UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Final Project Report
Statistical Methods for
Climate Sciences I & II
FS2023 & HS2024

Analysis of Spatial Precipitation Patterns of Switzerland

Sophie Spelsberg
Max Shore

May 31, 2024

Abstract

In this report, we identify and compare 4 hydroclimatic regions in Switzerland. For this purpose, clusters are computed based on the governing precipitation regimes of individual stations. The analysis comprises precipitation data (2000 - 2023) from 19 stations across Switzerland. In a first step, we assess patterns in the variance of the 19 time series through Principal Component Analysis and analyze the underlying meteorological mechanisms. Furthermore, hierarchical clustering is performed on the correlations of the individual stations with the Principal Components. Finally, we conduct an analysis of variance to assess the differences in total annual precipitation between the clusters. Through our analysis, we determine 4 regions - high Alps, Jura and western Plateau, eastern Plateau and eastern Alps, and south Alps – which are characterized by different precipitation regimes and show significant differences in annual precipitation and their seasonal cycle.

1 Introduction

Precipitation is one of the central variables of our climate system, and understanding its spatial and temporal distribution is key for applications such as agricultural planning, modelling the hydrological cycle or flood and drought prediction. However, those tasks can be challenging due to the high variability of precipitation both in space and time (Johansson & Chen, 2003). Orographic effects and small scale atmospheric dynamics can lead to large differences in precipitation: during a single rainfall event, precipitation amounts can vary significantly within less than a kilometer (Pedersen et al., 2010).

Switzerland, with its complex topography and diverse climate systems exemplifies these challenges. The country is characterized by a remarkably varied landscape that includes the alps, the central plateau and the Jura mountains. These topographical features create significant orographic effects, where moist air masses are lifted upwards leading to increased precipitation on the windward slopes and reduced precipitation on the leeward shadows. This geographical diversity results in pronounced spatial variability in precipitation. For instance, climatic regions like the Rhone valley and the plateau are protected against moist air masses coming from the north and the south, and thus have rather dry conditions (Fig. 1) (MeteoSwiss, 2024).

Additionally, Switzerland's climate system is influenced by a range of atmospheric circulation patterns. The interaction of Atlantic, Mediterranean, and continental air masses creates a complex climatic environment that contributes to the spatio-temporal variability of precipitation. Seasonal variations are also prominent, with extra-tropical cyclones bringing heavy snowfall to the alpine regions in winter, while summer convective storms can lead to intense but localized rainfall events (MeteoSwiss, 2024). Understanding patterns in these complex precipitation signals is crucial for managing water resources, predicting natural hazards, and planning for agricultural activities in a country where the weather can change dramatically over short distances and timescales.

The objective of this report is therefore to identify spatial patterns in the distribution of precipitation in Switzerland: we will extract the main sources of precipitation variability from our data and cluster measurement stations based on their governing precipitation regimes. For this purpose, we analyze a dataset of monthly precipitation measurements which spans 24 years between January 2000 and December 2023 and was provided by MeteoSwiss. The data consists of measurements from 19 weather stations

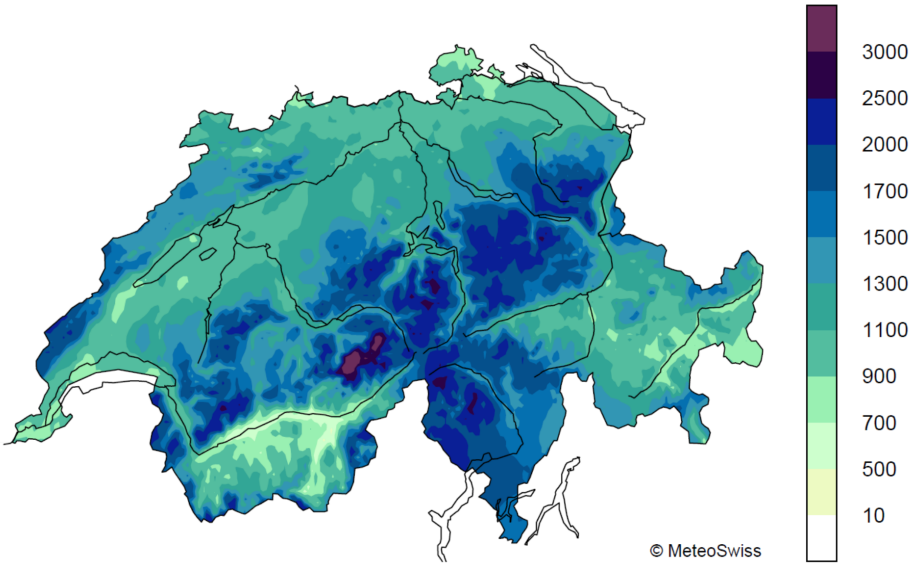

Figure 1: Annual total precipitation (mm) for the period 1991-2020 (MeteoSwiss, 2024).

across the country (Fig. 2). Our goal is to cluster these stations into groups with similar precipitation patterns. Since the different precipitation regimes do not necessarily imply that the total amount of precipitation differs between the clusters, we will then check whether the amount of annual precipitation also varies significantly between the clusters. In the following three steps, we will therefore:

1. analyze the main precipitation regimes across Switzerland via Principal Component Analysis (PCA)

2. identify different hydroclimatic regions based on the Principal Components (PC) through Clustering and

3. compare the annual mean precipitation between the clusters by performing an Analysis of Variance (ANOVA).

2 Methods

The following section provides an overview of the three methodologies utilized in this study and gives a brief introduction to the dataset employed.

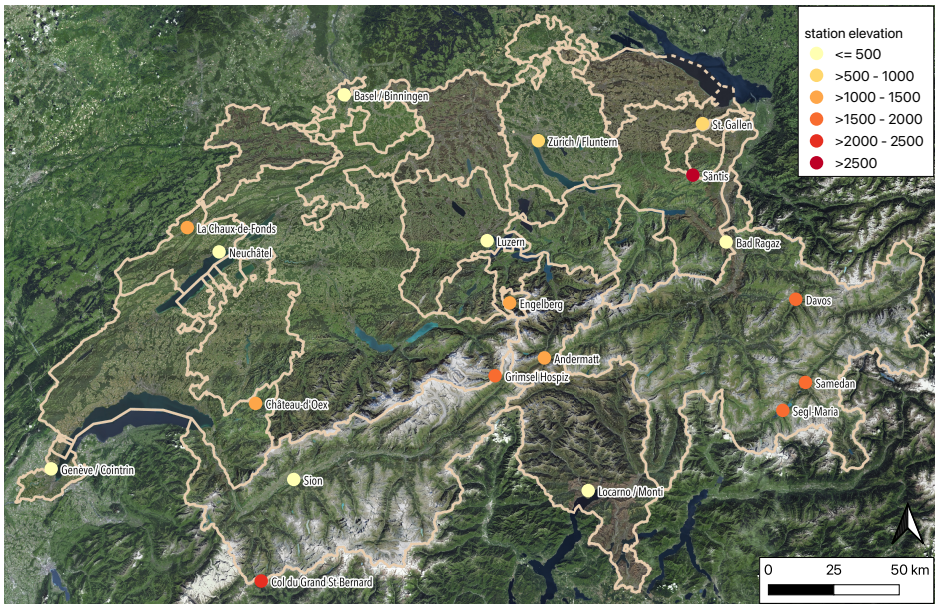

Figure 2: Map of all meteorological stations included in the analysis. Colors indicate elevation in m.a.s.l.

Each approach is presented within the context of the research, along with an explanation of its features and capabilities. All data analysis was conducted using methodologies implemented in R (R Core Team, 2022).

2.1 Data

The dataset used for this study is a subset of a meteorological dataset provided by MeteoSwiss. The original dataset spans 24 years between January 2000 and January 2024 and includes monthly measurements of 18 meteorological variables from 20 stations. For our analysis, we extracted the precipitation data until December 2023 and used all 19 stations that provided measurements for this variable. Additionally, we extracted the metadata available for all stations, including their position and elevation. Their elevations are between 316 and 2501 m.a.s.l. (Fig. 2). An overview of the stations and their codes can be found in Table 1. From the initial dataset we calculated additional measures such as total annual, mean monthly and mean annual precipitation.

Table 1: List of all stations

Station name	**Station code**
Andermatt	ANT
Basel / Binningen	BAS
La Chaux-de-Fonds	CDF
Château-d'Oex	CHD
Davos	DAV
Engelberg	ENG
Grimsel Hospiz	GRH
Col du Grand St-Bernard	GSB
Genève / Cointrin	GVE
Luzern	LUZ
Neuchâtel	NEU
Locarno / Monti	OTL
Bad Ragaz	RAG
Säntis	SAE
Samedan	SAM
Segl-Maria	SIA
Sion	SIO
Zürich / Fluntern	SMA
St. Gallen	STG

2.2 Descriptive statistics

The dataset was initially examined to gain a better understanding of its basic characteristics. This initial phase is crucial as it provides direction for the following analysis stages. We calculated a correlation matrix between the precipitation records of all 19 stations and visually assessed the data using a scatterplot matrix. After conducting our initial analysis, we proceeded to assess the multivariate normality of the data using a Henze-Zirkler test. This test is commonly used to determine whether the data follows a multivariate normal distribution, which is an important assumption in many statistical analyses. Moreover, the assessment of univariate normality was carried out individually for each station through an Anderson-Darling test. This approach allows us to identify any potential deviations from normality that may be specific to certain stations.

2.3 Principal Component Analysis

Principal Component Analysis (PCA) stands as a foundational tool in the realms of meteorology and climatology, especially when dealing with vast datasets. Its primary objective is to streamline data, thereby reducing its dimensions while retaining as much relevant information as possible. This process facilitates a clearer comprehension and interpretation of the inherent data structure.

In the context of precipitation data analysis, a typical dataset may comprise numerous observations (n) across various variables (p). For instance, if we consider monthly precipitation records spanning 24 years from 19 distinct weather stations in Switzerland, we would have $n = 288$ observations for $p = 19$ variables. Often, these variables exhibit strong correlations, particularly when stations are geographically proximate. Consequently, the aim is to distill the original dataset into a smaller subset of variables ($m < p$) that encapsulates as much of the information as possible. This entails crafting new variables that, while distinct from the originals, are derived from them. PCA offers an elegant solution to this challenge, leveraging linear transformations of the original variables to extract essential patterns and streamline the dataset effectively. Hereby, PCA is a powerful tool for reducing noise in the data, which ensures more stable clustering (2.4) results in the second step (Ben-Hur & Guyon, 2003).

The method can be described as follows: we define principal components $Y_i,\ i = 1, ..., p$ that can be expressed as linear combinations of the original dataset. Let $\mathbf{X}(n \times p)$ be the original dataset matrix with a covariance matrix $\mathbf{\Sigma}$.

$$\begin{aligned} Y_1 &= \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p &&= \boldsymbol{\alpha_1}\mathbf{X} \\ &\vdots && \vdots \\ Y_p &= \alpha_{p1}X_1 + \alpha_{p2}X_2 + \cdots + \alpha_{pp}X_p &&= \boldsymbol{\alpha_p}\mathbf{X} \end{aligned} \tag{1}$$

From eq.1, we can compute the variances and covariances of the newly defined linear combinations.

$$\begin{aligned} Var(Y_i) &= \boldsymbol{\alpha_i'}\mathbf{\Sigma}\boldsymbol{\alpha_i} \\ Cov(Y_i, Y_k) &= \boldsymbol{\alpha_i'}\mathbf{\Sigma}\boldsymbol{\alpha_k}, \quad k = 1, ..., p \end{aligned}$$

PCA aims at having each principal component explaining the maximum amount of variance while being orthogonal to each other. It follows that

for each Y_i, we are looking for the linear combination $\boldsymbol{\alpha_i}\mathbf{X}$ that maximizes the variance. To ensure orthogonality between the PCs, the condition $Cov(Y_i, Y_k) = 0$ must be satisfied. It is important to note that $\boldsymbol{\alpha_i'}\boldsymbol{\alpha_i} = 1$ in order to keep the solution that conserves variance. It emerges that the eigenvectors of the covariance matrix $(\mathbf{e_i}, i = i, ..., p)$ satisfy these conditions. Thus, we can compute the PCs with the following relationship:

$$Y_i = \mathbf{e_i}\mathbf{X} \tag{2}$$

Each Y_i defines a new axis for the dataset and has coefficients that represent the contributions of the original variables to that component. Additionally, we define the loadings of each variable as the coefficients (α) multiplying the original variables in the linear combination (eq.1) that defines each PC. They provide a measure of the contribution or weight of each original variable to a PC, indicating the degree of similarity between a variable and a PC. These loadings are represented by the eigenvectors of the covariance matrix (eq.2). Every eigenvector forms a pair with an eigenvalue. The latter represent the amount of variance explained by each PC. Consequently, the eigenvalues serve as indicators of the extent to which a particular principal component accounts for the overall variability.

One of the primary difficulties in PCA involves choosing the appropriate number of PCs to proceed with the analysis. There is a risk of selecting a too small amount of components that fail to adequately explain the variance present in the data. Conversely, opting for a large number of PCs to describe the majority of the variance can impede the initial objective of this statistical tool, which is to reduce as much as possible the number of variables to explain the data's variance. To select the optimal number of PCs, there are several empirical methods that can be applied. We used two commonly employed techniques:

- we assessed the scree plot visually and omitted irrelevant PCs by finding the "elbow" in the plot and
- we applied Kaiser's rule and selected all PCs with eigenvalues greater than one (Kaiser, 1960).

When conducting PCA on variables that exhibit variations in scales or magnitudes, it is recommended to standardize the data prior to utilizing the technique. Although this compresses (or expands) the variation to a unit

variance of one, it avoids having PCs dominated by variables with larger magnitudes. In order to address the significant differences in precipitation patterns and variance across stations, we normalized the precipitation data for this study. Failure to do so could lead to an over-representation of stations with higher precipitation levels within the first PCs. This allows us to understand the underlying spatio-temporal precipitation patterns and mechanisms independently of differences in absolute precipitations amounts between the stations.

With scaled data, the procedure is similar to what is described above, but instead of the covariance matrix $\mathbf{\Sigma}$, the correlation matrix $\boldsymbol{\rho}$ is used. We therefore based the PCA on the correlation matrix computed previously (section 2.2). A useful property that arises in this case is that the loadings of the PCs are equivalent to the correlations between the original variables and the PCs. We used those correlations for clustering the data in the second part of our analysis.

2.4 Clustering

Clustering is a technique that allows grouping objects based on their similarity with the aim to find the underlying structure in a dataset. It hereby helps to identify patterns and similarities in the data. For this report, we used an agglomerative hierarchical clustering (AHC) approach. AHC starts with the individual objects (in this case stations) and calculates the distance between all of them to find the closest two objects for clustering (Jolliffe & Philipp, 2010). It follows a simple algorithm:

1. Start with the n individual objects for clustering
2. Calculate the distances between all of them to find the two closest objects
3. Aggregate the two objects to a single one
4. Calculate the distances between the new object and the remaining objects and find the two closest objects again
5. Repeat step 3 and 4 until all objects are aggregated and record the current stage of the clusters after each repetition

This process will result in one big cluster at the end; the stages in between can be displayed in the form of a dendrogram. The dendrogram can then

be used to partition the data into k individual clusters by cutting it at a certain position. In contrast to non-hierarchical clustering methods such as k-means, AHC does not include a random element and will therefore reproduce the same cluster structure under the same initial conditions. We chose this approach to be able to assess the distances between the elements of a single cluster through the dendrogram and to make the clustering process and the composition of the clusters more transparent.

For the hierarchical clustering algorithm, it is necessary to define

- a **distance measure** between the objects in the multivariate space and
- a **linkage method** between the clusters, i.e. a rule that defines the position of a cluster from which the distance should be computed

Possible distance measures include the Euclidean distance, the statistical distance or the Minkowski metric. For this report, we used the **Euclidean distance**, which is defined as

$$d(x, y) = \sqrt{(x - y)'(x - y)}$$

and can be described as the length of a straight line between two points in a multidimensional space. Euclidean distance is the most commonly used distance measure for clustering.

As for the distance, there are multiple possible options for the linkage method: single linkage (the minimum distance between two clusters), complete linkage (the maximum distance), average linkage, or Ward's method. The goal is to define a linkage method that minimizes variance within the clusters and maximizes variance between them. Various techniques were evaluated, and the average linkage method was selected due to its ability to generate more consistent clusters compared to, e.g., the single linkage method. For this method, the distance between all pairs of individual objects within two clusters is calculated and then averaged (Govender & Sivakumar, 2019).

Furthermore, we decided to perform the clustering not on the initial data set, but on the correlations of the individual stations with the most important PCs (i.e. the loadings) as in Baeriswyl & Rebetez (1997). This step offers two important advantages over clustering on the monthly precipitation dataset:

- It reduces the number of dimensions of the dataset and helps to avoid the curse of dimensionality: In high-dimensional data, points tend to

be equidistant from each other when euclidean distance is measured. Cluster analysis on a 289-dimensional set of precipitation data is not advisable, as the distances between the points tend to become meaningless (Aggarwal et al., 2001).

- The goal of this report is to identify clusters with similar precipitation regimes, it is therefore important to perform the clustering after extracting only the most important precipitation patterns from the 19 time series. As precipitation is influenced by a variety of factors (as described in section 1), the data tends to be noisy. However, random noise has a strong negative influence on the quality of clustering. A preceding PCA filters this noise out (Ben-Hur & Guyon, 2003).

To determine the ideal number of clusters, we visually assessed the dendrogram and compared the structure of our clusters with the results of other algorithms. Our aim was to find a stable configuration of clusters with multiple elements per cluster. To evaluate the stability of the clustering, we calculated different versions of the clusters using several algorithms and then compared the differences between them. The comparison included AHC with average linkage, AHC with complete linkage, k-means clustering and a modified version of the dataset in which we weighted the PC loadings according to their proportion of explained variance.

For the clustering, there were no further prerequisites. While it can make sense to scale data previous to applying a clustering algorithm when the dimensions of the data are measured in different scales, scaling was not necessary for this report. As the clustering was performed on the correlations of the individual stations with the first four PCs, the scales of the four dimensions were the same.

2.5 Analysis of Variance

Analysis of Variance (ANOVA) is a statistical method used to compare the means of two or more groups to determine whether there are statistically significant differences among them. It accomplishes this by examining the variance between groups relative to the variance within groups.

In this study, we utilized ANOVA to investigate the differences in yearly mean precipitation between previously defined clusters (2.4) of weather stations in Switzerland. This method would allow to further assess the separation between climatic zones of Switzerland by delving into the differences in

magnitude of precipitation between the zones, while the previous methods looked into spatio-temporal discrepancies. We computed the yearly mean of precipitation for each cluster, resulting in 96 samples split between 4 clusters. The yearly mean precipitation values for each cluster served as the dependent variable, while the cluster assignments acted as the independent variable. To apply ANOVA, the dataset must satisfy specific conditions.

- **Independence**: The observations within each group are independent of each other. In this study, this assumption holds as the yearly mean precipitation values for each cluster are calculated independently of other clusters.
- **Normality**: The data within each cluster must follow a normal distribution. This assumption is important for accurate estimation of parameters and confidence intervals. However, ANOVA has been shown to be robust against minor departures from normality (Blanca Mena et al., 2017).
- **Homogeneity of Variances**: The variances of the groups being compared are equal (homoscedasticity). Violations of this assumption can affect the accuracy of the F-statistic and subsequent p-values. When this assumption is violated, Welch's ANOVA test can be conducted.

Initially, we applied traditional ANOVA. However, the assumption of homogeneity of variances was violated. Using traditional ANOVA in this scenario would result in poor and undesired power properties. Therefore, we employed Welch's ANOVA, a robust alternative that does not require the assumption of equal variances (Wilcox & Keselman, 2003). This test allows us to test the null hypothesis that the mean yearly precipitation is equal across all clusters ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$), against the alternative hypothesis (H_1) that at least one cluster mean differs from the others.

Welch's ANOVA defines a test statistic, similar to the traditional homoscedastic case, as follows:

Let n_k be the number of samples in the k^{th} cluster, we define the weights

$$w_1 = \frac{n_1}{s_1^2}, w_2 = \frac{n_2}{s_2^2}, ..., w_j = \frac{n_j}{s_j^2}$$

Then compute

$$
\begin{aligned}
U &= \sum w_j, \\
\tilde{X} &= \frac{1}{U} \sum w_j \bar{X}_j, \\
A &= \frac{1}{J-1} \sum w_j (\bar{X}_j - \tilde{X})^2, \\
B &= \frac{2(J-2)}{J^2-1} \sum \frac{(1-\frac{w_j}{U})^2}{n_j - 1}, \\
F_w &= \frac{A}{1+B}
\end{aligned}
$$

F_w is the test statistic used for Welch's ANOVA. Additionally, we must compute the degrees of freedom for this test under H_0

$$\nu_1 = J - 1$$

and

$$\nu_2 = \left(\frac{3}{J^2-1} \sum \frac{(1-\frac{w_j}{U})^2}{n_j - 1} \right)^{-1}$$

F_w approximately follows an F distribution (see Fig. 3) with degrees of freedom ν_1 and ν_2. We reject the null hypothesis H_0 if $F_w \geq f$, where f is the $1-\alpha$ quantile of the F-distribution with degrees of freedom ν_1 and ν_2. In this study, we conducted Welch's F-test using the previously defined $J = 4$ clusters as independent groups.

Following this combined test, we carried out a Games-Howell post-hoc test to investigate all possible pair-wise differences in means between the four clusters. This test is used here as it tolerates unequal variances between groups. Similarly to the Welch's test, Games-Howell uses a correction of the degrees of freedom to take into account the heteroskedastic nature of the data (Shingala & Rajyaguru, 2015)

$$\nu = \frac{(s_i^2/n_i + s_j^2/n_j)^2}{\frac{(s_i^2/n_i)^2}{n_i - 1} + \frac{(s_j^2/n_j)^2}{n_j - 1}}$$

The test statistic employed is akin to that of a student's two sample t-test

$$t_{GH} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$$

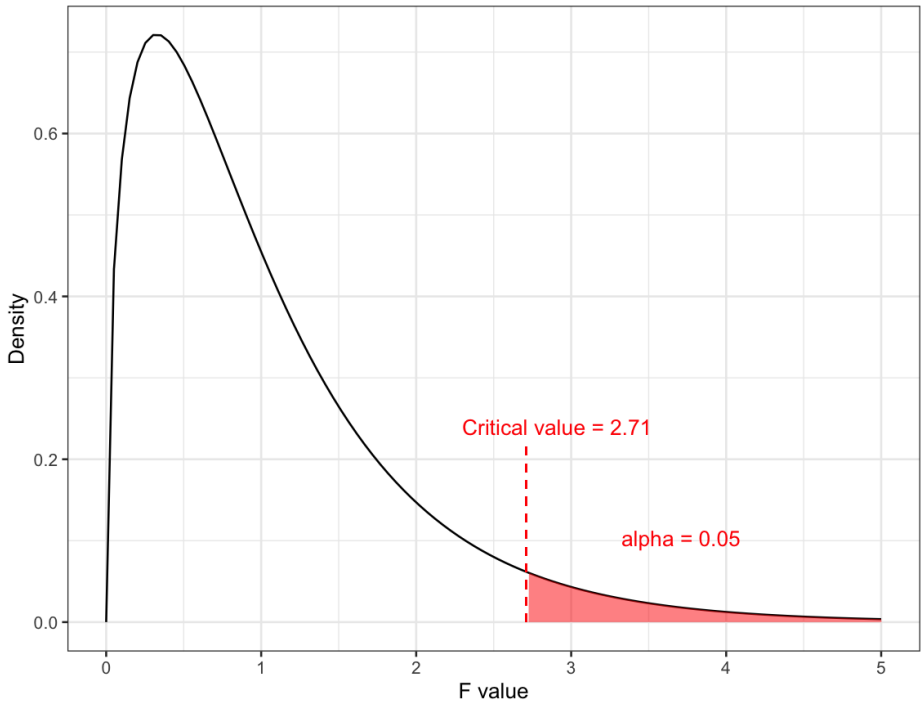

Figure 3: An F-distribution with 3 and 87 degrees of freedom. The probability that F is greater than 2.7 is .05, as depicted by the area of the shaded region

wheres s_i^2 and s_i^2 are the sample variances for the i^{th} and j^{th} groups and n_i and n_j are the respective sample sizes from population i and j. Under H_0 the test has the same assumptions than the Welch's test (iid).

3 Results and Discussion

The following section will present the results obtained from the three methods utilized to analyze the precipitation patterns in Switzerland between 2000 and 2023. Prior to that, we will display the findings from the preliminary examination of the dataset.

3.1 Exploration of the dataset

In a first step, we assessed multivariate normality of the dataset. The precipitation data does not follow a multivariate normal distribution, and none of the 19 individual station measurement series shows a univariate normal

distribution. This reflects the properties of precipitation data which are often not distributed evenly around the mean. Noticeably, factors such as seasonal variability, extreme events (e.g., heavy rainfall), and the presence of zero or near-zero monthly precipitation values produce skewed datasets with outliers.

Following assessment of (non-)normality, we computed correlations between all individual station time-series in the dataset. The objective was to identify stations with strong or weak correlations between their precipitation records as a first indicator of possible patterns in the dataset. Correlations between the time-series range between -0.01 (Locarno - Säntis) and 0.94 (Samedan - Segl-Maria), with most values showing moderate correlation around 0.55 (Fig. 4). For comparison, the temperature records of the 19 stations correlate with an average of 0.99. This is consistent with the concept of precipitation as a meteorological variable with high spatio-temporal variability: while some large scale atmospheric circulation mechanisms (such as the North Atlantic Oscillation, NAO) will influence precipitation across all stations in Switzerland, the individual precipitation signal at a single station will be influenced by a multitude of different factors, including small scale atmospheric dynamics, effects of local topography and random noise. The correlation matrix shows that some stations, e.g. Grimsel Hospiz or Col du Grand St-Bernard, show rather low correlations with most other stations. This could indicate that the hydro-climate of these locations is dictated by a differing precipitation regime. In order to test this hypothesis, we will apply a cluster analysis as a second step. However, since noise in the precipitation data would have a negative impact on the quality of the clustering, we will first reduce the noise by extracting the most important large-scale precipitation patterns using PCA.

3.2 PCA of spatial precipitation patterns

The precipitation regimes were assessed in a first step through Principal Component Analysis, which can be used to find common patterns within the 19 time series and correlations of the respective stations with those patterns.

3.2.1 PC selection

Before interpreting the PCs, we started with selecting the most relevant components. To obtain the ideal number of PCs for further analysis, we first assessed the scree plot visually (Fig. 5). The largest proportion of the variance (58%) is explained by PC_1, while PC_{2-4} each contribute significantly

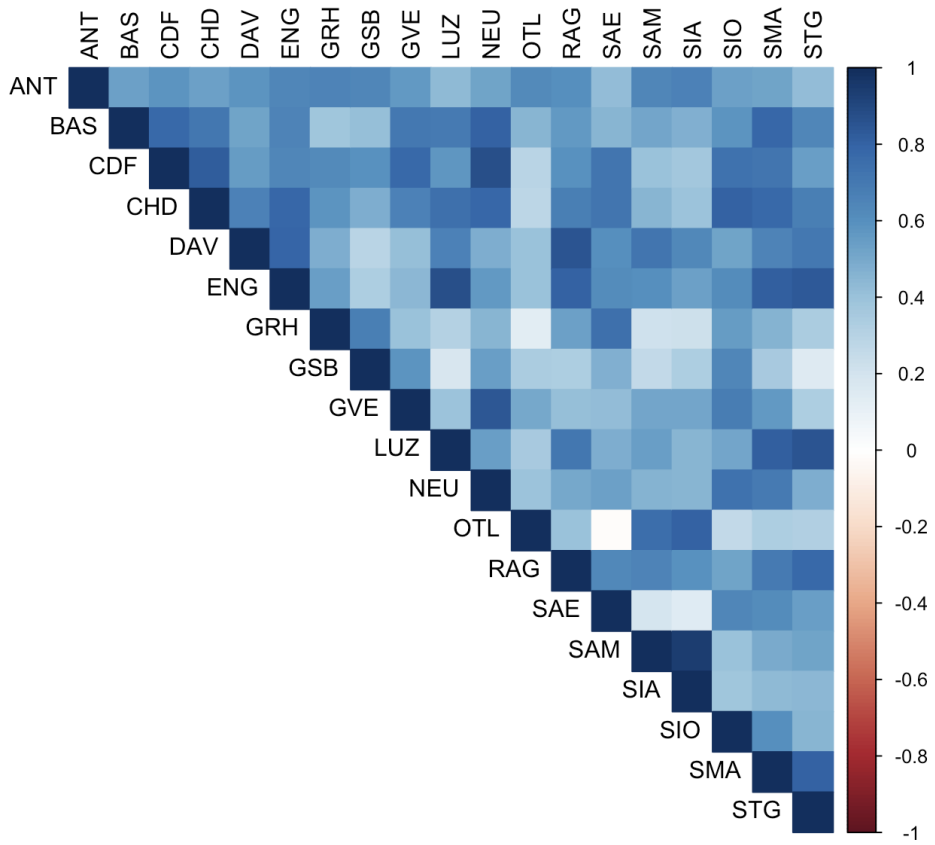

Figure 4: The correlation matrix between all individual precipitation records

(6-12%). From PC_5 onwards, the proportion of explained variance drops below 3% (Fig. 5). To validate our selection, we applied the Kaiser criterion in a second step. Since only the eigenvalues of the first 4 PCs are above 1, the criterion yields the selection of the same number of PCs as the analysis of the scree plot. We therefore chose the first 4 PCs for further analysis and omitted the other PCs. The total percentage of variance accounted for by the initial 4 principal components is 86%. Retaining extra PCs would not have significantly enhanced the ability to account for additional variability in the data, but rather would have increased the complexity of the interpretation.

3.2.2 Description and meteorological interpretation of the PCs

The loadings in Figure 6 represent the correlation between the time series of individual stations and the 4 principal components (PCs), which were scaled

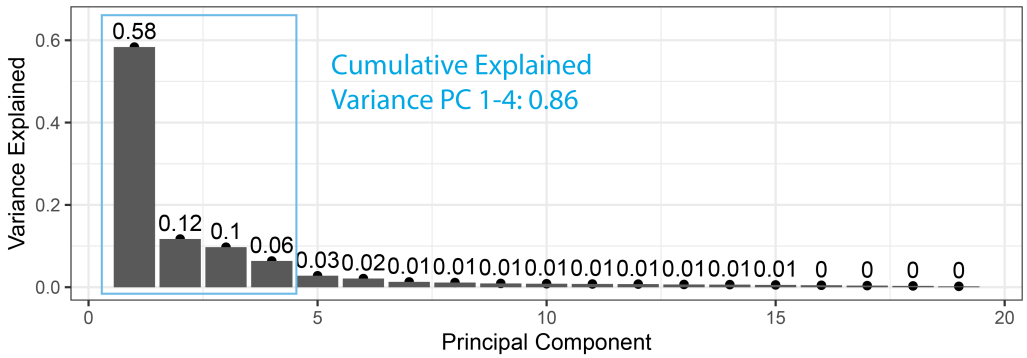

Figure 5: Scree plot showing the amount of variance that the individual PCs account for. The PCs chosen for clustering are marked by the blue box.

beforehand. Even though the PCs are not direct representations of climatic mechanisms, the loadings provide insight into how the PCs are linked to the atmospheric dynamics that influence precipitation patterns in Switzerland.

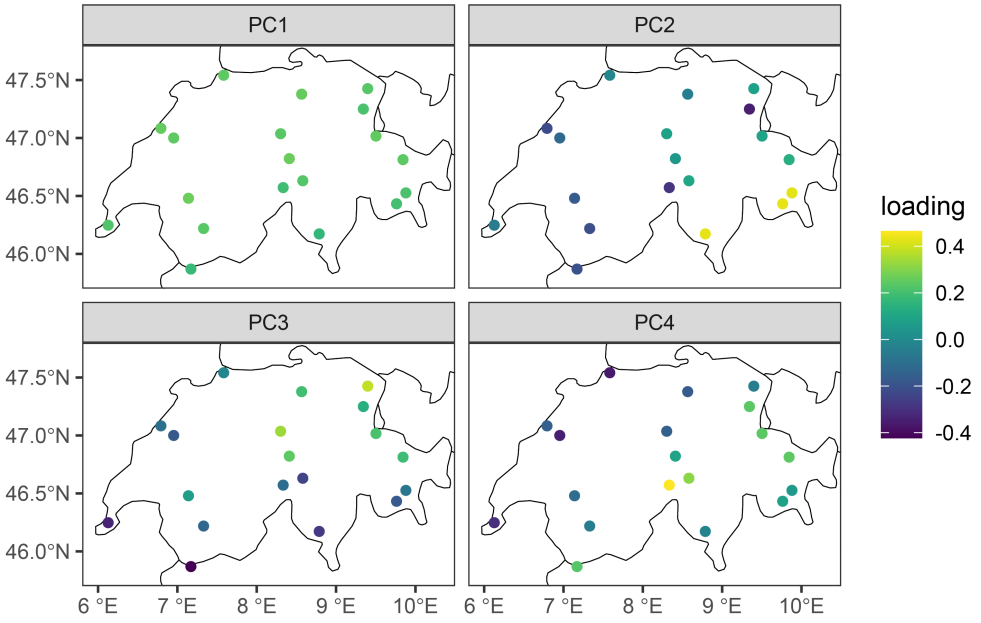

Figure 6: Loadings of the individual stations for PC1-4

- The first PC accounts for a large proportion of the total variance (58%) and shows very similar loadings for all stations. This PC is most likely representative of the overall weather patterns that influence Switzerland's climate. Large scale modes of climate variability as the North-Atlantic Oscillation (NAO) or the position of the jet stream change their state within weeks to months and control the weather and precipitation mechanisms in central Europe on these timescales. Thereby, these systems (almost) equally influence the precipitation at all 19 stations. This is consistent with both the similar loadings of all stations and the high amount of explained variance in PC_1. The outcome is not unexpected due to Switzerland's small size and the high probability of large-scale weather systems impacting the entire region, despite the presence of the Alps topographically dividing the country.

- The second PC shows high loadings for the three stations in Ticino/Engadin and neutral to negative loadings for all other stations. It displays the division between the precipitation patterns for the south alpine stations (which are already influenced by the Mediterranean climatic system) and the rest of the country. This separation is likely caused by the Alps, which act as a physical barrier for air masses and casts rain shadows downwind of the mountain range. This PC emphasizes the crucial role of orographical features on the distribution of precipitation across Switzerland.

- The third PC shows a less clear distinction, but a tendency towards positive loadings in the Northeastern part of the country and negative loadings in the South and the West. It could be associated with the increasingly continental precipitation patterns towards the east, where the influence of the maritime Atlantic climate and precipitation from low pressure systems is diminishing.

- The loadings of the fourth PC are correlated to the elevation of the stations ($r = 0.72$), high loadings can especially be found for the central alpine stations. It could therefore be interpreted as a PC accounting for variance in the data that is related to elevational differences and enhanced influence of orographic precipitation.

As all 4 PCs can be indicators of meteorologically relevant precipitation patterns in Switzerland, they are a valuable basis for clustering the data in

the next step. By reducing the number of variables from 19 to 4, the level of noise in the clustering process was considerably diminished. This means that the PCA algorithm was able to focus more on the most important variables, resulting in a more accurate and meaningful clustering of the precipitation patterns in Switzerland in the next step.

3.3 Clustering precipitation regions

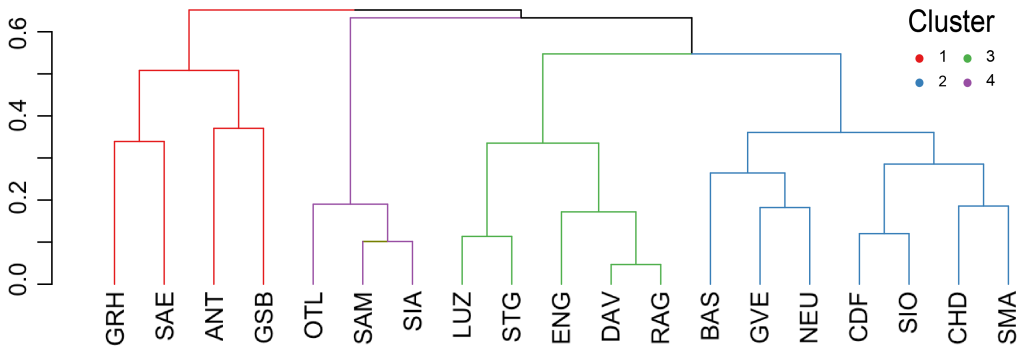

Figure 7: Dendrogram of the hierarchical clustering process, with clusters marked by color

We used hierarchical clustering to group the stations based on their correlations with the 4 PCs (i.e. the 4 most relevant precipitation patterns in Switzerland). Hereby, we split the stations into groups with similar precipitation regimes and determined the number of clusters through the dendrogram of the clustering process (Fig. 7). Within the 19 stations, we defined 4 clusters: (1) high Alps, (2) Jura and western Plateau, (3) eastern Plateau and eastern Alps and (4) south Alps (Fig. 8).

Within the clusters, the distance between some of the stations is very small, indicating a highly similar precipitation regime (e.g. Davos and Bad Ragaz). In contrast, the stations in the high alps cluster (e.g. Col du Grand St-Bernard, Säntis) don't show a consistent behaviour and are relatively distant from each other, despite forming a cluster (Fig. 7). This is also evident when comparing the clusters on the map: the stations in cluster 1 are geographically distant and will consequently be influenced by differing climate mechanisms. Including more high alpine stations with similar precipitation

characteristics into the analysis could therefore possibly refine the allocation of the clusters here and lead to a smaller variance within the clusters. Alternatively, this could lead to subdivisions of the alpine stations into separate clusters. However, on the current dataset, the division into clusters is very stable: all alternative clustering algorithms used for comparison returned the same or a very similar clusters. When comparing the average climate of the

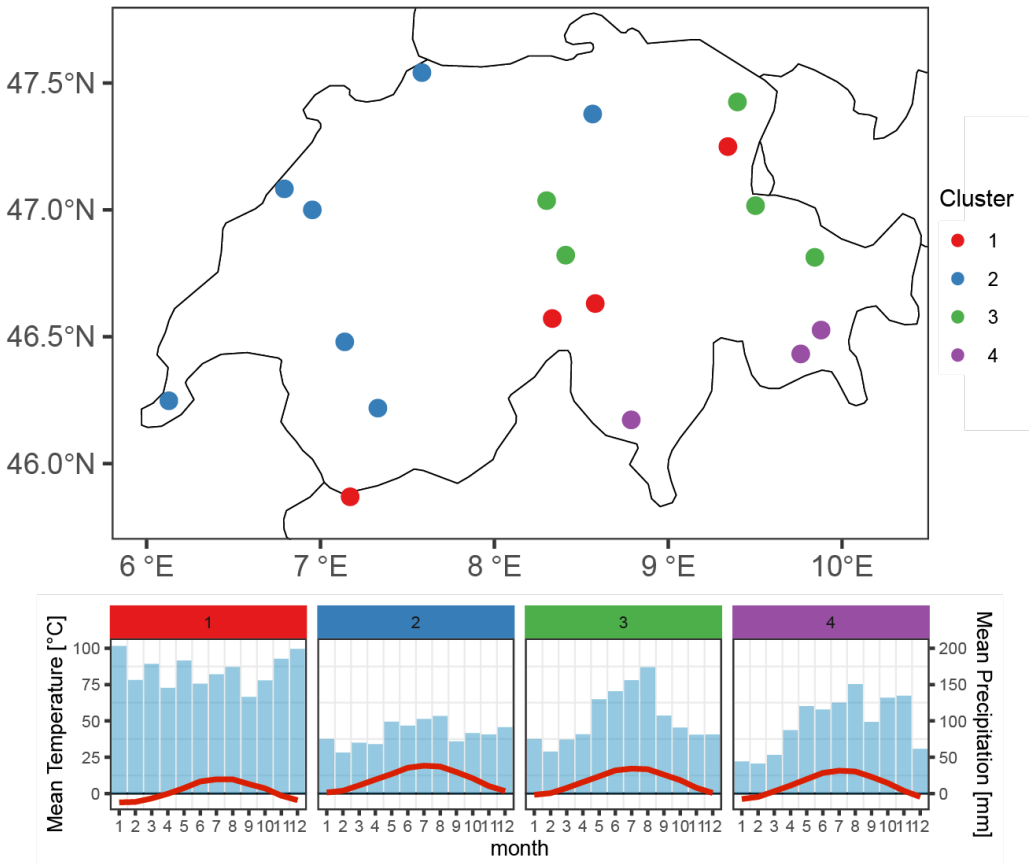

Figure 8: A) Map of all stations, color indicates cluster. B) Monthly mean precipitation and mean temperature by cluster.

4 clusters, they reveal clearly distinct patterns in the seasonal precipitation cycle:

- Cluster 1 – High Alps - receives extraordinarily high amounts of precipitation throughout the full course of the year. The climate diagram shows a clear offset in total precipitation in comparison to the other

three clusters. This is related to high amounts of orographic precipitation which the four stations receive due to their high elevation (between 1438 and 2501 m.a.s.l.).

- Cluster 2 - Jura and western Plateau - shows much lower amounts of precipitation. In contrast to most Swiss stations which receive more precipitation during the summer months (MeteoSwiss, 2024), the precipitation in cluster 2 is almost equally distributed to all months of the year. The equally high winter precipitation can mostly be imputed to the stations in the Jura. They receive considerable amounts of precipitation from low-pressure areas, which pass through Switzerland much more frequently in the winter months due to the position of the jet stream (Baeriswyl & Rebetez, 1997).

- Cluster 3 - eastern Plateau and eastern Alps - receives maximum precipitation during the summer months, caused by thunderstorms. The winter precipitations are less pronounced than for the Jura, as the intensity of low pressure systems decreases as they move eastwards.

- Cluster 4 - south Alps - shows a similar peak in summer precipitation as cluster 3, but receives additionally high amounts of precipitation during the autumn months. This can be attributed to the influence of foehn situations during this season, causing high amounts of precipitation (Baeriswyl & Rebetez, 1997).

The clear contrast between the hydroclimate of the 4 clusters indicates that the combined PCA and clustering process was successful in differentiating groups with distinct climate regimes.

3.4 Comparing cluster means using ANOVA

In order to determine whether the presence of distinct climatic regimes in Switzerland leads to variations in mean annual precipitation, we compared the yearly means of precipitation values for each cluster defined in the previous section (3.3). The examination revealed significant variations in average yearly precipitation among most of the different clusters. The diagram in Figure 9 illustrates the average yearly precipitation levels observed at various stations within a specific cluster. Among these, the alpine stations in cluster 1 stand out with the highest values (with a median of 1966.6), a figure matched only by the high end values in clusters 3 and 4. Clusters 2, 3,

and 4, on the other hand, have median values of 995.4, 1261.0, and 1108.2, respectively.

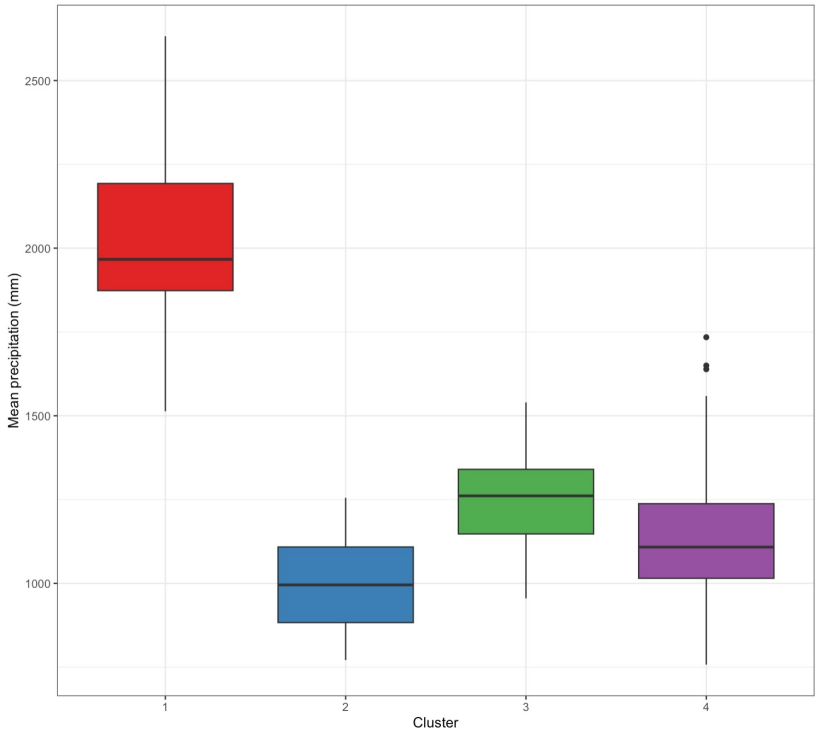

Figure 9: Boxplot of yearly means of precipitation. Grouping is made by clusters defined in section 3.3

.

3.4.1 Welch's test

As the conditions of normality and independence were satisfied, we could performed Welch's test on the yearly averaged precipitation data of each cluster. The null hypothesis (H_0) suggested that there is no difference in means between the groups, while the alternative hypothesis (H_1) proposed that there is a significant difference. The test resulted in a p-value of less than 0.005 ($p < 0.005$). This extremely small p-value provides strong evidence against the null hypothesis, allowing us to reject H_0 and conclude that there is a statistically significant difference between the means of the four groups at a significance level of 0.005. This result follows the conclusions drawn in section 3.3. Annual precipitation values between the clusters show significant

variation in magnitude (Fig. 9). This is expected for a country with high topographical variations and influence from different climatic regimes.

3.4.2 Games-Howell test

The null hypothesis (H_0) for the Games-Howell test suggests that there is no difference between the means of two given groups. The outcomes of the Games-Howell test reveal notable variances in yearly precipitation among the majority of group pairs that were compared (Table 2). Only the cluster

Table 2: Games-Howell test results for pairwise comparisons of cluster mean precipitation between groups. *P-value* adjusted using Tukey's method. *Estimate* corresponds to the estimated difference in means of ($2^{nd} - 1^{st}$ pair). * relate the significance level. "ns" = non-significant.

Pair	Estimate	Conf. Low	Conf. High	P-val	Significance
1 vs 2	-1029	-1198	-860	< 0.005	****
1 vs 3	-779	-949	-610	< 0.005	****
1 vs 4	-865	-1070	-660	< 0.005	****
2 vs 3	250	139	361	< 0.005	****
2 vs 4	165	1.22	328	0.048	*
3 vs 4	-85.4	-250	78.9	0.508	ns

pair 3 and 4 do not dismiss the null hypothesis (H_0) at a significance level of 0.05. Cluster 1 displays the most pronounced deviations from the other clusters with mean precipitation values close to double of the three other cluster. This observation is expected, considering that the mountainous areas within cluster 1 exhibit noticeably elevated precipitation levels compared to the remaining low to mid-altitude regions of Switzerland. Results for clusters 2 and 3 also reject H_0 at high significance level. High levels of precipitation during summer months affect cluster 3, while cluster 2 is less sensitive to the summer thunderstorms. This suggests that the difference in summer precipitations is the main discriminant between these two groups. Interestingly, while cluster 2 shows a statistically significant difference in precipitation compared to cluster 4 ($p = 0.048$, low significance), cluster 3 does not ($p = 0.508$). This suggests that while cluster 3 receives slightly more precipitation than cluster 4, there is no significant difference in precipitation levels between cluster 3 and cluster 4. This result highlights the similarity of the overall averaged precipitation rates between these two groups that show

distinctive climate regimes. A wider implication of this outcome is that despite variations in precipitation patterns over time between two climatic regions, the amount of precipitation can be equal, leading to the existence of distinct climatic systems with equivalent annual precipitation levels.

4 Conclusion

In this report, we successfully identified 4 hydroclimatic regions within our network of 19 stations: (1) the high Alps, (2) the Jura and western Plateau, (3) the eastern Plateau and eastern Alps, (4) and the south Alps. All 4 were classified by an approach that combined PCA and cluster analysis. In a first step, we performed a PCA on the 19 time series of precipitation. The goal was to identify and save only the most relevant precipitation patterns from the dataset, specifically those that explain the greatest amount of variability. The cluster analysis was then performed on the correlations of the individual stations with the PCs. As a third step, we compared the annual precipitation between the clusters to see whether they were also distinguishable by this variable.

Through PCA, we extracted 4 principal precipitation patterns that account for 86 % of the total variance in the time series. Our interpretation suggests that they can be related to four main mechanisms that control precipitation in Switzerland. Namely, they could potentially represent (1) the influence of NAO and the position of the jet stream, (2) the north-south division by the alps, (3) the east-west contrast in continentality and (4) elevation.

The precipitation in the four clusters is controlled by those components to a different extent. This leads to clearly distinct precipitation patterns over the course of a year:

- The high Alps are characterized by constant high amounts of orographic precipitation
- The Jura and western Plateau receive equal amounts of winter precipitation through low-pressure systems and summer precipitation through thunderstorms
- The eastern Plateau and eastern Alps are dominated by summer precipitation due to higher continentality

- The south Alps reveal a similar pattern as cluster 3, but with additional autumn precipitation through foehn situations

Additionaly, we carried out an ANOVA to assess whether also their mean annual precipitation varies significantly. Our analysis showed that the annual mean values are distinct among the clusters, except for clusters 3 and 4. Their annual precipitation levels are not statistically distinguishable, despite differing precipitation regimes and seasonal cycles. This result validates the combination of PCA and cluster analysis as an effective technique for grouping precipitation records based on the underlying precipitation regimes. The PCA is a powerful tool to reduce noise in the climate data, extract only the most relevant patterns and thereby retrieve meaningful clustering results – even within groups that are not distinguishable by mean value.

In order to even better recognise the spatial patterns in precipitation, it would however be advisable for further analyses to include additional stations from the MeteoSwiss network. This could contribute to a clearer extraction of the most relevant PCs and to a finer structure within the clusters, as it would fill spatial gaps within our dataset of 19 stations. As a consequence, it might be possible to retrieve a more precise and smaller scale division into clusters with consistent precipitation regimes, contributing to better prediction of hazards and understanding of the hydrological cycle in a world of a changing climate.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche & V. Vianu (Eds.), *Database theory — icdt 2001* (pp. 420–434). Springer Berlin Heidelberg.

Baeriswyl, P.-A., & Rebetez, M. (1997). Regionalization of precipitation in switzerland by means of principal component analysis. *Theoretical and Applied Climatology*, *58*(1), 31–41. https://doi.org/https://doi.org/10.1007/BF00867430

Ben-Hur, A., & Guyon, I. (2003). Detecting stable clusters using principal component analysis. In M. J. Brownstein & A. B. Khodursky (Eds.), *Functional genomics: Methods and protocols* (pp. 159–182). Humana Press. https://doi.org/10.1385/1-59259-364-X:159

Blanca Mena, M. J., Alarcón Postigo, R., Arnau Gras, J., Bono Cabré, R., & Bendayan, R. (2017). Non-normal data: Is anova still a valid option? *Psicothema*, *29*(4), 552–557.

Govender, P., & Sivakumar, V. (2019). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, *11*. https://doi.org/10.1016/j.apr.2019.09.009

Johansson, B., & Chen, D. (2003). The influence of wind and topography on precipitation distribution in sweden: Statistical analysis and modelling. *International Journal of Climatology*, *23*(12), 1523–1535. https://doi.org/https://doi.org/10.1002/joc.951

Jolliffe, I. T., & Philipp, A. (2010). Some recent developments in cluster analysis [Classifications of Atmospheric Circulation Patterns – Theory and Applications]. *Physics and Chemistry of the Earth, Parts A/B/C*, *35*(9), 309–315. https://doi.org/https://doi.org/10.1016/j.pce.2009.07.014

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

MeteoSwiss. (2024). The climate of switzerland. https://www.meteoswiss.admin.ch/climate/the-climate-of-switzerland.html

Pedersen, L., Jensen, N. E., Christensen, L. E., & Madsen, H. (2010). Quantification of the spatial variability of rainfall based on a dense net-

work of rain gauges. *Atmospheric Research*, *95*(4), 441–454. https://doi.org/https://doi.org/10.1016/j.atmosres.2009.11.007

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Shingala, M. C., & Rajyaguru, A. (2015). Comparison of post hoc tests for unequal variance. *International Journal of New Technologies in Science and Engineering*, *2*(5), 22–33.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychol. Methods*, *8*(3), 254–274.