of $Y/\mathbb{E}Y$) are common. This spatial pattern is captured by the scale parameter σ displayed in panel (b) of Figure 7. By construction, the parameters κ and ξ do not change within each cluster.

But they vary from clusters to clusters. More precisely, κ is higher in the low altitude area cluster (1.08) than in the alpine one (.6). As climatologically expected, ξ is higher in the alpine cluster (.17) than in the low altitude area one (.03).

To improve our understanding of the difference between our alpine and low altitude area clusters, we recall that EGPD p-return level of $\dfrac{Y}{\mathbb{E}[Y]}$ is given by

$$y_p = \frac{1}{\kappa B(\kappa, 1-\xi) - 1} \left[\left(1 - p^{1/\kappa} \right)^{-\xi} - 1 \right].$$

See proof in Appendix H. Applying this formula for Model (9), Figure 8 displays $y_{.99}$ and its associated .95 confidence intervals. The confidence intervals are obtained by bootstrapping. We remove the autocorrelation by extracting randomly a third of precipitation observation for each time series. The number of bootstrap replicates is chosen equal to 300. The dichotomy between the two clusters is clearly confirmed. But it is not clear if letting κ free clearly improves the fit. Hence, a remaining question is to determine if the

Algorithm 1 Semi-regional fit of Model (8) in cluster C

$cond = TRUE, eps = .001,$ and $u = 1$(mm)

procedure INPUT(Rainfall Matrix for cluster C)

Remove dry days by only taking $\{Y(s)|Y(s) > u\}$

Fit locally Model (7) at each location $s \in C$

Denote κ_0 and ξ_0 the cluster means of κ and ξ from Step 4

Compute $m(s)$ the sample mean at each $s \in C$

while $cond = TRUE$ **do**

Compute

$$\sigma_{new}(s) = \frac{\xi_0 \; m(s)}{\frac{\kappa_0}{F(u)} IB\left(H_{\xi_0}\left(\frac{u}{\sigma_0}\right), 1, \kappa_0, 1-\xi_0\right) - 1}$$

where $IB(.,.,.)$ is the incomplete Beta function

Compute mn the cluster mean over all $\frac{Y(s)}{\sigma_{new}(s)}$ ▷ The cdf $\frac{Y(s)}{\sigma(s)}$ does not depend on s in Model (8)

Calculate

$$\kappa_{new} = \frac{\xi_0 \; mn}{\frac{1}{F(u)} IB\left(H_{\xi}\left(\frac{u}{\sigma_{new}}\right), 1, \kappa_0, 1-\xi_0\right) - \frac{1}{\kappa_0}}$$

if $\max(|\kappa_{new} - \kappa_0|, |\sigma_{new} - \sigma_0|) < eps$ **then**

$cond = FALSE$

end if

$\kappa_0 \leftarrow \kappa_{new}$ and $\sigma_0 \leftarrow \sigma_{new}$

end while

Return $(\kappa_0, \sigma_0, \xi_0)^T$

end procedure

semi-regional and the regional models are truly different. In terms of the classical Akaike Information Criterion (Akaike 1987), the regional model, Model (8), has a lower value than the semi-regional one: 146 168 versus 146 532. Hence, from a parsimony perspective, it seems better to regionalize not only the shape parameter ξ but also the parameter κ, the spatial component being captured by $\sigma(s)$. To confirm this statement, Figure 9 compares the quantiles of models (9) and (8) with the local one, Model (7), (x-axis) at three different locations. The left and right columns correspond the two stations with lowest and highest silhouette coefficients, while the center column represents the cluster medoid. The first row corresponds to the alpine cluster and the second one to the low altitude area cluster. As pointed by the Aikaike criterion, the regional and semi-regional models provide similar quantiles. Concerning the comparison with the local model, one has to keep in mind that Model (7) has 191×3 different parameters compared to $2 + 191 \times 2$ parameters for Model (9) and $2 \times 2 + 191$ parameters for Model (8). Clearly, the second and third columns indicate that the strong reduction of parameters from Model (7) to Model (8) has a low impact for medoids and well classified stations. Concerning the alpine and low altitude area stations with the lowest silhouette coefficients, the first column indicates a departure for a few extreme values. This is certainly due to the estimation of ξ that is always difficult to estimate locally.

5. Conclusion

Our main goal in this work was to show that a simple and fast clustering approach based on an interpretable ratio could highlight climatologically coherent regions. This clustering algorithm[1] works through two steps: i) pointwise order statistic based estimation of the PWM ratio ω, see Eq. (3) and Eq. (6), on positive daily precipitation and ii) clustering of ω estimates with Manhattan distance in `PAM` algorithm.

One advantage is that this method is fully data driven and avoid the need of finding relevant covariates. The proposed approach was built on the main RFA idea, i.e. a normalizing factor that can capture well the spatial component in rainfall data. More specifically, by construction, ω is constant across an homogeneous region. All the inferential part was done by using probability weighted moments, simple quantities to estimate and interpret. We completely bypassed the delicate threshold selection step to define heavy rainfall by fitting the extended Pareto distribution.

Our analysis of Swiss daily precipitation data reveals an interesting point concerning model complexity. A relatively simple regional model with only two clusters and a spatially varying scale parameter can compete very well against complex models with various varying parameters. This highlights the strong variability of rainfall data and goes against the idea that complex marginal models have to be fitted. Still, one has

[1] Code is available in the GitHub repository `https://github.com/PhilomeneLeGall/RFA_regional_EGPDk.git`

to keep in mind that we do not model the spatial dependence, but only marginal behaviors. Our proposed approach is useful to infer at-site return levels, but irrelevant to infer ungauged locations. In addition, our data-driven `PAM` clustering algorithm also did not take into account the dependence between sites. One interesting perspective will be to combine our approach with the work of Saunders (2018), Bador et al. (2015) and Bernard et al. (2013) who only partitioned with respect to the spatial dependence, but not the marginal behaviors.

Acknowledgments

Within the CDP-Trajectories framework, this work is supported by the French National Research Agency in the framework of the " Investissements d'avenir" program (ANR-15-IDEX-02). We gratefully acknowledge financial support for this study provided by the Swiss Federal Office for Environment (FOEN), the Swiss Federal Nuclear Safety Inspectorate (ENSI), the Federal Office for Civil Protection (FOCP), and the Federal Office of Meteorology and Climatology, MeteoSwiss, through the project EXAR ("Evaluation of extreme Flooding Events within the Aare-Rhine hydrological system in Switzerland"). Part of this work was also supported by the French national programs (FRAISE-LEFE/INSU and 80 PRIME CNRS-INSU), and the European H2020 XAIDA (Grant agreement ID: 101003469). Philippe Naveau also acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project), and the ANR-Melody (ANR-19-CE46-0011) .

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. In *Selected papers of Hirotugu Akaike*, pages 371–386. Springer.

Asadi, P., Engelke, S., and Davison, A. C. (2018). Optimal regionalization of extreme value distributions for flood estimation. *Journal of Hydrology*, 556:182–193.

Bador, M., Naveau, P., Gilleland, E., Castellà, M., and Arivelo, T. (2015). Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe. *Weather and climate extremes*, 9:17–24.

Barton, Y., Giannakaki, P., Von Waldow, H., Chevalier, C., Pfahl, S., and Martius, O. (2016). Clustering of regional-scale extreme precipitation events in southern Switzerland. *Monthly Weather Review*, 144(1):347–369.

Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in France. *Journal of Climate*, 26(20):7929–7937.

Boucefiane, A. and Meddi, M. (2019). Regional growth curves and extreme precipitation events estimation in the steppe area of northwestern Algeria. *Atmósfera*, 32(4):287–303.

Burn, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research*, 26(10):2257–2265.

Carreau, J. and Bengio, Y. (2009). A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12(1):53–76.

Carreau, J., Naveau, P., and Neppel, L. (2017). Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation. *Water Resources Research*, 53(5):4407–4426.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of The American Statistical Association*, 102:824–840.

David, H. A. and Nagaraja, H. N. (2004). *Order statistics*. John Wiley & Sons.

Davison, A. C. and Huser, R. (2015). Statistics of extremes. *Annual Review of Statistics and its Application*, 2:203–235.

de Fondeville, R. and Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika*, 105(3):575–592.

Diebolt, J., Guillou, A., Naveau, P., and Ribereau, P. (2008). Improving probability-weighted moment methods for the generalized extreme value distribution. *REVSTAT-Statistical Journal*, 6(1):33–50.

Evin, G., Blanchet, J., Paquet, E., Garavaglia, F., and Penot, D. (2016). A regional model for extreme rainfall based on weather patterns subsampling. *Journal of Hydrology*, 541:1185–1198.

Evin, G., Favre, A.-C., and Hingray, B. (2018). Stochastic generation of multi-site daily precipitation focusing on extreme events. *Hydrology and Earth System Sciences*, 22(1):655–672.

Fawad, M., Ahmad, I., Nadeem, F. A., Yan, T., and Abbas, A. (2018). Estimation of wind speed using regional frequency analysis based on linear-moments. *International Journal of Climatology*, 38(12):4431–4444.

Ferreira, A. and de Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, 43(1):276–298.

Fougères, A.-L. (2004). Multivariate extremes. *Monographs on Statistics and Applied Probability*, 99:373–388.

Gottardi, F., Obled, C., Gailhard, J., and Paquet, E. (2012). Statistical reanalysis of precipitation fields based on ground network data and weather patterns: Application over French mountains. *Journal of Hydrology*, 432:154–167.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water resources research*, 15(5):1049–1054.

Guillou, A., Naveau, P., Diebolt, J., and Ribereau, P. (2009). Return level bounds for discrete and continuous random variables. *Test*, 18(3):584.

Hilker, N., Badoux, A., and Hegg, C. (2009). The Swiss flood and landslide damage database 1972-2007. *Natural Hazards and Earth System Sciences*, 9(3):913.

Hosking, J. and Wallis, J. (1993). Some statistics useful in regional frequency analysis. *Water resources research*, 29(2):271–281.

Hosking, J., Wallis, J., and Wood, E. (1985). An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. *Hydrological Sciences Journal*, 30(1):85–109.

Hosking, J. R. M. and Wallis, J. R. (2005). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data Clustering: A Review. *ACM Comput. Surv.*, 31(3):264–323.

Jalbert, J., Favre, A.-C., Bélisle, C., and Angers, J.-F. (2017). A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(5):941–962.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kojadinovic, I. and Naveau, P. (2017). Detecting distributional changes in samples of independent block maxima using probability weighted moments. *Extremes*, 20(2):417–450.

Li, D., Rao, M. B., and Tomkins, R. (2001). The law of the iterated logarithm and central limit theorem for L-statistics. *Journal of multivariate analysis*, 78(2):191–217.

Li, M., Li, X., and Ao, T. (2019). Comparative Study of Regional Frequency Analysis and Traditional At-Site Hydrological Frequency Analysis. *Water*, 11(3):486.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1-14, pages 281–297. Oakland, CA, USA.

Malekinezhad, H. and Zare-Garizi, A. (2014). Regional frequency analysis of daily rainfall extremes using L-moments approach. *Atmósfera*, 27(4):411 – 427.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Oehlert, G. W. (1992). A Note on the Delta Method. *The American Statistician*, 46(1):27–29.

Onibon, H., Ouarda, T. B., Barbet, M., St-Hilaire, A., Bobee, B., and Bruneau, P. (2004). Analyse fréquentielle régionale des précipitations journalières maximales annuelles au Québec, Canada/Regional frequency analysis of annual maximum daily precipitation in Quebec, Canada. *Hydrological sciences journal*, 49(4).

Ouarda, T., St-Hilaire, A., and Bobée, B. (2008). Synthèse des développements récents en analyse régionale des extrêmes hydrologiques. *Revue des sciences de l'eau/Journal of Water Science*, 21(2):219–232.

Ribereau, P., Guillou, A., and Naveau, P. (2008). Estimating return levels from maxima of non-stationary random sequences using the Generalized PWM method. *Nonlinear Processes in Geophysics*, 15(6):1033–1039.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Saunders, K. (2018). *An investigation of Australian rainfall using extreme value theory*. PhD thesis, Melbourne University.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A Review of Clustering Techniques and Developments. *Neurocomput.*, 267(C):664–681.

St-Hilaire, A., Ouarda, T., Lachance, M., Bobée, B., Barbet, M., and Bruneau, P. (2003). La régionalisation des précipitations: une revue bibliographique des développements récents. *Revue des sciences de l'eau/Journal of Water Science*, 16(1):27–54.

Stein, M. L. (2020). Parametric models for distributions when interest is in extremes with an application to daily temperature. *Extremes*. https://doi.org/10.1007/s10687-020-00378-z.

Tencaliec, P., Favre, A.-C., Naveau, P., Prieur, C., and Nicolet, G. (2020). Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount. *Environmetrics*, 31(2):e2582. https://doi.org/10.1002/env.2582.

Viglione, A., Laio, F., and Claps, P. (2007). A comparison of homogeneity tests for regional frequency analysis. *Water Resources Research*, 43(3). W03428, doi:10.1029/2006WR005095.

Zhang, Q., Xiao, M., Singh, V. P., and Li, J. (2012). Regionalization and spatial changing properties of droughts across the Pearl River basin, China. *Journal of Hydrology*, 472:355–366.

Zryd, A. (2008). *Les glaciers en mouvement: la population des Alpes face au changement climatique*, volume 47. Collection le savoir suisse.

Appendices

A. Probability weighted moments and L-moments

These moments were defined by Greenwood et al. (1979) as

$$\alpha_j(Z) = \mathbb{E}\left[ZF^j(Z)\right]. \tag{10}$$

where α_j denotes the PWM of order j for the random variable Z with cdf F. Extensions of PWMs have been proposed in the literature (see, e.g. Diebolt et al. 2008).

Straightforwardly from Hosking and Wallis (2005), we have:

$$\begin{aligned} \lambda_1 &= \alpha_0, \\ \lambda_2 &= \alpha_0 - 2\alpha_1, \\ \text{and} \quad \lambda_3 &= \alpha_0 - 6\alpha_1 + 6\alpha_2. \end{aligned}$$

B. PWM and dependence index

To make the link between PWMs and $\mathcal{L}^1$-distances, one can show that

$$\frac{1}{2}\mathbb{E}\left|Z_1 - Z_2\right| = \theta_{1:2} \cdot \alpha_{\theta_{1:2}-1} - \alpha_0, \tag{11}$$

whenever $\mathbb{P}\left[\max\left(Z_1, Z_2\right) \leq z\right)] = F^{\theta_{1:2}}(z)$ with the scalar $\theta_{1:2}$ representing a dependence index between Z_1 and Z_2. More generally, suppose that the multivariate vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)^t$ satisfies

$$\mathbb{P}\left[\max \mathbf{Z} \leq z\right] = F^{\theta_{1:k}}(z). \tag{12}$$

for some scalar $\theta_{1:k} \in [0, k]$. Note that this scalar can be interpreted as a well known measure of dependence in maxima and it is linked to temporal clustering in EVT and the concept of max-stability (e.g., see Davison and Huser 2015). Then one can see that

$$\mathbb{E}\left[\max \mathbf{Z}\right] = \theta_{1:k} \cdot \alpha_{\theta_{1:k}-1}(Z).$$

It follows that any affine transformation of the $\mathbf{Z}$ satisfies

$$\alpha_{\theta_{1:k}-1}(a + b\mathbf{Z}) = \frac{a}{\theta_{1:k}} + b \cdot \alpha_{\theta_{1:k}-1}(\mathbf{Z}), \text{ for any } b > 0.$$

C. Bounds of ω

It is obvious that ω is non-negative. So, we just need to show that it is smaller than one. The difference $|\max(z_1, z_2) - \max(z_1, z_3)|$ can only take two types of expressions, see Table below.

6 cases	$\lvert\max(z_1, z_2) - \max(z_1, z_3)\rvert$	Upper-bound
$z_1 \leq z_2 \leq z_3$	$\lvert z_2 - z_3\rvert$	$\lvert z_2 - z_3\rvert$
$z_1 \leq z_3 \leq z_2$	$\lvert z_2 - z_3\rvert$	$\lvert z_2 - z_3\rvert$
$z_2 \leq z_1 \leq z_3$	$\lvert z_1 - z_3\rvert$	$\lvert z_2 - z_3\rvert$
$z_2 \leq z_3 \leq z_1$	0	$\lvert z_2 - z_3\rvert$
$z_3 \leq z_1 \leq z_2$	$\lvert z_1 - z_3\rvert$	$\lvert z_2 - z_3\rvert$
$z_3 \leq z_2 \leq z_1$	0	$\lvert z_2 - z_3\rvert$

It is either equal to zero, when $z_1 = \max(z_1, z_2, z_3)$, or to $|z_i - z_j|$ for some $i \neq j$. In all cases, it is upper bounded by $|z_3 - z_2|$. So,

$$|\max(z_1, z_2) - \max(z_1, z_3)| \leq |z_3 - z_2|$$

and

$$\mathbb{E}\,|\max(Z_1, Z_2) - \max(Z_1, Z_3)| \leq \mathbb{E}\,|Z_3 - Z_2|\,.$$

If $\mathbb{E}\,|Z_1 - Z_2| = \mathbb{E}\,|Z_2 - Z_3|$ (e.g. if Z_i is stationary then $Z_3 - Z_2 \overset{d}{=} Z_2 - Z_1$ and this equality holds), then

$$\mathbb{E}\,|\max(Z_1, Z_2) - \max(Z_1, Z_3)| \leq \mathbb{E}\,|Z_1 - Z_2|\,.$$

It follows that $\omega \leq 1$. □

D. Expressing ω as a ratio of differences between maxima and means

By definition, $\omega = \frac{\mathbb{E}|\max(Z_1,Z_2)-\max(Z_1,Z_3)|}{\mathbb{E}|Z_1-Z_2|}$. Let's relate it to the ratio expressed in Kojadinovic and Naveau (2017)

$$\frac{\mathbb{E}\left[\max(Z_1, Z_2, Z_3) - (Z_1, Z_2, Z_3)/3\right]}{\mathbb{E}\left[\max(Z_1, Z_2) - (Z_1, Z_2)/2\right]}.$$

Denote

$$\begin{aligned} d_{i\vee j, i\vee k} &= \frac{1}{2}\mathbb{E}\,|\max(Z_i, Z_j) - \max(Z_i, Z_k)|\,, \\ \text{and } d_{i,j} &= \frac{1}{2}\mathbb{E}\,|Z_i - Z_j|\,. \end{aligned}$$

The $\mathcal{L}1$-distance between $\max(Z_1, Z_2)$ and $\max(Z_1, Z_3)$ is equal to

$$\begin{aligned} d_{1\vee2,1\vee3} &= \mathbb{E}\left\{\max\left(Z_1, Z_2, Z_3\right)\right. \\ &\quad \left. -\frac{1}{2}\left[\max(Z_1, Z_2) + \max(Z_1, Z_3)\right]\right\}, \end{aligned}$$

with $\mathbb{E}\left[\frac{1}{2}\left(\max(Z_i, Z_j)\right]\right. = \frac{1}{2}d_{i,j} + \frac{1}{4}\mathbb{E}\left[Z_i + Z_j\right]$. It follows that $d_{1\vee2,1\vee3}$ is equal to

$$\begin{aligned} &\mathbb{E}\left[\max\left(Z_1, Z_2, Z_3\right)\right] - \frac{1}{2}\left(d_{1,2} + d_{1,3}\right) \\ &\qquad - \frac{1}{4}\left(2\mathbb{E}Z_1 + \mathbb{E}Z_2 + \mathbb{E}Z_3\right). \end{aligned} \tag{13}$$

Hence,

$$\begin{aligned} d_{1\vee2,1\vee3} + d_{1\vee2,2\vee3} + d_{1\vee3,2\vee3} &= 3\mathbb{E}\left[\max\left(Z_1, Z_2, Z_3\right)\right] \\ &\quad - \left(d_{1,2} + d_{1,3} + d_{2,3}\right) \\ &\quad - \left(\mathbb{E}Z_1 + \mathbb{E}Z_2 + \mathbb{E}Z_3\right). \end{aligned}$$

Finally, we can write that

$$\begin{aligned} &\frac{\mathbb{E}\left[\max\left(Z_1, Z_2, Z_3\right) - \left(Z_1 + Z_2 + Z_3\right)/3\right]}{\left(d_{1,2} + d_{1,3} + d_{2,3}\right)/3} \\ = \quad &\frac{d_{1\vee2,1\vee3} + d_{1\vee2,2\vee3} + d_{1\vee3,2\vee3}}{d_{1,2} + d_{1,3} + d_{2,3}} + 1. \end{aligned}$$

If the increments $(Z_i - Z_j)$ are stationary, then

$d_{1,2} = d_{1,3} = d_{2,3}$. From Eq. (13), it follows that

$d_{1\vee2,1\vee3} = d_{1\vee2,2\vee3} = d_{1\vee3,2\vee3}$, then

$$\begin{aligned} 1 + \omega &= \frac{\mathbb{E}\left[\max\left(Z_1, Z_2, Z_3\right) - \left(Z_1 + Z_2 + Z_3\right)/3\right]}{\mathbb{E}\left[\max\left(Z_1, Z_2\right) - \left(Z_1 + Z_2\right)/2\right]} \\ &= \frac{\mathbb{E}\left|\max\left(Z_1, Z_2\right) - \max\left(Z_1, Z_3\right)\right|}{\mathbb{E}\left|Z_1 - Z_2\right|} + 1. \end{aligned}$$

□

By construction, the ratio ω has also a clear connection with PWMs and can be rewritten, under assumption (12), as

$$\omega = \frac{\theta_{1:3} \cdot \alpha_{\theta_{1:3}-1} - \alpha_0}{\theta_{1:2} \cdot \alpha_{\theta_{1:2}-1} - \alpha_0} - 1.$$

E. Convergence of PWM and ω estimators

The convergence of PWMs estimators, $\hat{\alpha}_i, i = 1, 2, 3$ is insured by Theorem 2.1 of Li et al. (2001), with $H = u, G = F^{-1}$ and $J = v$, under appropriate conditions on Y. □

Following proposition provides more details.

Proposition 1 *Let the random variable Y with c.d.f F s.t $\mathbb{E}Y^2$ is finite. The PWMs , α_k, of order $k = 0, 1, 2$ and their estimators, $\hat{\alpha}_k$, satisfy*

$$\sqrt{n}(\hat{\alpha}_k - \alpha_j) \xrightarrow{d} \mathcal{N}(0, \sigma_k^2), \quad k = 0, 1, 2 \tag{14}$$

with
$\sigma_k^2 = \mathbb{E}\left[-U^k F^{-1}(U) + \alpha_j - k\int_0^1 (\mathbb{1}(U \leq t) - t)t^{k-1}F^{-1}(t)\mathrm{dt}\right]^2$ *where* $U \sim \mathcal{U}\ (0, 1)$.

Classical delta-method (e.g., see Oehlert 1992) arguments lead to the convergence of $\hat{\omega}$.

Proposition 2 *Let Σ the covariance matrix of the PWM vector $(\alpha_0, \alpha_1, \alpha_2)$. We have*

$$\sqrt{n}\left(\hat{\omega} - \omega\right) \xrightarrow{d} \mathcal{N}\left(0, D(\alpha_0, \alpha_1, \alpha_2)\Sigma^t D(\alpha_0, \alpha_1, \alpha_2)\right),$$

where D is the Jacobian matrix of the trivariate function defined as $(x, y, z) \longmapsto \dfrac{3z - x}{2y - x}$.

We apply delta-method to the PWM estimators. Indeed, proving the convergence of $\hat{\omega}$ is equivalent to prove that $\sqrt{n}\left[g\left(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2\right) - g\left(\alpha_0, \alpha_1, \alpha_2\right)\right]$ converges in distribution to $\mathcal{N}\left(0, D\left(\alpha_0, \alpha_1, \alpha_2\right)^T \Sigma D\left(\alpha_0, \alpha_1, \alpha_2\right)\right)$.
where $g : \mathbb{R}^3 \setminus \mathcal{P} \longrightarrow \mathbb{R}$ where $\mathcal{P} : 2y - x = 0$ is defined by $g(x, y, z) = \dfrac{3z - x}{2y - x}$.

The random vector $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ converges to $(\alpha_0, \alpha_1, \alpha_2)$ with covariance matrix written Σ. In addition, the function g is differentiable in $(\alpha_0, \alpha_1, \alpha_2)$ with Jacobian matrix equal to $D(x, y, z) = \left(\dfrac{3z - 2y}{(2y - x)^2}, \dfrac{2x - 6z}{(2y - x)^2}, \dfrac{3}{2y - x}\right)$.
Eventually, the delta method ensures convergence in distribution of $\hat{\omega}$. □

F. `PAM` algorithm

To summarize, the practitioner has to provide a number of clusters k and a matrix containing all the pairwise dissimilarities, say $D = [d_{i,j}]$, where $d_{i,j}$ represents the dissimilarity between ω_i and ω_j of the weather stations i and j. Each non-medoid point, say j, of the data-set is associated to its closest medoid, i.e. it minimizes $\min_{m_1,\ldots,m_k} d_{j,m}$ where the k medoids set is denoted $\{m_1, \ldots, m_k\}$. The overall, `PAM` criterion is to find the group of medoids that minimizes the total cost

$$\sum_j \min_{m_1,\ldots,m_k} d_{j,m} \tag{15}$$

To solve this optimization problem, the first medoid is the solution to Eq. (15) with $k = 1$, that is to say the most centrally located point. The second medoid is the solution with $k = 2$ but with the first medoid fixed (to the one previously found). Still, every swap possible between a medoid and any point non-medoid is

tested. If the cost function decreases, then the swap is kept and the algorithm stops when no swap improves the total cost of the partition.

Given k the number of clusters, the silhouette coefficient for site i that belongs to the cluster j is defined as

$$s_i(k) = 1 - \left(\frac{d_{ij}}{\delta_{i,-j}}\right) \tag{16}$$

where $\delta_{i,-j}$ the smallest of the $j-1$ average distance between site i and all other sites associated with a cluster different from j. If $s_i(k) \approx 1$, station i is well classified since the intra-cluster distance is significantly smaller than the distance between clusters. On the contrary, if $s_i(k) \approx -1$, the station i should be in another cluster. If $s_i(k) \approx 0$, the point is as close to points in the medoid as to other points. The overall quality of the partitioning in k clusters is assessed by computing the average silhouette coefficient over all sites.

Intra-cluster inertia is defined as

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - G_i\|^2, \tag{17}$$

where C_i corresponds to the i^{th} cluster and G_i is the associated medoid. The tightness of clusters is obtained by computing the ratio of intra-cluster inertia over the total inertia.

G. PWM of order $0, 1$ and 2 for $EGPD(\kappa, \sigma, \xi = 0)$

The PWM of order 0,1 and 2, as defined in Eq. (10), for an EGPD with null shape parameter are given by:

$$\begin{aligned}
\alpha_0 &= \sigma\kappa\Gamma(\kappa), \\
\alpha_1 &= \sigma\kappa\Gamma(\kappa)\left[1 - 2^{-\kappa-1}\right], \\
\alpha_2 &= \sigma\kappa\Gamma(\kappa)\left[1 - 2^{-\kappa} + 3^{-\kappa-1}\right].
\end{aligned}$$

H. Normalized return level values

Let $Y \sim EGPD(\kappa, \sigma, \xi)$. By Naveau et al. (2016), expectation of Y (non-censored PWM of order 0) is

$$\mathbb{E}Y = \frac{\sigma}{\xi}(\kappa B(\kappa, 1-\xi) - 1).$$

Yet, $\dfrac{Y}{\mathbb{E}Y} \sim EGPD(\kappa, \sigma_N, \xi)$ and $\mathbb{E}\left(\dfrac{Y}{\mathbb{E}Y}\right) = 1$.
Hence, σ_N satisfies $\dfrac{\sigma_N}{\xi}(\kappa B(\kappa, 1-\xi) - 1) = 1$. Thus,

$$\sigma_N = \frac{\xi}{\kappa B(\kappa, 1-\xi) - 1} \tag{18}$$

Eventually, when $\xi > 0$, the p-return level of $\frac{Y}{\mathbb{E}Y}$ only depends on flexibility parameter κ and shape parameter ξ. More precisely, the p-return level is given by

$$y_p = \frac{1}{\kappa B(\kappa, 1-\xi) - 1} \left[\left(1 - p^{1/\kappa} \right)^{-\xi} - 1 \right] \tag{19}$$

□

I. Maps of clusters (traditional RFA and our data-driven algorithm)

List of Tables

PAM clustering approach type	Mean silhouette coefficient	intra-cluster inertia ratio
classical RFA	.05	.96
with $\hat{\omega}$, see Eq. (5)	.69	.32

Table 1: Swiss daily precipitation. A higher mean silhouette criterion and a lower intra-cluster inertia ratio indicate a better clustering performance.

Percentages of sites	Cluster names	Three RFA homogeneity tests		
10% (.63)	northern	.343	-2.08	-1.48
10% (.10)	southern	13.2	8.62	4.10
20% (.62)	northern	2.23	-1.62	-.948
20% (.025)	southern	21.7	18.7	15.9

Table 2: Traditional RFA Swiss daily precipitation analysis. First row: homogeneity tests calculated with only locations having a silhouette coefficient above .63, representing 10% of the RFA northern cluster. A departure from zero in the three test values indicate a lack of homogeneity. The other three rows show that the RFA southern cluster is less homogeneous than the northern one and that increasing the number of sites, say 20%, deteriorates homogeneity.

Percentages of sites	Cluster names	Three RFA homogeneity tests		
10% (.81)	low altitude area	.215	-2.07	-2.27
10% (.71)	alpine	2.76	.282	1.75
20% (.80)	low altitude area	5.36	.624	-.863
20% (.71)	alpine	2.21	.692	2.52
25% (.80)	low altitude area	6.00	2.21	1.23
25% (.70)	alpine	1.96	.568	2.65
30% (.79)	low altitude area	9.19	4.72	2.61
30% (.69)	alpine	3.15	1.69	3.34

Table 3: Same as Table 2 but for our `PAM` approach based on $\hat{\omega}$, see Eq. (5).

List of Figures

9 Comparison of quantile-quantile plots: The x-axis corresponds to the local quantiles from Model (7). The y-axis displays semi-regional and regional quantiles, i.e. from models (9) and (8), respectively. Rows indicate the cluster family, alpine or low altitude area, and the column corresponds to the station type: the worst (best) classified station in the first (third) column. The medoid station is represented by the middle column. 40

10 Partition of Swiss weather stations applying `PAM` algorithm to covariates (traditional RFA approach, see Hosking and Wallis 2005) or $\hat{\omega}$ (our method). Shape and color of the points indicate their cluster. 41

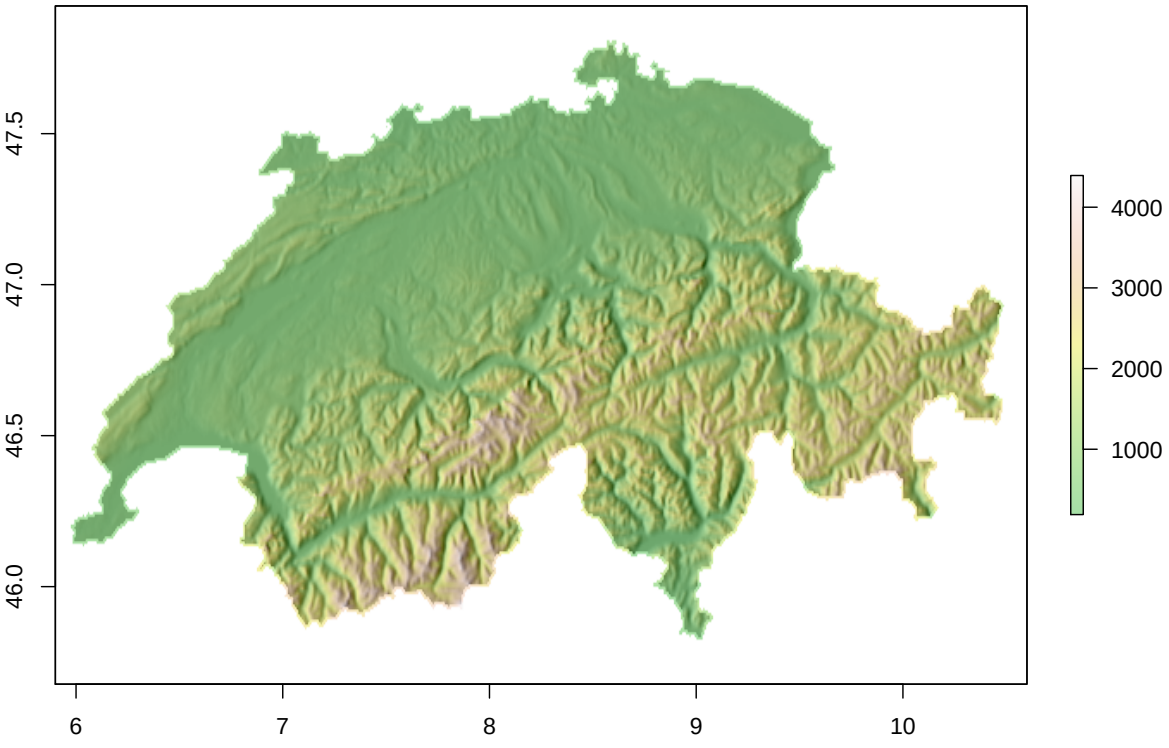


Figure 1: Switzerland elevation map (the scale is in meters)

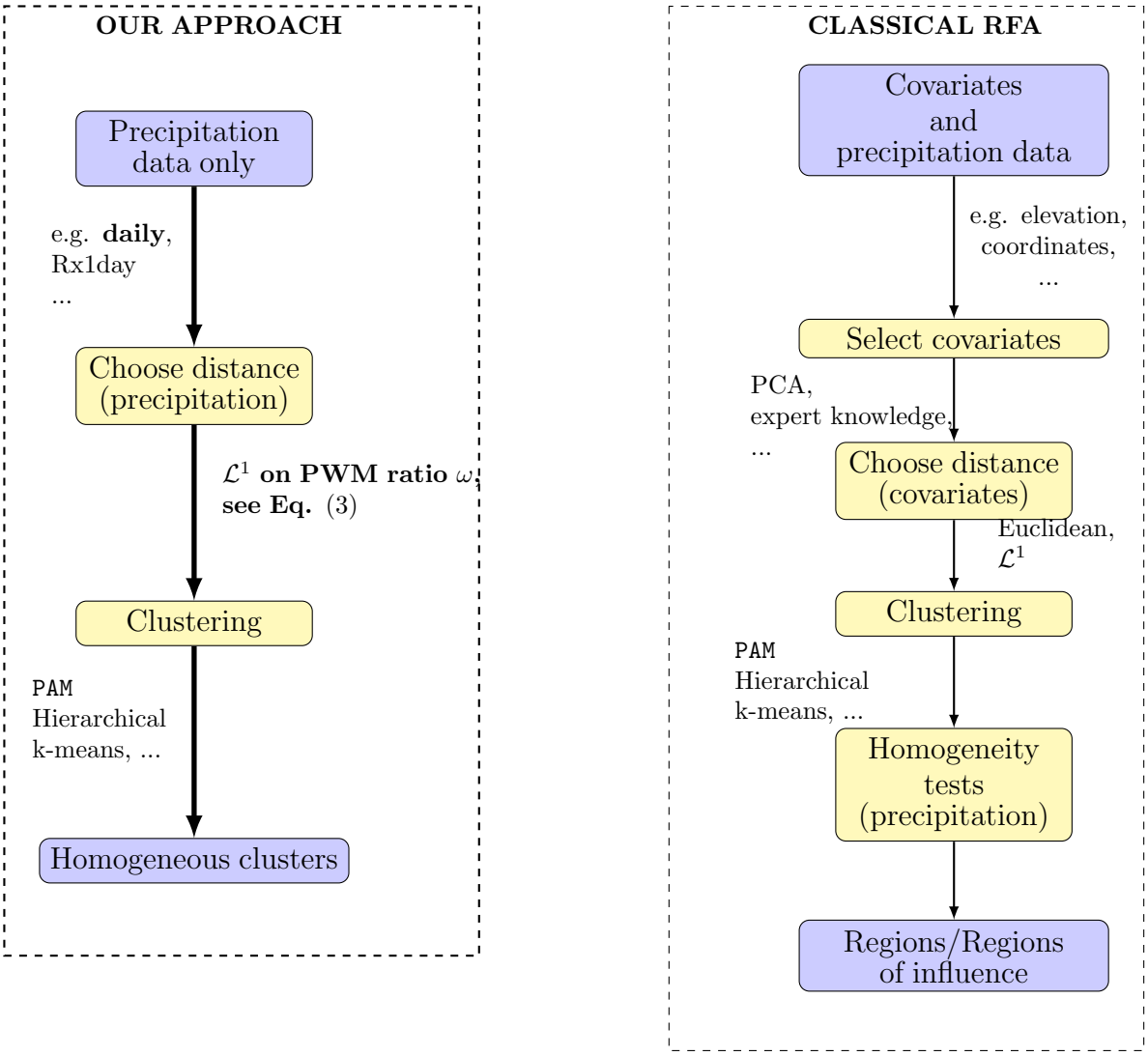


Figure 2: Steps to delineate homogeneous regions in RFA (traditional vs. method introduced in this paper). Rx1day is the annual maximum daily precipitation). The left method corresponds to the methods developed in this paper. The other way is the traditional RFA path with homogeneity tests, see e.g Hosking and Wallis (2005).

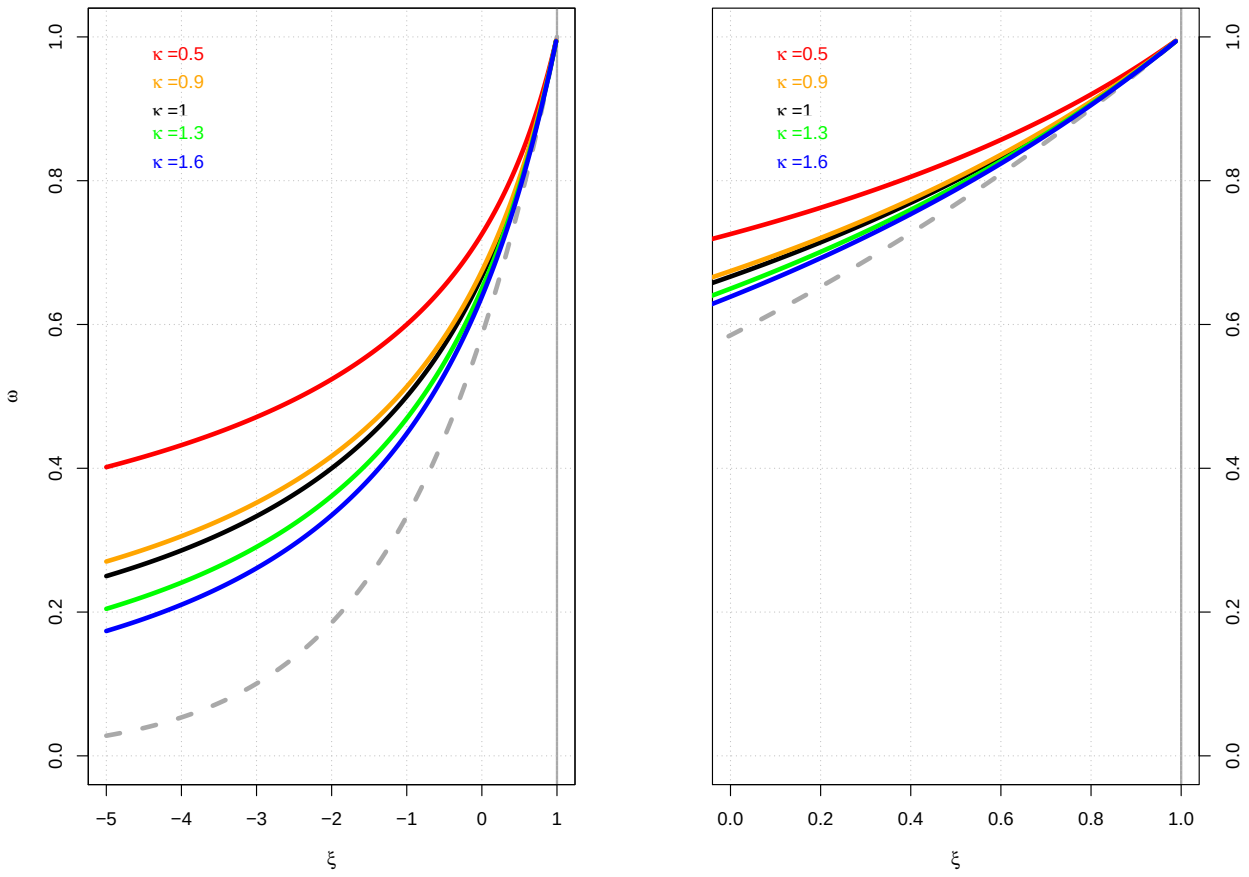


Figure 3: The y-axis represents the ratio ω for a EGPD(κ, σ, ξ), see Eq. (4). The x-axis corresponds to the upper tail shape parameter ξ. The left panel has $\xi \in (-5, 1)$ while the right panel provides a zoom on $\xi \in (0, 1)$. Each color represents a different value of $\kappa \in \{.5, .9, 1, 1.3, 1.6\}$ The gray dotted line corresponds to the GEV case. The black line with $\kappa = 1$ corresponds to the GP case.

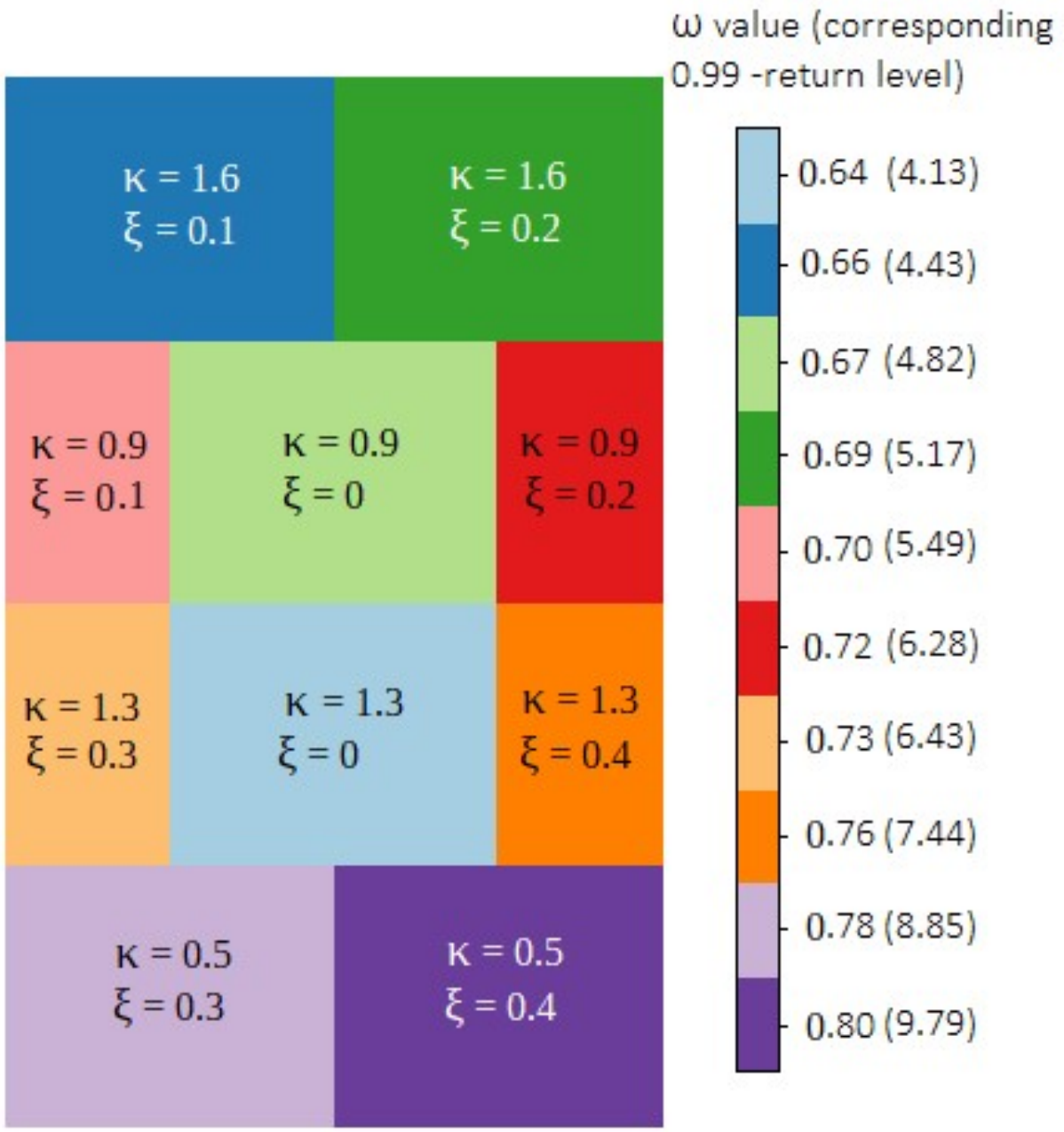
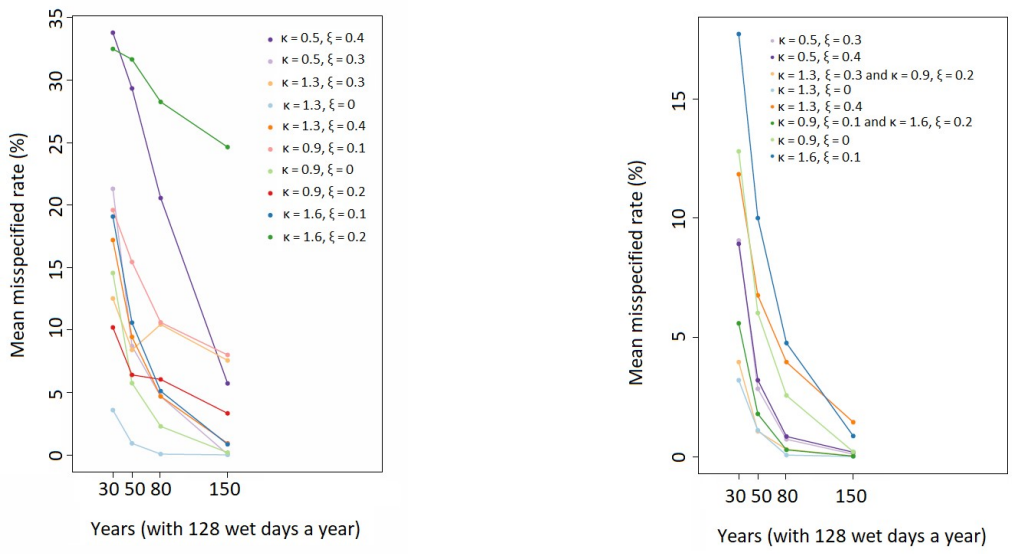


Figure 4: Experimental design setup based on EGPD(κ, σ, ξ), see Eq. (4). The colors correspond to 10 values of the ratio ω with their associated κ and ξ parameter values. A large (small) rectangle contains 40 (20) locations. The numbers in brackets in the color legend represent the 99% return level associated with each combination.



(a) $k = 10$ clusters

(b) $k = 8$ clusters

Figure 5: Misspecified PAM clustering rates with respect to the 10 regions shown in Figure 3 for k chosen as input in PAM algorithm. $k = 10$ (Panel (a)) corresponds to the number of simulated homogeneous regions. The case $k = 8$ (Panel (b)) corresponds to an underestimation of the number of homogeneous regions (e.g. for small samples).

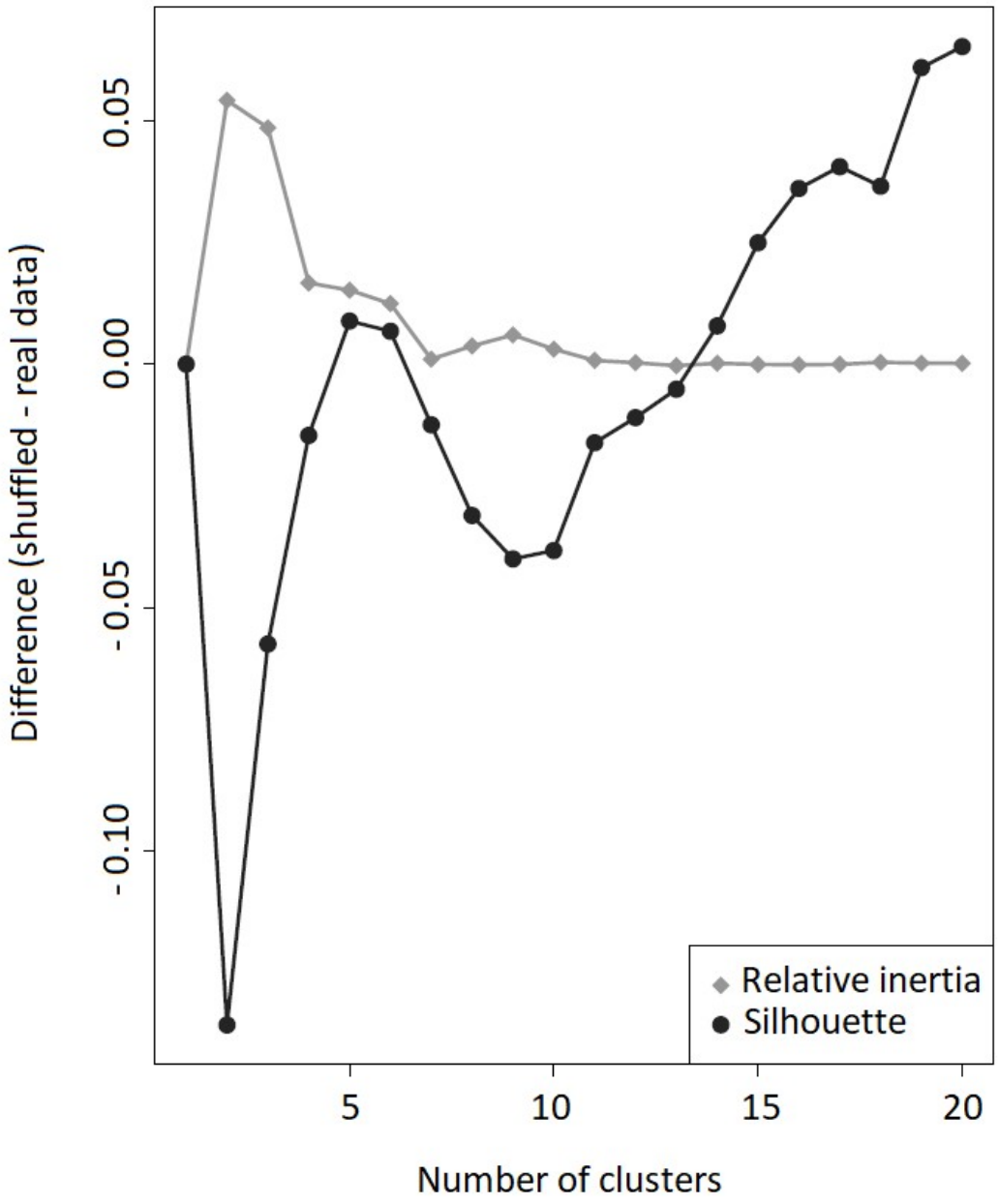
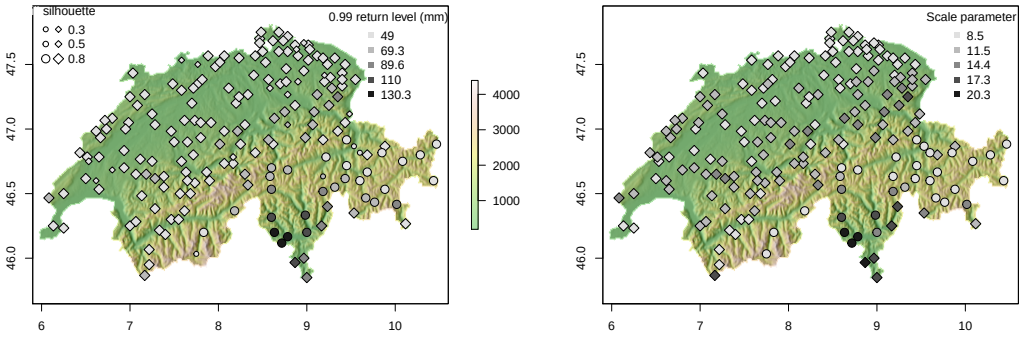


Figure 6: Swiss daily precipitation. Differences of relative inertia and silhouette coefficient between shuffled data and observed daily precipitation as a function of the number of clusters.



(a) Estimated .99 return-levels.

(b) Semi-regional σ estimates

Figure 7: Swiss daily precipitation modeled with the regional Model (8): Panel (a): `PAM` outputs in two clusters that are identified by circles (so-called "low altitude area cluster") and diamonds (so-called "alpine cluster"), see also Figure 10. The size of the points is proportional to the silhouette coefficient. The gray nuance color legend corresponds to .99 return level fits from Model (8). Panel (b): Estimates of σ in Model (8).

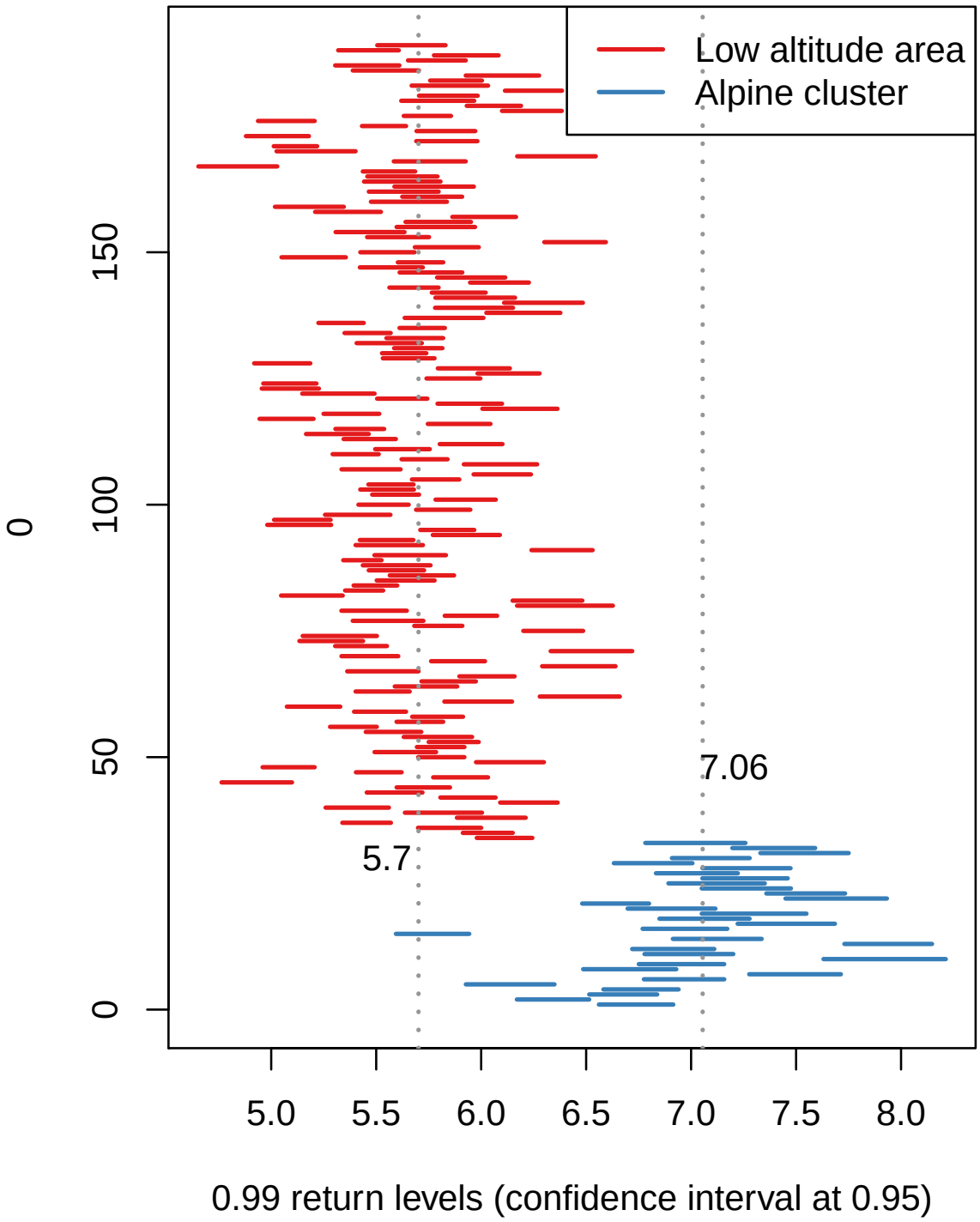


Figure 8: Confidence intervals of normalized .99-return level estimates at the 95% level from Model (9). Each horizontal line corresponds to a station from either the low altitude area or alpine clusters, indicated by blue diamonds or red circles respectively in Panel (b) of Figure 10. The dotted vertical lines indicate mean estimates within each cluster.

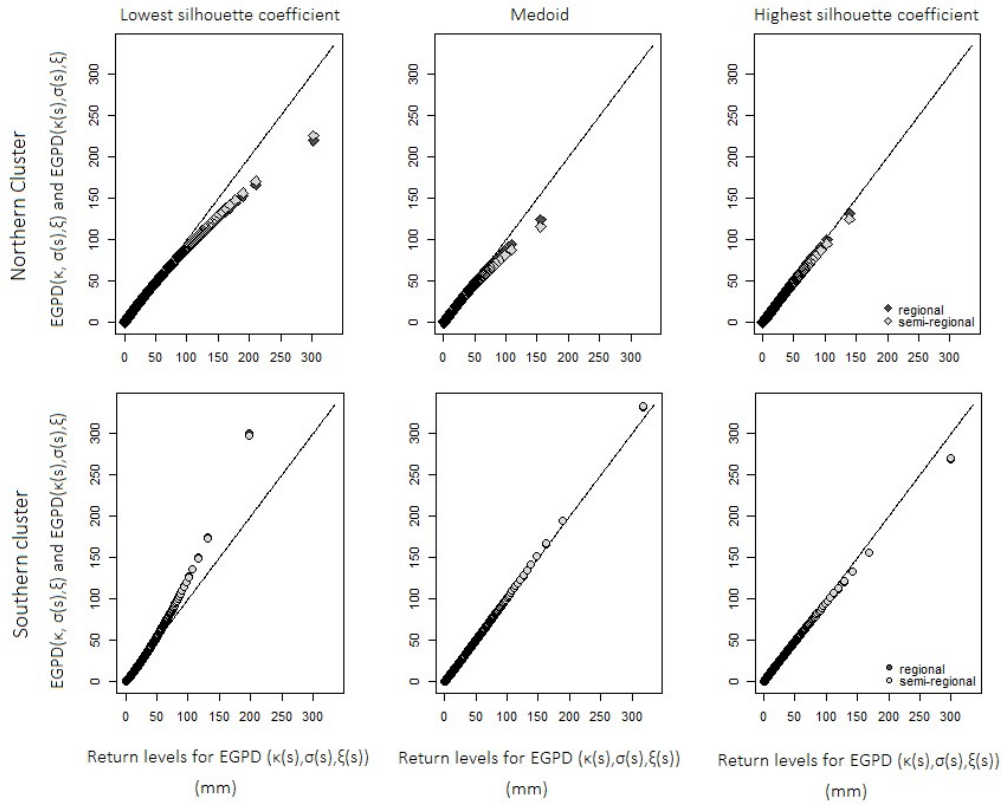
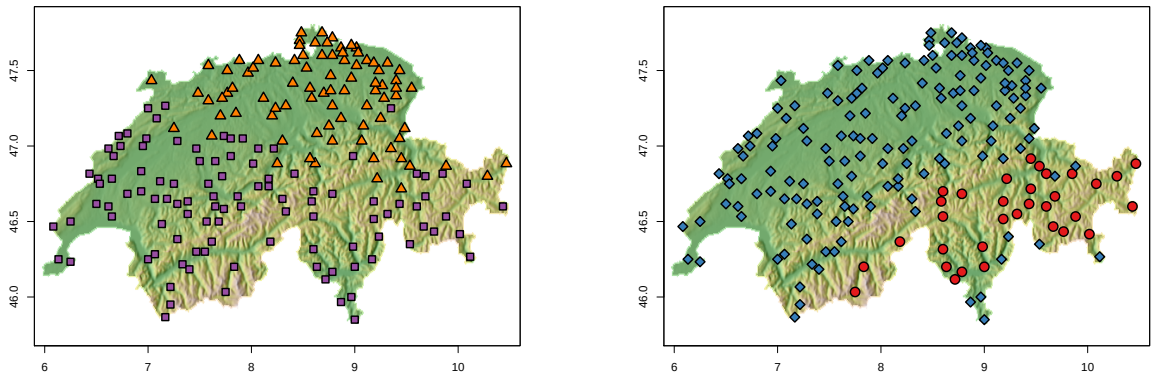


Figure 9: Comparison of quantile-quantile plots: The x-axis corresponds to the local quantiles from Model (7). The y-axis displays semi-regional and regional quantiles, i.e. from models (9) and (8), respectively. Rows indicate the cluster family, alpine or low altitude area, and the column corresponds to the station type: the worst (best) classified station in the first (third) column. The medoid station is represented by the middle column.



(a) Normalized elevation and coordinates.

(b) $\hat{\omega}$

Figure 10: Partition of Swiss weather stations applying `PAM` algorithm to covariates (traditional RFA approach, see Hosking and Wallis 2005) or $\hat{\omega}$ (our method). Shape and color of the points indicate their cluster.