UNIVERSITÄT BERN
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

FINAL PROJECT REPORT

Statistics for Climate Sciences I & II
Dr. M. Piot & Dr. S. Allen
HS2021-FS2022
June 7th, 2022

Comparison of Recent Climate Normal Periods in Switzerland (1980-2009 and 1990-2019)

Stella Bērziņa[1] | Violet Buxton-Walsh[1] | Sol Kislig[1]

[1]Oeschger Centre for Climate Change Research (OCCR) , Graduate School of Climate Sciences, University of Bern, Switzerland

We compare and contrast the two most recent Climate Normal periods (1980-2009 and 1990-2019) for Switzerland. Hotelling's T^2 test determines climate variables significantly differ between periods, and t-tests find sunshine duration and wind speed differ at many stations. We perform multiple linear regression to explore factors contributing to a change in sunshine duration across periods yielding inconclusive results. We use cluster analysis to explore which stations experience similar changes between normals, and meaningfully interpret the results based on station locations and uniquely regional climate changes.

KEYWORDS
Climate Normals, Hotelling's T^2 Test, Multiple Linear Regression, Clustering Analysis

Abbreviations: CLT, Central Limit Theorem; WMO, World Meteorological Organization; CN, Climate Normal; (M)LR, (Multiple) Linear Regression; MVN, Multivariate Normal(ity); LSE, Least Sqaures Estimation; MSE, Mean Squared Error; ML, Machine Learning; PCA, Principal Component Analysis, EDA; Exploratory Data Analysis.

1 | INTRODUCTION

Climate normals are long-term averages of meteorological parameters which reflect the typical climate of a particular region over a certain time frame. Traditionally, current weather events are compared to thirty-year-long normals to make statements about how they compare to typical climate. Due to an increased rate of change in climatic variables as the result of anthropogenic climate change, the World Meteorological Organization (WMO) recently began updating the recommended dates to use for calculating normals every ten years, instead of every thirty years (MeteoSwiss, 2022). In January 2022 the newest normal was introduced to Switzerland by the Swiss Federal Office of Climatology and Meteorology (MeteoSwiss), shifting the old normal forwards to encompass the most complete thirty year dataset available. Previously, reference periods were compared to the period from 1980-2009 while they are now compared to the period from 1990-2019.

To compare statements about climate change in Switzerland made before and after the shift, it is helpful to understand how these periods differ. Upon investigation, MeteoSwiss found that although two thirds of the data used (data from 1990-2010) overlaps between the two periods, the climate variables measured still show significant differences (MeteoSwiss, 2021). Specifically, mean temperature increased by 0.4-0.5C and sunshine duration increased by 5-10% in central Switzerland between the periods, while there was no significant change in precipitation. Thus difference matters for the interpretation of current weather, because the "normality" of current weather depends on which climate period it is compared to. For example, the rainy Summer of 2021 would be considered 0.5C warmer than average in reference to 1980-2009, but 0.1C cooler than average when compared to 1990-2019 (MeteoSwiss, 2021).

This analysis explores the question *How do climate variables differ between the two climate normal periods (1980-2019 and 1990-2019) in Switzerland?*. To answer this, we ask (1) Are the climate variables significantly different between the two periods? (2) What are the causal drivers of any significant differences between the periods? (3) Which weather stations experience similar changes, and are these what we would expect based on their geography? Establishing these differences is critical for contextualizing current weather data, and poses an interesting real-life question which can be explored in spite of the limitations of this analysis.

2 | METHODS

2.1 | Data

The multivariate dataset for this analysis includes monthly averages of nine climatic variables from twelve Swiss monitoring stations from 1971 to 2021, and is provided by the Swiss Federal Office of Meteorology and Climatology. We assume the stations represent the full range of potential climatic conditions across Switzerland. Analysis is performed in R.

2.1.1 | Manipulations

To compare old and new normals, we subset the data into periods from 1980 through the end of 2009 and 1990 through the end of 2019. We make the simplifying assumption that the ten years of unique data from the subsamples of 1980 through the end of 1989 and 2010 through the end of 2019 are representative of the overall difference between periods, and use these subsamples for our analysis to avoid the potential for linear dependence which arises when the

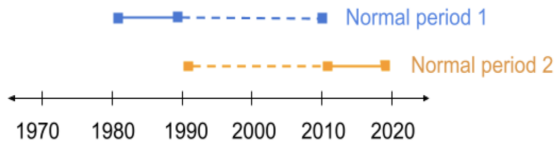

FIGURE 1 The two climate normal periods. The dotted line indicates overlap, and the solid line indicates the ten year subsamples CN1 and CN2 used for this analysis.

samples share two thirds of their data. This step was necessary to meet the requirement of statistical tests conducted that samples be independent. Missing data was in-filled with mean variable values unless otherwise specified. This limits those conclusions which rely on missing data, e.g. from pressure measurements in St. Gallen and Luzern, which are absent for seven years in CN1. Unless otherwise specified, any reference to subsamples, CN/CN1/CN2/CNs refers to these ten year periods. We exclude the cloud cover as a variable at every point in the analysis based on its absence from multiple stations for the second period.

2.1.2 | Descriptive Statistics

To visually characterize relationships between climate variables and stations we first perform exploratory data analysis and generate summary statistics for each CN subsample (Manipulations 2.1.1). We generate scatter, density, and box plots as well as correlation metrics, and check for multivariate normality (MVN).

First, scatter plots show the relationship between two variables and list a correlation coefficient indicating the strength and sign of a linear correlation. These plots are useful for identifying outliers and understanding how variables co-vary. Next, density plots visualize the variance of observations around their mean helping to visualize the spread and skew of data. For normally distributed variables, density plots should resemble a normal distribution. Lastly, box and whisker plots compare median values, showing the skew of data by marking the median with a central line and giving quartiles at the ends of the "box" for 25% and 75% intervals. "Whiskers" extending from the box mark minimum and maximum values (range), making the plots especially useful for comparing the distributions across groups (or in our analysis, stations).

Because many multivariate statistical tests depend on MVN, we check that this condition holds for our subsamples. MVN can be verified graphically with quantile-quantile (Q-Q) plots, which plot the quantile standardized, ranked data against that of a normal distribution so that plots with straight diagonal lines indicate normally distributed variables. However, while variables in a multivariate normal dataset are also univariate normal, this is not necessarily true in reverse, and we cannot verify MVN on this basis. We therefore implement the less subjective, non-graphical Henze-Zirkler's test, "computes non-negative functional distance under the null hypothesis that the test statistic is approximately log-normally distributed" to test for MVN (R Project 2021).

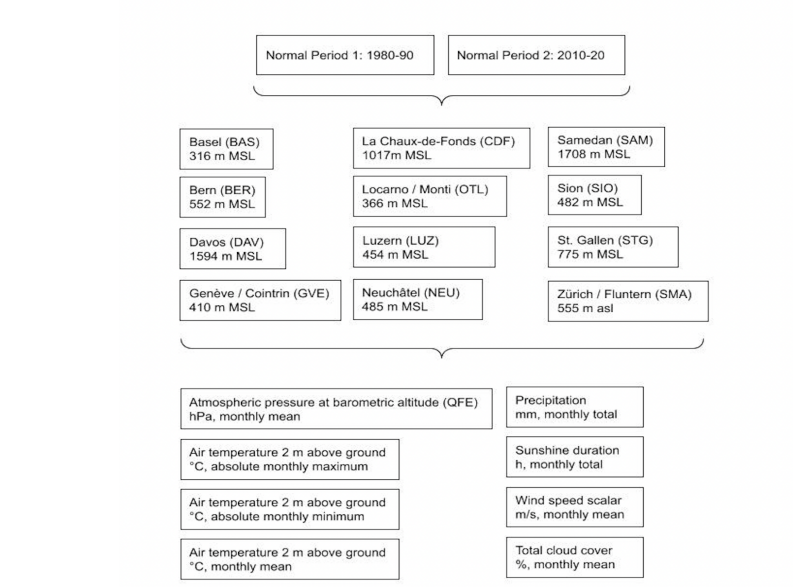

FIGURE 2 Visual overview of data structure, measurement units and frequencies used for this analysis

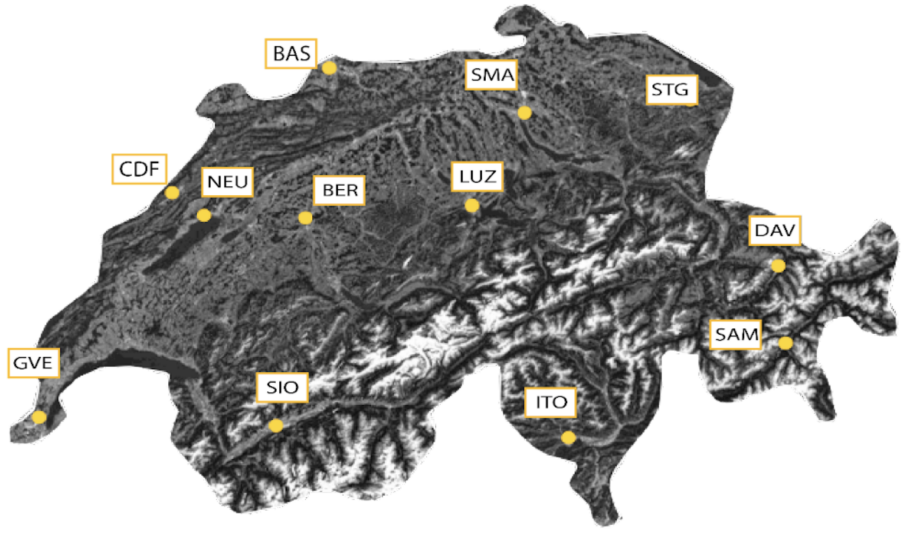

FIGURE 3 Map of measurement stations over a satellite map of Switzerland.

2.2 | Analysis

2.2.1 | Hotelling's T^2 Test

We explore the first sub-component of our research question, which aims to identify which variables differ significantly between periods. Answering this requires comparing eight climatic variables for twelve stations, and done via a multivariate comparison of the mean vectors for each CN. We compare the periods with a Hotelling's T^2 test, which uses a T^2 test statistic to compare multiple unpaired variables within a sample, or multiple paired or unpaired variables across two samples. In our case, each ten year period is a sample and all data is unpaired.

We also verify the formal statistical requirements of normally distributed data and equal variances. MVN is assumed based on a large sample size (n=120) which makes the CLT applicable (see explanation in 3.1). Equal variance is established by comparing the variance matrices of the samples to demonstrate a common variance-covariance matrix Σ. In our analysis the matrices agreed in terms of order and sign, leading us to assume equality.

Hotelling's T^2 test assumes two independent samples $X_{11}, \ldots X_{1n}$ and $X_{21}, \ldots X_{2n}$ with size n_1 and n_2 from p-variate populations and respective mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2, it is a essentially a multivariate generalization of the t-test (Piot, 2020). The test requires (1) defining the sample mean vectors, (2) calculating sample covariance matrices, (3) pooling into an estimator of a common variance-covariance matrix S_{pooled}, and using this to define a T^2 test statistic which approximately follows an F-distribution with n and $n - p$ degrees of freedom. If the condition in (4) holds, the null hypothesis is rejected.

$$\bar{x}_1 = \Sigma_{i=1}^{n_1} \frac{x_1}{n_1}, \bar{x}_2 = \Sigma_{i=1}^{n_2} \frac{x_2}{n_2} \tag{1}$$

The covariance matrices for each sample are given by:

$$S_1 = \frac{1}{n_1 - 1} \Sigma_{i=1}^{n_1} x_{1j} - \bar{x_1} - x_{1j} - \bar{x_1}, S_2 = \frac{1}{n_2 - 1} \Sigma_{i=1}^{n_2} x_{2j} - \bar{x_2} - x_{2j} - \bar{x_2}\prime \tag{2}$$

These matrices are pooled into a common covariance S_{pooled} as an estimator of Σ. This gives the T^2 statistics as follows:

$$T^2 = \frac{\bar{X} - \bar{Y} - \mu_X - \mu_Y}{S_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3}$$

Ultimately the null hypothesis is rejected in favor of the alternative hypothesis if:

$$T^2 > \frac{n_1 * n_2}{n_1 + n_2} * F_{p\ldots} \tag{4}$$

In this report, p-values are reported in favor of test statistics for ease of interpretation. Generally speaking, p-values describe the probability of obtaining the observed results given a true null hypothesis, and lower values imply greater statistical significance. The calculation of p-values requires we know the sampling distribution of the test statistic under H0, a sample of data, and the type of test being conducted (lower-tailed, upper-tailed or two-tailed). For Hotelling's T^2 test, H0 states there is no statistically significant difference between any pair of values taken from the two mean vectors at confidence level α, and is rejected in favor of H1 if any pair of values significantly differs between the mean vectors.

An advantage of this method is that running one multivariate test instead of several univariate comparisons avoids the inflation of type one errors, where the null hypothesis is rejected purely by chance (Rencher, 2002). Running eight separate univariate tests at the significance level =.05 would make the probability of at least one false rejection of the null hypothesis greater than =.05, and is avoided by instead conducting one multivariate test.

Broadly speaking, multivariate tests take advantage of correlations between variables and have greater statistical power over univariate tests. For climate data where variables are often correlated, this is especially useful. Multivariate testing also accounts for the inflation of α resulting from univariate tests as per Rao's paradox, where points inside the ellipse but outside the rectangle will be rejected in at least one univariate dimension but will be accepted multivariately (Figure 4). This figure also illustrates the greater discriminatory power of multivariate tests, since points inside the rectangle but outside the ellipse (dark gray area) are rejected in the multivariate case but accepted in the univariate case, resulting in a higher probability of incorrectly rejecting the null hypothesis in the univariate case.

Generally, conducting univariate tests with multivariate data becomes acceptable after the multivariate testing yields significant results. We therefore apply univariate comparisons in the form of individual t-tests on the basis of significant results from Hotelling's T^2 tests. Pairwise tests help determine more specifically which variables differ between the compared samples, and supplements the findings of Hotelling's T^2 , which only identifies the presence or absence of a significant change and not its composition. In pairwise testing, the Holm-Bonferroni correction helps control for the probability of inappropriately rejecting one or more null hypotheses where they should be actually retained (type one error). We report both uncorrected and corrected results. The significance level α must be set prior to a hypothesis test, and gives the maximum acceptable probability of incorrectly rejecting H0 (type I error). When α is uncorrected, we use the standard value of 5%.

The Holm-Bonferroni correction is performed as follows: p-values of m multiple comparisons are to be ordered by increasing size, and each p-value p_k is compared to $\frac{\alpha}{m+1-k}$. H0 is rejected when $p_k < \frac{\alpha}{m+1-k}$, and the next greatest p-value is examined with every iteration until H0 is no longer rejected.

Similarly, the Bonferroni correction divides alpha by the number of tests m to be conducted and compares each p-value to $\frac{\alpha}{m}$. Although this protects against type one error, it increases the probability of type two error in which a null hypothesis fails to be rejected, and is therefore regarded a conservative correction. The Holm-Bonferroni method results in a smaller increase of type two error because the alpha against which the test is compared is unique to each test, and is therefore preferred for this analysis.

2.2.2 | Multiple Linear Regression

Based on the results of variables shown to change significantly between CNs in Hotelling's T^2 test (section 2.2.1), we are able to answer the second sub-question guiding our report, which explores the drivers of significant differences between CNs. Linear regression (LR) helps answer this question by assessing the relative contributions of independent variables as predictors of significant changes between CNs, and understanding the relationships between variables. MLR is the multivariate case of single linear regression, which models the relationship between the expected value of a continuous response variable and independent (explanatory) variables (Piot 1.3.3). Although MLR covariates from a randomized controlled experiment can sometimes imply causality, our empirical analysis only demonstrates correlation. A model with r independent variables for a dataset with n observations with the error term ϵ and the predicted outcome

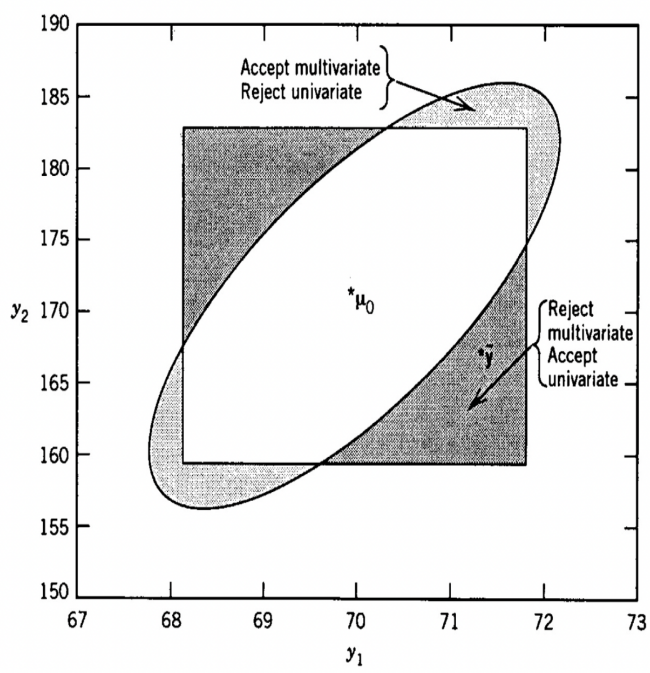

FIGURE 4 From Rencher (2002, p.116): Acceptance and rejection regions for univariate and multivariate tests. *Box*: Confidence region of univariate test statistic. *Ellipse*: Confidence region of multivariate test. *Dark grey areas*: Show that alpha is unnecessarily inflated when univariate tests are applied. *Light grey areas*: Show that there is an area of significance covered by the multivariate confidence region that is not captured by the univariate one.

y_n is given linearly (1) and in matrix notation (2, 3). represents the difference between an observed value of y and the value predicted by the model as driven by measurement error or missing explanatory variables.

$$y_n = \beta_0 + \beta_1 * x_{n1} + \beta_2 * x_{n2} + ...\beta_n * x_{nr} + \epsilon_n \tag{5}$$

$$\mathbf{Y} = \mathbf{Z}\beta + \epsilon \tag{6}$$

$$\begin{bmatrix} Y_1 \\ .. \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & .. & z_{1r} \\ .. & .. & .. & .. \\ 1 & z_{n1} & .. & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ .. \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ .. \\ \epsilon_n \end{bmatrix} \tag{7}$$

The vector of -values holds unknown regression coefficients representing the change in yi changes if the independent variable to which corresponds increases by one unit, and other values are zero. -values do not represent the influence of a particular independent variable on the outcome when independent variables have different units. In most cases the value of 0 is only meaningful as the predicted outcome when independent variables are all zero, and it primarily acts to minimize residuals by shifting the regression line vertically.

MLR aims to compute the vector of predicted values Y, and therefore requires regression coefficients and error terms to be calculated with Least Squares Estimation (LSE) (Piot 7.2). LSE approximates the solution to a set of MLR equations by minimizing the residual sum of the squares for every individual equation. Minimizing the sum of the squared residuals S estimates the terms as follows if Z has full rank (Piot Chp.7).

$$S = \Sigma_{j=1}^{n} y_j - \beta_0 - \beta_1 z_{j1} - ... - \beta_r z_{jr} \hat{} 2 = \left(y - \mathbf{Z}\beta\right) \prime \left(y - \mathbf{Z}\beta\right) \tag{8}$$

$$\hat{\epsilon} = y - \mathbf{Z}\beta \tag{9}$$

$$\hat{\beta} = \left(\mathbf{Z}\prime\mathbf{Z}\right) \hat{} -1 \mathbf{Z}\prime\mathbf{y} \tag{10}$$

Once calculated, these values are used to fit a linear regression model for predictive use. Model fit is evaluated with R^2 values which describe the proportion of the total variation in the y-values explained by the predictor variables, larger R^2 values imply better fit.

$$R \hat{} 2 = 1 - \frac{\Sigma_{j=1}^{n} \hat{\epsilon_j} \hat{} 2}{\Sigma_{j=1}^{n} \left(y_j - \bar{y}\right)^2} = \frac{\Sigma_{j=1}^{n} \left(\hat{y_j} - \bar{y}\right)^2}{\Sigma_{j=1}^{n} \left(y_j - \bar{y}\right)^2} \tag{11}$$

In addition to MVN of the data to which it is applied, linear regression assumes normally distributed independent errors with constant variance whose expected value should be equal to zero. Intuitively, unbiased model error is zero because predicted y-values under and overestimate the observed outcome the same rate and magnitude. Independent

variables should be uncorrelated with each other and with the error term, and observations of the error term should also be uncorrelated with each other.

$$\mathbb{E}\left(\epsilon_i\right) = 0, \mathbb{V}\mathrm{ar}\left(\epsilon_i\right) = \sigma\hat{}2, \mathbb{C}\mathrm{ov}\left(\epsilon_i, \epsilon_j\right) = 0, i \neq j \tag{12}$$

$$\mathbb{E}\left(\epsilon\right) = 0, \mathbb{C}\mathrm{ov}\left(\epsilon\right) = \mathbb{E}\left(\epsilon\epsilon I\right) = \sigma^2\mathbf{I} \tag{13}$$

Several techniques exist for improving MLR accuracy and identifying biases. Plotting residuals against predicted Y-values can indicate non-constant variance, and the need for transformation of the response value to better approximate a linear function. Variance stabilizing transformations can also be applied to independent variables. Correlation between residuals indicates the need for more explanatory variables, and time dependence would recommend using other analysis methods. Resampling methods, cross validation, and feature engineering can also improve prediction accuracy for limited datasets (Piot 7.5.3).

The accuracy of MLR is tested via an analysis of variance F-test which finds the proportion of variability explained by a model. When the following F-statistic is larger than the 1α quantile of an F-distribution with r and $nr + 1$ degrees of freedom, H0 is rejected and a linear relationship between the response and at least one independent variable is assumed. Individual regression parameters can be investigated similarly with a t-statistic, where H0 is rejected if the T-statistic is larger than the 1α quantile of a T-distribution with $nr + 1$ degrees of freedom.

MLR is limited by assuming a linear relationship between variables. Real data is rarely linearly separable, and underfitting of complex datasets is common. Overfitted models can also arise (usually due to multicollinearity, or having fewer observations than features), but this can be reduced by removing correlated predictor variables and regularization. MLR is also sensitive to outliers, though this matters less when applied to averages (as in this implementation).

2.2.3 | Clustering

Clustering is a method used to form sub-groups with meaningfully different characteristics from a pool of observations. To address the final sub-component of our research question, which asks which stations experience similar changes between CNs and if this might be explained by geographical location or altitude, we cluster by a "difference matrix" containing the difference between mean vectors of the two CNs. By averaging all variables at every station in each CN we create two mean vector matrices, which when subtracted give the difference in average variable values between periods for all stations, allowing us to assess changes between CNs by station. The resulting matrix is scaled as required for clustering.

We primarily rely on non-hierarchical clustering, which differs from hierarchical clustering in that it does rely on a similarity metric and begins with a pre-specified number of K clusters (or a random set of seed points for K initial cluster nuclei). K-means clustering assigns points to the cluster with the nearest centroid based on the (typically Euclidean) distance, while recalculating the centroid for every cluster when a new point is assigned, and repeating until all data is clustered.

Hierarchical clustering is a class of techniques which can be either agglomerative (grouping individual observations by

similarity) or divisive (splitting up a pool based on similarity), and can be conducted either for groups of items (using a distance metric) or variables (using a measure of association). The method operates by creating initial N clusters, calculating an NxN symmetric matrix of distances between all clusters, regrouping based on the "nearest" cluster seen in the matrix, updating the distance matrix, and repeating N-1 times until all objects are clustered. Dendrogram plots display the results of hierarchical clustering methods for every nth iteration, marking wherever a merger or division was made. Hierarchical methods are computationally intensive because they do require the re-calculation and storage of distance matrices. Therefore non-hierarchical methods are preferred for large datasets. Importantly, a distance metric is required to group the "difference matrix" by stations (items), for which we use average linkage. Other possible metrics include single linkage (minimum distance), complete linkage (maximum distance), and Ward's method (minimizing loss of information). We prefer average linkage (average distance between all pairs of items) because it compromises between the overgroupping tendency of single linkage and sensitivity to outliers from complete linkage.

In K-means clustering, final clusters can vary based on the initial partitioning of K clusters, so best practice suggests rerunning the clustering with multiple initial partitions to check their stability. This helps avoid poorly distinguished groups caused by multiple seed points inadvertently lying in a single cluster. For datasets where K has real meaning, certain sampling methods may also cause rare groups to not appear in the sample, making the value of K a nonsensical. Understandably, K-means is also affected by outliers which produce clusters with disparate points, but this is not a major concern for our data. In general, cluster analysis is designed to group data but is limited in its inability to explain particular grouping.

We can justify clustering by all variables even if they are strongly correlated in the original dataset, because they are not strongly correlated in the distance matrix by which we cluster (described further in 3.4).

3 | RESULTS AND DISCUSSION

3.1 | Descriptive Statistics

Descriptive statistics tables were constructed for all twelve stations yielding similar results. Figure 5 displays results for Bern, as this was the station for which we conducted MLR.

Correlation coefficients and scatter plots indicate some variables (e.g. global radiance and sunshine hours) are closely correlated while others are largely uncorrelated (e.g. mean scalar wind and maximum air temperature) (Figure 5). Based on visual inspection of figure 5 we design our regression model to avoid multicollinearity by using air pressure, mean air temperature, precipitation, and mean wind speed as covariates. Overall, similar correlation coefficients and overlapping scatter plots suggest similarity between CNs. However, density plots also illustrate differences in variables such as sunshine and wind. As a result we hypothesize only a few variables are significantly different between CNs, and proceed with a comparison of means (Hotelling's T^2 test) to quantify likely differences.

From the box plot comparison of stations in both periods we also note that not all stations record similar changes in variables between the CNs (Figure 6). In Basel for example, median wind speed is noticeably lower in CN2, while the reverse is true for St.Gallen. This prompts us to perform clustering analysis on the differences between CNs to identify which stations have changed similarly.

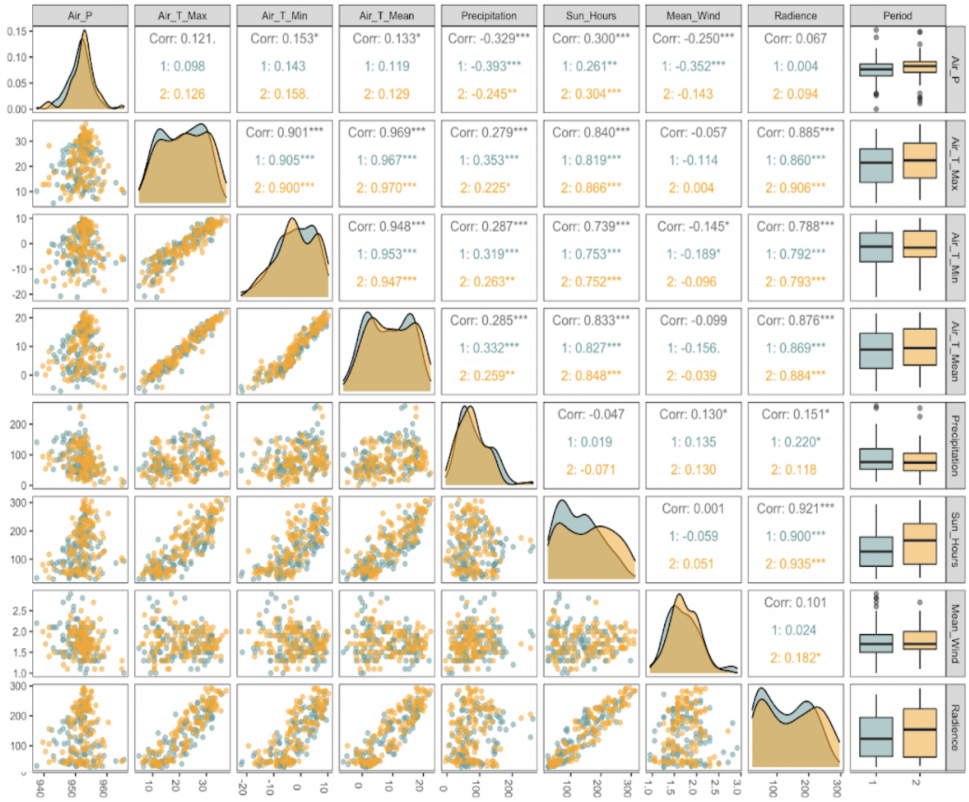

FIGURE 5 Summary descriptive statistics table for Bern. The blue colour indicates the first CN period (1980-1989), orange colour indicates the second CN period (2010-2019). The lower diagonal region shows scatter plots between every two climate variables. The main diagonal shows variable distributions in the 2 reference periods. The upper diagonal region shows three linear correlation coefficients, grey correlation indicates the correlation between variables in both periods together, blue indicates the correlation between variables in the first CN, orange indicates the correlation in the second CN. The last column shows a box plot comparison of variables between the two CN periods.

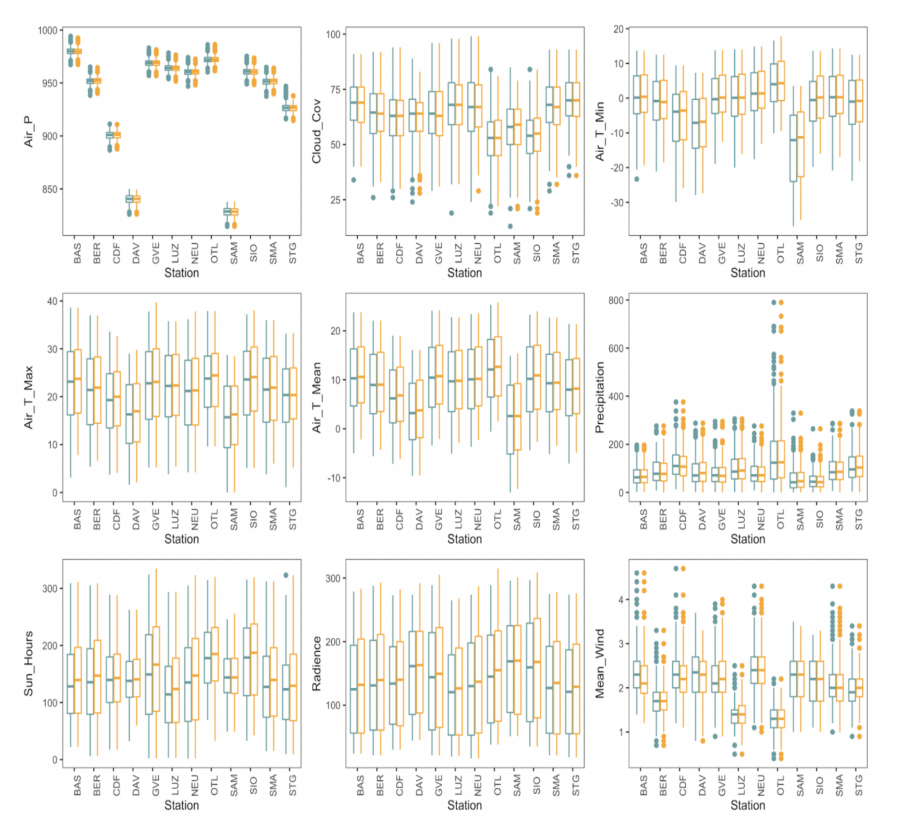

FIGURE 6 A box plot comparison of variables between stations and reference periods. The blue colour indicates the first reference period (1980-1990), orange colour indicates the second reference period (2010-2019). Y axis states the compared variable.

TABLE 1 Stations ordered by decreasing p-value and corresponding alpha level of comparison for the corrected Hotelling T^2 test statistic.

Station	p=value	alpha (corrected)
Zürich / Fluntern (SMA)	0.033193	0.05
Neuchâtel (NEU)	0.02312	0.025
Luzern (LUZ)	0.012642	0.016667
Bern / Zollikofen (BER)	0.00940	0.0125
Sion (SIO)	0.00090	0.01
Locarno / Monti (OTL)	0.000490	0.00833
Genève / Cointrin (GVE)	0.000184	0.007143
La Chaux-de-Fonds (CDF)	1.44E-04	0.00625
Samedan (SAM)	7.59E-06	0.005556
Davos (DAV)	7.45E-13	0.005
Basel / Binnigen (BAS)	1.21E-13	0.004546
St. Gallen (STG)	3.65E-14	0.004167

Testing for MVN indicates no station in any CN has normally distributed climate variables, which is corroborated by density plots (Figure 5). However, by application of the Central Limit Theorem (CLT) we fairly conduct statistical tests which assume MVN. The CLT states that for large samples (high values of n) mean vectors are approximately normal regardless of a variable's original distribution in the sample. More precisely, given a collection of random vectors $X_1, X_2, \ldots X_k$ that are independent and identically distributed, the sample mean vector $\bar{X}$, is approximately multivariate normally distributed for sufficiently large samples. If the $X_1, X_2, \ldots X_k$ are independently sampled from a population with mean vector μ and covariance matrix Σ, then the sample mean vector $\bar{X}$ is approximately multivariate normally distributed with the mean vector and covariance matrix. We assume that n=120 is sufficient to apply CLT.

3.2 | Hotelling's T^2 Test

We test for the variables which differ significantly between CNs by conducting Hotelling's T^2 tests for every station with Holm-Bonferroni correction. We test the null hypothesis H0 that the mean vectors of the eight climate variables in each station do not significantly differ between CNs against the alternative hypothesis H1 that at least one pair of values differs between the mean vectors. Missing data was skipped using na.rm = TRUE in R.

Hotelling's T^2 test indicates H0 can be rejected for all twelve stations with 95% confidence, meaning at least one weather variable significantly differs between CNs per station. This result holds when repeating the test with the Holm-Bonferroni correction to account for the problem of multiple comparisons (see section 2.2.1). Under the correction, p-values for each test are compared to a corrected level of significance α, and when p-values are smaller than alpha that leads us to reject the null hypothesis that mean vectors have no significant difference between CNs. In Table 1, all p-values are smaller than their corresponding corrected alpha, meaning H0 is rejected in every case and at least one weather variable is significantly different for each station after correction. To identify which variables

	BAS	BER	DAV	GVE	CDF	OTL	LUZ	NEU	SAM	SIO	STG	SMA
Atmospheric pressure	3.41E-01	0.06114 4454	0.47381 30196	0.81732 1183	6.26E-01	0.81643 163	0.20187 1469	0.89174 619	0.81643 163	0.20187 1469	0.89174 619	0.41728 794
Air temperature (max.)	8.74E-02	0.23315 4686	0.13237 52055	0.13201 9843	1.10E-01	0.07605 9	0.12762 9275	0.21283 989	0.07605 9	0.12762 9275	0.21283 989	0.41691 965
Air temperature (min.)	2.37E-01	0.69302 0704	0.30944 58135	0.05985 0061	1.62E-01	0.09543 779	0.13769 1495	0.15619 844	0.09543 779	0.13769 1495	0.15619 844	0.01586 478
Air temperature (mean)	8.80E-02	0.18392 5915	0.13589 72248	0.10639 3001	6.97E-02	0.06756 152	0.12524 0766	0.14570 657	0.06756 152	0.12524 0766	0.14570 657	0.12793 4
Precipitation	6.53E-01	0.12510 791	0.14479 779	0.38195 3184	1.40E-01	0.68555 118	0.22192 2228	0.20785 242	0.68555 118	0.22192 2228	0.20785 242	0.39710 943
Sunshine duration	5.77E-02	0.00230 3091	0.13126 84881	0.01512 6369	5.12E-02	0.04952 891	0.00296 0637	0.00615 18	0.04952 891	0.00296 0637	0.00615 18	0.79152 04
Wind speed	3.04E-16	0.73523 4159	0.00015 99742	0.00357 7643	4.46E-08	0.19469 239	0.03522 4592	0.52220 931	0.19469 239	0.03522 4592	0.52220 931	0.52117 364
Global radiation	1.35E-01	0.06149 9246	0.46117 04672	0.15225 6896	8.39E-02	0.05979 937	0.07672 6393	0.06797 755	0.05979 937	0.07672 6393	0.06797 755	0.42713 648

FIGURE 7 P-values for unpaired two-samples t-tests. Blue indicates significant p-values only before correction. Orange indicates p-values that remain significant after applying the Holm-Bonferroni correction.

differ between CNs we conduct 8 unpaired, univariate, two-sample t-tests per station (assuming independence and equal variances based on multivariate equality of variances). Because type one error increases with multiple pairwise tests, the Holm-Bonferroni correction was applied as described for the multivariate comparison. The p-values of the 96 t-tests (twelve stations * eight variables) are summarized as follows. The results of unpaired two-sample t-tests before correction showed significant differences in wind speed for Basel, Davos, Genève, la Chaux-de-Fonds, Luzern, and Sion. Sunshine duration was significantly different for Bern, Genève, Locarno, Luzern, Neuchâtel, Samedan, Sion and St. Gallen. In Zurich, maximum air temperature differed significantly between periods. The results from the unpaired two-sample t-tests with Holm-Bonferroni correction showed that sunshine duration was significantly different between CNs for Bern, Luzern, Neuchâtel, Sion, and St. Gallen. Wind speed was significantly different for Basel, Davos, Genève, and la Chaux-de-Fonds.

Our analysis was unable to reproduce a significant difference in mean temperature found by MeteoSwiss, but does share the finding of no significant difference in precipitation (MeteoSwiss 2022). The significant values pre and post correction show that if multiple testing is not corrected for, we reject H0 in more cases than if the correction is carried out. This occurs due to changes resulting from the size of confidence interval used.

Interestingly, no significant differences were found for Locarno, Samedan and Zurich after correction, meaning these stations yield insignificant results in univariate tests but significant results in a multivariate test. This may be explained by type one errors occurring in the multivariate Hotelling's T^2 test for these three stations. Because we are only 95% confident that we are correctly rejecting H0, there is still a 5% chance of a true null hypothesis being wrongly rejected due to T^2 values falling into the univariate region of rejection and multivariate region of acceptance (see Figure 4 in 2.2.1).

TABLE 2 Results of MLR with sunshine duration as the dependent variable for the two CN periods, yellow values are significant. MLR is done for station Bern. Last four columns are the regression coefficients, all significant at the level of 5%.

Period	F-statistic	Adjusted R^2	β_0	$\beta_1 airP$	$\beta_2 avgTemp$	$\beta_2 precip$	$\beta_3 Wind$
1	102.8	0.77	1602	1.723	9.034	-0.368	27.70
2	153.7	0.83	2619	2.768	10.06	-0.531	37.56

3.3 | Multiple Linear Regression

Based on the finding of significant changes in sunshine duration across the highest number of stations of any variable stations in section 3.2, we use MLR to assess the relative contribution of different variables to sunshine duration. We construct two models (one for each CN) with sunshine hours as the dependent variable and air pressure, total precipitation, mean scalar wind and mean air temperature based as independent variables. Predictors are selected based on their relatively low correlation coefficients (values < 0.7, Figure 4). We look for differences in the regression coefficients between CNs to assess the relative contributions of predictors to the dependent variable, and to explore what might drive significant differences.

Although we intended to implement MLR for values averaged over all stations with significant differences between periods, this was not possible due to missing pressure data in St.Gallen and Luzern. We therefore take Bern as an example station based on cluster analysis which groups Bern with other stations where sunshine hours also significantly differ between periods (section 3.4). We expect the drivers of significant differences in sunshine duration between CNs for Bern to be similar for this group, and interpret the results as likely representative of all stations with significant changes in sunshine hours. We fit the following regression model twice for Bern, once for each CN.

$$y_n = \beta_0 + \beta_1 * airP_n + \beta_2 * avgTemp_n + \beta_3 * precip_n + \beta_4 * wind + \epsilon_n \tag{14}$$

The adjusted R^2 value for both CN periods is high (0.77 and 0.83 respectively), meaning approximately 80% of the variability in sunshine duration can be explained by the independent variables. Using R, we also calculated a significant F-statistic, meaning at least one regression coefficient is significantly different from 0. Based on residual plots with no clear trend and median close to zero, a significant F-statistic value, and high R^2 values we assume high model accuracy. Based on its higher R^2 value, more of the second CN's variability is explainable by the model.

From Table 2 we can make statements such as "In the first CN, a one hour increase in mean sunshine duration will result from a ~28 unit increase in mean wind speed and a ~0.4 unit decrease in precipitation." However, we largely cannot interpret the regression coefficients due to differences in independent variable data units. Helpfully though, we can compare the coefficients between the two models. If a single coefficient exhibits large changes between CNs, we can infer that it is the likely driver of the difference in sunshine duration between periods. Although table 2 does show an increase in the coefficients of the second CN over the first, we could not identify a single variable which seems responsible for the difference between periods, and the difference may instead be driven by a combination of variables or unexplainable from our predictors. From the sign of the regression coefficients, we note that temperature, wind, and air pressure are positively correlated with sunshine hours while precipitation is negatively correlated. This makes intuitive sense because rain necessitates clouds, which decrease sunshine.

The regression coefficients indicate a statistically significant linear dependence of the mean of sunshine duration on air temperature, total precipitation and mean scalar wind speed in both CNs. Air pressure is a statistically significant predictor only in the second CN, for which we do not have a clear explanation. One possibility is that an air pressure difference between CNs is linked to the change in sunshine duration, which can be climatologically interpreted as a change in flow regime that drives differences in pressure and sunshine. Although sunshine duration is well represented by our linear model, we have no conclusive findings as to the drivers of a significant change in sunshine duration between CNs.

3.4 | Clustering

We perform cluster analysis on the "difference matrix" containing the change in period means between CNs for all included variables and stations to evaluate which stations have changed similarly between CNs (see 2.2.3 Methods; Clustering). Because higher altitudes warm faster under climate change, we expect the meteorological stations in our data to experience heterogeneous changes between CNs, which might be elucidated by clustering based on mean differences (CH2018). Clusters should match groups of stations for which we expect larger or smaller changes in variable means between CNs as the result of global warming, and be explainable by factors like location or altitude.

We use non-hierarchical K-means clustering to group the dataset, and display three clusters (Figure 8). Because our dataset has as many dimensions as variables (8) we are unable to plot the clusters across all dimensions. Therefore, our representation of clusters is based on the first two principal components of an eight dimensional dataset, which was calculated while plotting by the R function fbiz() from the factoextra package, and represents the largest axis of variability in the dataset.

In clustering, high correlation among variables should be avoided as it can overly weight arbitrary clusters. Correlated variables can be removed by clustering only the principal components of a dataset. We did not follow this approach, and instead clustered by all eight climatic variables. Yet despite strong correlations between variables in the original dataset, our results hold because the "difference matrix" by which we cluster contains mean differences between CNs, and not the original data. Literature suggests mean changes in climate variables are highly dependent on the station and not especially correlated with one another, meaning that variables exhibiting strong correlations in the raw dataset are not strongly correlated in the difference matrix (CH2018). This makes clustering by all variables without first reducing dimensionality via PCA valid, and creates the possibility of examining the similarity of changes between CNs across stations, instead of only the changes in principal components or particular variables.

Because the number of clusters for K-means clustering is subjective, we explored methods to improve our choice of K. An "elbow plot" of cluster number versus sum of squares error illustrated the fraction of variation explained by any given number of clusters. However, at no x-value did the slope flatten abruptly, indicating no clear point at which additional clusters are unhelpful and at which to set K. We therefore use three clusters based on our ability to meaningfully interpret the results for K = 3.

We check the stability of our clusters by performing both agglomerative hierarchical and non-hierarchical K-means clustering (figure 2 and 1 respectively). The methods yield similar results, which we take to indicate stability. Between the two methods only St. Gallen changes clusters from green in figure x to red in figure x. This is likely explained by its

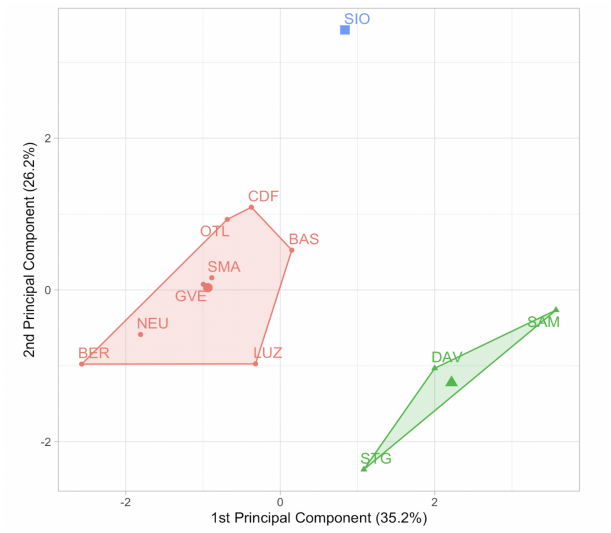

FIGURE 8 Non-hierarchical K-means clustering of stations with K= 3. The X and Y axis represent the first and second principal components of the eight dimensional difference matrix, and must be computed to visualize the clusters in two-dimensions. Red, green, and blue correspond to three different clusters.

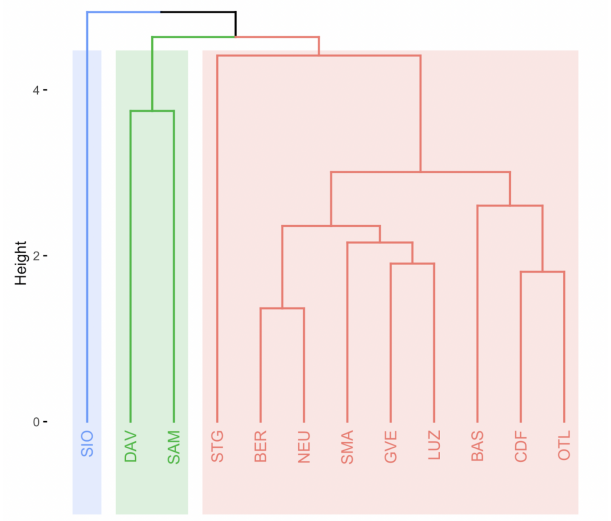

FIGURE 9 Agglomerative hierarchical clustering of stations. Red, green, and blue correspond to three different clusters, which are similar but not identical to the three clusters in Figure x. We find agglomerative and divisive methods to yield similar results.

unique geography as the only low-altitude station within a cluster encompassing the eastern part of Switzerland (green cluster).

K-means clustering (figure x) groups the stations as follows: Cluster 1 - Sion, Cluster 2 - St. Gallen, Davos, Samedan, Cluster 3 - Bern, Basel, Locarno, Luzern, Geneva, La Chaux de Fonds, Neuchâtel, Zürich. Because clustering divides the stations in groups of similar attributes but does not provide an explanation for these groupings, this result is open to our interpretation.

Notably, the first cluster (blue) identifies Sion individually. This might be explained by its location in the South-West alpine region, which is known for relatively low precipitation and high sunshine duration (CH2018). We cannot definitively discern the reason for Sion's separation when clustering by eight variables, but CH2018 does suggest uniquely pronounced warming in this region. Based on the difference matrix, we speculate that an increase in temperature is likely a contributing factor to this result.

The second cluster (green) includes St. Gallen, Davos and Samedan, which are all located in Eastern Switzerland, relatively inland of the third cluster (red). CH2018 groups Davos and Samedan together in an Eastern alpine region, and our clusters also group these stations (presumably due to their high altitude). However, St. Gallen is also included despite lying to the North of the other stations and not being high elevation (CH2018 labels St. Gallen as pre-alpine rather than alpine). This finding is interesting because Switzerland is typically divided by elevation (plateau, pre-alps, alps), but our results show that changes between CNs are similar in Eastern Switzerland regardless of elevation. Although we cannot definitively explain which variables drive this grouping, we take this as evidence of an East/West component to the clusters in addition to elevation.

The third cluster (red) contains the remaining stations. In summary, we find that (1) stations in Eastern Switzerland experienced similar changes between CNs, and (2) changes in the Western Alpine region differ from the rest of Switzerland, likely due to a unique increase in temperature. Overall this indicates detectable geographical variation in the effects of climate change in Switzerland, and is a valuable insight for climate change mitigation and adaptation projects.

4 | CONCLUSIONS

We conclude by remarking on our final answers to the guiding question "How do climate variables differ across the two most recent Climate Normals for Switzerland?" and its subcomponents: (1) Are the climate variables significantly different between the two periods? (2) What are the causal drivers of the significant differences between the periods? (3) Which weather stations experience similar changes, and are these what we would expect based on their geography?

To answer whether climate variables significantly differed between CNs, we compared each station in CN1 and CN2 with Hotelling's T^2 test and subsequent t-tests. The multivariate T^2 tests indicated at least one significant climate variable per station, leading us to conduct univariate t-tests on every climate variable at every station in order to identify which variables significantly differ. The Holm-Bonferroni correction was applied to account for multiple comparisons, and sunshine hours and/or wind speed was found to significantly differ for all 9 stations. While the difference in sunshine duration was previously reported by MeteoSwiss, the difference in wind speeds is novel.

To investigate the causal drivers of changes between CNs we modeled sunshine hours with MLR. Sunshine was selected based on the result of t-tests, which indicated significant changes in sunshine for most stations. Although we report high model accuracy, this analysis was inconclusive. Because all regression coefficients have greater absolute value for CN2 due to an overall higher mean sunshine duration, we could not identify a particular variable driving the change. Likely a combination of all variables is responsible, including ones not captured by our model such as solar fluctuations (CH2018). Due to lack of data, we report results only for Bern, which was selected based on clustering results which group Bern with some of the other stations for which t-tests also indicated significant changes in sunshine.

To address the question of which weather stations experience similar changes we clustered the change in mean vectors between periods by station. This identified Sion as experiencing different changes than the rest of Switzerland, likely due to unique climatic conditions corroborated by CH2018. An Eastern group was composed of St. Gallen, Davos, and Samedan, and indicates our clusters are not purely a function of elevation. Although Switzerland is typically climatically grouped by altitude, our analysis suggests location is actually more important. Collectively our analyses were able to meaningfully answer sub-questions one and three.

5 | LIMITATIONS

Our analysis is subject to several limitations, with further limitations of each method are discussed in the corresponding results sections. (1) We assume the ten year CNs evaluated are independent: Because the two climate normals studied have a twenty year overlap, we assumed they are partially linearly dependent. To meet the statistical requirement of independence, we therefore compare only the non-overlapping ten year part of each period (see 2.2.1). However, it is still unlikely these datasets are fully independent of each other. (2) We assume CNs are representative of thirty year climate normals: We are required to assume the ten year subsamples are representative of the entire thirty year climate normal in order to make statements about changes between the two, and it is not clear that this is the case. Specifically, our results may overstate the differences between periods because they rely on the ten year intervals with the furthest temporal distance from each other (first ten years of the old normal and last ten years of the new normal). (3) Multicollinearity: The predictors for our regression model were chosen due to being comparatively uncorrelated, in reference to other highly correlated variables. In our analysis "relatively low correlation coefficients" are still > 0.5, so despite selection of covariates using best practices, the prevalence of correlations across most variables impacts the precision of the model. (4) Data Availability: Cloud cover was excluded from our analysis due to missing data, so potential effects are not accounted for. Seven years of pressure measurements are also missing from Luzern and St.Gallen within the subsamples, limiting the strength of our conclusions.

6 | FUTURE WORK

Our suggestions for future work include examining seasonally disaggregated changes for stations between periods, and evaluating whether there is more change between periods for summer or winter. Additionally, it may be valuable to explore the influence of air pressure on sunshine duration to better understand why it acts as a significant predictor only in CN2, and understand if this has physical meaning or is only an artifact. Of course, our analysis could be easily expanded by considering wind speed, which (in addition to sunshine) was found to change significantly between CNs for many stations.

This report expands an existing investigation from MeteoSwiss into how the newly introduced climate normal differs from the previous one. We find several interesting results including an unreported change in wind speeds, and clusters that seem to group stations geographically in a way which differs from the climatic regions modeled in CH2018. Our finding of differences between the periods supports the decision of the WMO to update and verify climate normals on a more regular basis in light of rapid climate change. Further research into climate normals is likely to support this finding, as evidence of accelerating climate change provides support for stringent mitigation.

ACKNOWLEDGEMENTS

We would like to thank Dr. Michel Piot and Dr. Sam Allen for two semesters of lectures in statistics for climate sciences. We particularly appreciated Dr. Allen's email rapid response time, and helpful comments. We would also like to acknowledge Nour El-Ajou, Markella Bouchorikou, and Henrique Traeger for their collaboration. We would also like to acknowledge the R project for Statistical Computing.

REFERENCES

Croci-Maspoli, M., Schär, C., Fischer, A., Strassmann, K., Scherrer, S., Schwierz, C., Knutti, R., Kotlarski, S., Rajczak, J., Fischer, E. M., Fischer, E. (2018). CH2018 - Climate Scenarios for Switzerland - Technical Report. National Centre for Climate Services, Zurich (Switzerland).

Federal Office of Meteorology and Climatology MeteoSwiss. (2021, October 26). Klima-Normwerte 1991-2020 für die Schweiz.

Federal Office of Meteorology and Climatology MeteoSwiss. (2022, January 20). Einführung neue Klimanormwerte 1991-2020.

Piot, M. (2020). Statistics in Climate Sciences I + II. University of Bern.

Rencher, A. C. and W. F. Christensen (2012). Methods of Multivariate Analysis (Third ed.). New York: John Wiley Sons.