Homogeneity assessment of Swiss snow depth series: Comparison of break detection capabilities of (semi-) automatic homogenisation methods

Moritz Buchmann[1,2,3], John Coll[4], Johannes Aschauer[1], Michael Begert[5], Stefan Brönnimann[2,3], Barbara Chimani[6], Gernot Resch[7], Wolfgang Schöner[7], and Christoph Marty[1]

[1]WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland
[2]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[3]Institute of Geography, University of Bern, Bern, Switzerland
[4]CGG, Crawley, United Kingdom
[5]Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich Airport, Switzerland
[6]Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Vienna, Austria
[7]Institute of Geography, University of Graz, Austria

Correspondence: Moritz Buchmann (moritz.buchmann@slf.ch)

Abstract. Knowledge concerning possible inhomogeneities in a data set is of key importance for any subsequent climatological analyses. Well-established relative homogenization methods developed for temperature and precipitation exist, but ~~with only little experience for snow~~ have rarely been applied to snow cover related time series. We undertook a homogeneity assessment of Swiss monthly snow depth series by running and comparing the results from three well-established semi-automatic break point detection methods (ACMANT, Climatol, and HOMER). ~~Break points identified by each method~~ The multi-method approach allowed us to compare the ~~results of the different methods , and by only treating break points as valid if detected in reasonably close proximity by at least two methods, we increased the robustness of the results~~different methods and to establish more robust results using a consensus of at least two change points in close proximity to each other. We investigated 184 series, of various length between 1930 and 2021 and ranging from 200 to 2500 m a.s.l. and found 45 valid break points ~~. Of those 45,~~ in 41 of the 184 series investigated of which 71% could be attributed to relocations or observer changes. Metadata are helpful, but not sufficient for break point verification as more than 90% of recorded events (relocation or observer change) did not lead to valid break points. Using ~~such~~ a combined approach (2 out of 3 methods) is highly beneficial, as it increases the confidence in identified break points in contrast to any single method, with or without metadata.

1 Introduction

The quality of climate data time series analyses relies heavily on homogeneous input data, and such quality controlled and homogenized climate data are needed to improve climate-related decision making. Most decade to century-scale meteorological time series are affected by inhomogeneities ~~or break points~~ due to e.g.; changes of instrumentation, changes to station location and observer practices, or changes in the local environment such as urbanisation or plant growth (Tuomenvirta, 2001). Disentangling these break points from the underlying noise and variability in the data is challenging, but crucial for improving

confidence in any further analyses (e.g. Vertačnik et al., 2015). Accompanying metadata, if available, can be useful in helping to corroborate and verify breaks identified by statistical methods (Aguilar and Llanso, 2003). However, not every relocation or change in the station history necessarily leads to a break point in the data series in the first place.

There is no "one-method" solution when it comes to the detection of break points, but rather a collection of statistical tools. Homogeneity tests can be broadly divided into 'absolute' and 'relative' methods. The former are applied directly to individual candidate stations to identify statistically significant shifts in the section means (referred to as breaks or change points), while relative methods entail comparison of correlated neighbouring stations with a candidate station to test for homogeneity. If reference series exist, a 'relative' (rather than an 'absolute') approach where candidate series are compared to reference series is considered state-of-the-art (Venema et al., 2012) in ~~contemporay~~ contemporary climate sciences, as it allows the practitioner to eliminate any erroneous climatological shifts (Della-Marta and Wanner, 2006). Another advantage ~~is that relative methods~~ of some relative methods is that they do not require the reference series to be homogeneous themselves (Szentimrey, 1999; Caussinus and Mestre, 2004). Given the frequent occurrence of inhomogeneities in many climate time series, considerable efforts have been made to address the issue. These efforts by the community have produced a number of relative homogenization methods and toolboxes with varying degrees of user-interaction and ease of application to choose from. PRODIGE (French for miracle) (Caussinus and Mestre, 2004) proved to be among the best-performing methods evaluated in the COST (European Cooperation in Science and Technology) Action HOME alongside ACMANT (Adapted Caussinus-Mestre Algorithm for Networks of Temperature series) (Domonkos, 2011), Climatol (Climate Tools) (Guijarro, 2018), USHCN (US Historical Climatology Network) (Menne and Williams, 2009), and MASH (Multiple Analysis of Series for Homogenization) (Szentimrey, 1999).

Efforts towards efficient break detection have been made for many meteorological variables such as temperature (e.g. Kuglitsch et al., 2012; Begert et al., 2008), precipitation (e.g. Begert et al., 2005; Coll et al., 2020), and phenological series (Brugnara et al., 2020) using various methods and tools. In the case of snow time series however, only a few studies exist: Marcolini et al. (2017) investigated the use of SNHT (Alexandersson (1986); Alexandersson and Moberg (1997)) for detecting breaks in mean annual snow depth. Marcolini et al. (2019) compared the use of SNHT and PRODIGE for the break detection and subsequent homogenization of mean seasonal snow depth. Schöner et al. (2019) focused on trend analysis of seasonal mean snow depth in the Swiss-Austrian domain; using PRODIGE to identify inhomogeneous series in the records analysed. In our approach we choose to use multiple reference series processed using several modern relative methods, thereby increasing confidence in the results. As information concerning potential inhomogeneities is crucial in providing an accurate and reliable snow time series record, an in-depth homogeneity assessment of Swiss snow depth series is necessary.

For our study we used ACMANT (Domonkos, 2011, 2020), Climatol (Guijarro, 2018), and the semi-automatic tool HOMER (Caussinus and Mestre, 2004; Domonkos, 2011; Guijarro, 2018; Picard et al., 2011) as they were all used for break detection purposes in recent studies: Kuya et al. (2021a) used Climatol for precipitation and HOMER for temperature (Kuya et al., 2021b), Noone et al. (2016) used HOMER for precipitation and Coll et al. (2020) compared break detection performance of various methods including ACMANT, Climatol, and HOMER. Climatol is based on SNHT (Standard Normal Homogeneity Test) (Alexandersson and Moberg, 1997) and recommended by Coll et al. (2020) for detecting breaks in precipitation series.

55 HOMER (HOMogenizaton softwarE in R) is the extension development of and direct successor to PRODIGE and therefore is one of the most-used homogenization methods in climate sciences. In addition it provides the homogenization practitioner with more operational freedom (in terms of configuration possibilities) than Climatol or ACMANT. Applying the different methods to Swiss snow depth time series, and experimenting with various configurations allows us to investigate the suitability of the different set-ups and to provide a homogeneity assessment of the manual Swiss snow observation network.

60 Benchmark analyses of various methods exist for temperature and precipitation at both monthly and daily resolutions (e.g. Venema et al., 2012; Killick et al., 2021), however, no clear favourite has emerged. In addition, the use of Climatol (based on SNHT) and HOMER (based on PRODIGE) allows for a more direct comparison with work already undertaken to asses homogeneity and break detection for snow time series in the Alps (Marcolini et al., 2017, 2019; Schöner et al., 2019). Furthermore, this in-depth analysis of Swiss snow series allows for the identification of suspicious or erroneous data in the series analysed

65 which would otherwise have escaped detection. The process of homogenisation can be described in 3 steps: break detection, attribution (verification of break points), and correction. In this study, we focus on the first step and touch upon the second. To focus our study, our research questions are:

1. Which method or set-up works for break point detection in (Swiss) snow depth time series?

2. Is there any elevation dependence affecting the capability of the methods for break point detection?

70 3. ~~How do results from the detection methods agree with~~ Are the detected break points consistent with changes inferred from available metadata?

4. How homogeneous (in terms of detected break points) are the Swiss snow depth series investigated in our approach?

5. ~~How does the use of different~~ Are break point results similar for different snow cover variables (average snow depth and days with snow cover)~~affect break point detection~~?

75 The paper is organized as follows: Sect. 2 introduces the data set, while Sect. 3 details the methods used for the analyses. Results are presented in Sect. 4 and discussed in Sect. 5, and conclusions are drawn in Sect. 6.

2 Data

Our data consists of a newly compiled set of manually measured Swiss snow depth (HS) series obtained by the Federal Office of Meteorology and Climatology (MeteoSwiss) and the WSL-Institute for Snow and Avalanche Research SLF. Manual snow

80 measurements are conducted every morning between November and April at designated measurement fields with an observer reading off the snow depth from a graduated fixed stake, see Buchmann et al. (2021b) for more information. A favourable unique feature of using these manual snow depth measurements in Switzerland is that the instrument (graduated snow stake) and the general measurement procedure have not changed (Haberkorn, 2019), thereby eliminating one potential source of inhomogeneity in the records. We evaluated all the available series in the archives with data recorded between 1931 and 2021.

85 Selection criteria were that time series have to be longer than 30 years and have at least 80 % complete data between November
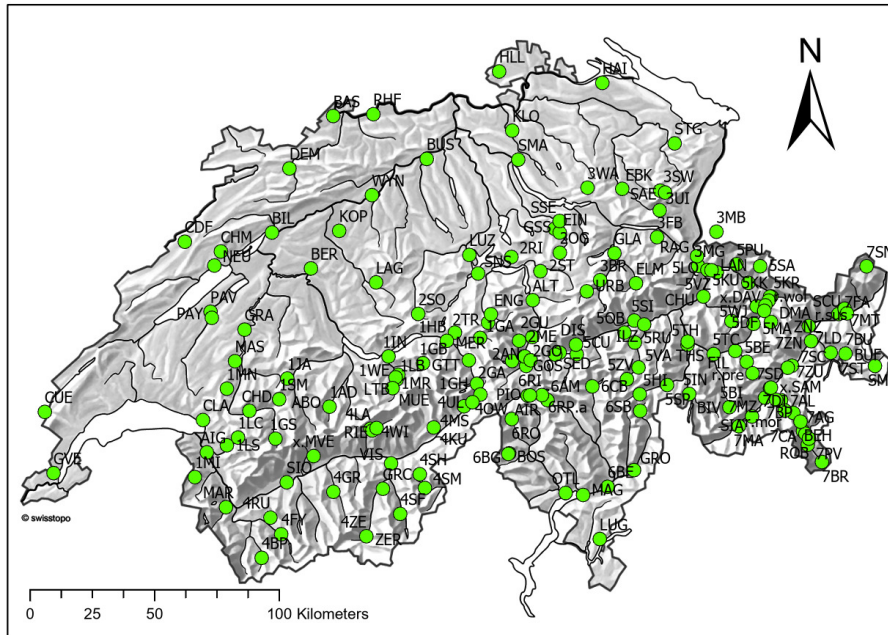
Figure 1. Map of Switzerland showing the distribution of the 184 stations used in this study.

and April, and entirely missing single years were also allowed. Applying these criteria resulted in a data set of 184 station time series, distributed widely over Switzerland; from 200 m a.s.l. to 2500 m a.s.l. Figure 1 shows the location and distribution of the stations used in this study (a list of all stations can be found in Table ~~??~~S1). We use monthly (Nov-Apr) mean snow depth and monthly sums of days with snow on the ground for every hydrological year (November of the last year to April of the

90 current year) as input data for the application of the break detection methods, and set the remaining six months to zero. Days with snow on the ground are defined as days with snow depth equal or greater than 1cm. ~~All subsequent outputs are annual values named HSavg (annual~~ The final outputs used in the break point analysis were annual time series of winter mean snow depth ~~) and~~ (HSavg) and snow cover duration (dHS1~~(annual sum of days with snow on the ground~~).

2.1 Metadata

95 Since the instruments for manually measuring snow depth have not changed over our period of analysis, this allowed a clearer focus on other metadata components, such as coordinate changes as a proxy for station relocation and observer changes. Such information is generally available and was compiled from various sources (station records, archives, and operational databases). Unfortunately, for the MeteoSwiss network the exact locations of the snow measurements can differ from the coordinates of the associated meteorological station, and especially in the early days, MeteoSwiss did not record snow specific coordinates.

100 Moreover, the quality differs from station to station and is generally more vague for older (greater than 30 years) periods in the station records (Aschauer and Marty, 2020). Our metadata is not perfect (i.e. there is missing or incomplete information for

some records), however, it is unlikely to be completely wrong and may offer some corroborative information for any breaks detected. Metadata is therefore used as additional verification for the identified break points where applicable/available with a tolerance of ± 2 years.

3 Methods

3.1 Break detection algorithms

3.1.1 HOMER

HOMER is a collection of functions for break point detection and homogenization. The pairwise comparison (PRODIGE) is based on a penalized likelihood criteria (Caussinus and Lyazrhi, 1997) and is composed of optimal segmentation in conjunction with dynamic programming (Hawkins, 2001). Further methods are joint-segmentation based on (Picard et al., 2011) and ACMANT-detection (Domonkos, 2011). We used pairwise comparison for our analyses because ACMANT-detection depends on a seasonal cycle and Gubler et al. (2017) reported issues with HOMER's joint-segmentation-module. Coll et al. (2020) highlight that HOMER has its disadvantages when dealing with incomplete data, particularly when the missing data comprise contiguous blocks earlier or later in a series. For this reason the WMO Task Team on homogenization recommends a missing data tolerance of 15 years for HOMER (WMO, 2017). However, as we are using mainly complete series and solely focus on pairwise detection, our analyses should not be affected. Network neighbourhoods in HOMER are constructed using station selection criteria, based either on distance or first difference correlations. Here we used a minimum correlation threshold of 0.8 (empirical values) and a minimum number of 5 reference series (similar to PRODIGE). In practice this means that if no reference series with correlations ≥0.8 are found for a particular candidate series, the 5 next best correlated stations are returned instead, and if more than 5 series exist with correlations ≥0.8, all are displayed.

HOMER is semi-automatic insofar as it provides the user with a set of graphical difference-series whereby the candidate series are compared to the reference series in each of the sub-networks based on the selection criteria applied. For each candidate and reference series in the derived sub-networks, any breaks detected are displayed to help inform subsequent adjustment decisions by the user. Difference-series in HOMER can be defined in two ways; either the candidate minus reference (Diff-mode), or candidate divided by reference (Ratio-mode). As our data is skewed due to the seasonal nature of snow cover data and limited at zero (no negative snow depth), we used ratio-mode instead of differences (details in Table 1). ~~We ran HOMER's pairwise-detection with fixed values for correlation and having selected a minimum number of 5 reference series (see Table 1).~~ Break points are interpreted as valid if they occur in more than half of the first 5 valid (with standard deviation of the noise (sigma) smaller than 0.3, similar length and geographical origin than candidate series) reference series (i.e. 3-out-of-5), within an uncertainty of ± 2 years, meaning that breaks are accepted as valid if they occur within ± 2 years of each other in at least 3 reference series. The ± 2 years are adopted from Kuglitsch et al. (2012). Venema et al. (2020) pointed out that parallel analyses of statistically or geographically relevant station data are the best solution for identifying breaks in time series. Some of the investigated snow time series are parallel series (Buchmann et al., 2021b) and if a break point is suggested in such a parallel

candidate-reference series, it is double counted. In cases where some of the first 5 reference series selected do not cover the whole time period of the candidate series, the next longest series with sigma smaller 0.3 might be included in the analysis. Furthermore, break detection was set to 'annual'. Once the potential break points are adjusted based on operator interpretation and re-entered, HOMER is run again to calculate the break magnitude correction factors for each station.

3.1.2 ACMANT

ACMANT (Domonkos, 2011; Domonkos and Coll, 2017) is the most automatic of the homogenization methods used in this study, which means hardly any parameters can be changed by the user. Reference series are constructed via correlations between candidate and reference series with a fixed minimal correlation of 0.4. For this study we used ACMANTv4.3 (Domonkos, 2020) run in standard precipitation mode (see Table 1). Break points and the associated break magnitude corrections applied by ACMANT are retrieved from an automatically produced file. Automatic networking (AN) has been developed (Domonkos and Coll, 2019) as a preparatory operation for homogenizing data sets of larger than 40 series (as is the case with the network here) with the ACMANT method. In AN, a specific network is constructed for each candidate series which provides optimal spatial comparison with the candidate series always in the centre of the network (Domonkos and Coll, 2017).

3.1.3 Climatol

In contrast to HOMER and ACMANT, Climatol uses composite reference series and detects breaks one-by-one with the Standard Normal Homogenization Test (SNHT; Alexandersson and Moberg (1997)) applied to anomaly series between candidate and reference series to detect break points. Anomalies are normalized either through the subtraction of or division by series averages, with the latter approach being recommended for skewed data series such as precipitation. Missing values are automatically infilled using values from neighbouring stations, thus allowing the method to compare series that do not share an intact data period prior to this adjustment. There are no pre-defined neighbourhoods as in HOMER, but a distance criteria is available to geographically confine the potential reference series (see Guijarro (2018); Coll et al. (2020) for more detail). Reference series are defined based on geographical proximity (Luna et al., 2012) using euclidean distances. By default, the vertical coordinates in metres carry the same weight as the horizontal coordinates in km. To account for the influence of elevation as a key control on snow, the scale parameter (wz) was adjusted, so that elevation counts 100 times more; in practice meaning an elevation difference of 500 m is equivalent to a horizontal distance of 50 km in the approach used here. Climatol also provides an exploratory analysis of the data which is necessary to properly set the parameters listed in Table 1. Once the parameters are set, Climatol can be run to produce an output file with suggested break points for each station. Break magnitude corrections are calculated as the change of mean before and after homogenisation as Climatol does not automatically provide these values.

3.2 Evaluation of break detection algorithms

We define valid break points as break points detected at concordant times for any given series by at least 2 out of 3 independent methods within a tolerance of $\pm$ 2 years regardless of metadata. As an additional measure, we only accepted as valid those

Table 1. Settings for HOMER, ACMANT, and Climatol

Method	Variable	Value	Comment
HOMER	Mode	Ratio, pairwise	
~~HOMER~~	Min. correlation	0.8	
~~HOMER~~	Min. reference series	5	
~~HOMER~~	Break detection	Annual	
~~HOMER~~	Break	3/5	3-out-of-5
~~HOMER~~	Tolerance for breaks	±2 years	
ACMANT	Mode	RR	Precipitation
~~ACMANT~~	min correlation	0.4	fixed value
~~ACMANT~~	Output	default	
Climatol	std	2	Normalisation: Divide by the mean; preferred if data is skewed
~~Climatol~~	wz	0.1	Scale parameter for the vertical coordinate
~~Climatol~~	wd	0,0,100	Distance weight
~~Climatol~~	dz.max	22	Value from expl-results
~~Climatol~~	snht1	35	determined by expl-results (left to default 25 for dHS1)
~~Climatol~~	snht2	45	determined by expl-results (left to default 25 for dHS1)

165   break points identified within 5 years from either the beginning or the end of a series, this to agree with procedures used in Kuya et al. (2021a). The break points detected were then compared to available metadata where applicable; this metadata information was the decisive factor used to attribute the exact year of break points. In the case of those series where no metadata information was available, either the common year detected by a majority of the methods or the first occurrence of a concordant break point within the defined tolerance threshold was used.

170   To evaluate the performance and set-up of each method, their suggested break magnitude corrections and their contributions to concordant break points (amount of valid break points compared to total number of detected break points) was measured and compared. To test whether break point detection depends on elevation or 'amount of snow', detected breaks are compared to both station elevation and climatological (calculated for the entire period) snow values (i.e. mean HSavg). To assess how many valid break points can be explained by metadata, and to see whether either station relocations or observer changes are

175   more prone to cause detectable breaks, we compared the concordant break points to records in the station history.

Using this combined approach of integrating the break point information from the various methods alongside the information from metadata allow us to more confidently estimate the homogeneity of the Swiss snow network. The main focus is on breaks in HSavg, however, the opportunity to use dHS1 alongside as a complementary break point detection approach is discussed in Section 4.3. Series without any detected break points or break points detected by only one method are considered homogeneous;

series with break points detected by at least 2 out of 3 methods (with or without metadata support based on our criteria) are considered inhomogeneous.

We are using HOMER 2.6, ACMANTv4.3, and Climatol 3.1.1, all analyses were run on a Windows system with R 4.1.1 (R Core Team, 2021).

4 Results

4.1 Comparison of method performances

4.1.1 Number of detected break points

To assess and be able to compare the three methods, we summed up all the detected break points separately for each method and grouped them into time periods to investigate the temporal distribution of the break occurrences between the methods. Here we found ACMANT returning the largest number of detected break points with 170 breaks in 98 of the 184 series. The total number of break detections for Climatol (61 in 54) and HOMER (32 in 30) are significantly smaller than the ACMANT break detections. Table 2 and Figure 2 panel A summarise the overall break detection frequencies between the three methods. Figure 2 panel B shows the temporal distribution of the detected break points between the methods. Based on this summary ACMANT detects the maximum number of break points between 1970 and 1980, whereas for Climatol and HOMER the 1980s is the decade associated with most detections. This suggests that ACMANT seems to be more sensitive in detecting changes than Climatol and HOMER, and the time period associated with the maximum number of break detections is coincident with the period where the maximum number of station records are available.

Table 2. Comparison of detected break points for each individual method. ~~Valid~~ A valid break point means it was detected by at least 2 out of 3 methods~~and~~. The complementary category uses break points from both dHS1 and HSavg (six break points are identical).

Method	Variable	Stations	Valid breaks	# Breaks	Detected valid breaks	Efficiency (%)
ACMANT	HSavg	184	31	170	31	18
Climatol	HSavg	184	31	61	17	28
HOMER	HSavg	184	31	32	23	72
Valid	HSavg	184	31	31	31	100
ACMANT	dHS1	184	20	177	17	10
Climatol	dHS1	184	20	43	16	37
HOMER	dHS1	184	20	30	11	37
Valid	dHS1	184	20	20	20	100
Complementary	both	184	45	45	45	100
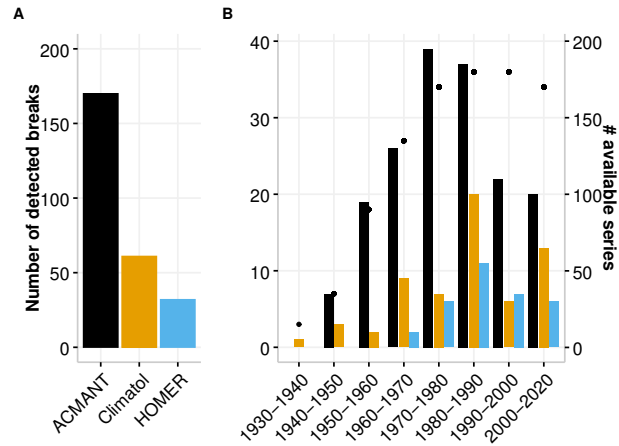
Figure 2. (A): Number of break points identified with ACMANT, Climatol, and HOMER for a total of 184 stations. (B): Number of break points and year of detection retrieved from ACMANT, Climatol, and HOMER for a total of 184 stations. The number of available series per decade is indicates with black dots and a corresponding right y-axis. All results are for HSavg.

4.1.2 Break magnitudes and detection capabilities

To investigate the sensitivity of the break detection capacity the break magnitude corrections for all identified breaks are analysed across the methods. Figure 3 panel A shows the number of detected breaks and their corresponding break magnitude cor-

200 rection categories. The majority of break magnitude corrections are distributed between 10-19% ~~for all three methods~~(ACMANT and Climatol) and 20-29% for HOMER. In contrast to ACMANT and Climatol, HOMER detects very few breaks with magnitude corrections below 10 %. This again indicates that ACMANT detects the most break points, including those of lower magnitude, whereas HOMER identifies fewer but higher magnitude breaks. Climatol tends to mirror the pattern of ACMANT break point detection, but with fewer detections overall. Figure 4 provides density distribution plots of the break magnitude

205 corrections for the three methods separately. Here we found almost normal distributions for the Climatol and ACMANT break magnitude density distributions with a peak around 0 (suggesting no or very small corrections). The break magnitude distribution for ACMANT is bimodal with a local minimum at 0, reflecting more low magnitude negative corrections than positive ones. The distribution of the break magnitude corrections for HOMER on the other hand is more uniform with no obvious peak. This reinforces not only that HOMER detects less break points than Climatol or ACMANT overall, but generally detects

210 larger magnitude breaks, a situation reflected in the broader distribution of break magnitude corrections.

4.1.3 Elevation and amount-of-snow dependencies

~~To investigate a possible~~ The availability (and quality) of suitable reference series for each candidate station is key for a proper break point detection. Snow distribution in Switzerland is highly elevation-dependent and our stations range from 200 to 2500 m a.s.l. To test the hypothesis that lower stations might not have enough suitable reference series for proper break point detections
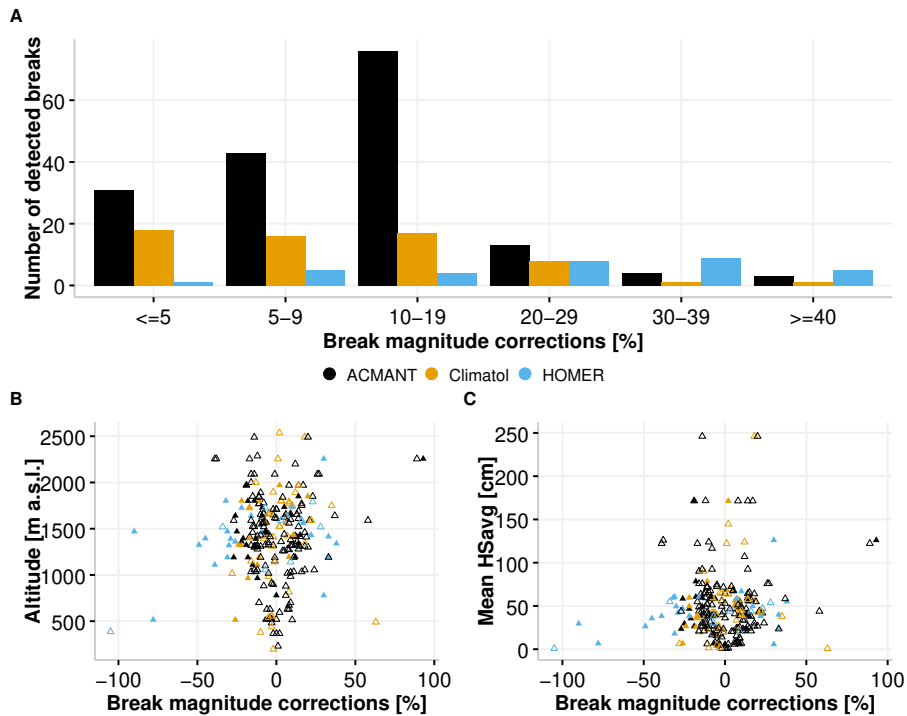
Figure 3. Plotted are in (A): Absolute break magnitude corrections and their distribution for ACMANT, HOMER, and Climatol. In (B): Break magnitude corrections versus station altitude. In (C): Break magnitude corrections against mean HSavg. ~~Subsequently verified~~ Valid break points are bold.
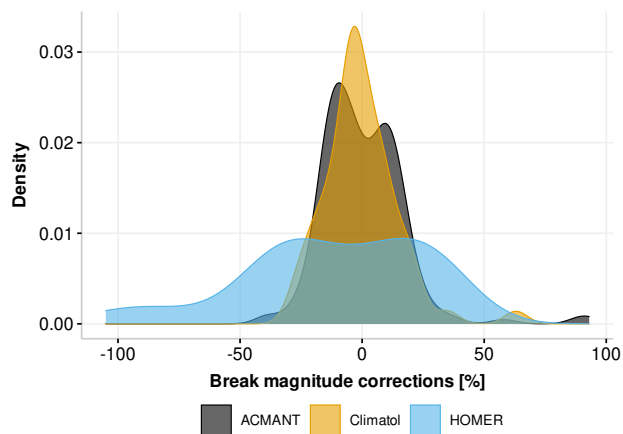
Figure 4. Density plots for break magnitude corrections for ACMANT, Climatol, and HOMER.
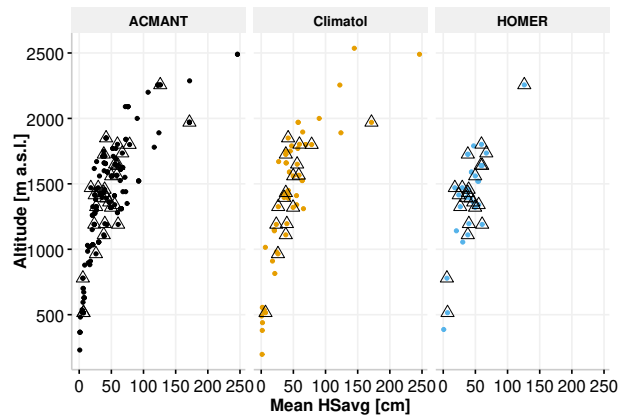
Figure 5. Break points in relation to station altitude and mean HSavg. Valid break points, identified by at least two methods, are marked with black triangles.

a possible elevation dependence is investigated. To explore a possible altitudinal or amount-of-snow influence on the break detection capability of the methods, the break magnitude corrections of the identified break points were compared to station elevation and climatological HSavg (mean over the entire period of the available records). This comparison of break detections with elevation found fewer break points below 500 and above 1800 m a.s.l. than between 1300 and 1700 m a.s.l. Whereas in comparison for the climatological HSavg, no clear pattern is visible. For these comparisons all three methods exhibited similar break detection patterns. Figure 3 panel B and C plots these relationships separately, while Figure 5 summarises the relationship between station elevation and climatological HSavg for every station where break points were detected~~by all three methods~~.

4.2 Concordant break points in HSavg

As the location and magnitude of break points obtained from the individual methods differed, for the remainder of the analysis only concordant break points (identified across all three methods) are considered. In this approach only breaks identified by at least 2 of the methods and obtained within ± 2 years are considered to be valid. The set of valid break points based on these criteria is shown in Table 3. This approach identified 31 break points in 30 series for HSavg, with only one series showing multiple break points detected across all three methods. The majority of valid break points were found with a combination of ACMANT and HOMER, whereas no break points in HSavg were found with only Climatol and HOMER. Figure 6 summarises the method combinations which led to break points being assessed as valid based on the criteria, while Figure 7 summarises the stations with identified inhomogeneities based on these same criteria. Multiple detections for series based on these criteria being applied across all three methods indicate that 83 % of the 184 Swiss snow series analysed can be considered homogeneous. However, the individual contributions to valid break points based on these selection criteria is different for each method, and these are summarised in Table 2. ACMANT and Climatol contribute to the detection of 18 % and 28 % respectively of valid
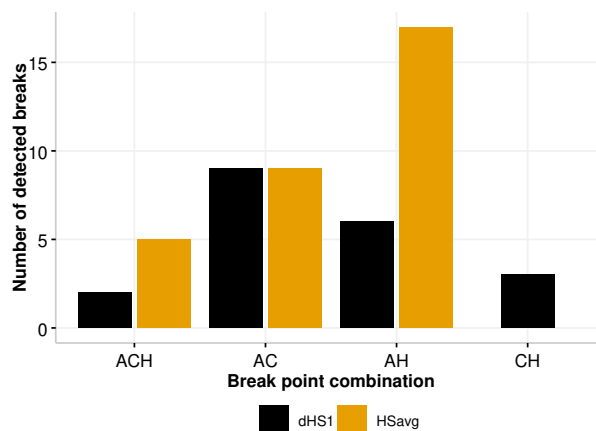
Figure 6. Method combinations for valid break points: ACMANT-Climatol-HOMER (ACH), ACMANT-Climatol (AC), ACMANT-HOMER (AH), and Climatol-HOMER (CH).

break points based on our selection criteria, whereas HOMER accounts for 72 % of the detections. To further support the break points with reference to the available metadata (either station relocation or observer changes), the break points obtained across the methods were compared to any recorded changes in the station history. This comparison with the available metadata found that 22 of the 31 detected break points were supported by metadata, and that of these 19 were due to station relocation, while three could be attributed to a change of observer. The remaining 11 break points had no metadata support in the station histories. Figure 7 (and Fig. S1 in the supplement) summarises the metadata information supporting the valid break points. This strongly suggests that station relocation is a likely explanation for the majority of the supported inhomogeneities detected in the snow series records analysed here.

4.3 dHS1 as complementary information

The same procedures as outlined in Sections 4.1 and 4.2 above were applied to another important and commonly used snow indicator: days with snow on the ground (dHS1). Here we found that for all methods except ACMANT, fewer break points are identified compared to the other measures (see Tab. 2). Here we found the majority of valid break points with the combination of ACMANT and Climatol (compared to AH for HSavg). From a total of 20 valid break points, ACMANT finds 17 followed by Climatol with 16 and HOMER with 11. This suggests that the dHS1 series appear to be more robust in terms of there being less breaks than for the HSavg series. Comparing the mean absolute break magnitude corrections for HSavg (14%) with the ones for dHS1 (5%) also reveals that the breaks identified in dHS1 are on average smaller than the ones found in HSavg. Table 2 (bottom) summarises the number of identified breaks for the two indicators. Figure 8 compares the valid break points found for HSavg and dHS1. Six (out of 45) of the break points (13%) agree, indicating that a complementary approach may be beneficial
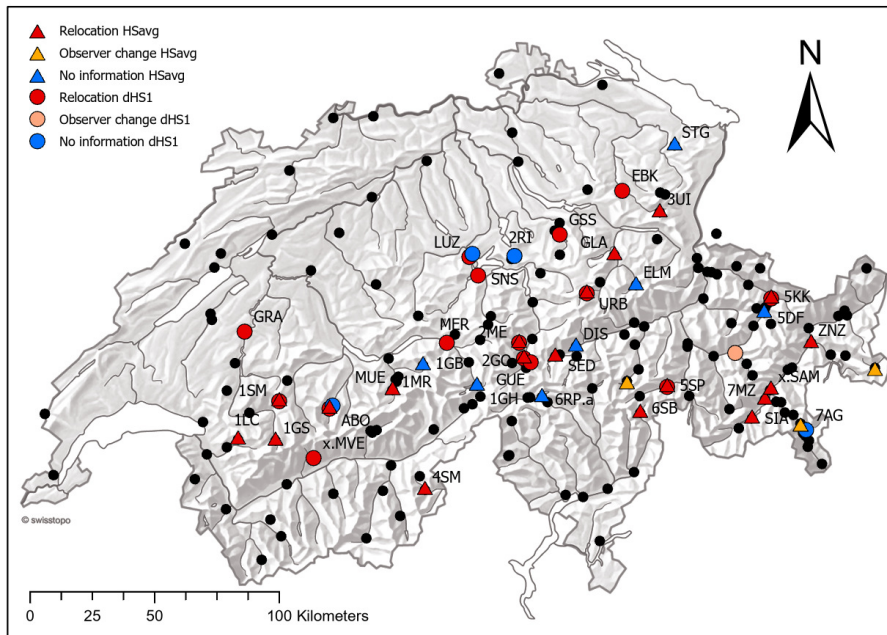
Figure 7. Map highlighting the location of series with identified valid break points and information from metadata where applicable. Series where no break points are detected are marked with black circles.

for detecting valid break points. Figure 7 provides an overview of the valid breaks and metadata information available for both HSavg and dHS1.
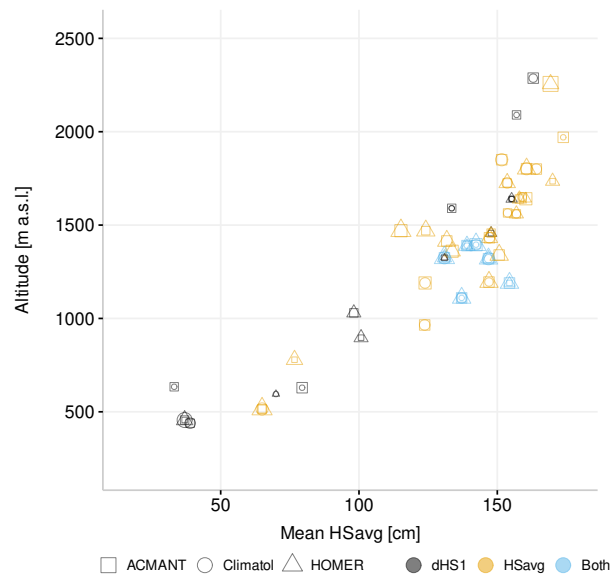
Figure 8. Comparison of valid break points found for HSavg and dHS1. Valid break points, detected by at least two methods, are coloured grey (dHS1), yellow (HSavg), or blue (both)~~, breaks only detected by one method are shown in grey~~. The shape indicates which method detected the break and the size is the corresponding break magnitude correction.

Code	ACMANT	Climatol	HOMER	Metadata	Metadata type	Combination	Variable	Break year	Altitude [m a.s.l.]
6RP.a	1955	1953			No information	AC	HSavg	1953	1800
x.MVE	1957	1957		1957	Relocation	AC	dHS1	1957	1590
1GH	1958	1960			No information	AC	HSavg	1958	1970
1GS	1966	1966	1965	1964	Relocation	ACH	HSavg	1964	1195
7ST	1967	1967	1968	1968	Observer change	AH	HSavg	1968	1415
SNS	1968	1967		1968	Relocation	AC	dHS1	1968	439
1GB	1970	1970			No information	AC	HSavg	1970	1565
DIS	1971	1970			No information	AC	HSavg	1970	1190
GLA	1970	1970			No information	AC	HSavg	1970	515
STG	1970		1971		No information	AH	HSavg	1970	779
GRA	1970	1969		1971	Relocation	AC	dHS1	1971	634
LUZ	1971	1971	1971	1971	Relocation	AH	dHS1	1971	456
5DF	1972	1972	1975		No information	AC	HSavg	1972	1560
GLA	1975		1975	1975	Relocation	AH	HSavg	1975	515
EBK	1974	1974		1976	Relocation	AC	dHS1	1976	629
6SB	1979		1981		Relocation	AH	HSavg	1979	1640
4SM	1979		1979	1980	Relocation	AH	HSavg	1980	1470
x.SAM	1980	1982	1979	1980	Relocation	AH	HSavg	1980	1726
GUE	1980	1980		1981	Relocation	AC	dHS1	1981	2287
URB	1980	1979		1981	Relocation	AC	dHS1	1981	1395
URB	1980	1980	1980	1981	Relocation	ACH	HSavg	1981	1395
BEH	1982		1982	1983	Observer change	AH	HSavg	1983	2256
ELM	1983	1984			No information	AC	HSavg	1983	965
ABO		1985	1985	1984	Relocation	CH	dHS1	1984	1325
ABO	1984	1981	1984	1984	Relocation	AH	HSavg	1984	1325
SED	1985	1982	1983	1985	Relocation	ACH	HSavg	1985	1432
1LC	1988		1987	1987	Relocation	AH	HSavg	1987	1360
1SM	1989	1990	1989	1988	Relocation	ACH	dHS1	1988	1390

Table 3: List of valid break points concordantly identified in 184 investigated Swiss snow time series. Stations are ordered according to break year and series with break points identified in both HSavg and dHS1 are marked bold. Combination code refers to ACMANT (A), Climatol (C), and HOMER (H).

Station	ACMANT	Climatol	HOMER	Remark	Break year	Combination	Series	Altitude
1SM	1987		1988	Relocation	1988	AH	HSavg	1390
2GO	1990	1991		Relocation	1989	AC	dHS1	1110
2GO	1988	1988	1988	Relocation	1989	ACH	HSavg	1110
2RI	1990	1990	1989	No information	1989	ACH	dHS1	1640
5ZV	1990		1989	Observer change	1989	AH	HSavg	1735
ABO	1990		1990	No information	1990	AH	dHS1	1325
FIL	1990		1990	Observer change	1990	AH	dHS1	1030
MUE	1989		1988	Relocation	1990	AH	HSavg	1641
1MR	1991	1991		No information	1991	AC	HSavg	1650
MER		1989	1989	Relocation	1992	CH	dHS1	595
5SP	1994		1993	Relocation	1994	AH	HSavg	1457
3UI	1995		1995	Relocation	1996	AH	HSavg	1340
2ME		1997	1996	Relocation	1997	CH	dHS1	1320
2ME	1997	1997	1995	Relocation	1997	ACH	HSavg	1320
5KK	1998		2000	Relocation	1997	AH	dHS1	1190
5KK	2000		2000	Relocation	1997	AH	HSavg	1190
7MZ	1997	1996		Relocation	1998	AC	HSavg	1850
7AG	2000	1999		No information	1999	AC	dHS1	2090
ZNZ	2000		2001	Relocation	2001	AH	HSavg	1470
SIA	2001		2001	Relocation	2001	AH	HSavg	1802
5SP	2003		2001	Relocation	2004	AH	dHS1	1457
GSS	2008		2008	Relocation	2009	AH	dHS1	897
LUZ	2012	2012		No information	2012	AC	dHS1	456

5 Discussions

5.1 Break detection

The main differences between the three methods arises from both the way reference series are constructed and in how breaks are classified as valid. Reference series in ACMANT and HOMER are constructed based on correlation, but with different correlation coefficient thresholds applied for the selection of reference series in the networks, whereas in Climatol, proximity (defined by euclidean distances), is used. As these different configuration parameters allow for a large number of reference series combinations, as well as for how breaks are classified, it is no surprise that the results differ between the methods. As the construction of reference series is an intrinsic part of the methods, a comparison with the same input-network is not possible.

The detection capability of Climatol depends on the choice of wz (vertical scale parameter) and the SNHT-thresholds as outlined by Guijarro (2018). However, the determination of appropriate SNHT-thresholds is less straightforward, an issue which has been reported by Kuya et al. (2021a). Lowering the thresholds, increased the number of identified break points. However, if the thresholds are too low, multiple breaks within the same years are detected which implies that the method is too sensitive. Conversely, if the thresholds are too high, no breaks are detected at all. As suggested by Kuya et al. (2021a), the optimal threshold choice is one which prevents multiple break detections within the same year. The weight of the vertical coordinate (wz) is especially important for snow, as elevation in conjunction with small-scale topographic variation is an important control on snow depth.

ACMANT on account of the automation (and hence no user interaction) is the most objective method, but consequently there are no settings to optimise based on the judgement/experience of a practitioner (Pérez-Zanón et al., 2015). However, there seems to be a sensitivity issue with ACMANT: If using 188 instead of 184 series, the break points for some stations shifted slightly (~~2years~~$\pm$ 2 years) due to different station combinations being used as reference series. This indicates that the break detection algorithm is less robust, or more dependent on available stations, especially as the correlation threshold used (0.4) is low. For comparison reasons, we did not perform any pre-treatment outside the actual methods (such as pre-defined neighbourhood networks), however, this might be beneficial for ACMANT (Domonkos and Coll, 2019).

Break detection in HOMER relies on multiple reference series in conjunction with support decisions from an experienced user. While this makes it the most robust application, it is also the most subjective of the three methods used here.

While there are many benchmark data sets for other climate variables (e.g. Willett et al., 2014; Venema et al., 2012), to date there are no such benchmarks for snow so far, which is unfortunate as these could have helped to thoroughly benchmark the methods. As the method with the highest number of detected breaks, ACMANT seems to be the most sensitive of the methods investigated here (Pérez-Zanón et al., 2015; Fioravanti et al., 2019) followed by Climatol and HOMER with smaller numbers of identified break points. Coll et al. (2020) found similar increments when analysing the break detection capabilities of ACMANT, Climatol, and HOMER for a large network of precipitation series in Ireland. Figure 5 shows no clear relationship between altitude, meanHSavg and valid break points. Most valid break points are however, detected at stations between 1000 and 1800 m a.s.l. A possible explanation is the large number of available stations in that particular elevation band, thereby ensuring that enough reference series are available. However, the network density for stations above 1800 m a.s.l. and below

1000 m a.s.l. is sparse by comparison shown by Figure 5. According to Gubler et al. (2017) the number of detected break points

can be reduced by up to 50% in sparse networks combined with low signal-to-noise ratio (SNR).

Analysing the standard deviation of the ratio series (sigma; and used as a proxy for noise) as well as the median correlations with HOMER reveals that for stations with altitude below 800 and above 1800 m a.s.l. the range of the median sigma of the five reference series with lowest sigma per candidate series is larger than for stations between 800 and and 1800 m a.s.l. Fig. S2 in the appendix shows that relationship. A similar situation is clear when the median correlations of these subsets per candidate series are considered (Fig. S3). The higher noise (due to not enough suitable reference series available) associated with these lower (upper) stations may explain the different results between the networks at different densities, and hence a possible explanation for none of the methods detecting as many break points for stations in those elevation bands.

5.1.1 Break magnitude corrections

Break magnitude corrections and their corresponding density plots (Fig. 4) are similar to the ones found by Coll et al. (2020) working on Irish precipitation series. Differences in break magnitude corrections are also affected by different station neighbourhoods and arising from this the different subsets of stations used for the homogenisation process. With a default internal correlation threshold of 0.4, ACMANT's neighbour selections cannot be the same as those of HOMER based on the 0.8 correlation threshold used.

5.2 Choice of method

According to our analyses, HOMER performs better than ACMANT or Climatol when the number of detected valid break points are compared to the number of break points detected overall. However, this comes at a cost based on the time input of the user and also requires extensive expert knowledge. In terms of ease of use based on automation, ACMANT would be the method of choice. However, sifting through the large number of break points detected involves a lot of post-hoc processing, and also requires in-depth knowledge about the network. Climatol provides the second best efficiency (in terms of the ratio of valid breaks to total identified break points), however with 17 valid break points detected, 14 would have gone undetected without a combined approach, see Table 2. As outlined in Section 5.1 and reported by Kuya et al. (2021a), the results from Climatol heavily depend on the initial set-up. Coll et al. (2020) recommended Climatol as the method to use for precipitation because the identified break points appeared likely to be more realistic than the ones found with either ACMANT or HOMER in the analysis of their network. However, our analysis shows that this might not be the same when applied to snow series in a more topographically complex region. We found rather that a combination of the three methods works best, and that HOMER performs significantly better than Climatol or ACMANT in the context of our network setting. This application of multiple methods was also used and recommended by Kuglitsch et al. (2012); Toreti et al. (2012) for Swiss temperature series, as well as by Marcolini et al. (2019) for Austrian snow depth series. A recent study by Brugnara et al. (2020) applying homogenization methods to the phenological network in Switzerland also recommended this application of multiple methods.

5.2.1 Choice of variable: HSavg versus dHS1

Results from Buchmann et al. (2021b) and Buchmann et al. (2021a) concerning the stability, robustness, and variability of the snow variables HSavg and dHS1 investigated here, suggest that HSavg is on average less stable and shows more variation than dHS1. The larger variability associated with HSavg when compared to dHS1 would be expected to lead to larger breaks and thus result in an increased likelihood of break point detection in HSavg. However, associated with these larger variations, there is also increased noise across the series which has the effect of generally reducing the detection capability for lower magnitude break points altogether. dHS1, on the other hand, is more stable and shows less variability across the series, suggesting less noise for this variable, and therefore break points with lower amplitudes overall. However, only looking at an average variability (or standard deviation) does not necessarily improve our understanding, as the temporal evolution of variation of any single station is more important, but is also a property affected by inhomogeneities. Analysing the mean absolute break magnitude corrections for ~~these~~ the 6 break points (~~identified in~~ identified in both HSavg and dHS1 ~~)~~ reveals that the amplitudes of those detected for dHS1 are significantly smaller than the ones for HSavg. Break magnitude corrections retrieved from HOMER using breaks detected in dHS1 and inserted into HSavg do not differ from the ones obtained through breaks detected purely in HSavg. Based on this finding, break points detected in dHS1 may be used to calculate corrections for HSavg. These results tend to corroborate the hypothesis that the three methods are detecting smaller break points with dHS1, than with HSavg. For stations below 1000 m a.s.l. the majority of valid break points are detected in dHS1, hence these results tend to support the benefits of a complementary (dHS1 and HSavg) approach. As only 6 break points are identified with both variables, a complementary approach seems beneficial. Combining the results from both HSavg and dHS1 returns 45 (Table 2) valid break points in 184 Swiss snow time series.

5.3 Homogeneity of Swiss snow series

Our analysis shows the need for a combined use of some of the available methods in order to retrieve a set of break points for Switzerland where we can have higher confidence in their validity. In the majority of cases the methods agree well, with HOMER and Climatol returning the highest proportion of break points deemed valid based on the criteria we applied. Kuglitsch et al. (2012) could explain most of their break points for Swiss temperature series based on ~~combination of a number of problems~~ a combination of reasons, whereas our results tend to indicate that station relocation is the most likely source of inhomogeneities for the snow series analysed (Coll et al., 2020; Kuya et al., 2021a). However, due to the incomplete nature of the metadata available to support our study, we are unable to investigate this in more detail. For example, the entry 'observer change' can mean change of observer only, but may unfortunately also imply a combination of relocation and observer change. In spite of having identified relocation as the main explanation for the majority of break points, this is not consistently the case, since from a total of 519 recorded location changes in the station histories across the network, only 45 relocated stations (9%) produced a valid break point across all three methods. The majority of our identified valid break points are located between 1000 and 1800 m a.s.l. at stations which normally experience a continuous snow cover from November to April of the following year. From Figure 7, the main geographical locations of series with breaks are the northern Prealps, Bernese Alps,

and the Engadine. The lack of inhomogeneity detections for series South of the Alps (Ticino) can largely be explained by a lack of suitable reference series available. The southern parts of Switzerland are dominated by low valleys surrounded by high mountains. This steep gradient in conjunction with small-scale varying climatic conditions seems to be a reason for not having enough reference series and subsequently inhibit any suitable break detection due to a lack of appropriate reference series.

5.3.1 Comparison with Italy, Austria, and the Alps

Similar break detection investigations have been performed for Italy, Austria, and the Austrian-Swiss domain: Marcolini et al. (2017) applied the SNHT to 106 closely adjacent snow time series for the same region of Italy, and in that work they reported 20% of the series to be inhomogeneous. In terms of altitudinal and temporal extent, the data set used in that work is similar to ours, however, the 184 stations comprising our network are not from such a climatically coherent region.

Marcolini et al. (2019) investigated only 25 series between 200 and 1600 m a.s.l. whereas we analysed 184 series between 200 and 2500 m a.s.l., but only found valid break points in series between 500 (400 dHS1) and 2300 m a.s.l. (and the majority of these between 1000 and 1800 m a.s.l.). Moreover, Marcolini et al. (2019) found 11 (5 with SNHT and PRODIGE) breaks in 25 series. They also reported an agreement of 45% between SNHT and PRODIGE in identifying the same break points. We only found 20 % (25% for HSavg and 16% for dHS1) agreement between Climatol and HOMER. In other work, Schöner et al. (2019) performed a break detection analysis on 96 Swiss snow series between 1961 and 2012 using PRODIGE and reported 25 series (26%) as inhomogeneous. We found similar numbers: 17% inhomogeneous time series when only HSavg-breaks are considered and 25 % with the complementary approach (HSavg and dHS1).

Using the complementary approach (HSavg and dHS1), we found similar numbers of potential break points as have been reported in previous studies (Marcolini et al., 2017, 2019; Schöner et al., 2019), this in-spite of the time periods and data sets used in our study not being identical. However, if we only compare the break points found in HSavg, our results show fewer break points than have been identified by the previous studies. Possible explanations for the differences are the combined use of three break detection methods in our work (rather than reliance on one method), as well as our fairly strict criteria for defining valid break points.

5.4 Further issues

The attribution of each valid break point to a single hydrological year is not straightforward, even when combined with the information from the available metadata, as not all series we defined as having valid break points had metadata available, or when metadata was available, it was not necessarily complete or entirely correct.

Furthermore, as the number and location of true break points in our data set remain unknown, statements about the probability of detection and false alarm rates (Brugnara et al., 2019) are not possible.

Stations below 500 m a.s.l. regularly experience winter months with little snow and only few snow days, thus for a number of years in these records the recorded monthly means can be near zero. This lack of consistent snow records and the large variability (noise) associated with lower stations makes it virtually impossible to have sufficient suitable reference series.

385   The next step to obtain homogenised snow series is to find the best method to use and further validate and subsequently correct the identified break points.

6   Conclusions

We present the first in-depth break point detection comparison of three homogenisation methods for snow depth time series. In addition, we present the first detailed homogeneity assessment of Swiss snow depth series using state-of-the-art homogeni-

390   sation methods. Our analyses suggest that using ratios, working with monthly input data and annual detection, and combining the results from three methods (ACMANT, Climatol, HOMER) and two variables (HSavg and dHS1) offers a promising configuration to more accurately identify inhomogeneities in (Swiss) snow depth series.

     By treating break points as valid when identified by at least 2 out of 3 methods, and by the application of strict criteria increases the robustness of, and the confidence in, the results compared to the use of a single method. For all the methods

395   applied, expert knowledge about the network in question is indispensable. If however, the practitioner is limited to only one method, based on the data and analysis here the method of choice would be HOMER. However, the approach combining multiple methods introduced here for application to snow depth series is more rigorous as it provides more confidence in the results. Concerning the total set of valid break points ACMANT and Climatol appear to overestimate, whereas HOMER underestimates the number of valid break points, both for HSavg and dHS1 based on the criteria used.

400   We identified 45 valid break points ~~(25%)~~ in ~~184~~ 41 of 184 (22%) series investigated using a complementary approach of HSavg and dHS1, of these 71 % could be explained by metadata, and based on the metadata 88 % of the identified breaks could be attributed to station relocation. At low elevations, we identified a lack of suitable reference series due to many stations having inconsistent snow-lie and are associated with large year-to-year variations, possibly masking the signal within the noise. However, further work is required, especially in view of the growing research effort in relation to attribution (break point

405   verification) and homogenisation (correction) efforts.

Data availability.   Daily manual snow depth measurements in Switzerland are subject to copyright but can be obtained on request directly from the two sources: MeteoSwiss and SLF. The basis for our analysis, the monthly input values are available on request through EnviDat https://doi.org/10.16904/envidat.297, as is the corresponding metadata.

Author contributions.   MBU performed the analyses, produced the figures and wrote the draft. The manuscript was written by MBU and JC
410   with inputs from all authors. JA, CM and MBT compiled the metadata. The study was devised by MBU, SB, MBT and CM, and supervised by CM.

Competing interests.   We do not have any competing interests.

Acknowledgements. Thanks are due to Elinah Khasandi Kuya, Peter Domonkos, and Jose Guijarro for tips regarding HOMER, ACMANT and Climatol. Thanks are also due to Ross Brown and two anonymous reviewers whose comments helped to improve the manuscript.

References

Aguilar, E. and Llanso, P.: Guidelines on climate metadata and homogenization, WMO, 2003.

Alexandersson, H.: A homogeneity test applied to precipitation data., 1986.

Alexandersson, H. and Moberg, A.: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends, International Journal of Climatology, 17, 25–34, https://doi.org/10.1002/(sici)1097-0088(199701)17:1<25::aid-joc103>3.0.co;2-j, 1997.

Aschauer, J. and Marty, C.: Providing Data Provision for a Sensitivity Analysis of Snow Time Series, resreport, WSL Institute for Snow and Avalanche Research SLF, https://www.meteoschweiz.admin.ch/content/dam/meteoswiss/en/Forschung-und-Zusammenarbeit/Internationale-Zusammenarbeit/GCOS/doc/Final_report_Poviding_Data_Provision_for_a_Sensitivity_Analysis_of_Snow_Time_Series.pdf, research Report for GCOS Switzerland, 2020.

Begert, M., Schlegel, T., and Kirchhofer, W.: Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000, International Journal of Climatology, 25, 65–80, https://doi.org/10.1002/joc.1118, 2005.

Begert, M., Zenklusen, E., Häberli, C., Appenzeller, C., and Klok, L.: An automated procedure to detect discontinuities; performance assessment and application to a large European climate data set, 17, 663–672, https://doi.org/10.1127/0941-2948/2008/0314, 2008.

Brugnara, Y., Good, E., Squintu, A. A., van der Schrier, G., and Brönnimann, S.: The EUSTACE global land station daily air temperature dataset, Geoscience Data Journal, 6, 189–204, https://doi.org/https://doi.org/10.1002/gdj3.81, 2019.

Brugnara, Y., Auchmann, R., Rutishauser, T., Gehrig, R., Pietragalla, B., Begert, M., Sigg, C., Knechtl, V., Konzelmann, T., Calpini, B., and Brönnimann, S.: Homogeneity assessment of phenological records from the Swiss Phenology Network, 64, 71–81, https://doi.org/10.1007/s00484-019-01794-y, 2020.

Buchmann, M., Begert, M., Brönnimann, S., and Marty, C.: Local-scale variability of seasonal mean and extreme values of in situ snow depth and snowfall measurements, The Cryosphere, 15, 4625–4636, https://doi.org/10.5194/tc-15-4625-2021, 2021a.

Buchmann, M., Begert, M., Brönnimann, S., and Marty, C.: Evaluating the robustness of snow climate indicators using a unique set of parallel snow measurement series, International Journal of Climatology, 41, E2553–E2563, https://doi.org/10.1002/joc.6863, 2021b.

Caussinus, H. and Lyazrhi, F.: Choosing a Linear Model with a Random Number of Change-Points and Outliers, 49, 761–775, https://doi.org/10.1023/a:1003230713770, 1997.

Caussinus, H. and Mestre, O.: Detection and correction of artificial shifts in climate series, 53, 405–425, https://doi.org/10.1111/j.1467-9876.2004.05155.x, _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2004.05155.x, 2004.

Coll, J., Domonkos, P., Guijarro, J., Curley, M., Rustemeier, E., Aguilar, E., Walsh, S., and Sweeney, J.: Application of homogenization methods for Ireland's monthly precipitation records: Comparison of break detection results, International Journal of Climatology, 40, 6169–6188, https://doi.org/10.1002/joc.6575, 2020.

Della-Marta, P. M. and Wanner, H.: A Method of Homogenizing the Extremes and Mean of Daily Temperature Measurements, Journal of Climate, 19, 4179 – 4197, https://doi.org/10.1175/JCLI3855.1, 2006.

Domonkos, P.: Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT), International Journal of Geosciences, 02, 293–309, https://doi.org/10.4236/ijg.2011.23032, 2011.

Domonkos, P.: ACMANTv4: Scientific content and operation of the software, 2020.

Domonkos, P. and Coll, J.: Homogenisation of temperature and precipitation time series with ACMANT3: method description and efficiency tests, International Journal of Climatology, 37, 1910–1921, https://doi.org/10.1002/joc.4822, 2017.

Domonkos, P. and Coll, J.: Impact of missing data on the efficiency of homogenisation: experiments with ACMANTv3, Theoretical and Applied Climatology, 136, 287–299, https://doi.org/10.1007/s00704-018-2488-3, 2019.

Fioravanti, G., Piervitali, E., and Desiato, F.: A new homogenized daily data set for temperature variability assessment in Italy, International Journal of Climatology, 39, 5635–5654, https://doi.org/https://doi.org/10.1002/joc.6177, 2019.

455 Gubler, S., Hunziker, S., Begert, M., Croci-Maspoli, M., Konzelmann, T., Brönnimann, S., Schwierz, C., Oria, C., and Rosas, G.: The influence of station density on climate data homogenization, International Journal of Climatology, 37, 4670–4683, 2017.

Guijarro, J. A.: Homogenization of climatic series with Climatol, https://doi.org/10.13140/RG.2.2.27020.41604, 2018.

Haberkorn, A.: European Snow Booklet – an Inventory of Snow Measurements in Europe, https://doi.org/10.16904/envidat.59, 2019.

Hawkins, D. M.: Fitting multiple change-point models to data, Computational Statistics and Data Analysis, 37, 323–341,
460 https://doi.org/10.1016/S0167-9473(00)00068-2, 2001.

Killick, R. E., Jolliffe, I. T., and Willett, K. M.: Benchmarking the performance of homogenisation algorithms on synthetic daily temperature data, International Journal of Climatology, n/a, https://doi.org/https://doi.org/10.1002/joc.7462, 2021.

Kuglitsch, F. G., Auchmann, R., Bleisch, R., Brönnimann, S., Martius, O., and Stewart, M.: Break detection of annual Swiss temperature series, 117, https://doi.org/10.1029/2012JD017729, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012JD017729, 2012.

465 Kuya, E. K., Gjelten, H. M., and Tveito, O. E.: Homogenization of Norwegian monthly precipitation series for the period 1961-2018, Tech. rep., Norwegian Meteorological Institute, https://www.met.no/publikasjoner/met-report/_/attachment/download/829d6a8e-1ec8-417d-8231-0268bddac4bb:7eb30ca105a6cf1d470453976bb8d729a64b25a4/METreport%2004-21-Homogenization%20of%20precipitation%201961-2018.pdf, mETreport, 2021a.

Kuya, E. K., Gjelten, H. M., and Tveito, O. E.: Homogenization of Norway's mean monthly temperature series, Tech. rep., Norwegian Me-
470 teorological Institute, https://www.met.no/publikasjoner/met-report/_/attachment/download/94481ff9-6366-4569-84f5-881d8e671245:e83e7420a5810019bd2241684070d31d74a35c81/METreport03-2020.pdf, mETreport, 2021b.

Luna, M. Y., Guijarro, J. A., and López, J. A.: A monthly precipitation database for Spain (1851–2008): reconstruction, homogeneity and trends, Advances in Science and Research, 8, 1–4, https://doi.org/10.5194/asr-8-1-2012, 2012.

Marcolini, G., Bellin, A., and Chiogna, G.: Performance of the Standard Normal Homogeneity Test for the homogenization of
475 mean seasonal snow depth time series: PERFORMANCE OF SNHT FOR SNOW DEPTH TIME SERIES, 37, 1267–1277, https://doi.org/10.1002/joc.4977, 2017.

Marcolini, G., Koch, R., Chimani, B., Schöner, W., Bellin, A., Disse, M., and Chiogna, G.: Evaluation of homogenization methods for seasonal snow depth data in the Austrian Alps, 1930–2010, 39, 4514–4530, https://doi.org/10.1002/joc.6095, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.6095, 2019.

480 Menne, M. J. and Williams, C. N.: Homogenization of Temperature Series via Pairwise Comparisons, Journal of Climate, 22, 1700–1717, https://doi.org/10.1175/2008jcli2263.1, 2009.

Noone, S., Murphy, C., Coll, J., Matthews, T., Mullan, D., Wilby, R. L., and Walsh, S.: Homogenization and analysis of an expanded long-term monthly rainfall network for the Island of Ireland (1850–2010), International Journal of Climatology, 36, 2837–2853, https://doi.org/10.1002/joc.4522, 2016.

485 Pérez-Zanón, N., Sigró, J., Domonkos, P., and Ashcroft, L.: Comparison of HOMER and ACMANT homogenization methods using a central Pyrenees temperature dataset, Advances in Science and Research, 12, 111–119, https://doi.org/10.5194/asr-12-111-2015, 2015.

Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B., and Robin, S.: Joint segmentation, calling, and normalization of multiple CGH profiles, Biostatistics, 12, 413–428, https://doi.org/10.1093/biostatistics/kxq076, 2011.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2021.

Schöner, W., Koch, R., Matulla, C., Marty, C., and Tilg, A.-M.: Spatiotemporal patterns of snow depth within the Swiss-Austrian Alps for the past half century (1961 to 2012) and linkages to climate change, International Journal of Climatology, 39, 1589–1603, https://doi.org/10.1002/joc.5902, 2019.

Szentimrey, T.: Proceedings of the Second Seminar for Homogenization of Surface Climatological Data, 1999.

Toreti, A., Kuglitsch, F. G., Xoplaki, E., and Luterbacher, J.: A Novel Approach for the Detection of Inhomogeneities Affecting Climate Time Series, 51, 317–326, https://doi.org/10.1175/JAMC-D-10-05033.1, 2012.

Tuomenvirta, H.: Homogeneity adjustments of temperature and precipitation series—Finnish and Nordic data, International Journal of Climatology, 21, 495–506, https://doi.org/https://doi.org/10.1002/joc.616, 2001.

Venema, V., Trewin, B., and Wang, X.: Guidelines on Homogenization 2020 edition, Tech. rep., World Meteorological Organization Issue WMO-No. 1245, https://library.wmo.int/?lvl=notice_display&id=21756, 2020.

Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P., and Brandsma, T.: Benchmarking homogenization algorithms for monthly data, 8, 89–115, https://doi.org/10.5194/cp-8-89-2012, publisher: Copernicus GmbH, 2012.

Vertačnik, G., Dolinar, M., Bertalanič, R., Klančar, M., Dvoršek, D., and Nadbath, M.: Ensemble homogenization of Slovenian monthly air temperature series, International Journal of Climatology, 35, 4015–4026, https://doi.org/10.1002/joc.4265, 2015.

Willett, K., Williams, C., Jolliffe, I. T., Lund, R., Alexander, L. V., Brönnimann, S., Vincent, L. A., Easterbrook, S., Venema, V. K. C., Berry, D., Warren, R. E., Lopardo, G., Auchmann, R., Aguilar, E., Menne, M. J., Gallagher, C., Hausfather, Z., Thorarinsdottir, T., and Thorne, P. W.: A framework for benchmarking of homogenisation algorithm performance on the global scale, Geoscientific Instrumentation, Methods and Data Systems, 3, 187–200, https://doi.org/10.5194/gi-3-187-2014, 2014.

WMO: World Meteorological Organization: Web site of the task team on homogenization, http://www.climatol.eu/tt-hom/, 2017.