Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

MeteoSwiss

Technical Report MeteoSwiss No. 271

Quality Analysis and Classification of Data Series from the Swiss Phenology Network

R. Auchmann, Y. Brugnara, T. Rutishauser, S. Brönnimann, R. Gehrig, B. Pietragalla, M. Begert, C. Sigg, V. Knechtl, B. Calpini , T. Konzelmann

ISSN: 2296-0058

Technical Report MeteoSwiss No. 271

Quality Analysis and Classification of Data Series from the Swiss Phenology Network

R. Auchmann[1], Y. Brugnara[1], T. Rutishauser[1], S. Brönnimann[1], R. Gehrig[2], B. Pietragalla[2], M. Begert[2], C. Sigg[2], V. Knechtl[2], B. Calpini[2], T. Konzelmann[2]

[1] Oeschger Center for Climate Change Research and Institute of Geography, University of Bern, Bern

[2] Federal Office of Meteorology and Climatology MeteoSwiss

Recommended citation:

Auchmann R, Brugnara Y, Rutishauser T, Brönnimann S, Gehrig R, Pietragalla B, Begert M, Sigg C, Knechtl V, Calpini B and Konzelmann T: Quality Analysis and Classification of Data Series from the
Swiss Phenology Network, *Technical Report MeteoSwiss,* **271,** 77 pp.

Editor:

Federal Office of Meteorology and Climatology, MeteoSwiss, © 2018

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Abstract

The **PhenoClass** project, financed by MeteoSwiss in the framework of GCOS Switzerland, comprised a quality control, break detection and development of a novel classification scheme for the phenological series of the Swiss Phenology Network (SPN). The overall aim of the project was the identification of the most valuable series and stations of the SPN. Therefore, the entire dataset of the SPN was subjected to quality control procedures, including a break detection. A novel classification scheme enabled the subsequent identification of highly valuable series and stations, using results from the quality control and the break detection.

The **data** used in the project encompass series from 167 different stations in Switzerland, 70 of which started in 1951 with the initiation of the SPN. A set of additional parameters started being observed in 1996; currently 69 different parameters are being observed. The dataset analysed comprises (until 2015) 9ˈ455 series with a total of 205ˈ808 single observations.

The project's overall **objectives** were to:

- develop quality control and homogeneity assessment procedures as input for the classification
- develop a novel classification scheme for ranking the series and stations
- identify the most valuable (i.e. high quality, homogeneous, long, complete) series of the SPN

The following objectives were achieved:

- A quality control procedure tailored to the SPN was developed and applied to all series. As a result each single observation has a (boolean) flag assigned.
- A novel break detection method was designed for phenological series.
- A novel classification scheme was developed in order to identify the most valuable Swiss phenological stations and series. The main input for the classification resulted from the quality control and the break detection, also information such as the length and completeness of each series were used as input for the classification.
- An R – package was produced containing the break detection and the classification.

Zusammenfassung

Das Projekt **PhenoClass**, finanziert durch MeteoSwiss im Rahmen von GCOS Schweiz, umfasst die Qualitätskontrolle, eine Homogenitätsprüfung und die Entwicklung eines neuartigen Klassifikationssystems für die Phänologischen Beobachtungsreihen des Schweizer Phänologie Beobachtungsnetzes (SPN).

Das Ziel des Projekts war es, die hochwertigsten, lange Reihen und Stationen des SPN zu identifizieren und hierarchisch einzustufen. Um dies zu erreichen wurden alle Reihen einer Qualitätskontrolle unterzogen, die auch eine Homogenitätskontrolle enthielt. Ein neues Klassifikationsschema ermöglichte schliesslich, hochqualitative Reihen zu identifizieren. Als Grundlage für das Klassifikationsschema wurden die erhoben Informationen aus der Qualitätskontrolle und der Homogenitätskontrolle herangezogen, wie auch Informationen zur Länge und Fehlwertrate der Datenreihen.

Die im Projekt verwendeten **Daten** umfassen Beobachtungsreihen von 169 Stationen des Schweiz, wovon ca. 70 auf das Gründungsjahr des SPN 1951 zurückgehen. Ab 1996 wurden zusätzliche Parameter beobachtet; heute werden 69 verschiedene Parameter beobachtet. Der analysierte Datensatz umfasst (bis 2015) 9ˈ455 Beobachtungsreihen mit 205ˈ808 Einzelbeobachtungen.

Die **Ziele** des Projektes waren:

- Die Entwicklung einer Methode zur Qualitätskontrolle und zur Homogenitätskontrolle als Grundlage für die Klassifikation der Reihen
- Die Entwicklung eines Klassifikationssystems um die Datenreihen und Stationen nach ihrer Qualität, Vollständigkeit und Länge einzustufen
- Die Identifikation der wertvollsten (i.e. hochqualitativen, homogenen, langen, vollständigen) Datenreihen des SPN

Die folgenden Ziele wurden erreicht:

- Eine Qualitätskontrolle wurde speziell für den SPN Datensatz entwickelt und für alle Reihen angewendet. Jede Einzelbeobachtung erhielt einen Qualitäts-Flag.
- Eine neuartige Methode zur Homogenitätsprüfung für phänologische Datenreihen wurde entwickelt und angewendet.
- Ein neuartiges Klassifikationsschema wurde entwickelt um hochwertige Datenreihen und Stationen zu identifizieren. Als Grundlage dienten hauptsächlich die Ergebnisse der Qualitätskontrolle und der Homogenitätsprüfung, es wurden aber auch Informationen zur Länge der Datenreihen und der Vollständigkeit der Datenreihen herangezogen.

- Ein R-Paket wurde erstellt, welches die automatische Homogenitätsprüfung und die Klassifikation enthält.

Contents

1 Introduction

1.1 Present Status

Observations of plant phenological phases not only constitute a monitoring of plant life in general, but serve the assessment of agricultural suitability, changes in habitat factors and others. Because of high sensitivity of many phenological phases towards temperature, plant phenology has become an important climate change impact indicator in Switzerland (Studer et al. 2005, Seiz and Foppa 2007, MeteoSwiss 2018), in Europe (Menzel et al. 2006, Fu et al. 2015) and globally (IPCC 2007). Its independence from instrumental temperature measurements makes phenology a particularly attractive indicator of global warming (Anderson et al. 2013). Because observations of plant phenological phases date back up to several centuries, phenological observations can be used as a proxy for climate reconstruction (Rutishauser et al. 2008, Ge et al. 2014). Conversely, the state of the vegetation influences physical and biophysical feedbacks of climate change (Peñuelas et al. 2008). However, warming effects are potentially not stable over time (Rutishauser et al. 2008, Fu et al. 2015), other factors such as day length or precipitation intermingle (Körner and Basler 2009, Stöckli et al. 2011), or differ depending on their origin from networks or experimental sites (Wolkovich et al. 2012). Consequently, continued high-quality observations of phenology are of outmost importance (Rutishauser et al. 2012) in order to provide observational data for further analyses and model verification, and at the same time the historical record needs to be preserved.

The Swiss Phenology Network (SPN), was founded in 1951 and is maintained by the Federal Office of Meteorology and Climatology MeteoSwiss. Phenology has been defined a relevant parameter for the National Climate Observing System - GCOS Switzerland (MeteoSwiss 2018), and is as such recognized an important factor in climate monitoring for terrestrial observations of the biosphere.

Twelve phenology stations of Switzerland with the longest data series are part of the inventory of the most important climate observations in Switzerland (MeteoSwiss 2018). Several phenological data bases have been built up, including the European PEP725 data base, in which the observations of SPN are integrated.

According to the GCOS Monitoring Principles, ensuring high data quality and homogeneity of series is of outmost importance for long term climate monitoring (WMO 2016). For climate analysis the use of high quality, long, and homogeneous series is crucial (Seiz and Foppa 2007) in order to perform reliable analysis for, e.g. climate change applications. For instance, inhomogeneities in a series due to changes in observers, can hamper a reliable analysis of long term trends of phenological phases. Furthermore, series from the same station network can differ substantially in their record characteristics (e.g. length, completeness) and quality issues (e.g. homogeneity, data quality). Therefore it is necessary to develop methods and algorithms to identify high quality, homogeneous data series in larger datasets.

In this project, financed by MeteoSwiss in the framework of GCOS Switzerland, we developed methods to assess the data quality and homogeneity of each series of the SPN. We then translated the results as well as other relevant record information into a ranking system which allowed for a hierarchical listing of all series of the SPN. As a final result all series were ranked according to the newly developed classification scheme and the most valuable Swiss series and stations were defined.

1.2 Objectives

The goal of the project PhenoClass is the thorough assessment and subsequent classification of all data series and stations of the Swiss Phenology Network. The 12 selected phenology stations in GCOS Switzerland will be reviewed and suggestions will be made for a possible update of the selection of the most valuable phenology data series and stations in Switzerland.

The PhenoClass project aims at:

- The assessment of the data quality and homogeneity of the series of the SPN as basic criteria for the classification
- The development of a novel classification scheme for all phenological series of the SPN maintained by MeteoSwiss according to the criteria data quality, homogeneity, completeness, and length
- The identification of the most valuable Swiss phenological series and stations according to quality, homogeneity, length, and completeness

Methods for determining all criteria were developed, their subsequent indicator levels/states quantified for each record, and criteria had been weighted and translated into the classification scheme using a point system. Subjecting all series of the SPN to the procedure resulted in a hierarchical point-based ranking of all series and the identification of the most valuable Swiss phenological series (series with the highest score).

1.3 Data

The SPN comprises today 167 stations across Switzerland (Fig. 1). The onset dates of up to 69 different phenological events for 26 different plant species are currently being observed. Observed phenological phases are the dates of start of flowering and full flowering, leaf unfolding, leaf coloring, leaf drop, fruit ripening and additionally the date of hay harvest. Thus, for each observed phase, plant species and location, annual time series are available. Observations are performed preferably at the same plant over several years. The data are recorded on observations sheets and sent to MeteoSwiss, where they are entered into a data base. Possible sources of errors or inhomogeneity thus include changes of the observer, changes of the observed plants, changes in the surroundings as well as the age of the plant. The earliest observations were recorded in 1951 at 70 stations, 37 species had been observed (Defila and Clot 2001). A set of additional parameters started being observed since 1996; currently 69 different parameters are being observed. For the period until 2015, 9'455 series with 205'808 single observations exist.

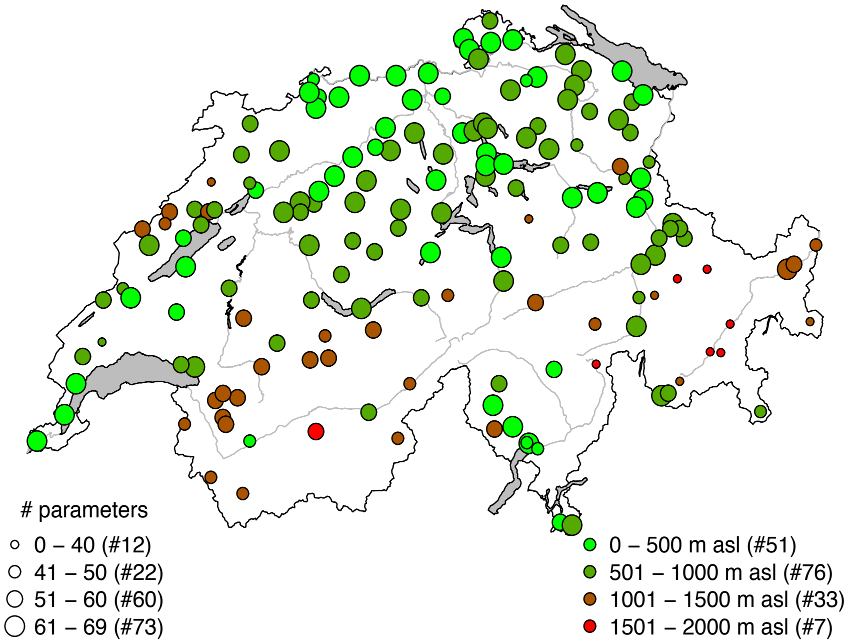


Figure 1: Location of phenology stations by height (colors) and number of parameters per station (point size). The legends show in brackets the number of stations for each category.

2 Methods

To identify the most valuable series of the SPN, criteria had to be defined. For each criterion indicators had to be specified. In accordance with the GCOS Climate Monitoring Principles (WMO 2016) the following criteria had been used in particular, with their subsequent indicators (in the format*:* criterion - indicator):

- Temporal coverage – length of record (Chapter 2.1.1, Level 0)
- Completeness I – missing values relative to length of record (Chapter 2.1.1, Level 0)
- Completeness II – number of data gaps >5 consecutive years (Chapter 2.1.1, Level 0)
- Reliability – number of quality flagged values relative to record length (Chapter 2.1.2.-2.1.6. Level 2-4)
- Stability over time – number of inhomogeneities in a record (Chapter 2.2)

At the level of series, statistics were compiled to assess the first three indicators (length, gaps, and number of single missing values). The fourth indicator was derived from the QC of the individual values (fraction of final flags per record). The fifth indicator is assessed based on breakpoint detection and an assessment of breaks with metadata. Additionally, at the level of stations, a statistics of diversity was compiled. It described the number of different species observed at the same stations. In the following we start with the statistics of the length and missing values per record and station (Chapter 2.1), the QC (Chapter 2.2) and the breakpoint detection (Chapter 2.3).

2.1 Summarizing Statistics of Length and Missing Values

For assessing the length, we assigned each series to one of six lengths categories, where length is defined as y_l - y_f + 1, where y_l and y_f are the last and first year of the series, respectively. Furthermore, the fraction of missing values as well as the number of gaps >5 years were calculated for each series.

2.2 Quality Control

The QC of phenological observations ideally detects defective observations (e.g. outliers) due to transmitting errors, typing errors, observation errors, transcription errors or similar. Such errors are usually of random nature, with some exceptions (e.g. continuous mix-up of species/columns).

Data quality control in phenological networks still awaits international standard procedures. For a few national networks, QC procedures have been developed. National data quality assurance and quality control reports are available from the website of the USA National Phenology Network or from the DWD (K. Zimmermann, pers. comm.). In Switzerland a QC procedure was developed for the SPN data and was applied after 2015 (Pietragalla et al. 2016), but not retroactively. A thorough QC and quality assessment of the entire SPN has never been undertaken.

For the QC of historical series, a procedure tailored to Swiss phenological series and stations has been developed, in close collaboration and coordination with existing routines and ongoing QC work at MeteoSwiss. Statistical methods, aided by expert knowledge at clearly indicated occasions, were combined for a reliable assessment.

Indicators refer to individual series, but QC was performed at the level of individual values. It consists of several automatic steps and an expert step. In Levels 1-3 automatic boolean flags were derived. Level 4 describes the expert inspection of all automatic flags by two experts, resulting in two code variables (one per expert) that pertain to the automatic flags. A final flag was then set based on combining the automatic flags and the expert codes. An overview of the QC test and their outcome (flags, codes) is provided in Table 1.

Table 1: Overview of the QC procedure.

QC Level	Test	Outcome
1a	Exceptional values (not in same year, consecutive identical values)	Flag (0/1)
1b	Implausible values	Flag (0/1)
1c	(Biologically) inconsistent values	Flag (0/1)
2a	Inconsistent with other parameter at same station	Flag (0/1)
2b	Inconsistent with same parameter at other stations	Flag (0/1)
3	Inconsistent with temperature	Flag (0/1)
4a	Expert control This Ruthishauser	Code
4b	Expert control Renate Auchmann	Code
5	Final flag	Flag (0/1)

2.2.1 Level 1: Absolute Value Tests

Level 1 is the first of three automatic QC Levels. Level 1 consists of absolute value tests. Three sublevels were defined:

- Level 1a: exceptional values. Observations that were observed in the year before or after the observation year (e.g. day-of-year >366 or <1), or three or more consecutive identical observations for a given parameter. Level 1a flags are not assumed to necessarily represent unreliable observations. The onset of a phenological phase may be outside the year considered (e.g. onset of hazel blossom in previous December). Likewise, three (or more) consecutive identical observations can appear by chance, however are very unlikely.
- Level 1b: implausible values. Observations that lie outside +/- 3 standard deviations of the record mean of each series.
- Level 1c: inconsistent values. Values that are inconsistent with the biological order of parameters. For assessing Level 1c flags a list with rules (original list provided by MeteoSwiss with adjustments) containing the biological order of parameter pairs was applied. For in-

stance, for all deciduous trees the start of flowering cannot be later than the full flowering. The list used contains 40 rules (Appendix A).

2.2.2 Level 2: Comparison with Neighbouring Stations and Within-Station Parameters

In Level 2 we compared each series (termed candidate series) either with single correlated series of the same station (Level 2a) or with correlated neighbouring series (Level 2b). These series are termed reference series and they were chosen according to the following rules:

- Level 2a: series of parameters at the same station with at least 10 overlapping observations and a correlation of r>0.6;
- Level 2b: series of the same parameter from other stations with at least 10 overlapping observations and a correlation of r>0.6.

Each of these series was used as independent variable in a simple linear least-squares regression to model the candidate series. A leave-one-out approach was then used such that each individual observation could be modeled from the remaining data. Depending on the number of available reference series, this procedure yields one or more standardized residuals ε_s (observed value minus modeled value divided by the standard deviation of model residuals) for each individual observation. This observation was flagged if $|\varepsilon_s|$ (if only one reference series was available) or the median of $|\varepsilon_s|$ (if several reference series were available) was larger than 3.

2.2.3 Level 3: Model Using Temperature

In Level 3 we used gridded temperature data from MeteoSwiss (Frei, 2014) to estimate the candidate observation. Three-monthly temperature means (of the closest grid point) were used to estimate the candidate observation (i.e. day-of-year; DOY). Note the temperature dataset by Frei (2014) only spans the years 1961-2011. Hence, for the Level 3, flags could only be derived inside that period.

We applied the method to all series that have 10 or more observations during 1961-2011. The candidate was left out in the estimation procedure. We used the 3-monthly temperatures before the mean onset date of the series (e.g. if mean onset date is July 10, April-May-June temperatures were used). Again, standardized residuals ε_s were calculated and values for which $|\varepsilon_s|$ was larger than 3 were flagged.

2.2.4 Level 4: Expert Control

In a last step (Level 4), all automatic flags from the Levels 1-3 were inspected by two experts, This Rutishauser (TR) and Renate Auchmann (RA). They inspected all automatic flags for each parameter at each station (~170 stations and 69 parameters) using a standardized inspection sheet (Appendix B).

Both persons individually assigned an additional variable number according to the following code:

- (1) Observation correct (overrules automatic flag)
- (2) Observation probably correct (overrules automatic flag)

- (3) Automatic flag probably ok, but cannot be ruled out that observation is correct (keep automatic flag)
- (4) Automatic flag correct, observation problematic (keep automatic flag)
- (5) No automatic flag attributed, but expert flag additionally added

The difference between variable number 1 and 2 or number 3 and 4 rests with the subjective assessment of the experts. Additionally, three new columns for comments were added:

- commentsFlag: comment on one specific flag (a large number of flags are added a comment)
- commentsPhase: comments on candidate parameter/series (if any)
- commentsStation: comments on candidate station (if any)

2.2.5 Level 5: Final Flags

Based on the automatic flags and the expert assessment of the flags, final flags were set in the following way: If either TR or RA set code 5 (expert flag), the observation was flagged. In all other cases, the lower of the two codes (TR and RA) was applied to the automatic flags. The thresholds for automatic flagging were chosen relatively rigorous, such that many values go through visual inspection. Choosing the lower of the two codes now allows many of the rigorous automatic flags to be overruled.

2.3 Homogeneity Assessment

Historical phenological series, just like meteorological series, are prone to inhomogeneities caused by factors that affect the way observations are carried out (e.g. a change of the observer, change of observed plant, change in the environment, among others). Inhomogeneities can express themselves as shifts in the mean, trends in the mean, changes in variance, or changes in other statistics. Shifts in the mean (e.g., due to a change of the observer or the change of the observed plant) can be detected using statistical methods, yielding statistics of so-called “breakpoints” (time points with a high probability for a shift in the mean). Other types of inhomogeneities (e.g., trends due to changes in the station surrounding or the age of the plant) are more difficult to detect. Series that are affected by shifts in the mean should not be used for trend analyses, hence it is important to identify these series.

We used an algorithm for the detection of changes in the mean similar to that used for Swiss temperature series in Kuglitsch et al. (2012). We independently applied three statistical tests to each phenological series that has at least 20 observations (shorter series constitute a too small sample for meaningful statistical testing). The agreement among the three tests determines which breakpoints are to be considered significant. Each test is applied to difference series between the candidate and well-correlated reference series.

The whole procedure is fully automatic and reproducible, the detection does not involve subjective decisions after the initial parameters are set.

2.3.1 Statistical Tests

The tests that we used are the following:

- Standard Normal Homogeneity Test (SNHT), as described in Alexandersson and Moberg (1997);
- Pettitt's test, as described in Pettitt (1979);
- penalized maximal *t* test, as described in Wang (2008), implemented in the RHtests software version 5 (http://etccdi.pacificclimate.org/software.shtml).

In principle, their purpose is to test a null-hypothesis of homogeneity, in which the normalized difference series are randomly distributed around zero. The p-value threshold for significance was set to 0.05 in all three tests. Using multiple tests reduces false detections (Kuglitsch et al. 2012). However, the detection power of the test decreases towards the ends of a series, i.e., breaks that are close to the beginning or end of a series have a smaller chance to be detected. Toreti et al. (2012) estimated a decrease of the probability of detection between 25-75% (depending on the method used) for a breakpoint located at the 20th year of a 100-year long series, compared to a breakpoint located at the middle of the series. SNHT resulted to be the method with the lowest decrease (i.e., hit rate less sensitive to the position of the breakpoint), but also the one that gives the largest number of false detections.

Another difficulty concerns multiple breaks in a series. In general, the closer to each other two breakpoints are, the more difficult it becomes to detect them, because of the reduced sample. Two breakpoints occurring in two consecutive years cannot be detected at all. Moreover, the nature of the tests, in particular of SNHT and Pettitt's, implies that single breakpoints are much more likely to be detected than multiple breakpoints. This is because the tests can only detect one breakpoint at a time. When a breakpoint is detected, the tests are then applied to the two sub-periods separated by the breakpoint to look for additional breakpoints, and so on recursively until the sub-periods become too short or no additional breakpoints are found. The penalized maximal *t* test does not have this shortcoming, however it is in general less powerful than the other tests (using it alone would have meant the non-detection of more than 60% of the single breakpoints that we found, see Sect. 3).

A breakpoint is considered detected by a certain test if the test finds it in at least three difference series (candidate minus three reference series). Due to the noise of the series, the year assigned to a breakpoint is affected by some uncertainty. For this reason we allowed a tolerance of one year (for example, if the first difference series has a breakpoint in 1979, the second in 1980, and the third in 1981, then these are considered to be the same break (occurring in 1980). An iterative procedure was developed to do that attribution that starts with 0 tolerance and then increases to 1 year. This way the year with more detections is preferred. After defining breakpoints for each tests, the results of all three tests were compared. If two or three tests detect the same break (again +/- 1 year) a breakpoint is set ("significant" breakpoint).

2.3.2 Differences Between Temperature and Phenological Series

The breakpoint detection method that we used has been validated for temperature. Phenological and temperature series have similar statistical properties. For instance, annual temperature means are normally distributed; 86% of the Swiss phenological series with at least 20 observations do not differ significantly (p=0.05) from a normal distribution, according to the Lilliefors (Kolmogorov-Smirnov) test (i.e., 14% instead of the expected 5% are not normally distributed).

However, phenological series are in general less correlated in space (Güsewell 2014) than temperature, which makes the detection of breakpoints less effective. This is mainly because the behaviour of plants is more complicated than that of temperature, being dependent on both meteorological and biological factors. Temperature series (annual means) usually have correlations larger than 0.9 when distances between stations are in the order of tens of kilometres, whereas many phenological series do not reach correlations above 0.7 for the same distances.

Phenological series also show larger inter-annual variability in comparison with temperature, which negatively affects the signal-to-noise ratio of inhomogeneities. They are in some way similar to precipitation series, for which the same statistical methods used for temperature are usually applied, although with significantly lower detection scores (Venema et al. 2012).

2.3.3 Selection of Reference Series

We tested five different approaches of how to select reference series. The approaches made use of information such as phase/species, altitude, correlation, overlap, and tolerance year (Table 2). The experiments differ in the combination of selection criteria (except correlation, which was always set to 0.6). Table 2 shows an overview of the experiments.

Experiments ALL1 and ALL2 did not use biological constraints. We used eight reference series with one (ALL1) or two (ALL2) years tolerance (Sect. 2.2.1). The use of any reference series that is well correlated with the candidate, independently from its nature had the advantage that enough reference series could be found for every record in every period. Inter-species correlations are often as large as those between the same species. Using other species as reference, however, would increase the risk of misinterpreting different biological reactions to forcings such as a rapid warming. This, however, raised concerns about the different ways that different species and phases can react to climate change and other forcings.

In experiment 8REF we introduced a biological constraint, in which only correlated series of the same phenological phase were accepted. However, we considered some phases to be enough biologically related for them to be used as references of other phases. These are the start of flowering with the full flowering and the leaf/needle colouring with the leaf/needle drop (always for the same species). Additionally a maximum difference in onset days of 30 days and again 8 references were used. In 8REF a reference series could not come from the same station of the candidate series unless no other alternatives were available, to avoid simultaneous inhomogeneities due to changes of observer. The number of tested series dropped in this experiment to 35% compared to ALL1 or ALL2.

To enlarge the number of series that could be tested and at the same time reach a low false detection rate, we tested experiment 5REF (marked in red) which needed only five reference series. A further experiment 5REF_NOQC (same as 5REF but using not quality controlled data) was performed as an indirect assessment of the QC. It was not considered further, but the fact that fewer breakpoints were found despite the larger number of series indicates the importance of QC prior to the breakpoint detection.

Table 2: Overview of the experimental break detection setups.

	Experiments				
	ALL1	**ALL2**	**8REF**	**5REF**	**5REF_NOQC**
biological constraint	No	No	Yes	Yes	Yes
statistical tests	2 / 3	2 / 3	2 / 3	2 / 3	2 / 3
reference series	3 / 8	3 / 8	3 / 8	3 / 5	3 / 5
min overlap	No	No	90%	90%	90%
min length	10 yrs	10 yrs	20 yrs	20 yrs	20 yrs
max elevation diff.	1000 m	1000 m	750 m	750 m	750 m
max onset diff.	No	No	30 days	30 days	30 days
min correlation	0.6	0.6	0.6	0.6	0.6
tolerance	1 year	2 yrs	1 year	1 year	1 year
quality controlled	Yes	Yes	Yes	Yes	No
series tested	7393	7393	2566	2925	2951
breakpoints	485	644	330	156	141

The main criteria of the references of experimental setting 5REF are:

- Pearson's correlation with candidate series must be at least 0.6;
- number of missing values in the years covered by candidate must be as close as possible to zero and no larger than 10% of candidate’s length;
- biological constraint: reference must be same species and phase of the candidate, with a few exceptions listed above;
- elevation difference between candidate and reference station cannot be larger than 750 meters;
- mean onset difference between candidate and reference record cannot be larger than 30 days.

If more than five series fulfil all requirements the five series with most overlapping observations were considered. Note that the reference series themselves might have breaks (i.e., a break detected from comparing two series could be in either of the series), hence we required that 3 out of 5 reference series must indicate the break (Table 2). Given the ratio of breaks to number of series (ca. 5%) it is very unlikely that two or three reference series have a break at the same time.

Kuglitsch et al. (2012) used ten reference series for temperature. Successive validations that were carried out using benchmark products (Venema et al. 2012) showed that the best scores are obtained with eight series. For phenological series, we found a potential issue of false detections, because plants are very sensible to local factors. For example, the phenological reaction to a rapid warming in a region with dry climate, such as the Wallis, might be different from that in a wetter region, like the Tessin, and this would create a discontinuity in some difference series that could be interpreted as an inhomogeneity. Therefore, we chose to use five references, and a breakpoint has to be seen by the majority of the references (three out of five) to be confirmed. This has of course the side effect of reducing the power of detection of the algorithm, particularly for mid-size breakpoints, but also has the advantage of increasing the quantity of data for which it is possible to perform the breakpoint detection (+16%). The 5REFprocedure guarantees some consistency in the method (i.e. the probability of finding a breakpoint is similar for each and every data point). If this is possible only for a sub-period of a record (i.e. by picking a later starting year), then the breakpoint detection is performed only on that sub-period (this affects 29% of the analysed series). Based on these considerations, we used the 5REF setting in this project (Table 2, marked red).

2.3.4 Use of Metadata

Possible reasons for inhomogeneities include changes in observers, changes in observed plant, changes in the environment or the age of the plant, among others. Of these, only observer changes are available in the metadata. The years when changes of observer occurred are used to adjust the position of detected breakpoints (metadata adjustment): if a breakpoint is detected one year before or after the year when the observer changed, it was moved to the year of the change. Metadata adjustment was done for each of the three tests separately and again on the final set of breakpoints.

In a similar fashion, breakpoints were forced to be at the year preceding a large gap (>=3 years) if a statistically detected breakpoint appears one or two years before the first gap year or if it appears in the first year of observation after a gap of more than 3 years.

2.4 Classification Scheme

In order to assess series and stations, we combined the statistics of the length and completeness of the series with the statistics from the QC of the individual values and the results from the breakpoint detection. For each of the five indicators (Sect. 2, first paragraph), a set of five thresholds were defined in order to partition the range of possible values into five or six classes (Table 3). Attributing each class a colour (from red to green) results in a "traffic light" classification system that gives a colour for each indicator for each record.

In order to obtain a single score summarizing all criteria for each series, the six classes were assigned points (0, 0.2, 0.4, 0.6, 0.8, 1), which were then combined by forming a weighted average (users can reweight the criteria according to their needs), termed score. We tested the weights to ensure a reasonable hierarchy of series.

This final weighted score was then again classified into a limited number of quality classes (e.g. “highly valuable series”) based on the “traffic light system”. This procedure was performed for each record as well as (with the additional criterion of diversity) for each station. This classification scheme could (for future applications) be overlaid by others (e.g. climate regions), if required.

2.4.1 Criteria, Indicators, Thresholds

The criteria used for the classification are shown in Table 3 (leftmost column), together with their indicators, units and information on whether the indicator states had been derived automatically (A) or by expert knowledge (E; given in brackets). Five of the six indicators are applied to single series (top five rows), one indicator is only applied to stations (last row of Table 3).

For each indicator five or six thresholds had been set. When possible, meaningful category names were added. For instance, for the criterion temporal coverage and its indicator “length of series” [yrs] the following thresholds/bins were defined (Table 3, first row):

- (0, 5] years (category name: “too short”)
- (5, 10] years (“very short”)
- (10, 20] years (“short”)
- (20, 30] years (“medium long”)
- (30, 50] years (“long”)
- (50, 100] years (“very long series”)

Thresholds were obtained by expert knowledge and visual inspection of distributions to obtain either almost equally spaced bins (e.g. Temporal Coverage, Completeness I, Completeness II, Long-term Stability) or bins that are based on the distribution of the indicator states (e.g. Reliability, Diversity; the distributions of the fraction of quality flags per series and the number of different series observed at one station are both highly skewed). Note that for the criteria “Reliability” an additional option is possible: If the series could not undergo the break detection (due to not meeting length requirements or not enough suitable reference stations) no class was assigned.

Table 3: Criteria and their indicators (A for automatically derived, E derived by expert knowledge) with thresholds (colored fields) and points assigned to each series (top part of table) and station (bottom part of table) according to its respective indicator state (traffic light system). Second last column: for Long-term Stability not all series underwent a break detection and are marked with “not tested”. Most right column: Weight for each criterion in the final classification from 0 to 1.

	0 points	0.2 points	0.4 points	0.6 points	0.8 points	1 point	no class	weight
Temporal coverage (Length/Period from first to last observed year [yrs]; A)	[0, 5) too short	[5, 10) very short	[10, 20) short	[20, 30) medium long	[30, 50) long	[50, 100) very long		[0-1]
Completeness I (Fraction of single missing values of series [%]; A)	[75, 100) too incomplete	[50, 75) very incomplete	[25, 50) many missVals	[10, 25) some missVals	[0.00001, 10) single missVals	[0, 0.00001) complete		[0-1]
Completeness II (Number of gaps >5 years per series, A)	>3	3	2	1		0		[0-1]
Reliability (Fraction of quality flags per series [%]; A, E)	(10, 85]	(5.5, 10]	(4, 5.5]	(2.5, 4]	(1, 2.5]	0		[0-1]
Long-term stability (Number of breakpoints per series; A)	more than 3 breaks	3 breaks	2 breaks	1 break		0 breaks	not tested	[0-1]
Criteria for Stations								
Diversity (Number of different species observed at station; A)	(1, 10]	(10, 19]	(19, 21]	(21, 23]	(23, 25]	(25, 27]		[0-1]

2.4.2 Weights and Definition of Classes for Data Series and Stations

For forming a score for each series, the criteria are further weighted. Equation 1 shows the calculation of the weighted average $\overline{x}$ for a single series, where x_i are the points assigned to each criteria i, w_i is the weight and the number of criteria used.

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i\ w_i)}{\sum_{i=1}^{n} w_i} \qquad \text{(Eq. 1)}$$

Criteria that are more important were assigned higher weights than criteria that are less important for assessing high quality series, with 0 being the lowest weight (i.e. criterion is not being accounted for; Table 3, last column). Thirty percent of the weight was attributed to long-term stability, which thus receives most weight. The remaining 70% were distributed equally among the factors “temporal coverage”, “completeness” (internally subdivided into 2 indicators), and “reliability” (Tests using other weights showed that equal weights for each criterion put emphasis on the completeness because of two completeness criteria). This resulted in the following weights for single series:

- Temporal coverage: 0.23
- Completeness I: 0.12
- Completeness II: 0.12

- Reliability: 0.23
- Long-term stability: 0.3

When all five criteria (for single series) are applied the weights directly represent percentages (i.e. the sum of all weights equals 1) for calculating a mean score for the series. If only four could be assessed (no breakpoint detection possible) the weights correspond to 17% (Completeness I and Completeness II) and 33% (temporal coverage, reliability).

Table 4: Definition of classes for series.

Class	Definition of quality class	Score
1	most valuable series: very long, complete, homogeneous, no quality flags (maximum weighted points)	(0.9999,1]
2	highly valuable series	(0.95, 0.999]
3	very valuable series	(0.9, 0.95]
4	valuable series	(0.85, 0.9]
5	medium valuable series	(0.75, 0.85]
6	low valuable series	(0.6, 0.75]
7	very low valuable series	(0, 0.6]

To obtain a classification of the series, the weighted series score was again partitioned into seven classes. Class 1 is reserved for series with a maximum score of 1. The remaining classes refer to bins of 0.05, with Class 7 comprising all series with scores below 0.6. The bins for the station ranking were defined with the exception of the last class with a uniform size from Class 1 to Class 6 (Table 4).

Table 5: Definition of classes for stations.

Class	Definition of quality class	Score
1	most complete stations	(0.9, 0.95]
2	highly valuable stations	(0.85, 0.9]
3	very valuable stations	(0.8, 0.85]
4	valuable stations	(0.75, 0.8]
5	medium valuable stations	(0.7, 0.75]
6	low valuable stations	(0.6, 0.7]

To obtain a final station score, the weighted scores of all series per station were averaged and weighted with 0.9. Additionally, the “Diversity” criterion was added with weight 0.1. Hence diversity has a much lower weight than all other criteria and only little effect on the final station ranking (which is intended). According to the station score, stations were assigned to a final quality class (Table 5), again using steps of 0.05.

The final result of our assessment is a plot for each series displaying the “traffic light”, a quality class for each series and a quality class for each station. In addition, we provide a separate recommendation for GCOS stations, i.e. the main sites for phenological observations in Switzerland as defined in the report ‘National Climate Observing System’ (MeteoSwiss, 2018), for which length and stability are the most important criteria. Our recommendation is based on the length of the series and the results from the breakpoint detection (this is introduced in Sect. 3.6).

3 Results

3.1 Quality Control

The assessment of the length of the series is shown in Figures 2 and 3. The majority (more than 50%) of the 9˙455 existing series are “short” to “medium long” (i.e. 0-30 years long). In all, 15.3% of the series are 50-65 years long, many of those “very long” series are located on the Swiss Plateau.

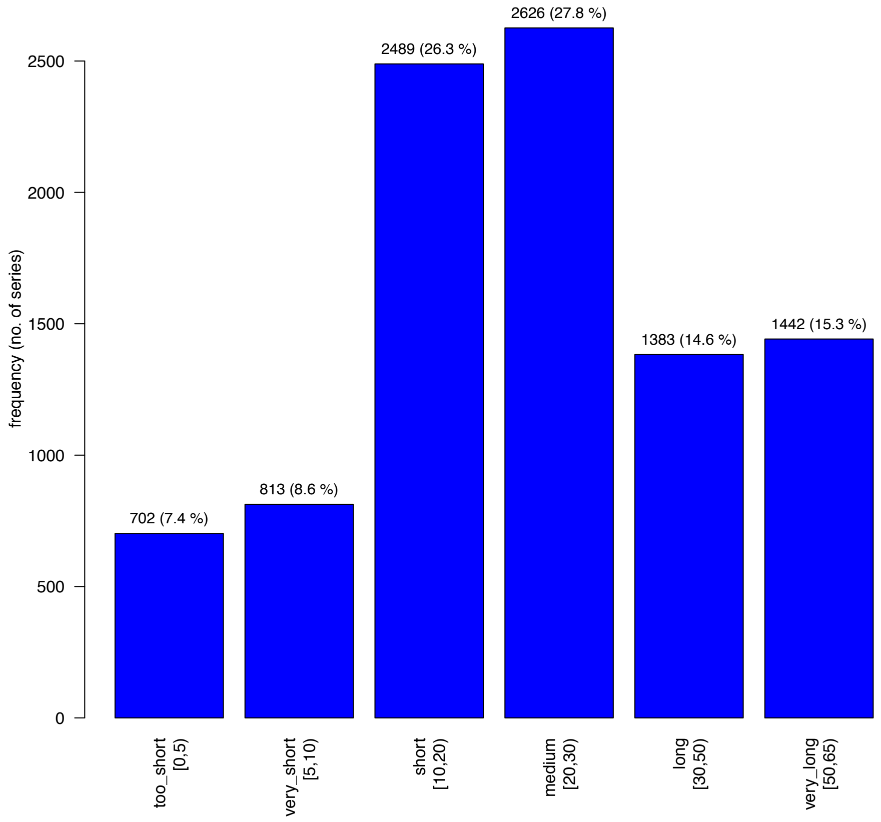


Figure 2: Absolute and relative (in brackets) frequency of single series per length category.

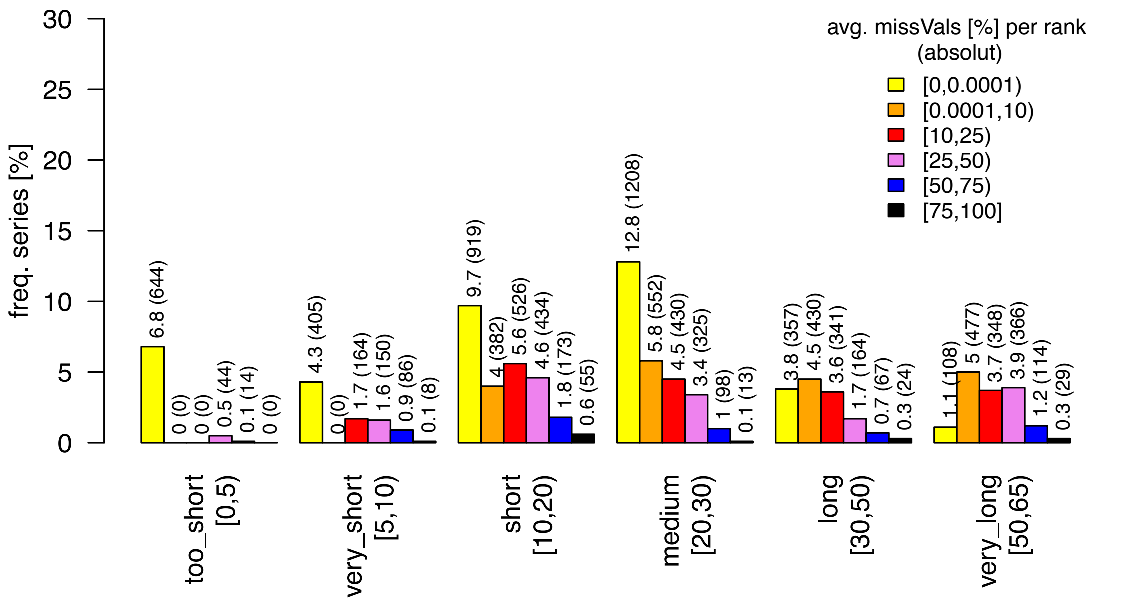


Figure 3: Relative frequency of series [%] for each length category (x-axis) and percentage of missing values (color, e.g. yellow indicates complete series without missing values). Absolut number of series are given in brackets.

For all series the number of missing values and the number of data gaps (data gaps are defined as consecutive missing observations >= 5 years) have been determined. In all, 82.0% of all series (=7'757 series) have no large gap >5 years. 14.5% of the series (=1'371 series) have one large gap, only ~3.5% of the series have two or more large gaps. Furthermore, 3'641 (38.5%) series are complete (i.e. do not have a single missing value); see yellow bars in Figure 3.

The automatic QC yielded 4'764 flags (overall, including multiple flags for a single observation from various levels of QC). 4'019 observations have at least one automatic flag, this is 1.96% of the observations. Of those 4'764 automatic flags 43% (absolute: 2'050) are set in Level 1, 32% (abs. 1'513) in Level 2 and 0.25% (abs. 1'191) in Level 3 (Fig. 7). The relative distribution of the expert flags of the two observers is shown in Figure 4. Expert inspection diverges mostly for code 3 and 4. Expert TR assigned 275 times code 3, expert RA 1'461 (they diverge by 1'186). For code 4, expert TR assigned 2'239 times code 4, expert RA 1'088 times (they diverge by 1'151). However totals of “overruled flags” (code 1 & 2) and “kept flags” (code 3 & 4) diverge very little. For the total of code 1 and 2 the divergence is 35 (TR assigned either code 1 or 2 1'505 times, RA 1'470 times), for the total of code 3 and 4 the divergence is also 35 (TR assigned 2'514 times either code 3 or 4, RA 2'549 times). The divergence in the expert inspection most probably stems from TR being more “sure” about keeping or removing flags, RA showing a “central tendency”.

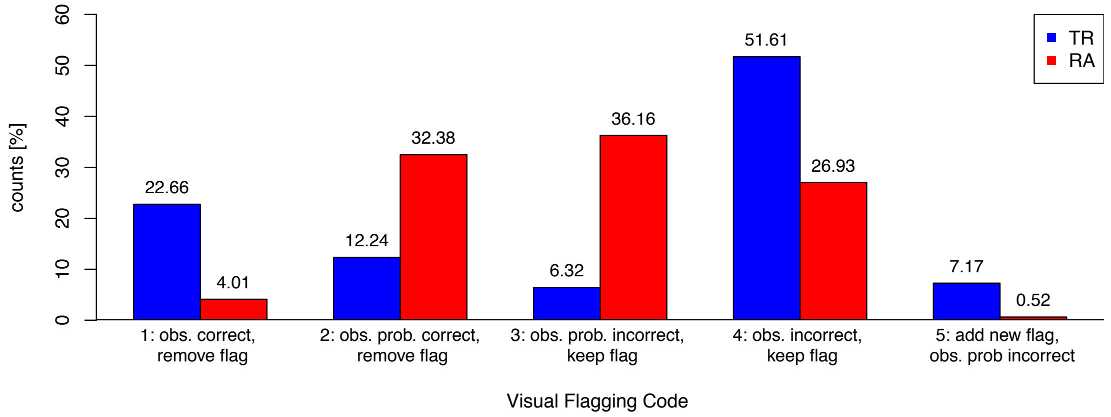
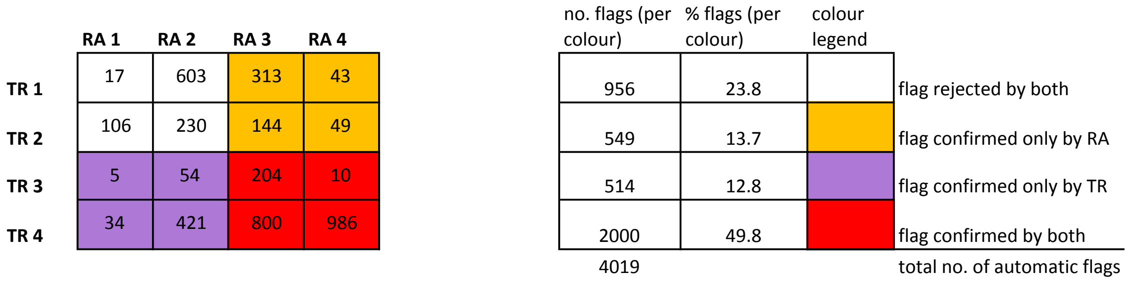


Figure 4: Relative distribution of expert codes of the two inspectors (TR: T. Rutishauser (blue), RA: R. Auchmann (red)). Sum of all red bars =100% (same for blue bars).

For single codes Figure 5 shows the distribution (in absolute numbers) of expert codes from the two inspectors in the evaluation of each flag. Overall, in 49.8% of the cases RA and TR agree for the category “keep flag” (Fig. 5, red fields), in 23.8% of the cases RA and TR agree about “remove flag” (Fig. 5, white fields). Hence in total, in 73.6% of all cases TR and RA agree about the general removing or keeping of an automatic flag. In 26.4% of the cases RA and TR disagree about the general category (for 12.8% RA assigned a “remove flag”, where TR assigned a “keep flag”; Fig. 5, violet fields, and for 13.6% the opposite case applies; Fig. 5, orange fields). In the final flag categories these cases with a disagreeing evaluation, the flag was cancelled for being of the safe side and not flagging data, which probably could be correct (see method of using the minimum code of the evaluation of both experts for attributing the final code).



	RA 1	RA 2	RA 3	RA 4
TR 1	17	603	313	43
TR 2	106	230	144	49
TR 3	5	54	204	10
TR 4	34	421	800	986

no. flags (per colour)	% flags (per colour)	colour legend	
956	23.8	⬜	flag rejected by both
549	13.7	🟧	flag confirmed only by RA
514	12.8	🟪	flag confirmed only by TR
2000	49.8	🟥	flag confirmed by both
4019			total no. of automatic flags

Figure 5: Distribution (absolute frequencies) of expert codes by code category and inspector (RA/TR 1 and 2 = remove flag, RA/TR 3 and 4 = keep flag).

Figure 6 shows the distribution of the final codes. Around 49.8% of the 4019 automatic flags were confirmed, 319 additional flags were introduced (code 5; this code was assigned when experts found a quality issue during the inspection that was not captured by the automatic QC). Final flags affect 2319 observations (= ~1.13% of all observations were attributed a final flag, 1.96% were attributed an automatic flag). Hence, a combination of automatic and expert quality control methods enabled a reduction of automatic flags by 49%.

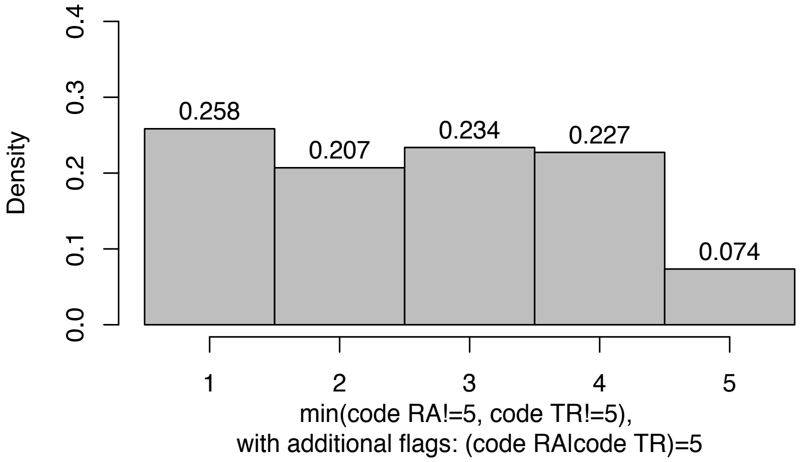


Figure 6: Distribution of the combined expert codes (min(code_TR, code_RA) of the two inspectors (TR: T. Rutishauser, RA: R. Auchmann), or 5 if any of the 2 inspectors added a new expert flag (code 5).

Figure 7 shows the distribution of final flags per QC Level. For Level 1, 1390 (67.8%) of the 2050 automatic flags were confirmed (Level 1: absolute value check). Of automatic flags of Levels 2 (Comparison with neighbouring stations and within-station parameters) and 3 (model using temperature), around 30-40% were confirmed. This information could potentially be used for e.g. a weighting (or probabilistic) procedure in an automatic-only quality control algorithm.

Figure 8 shows the Swiss map with all stations by their relative number of flags per station (in %). Sta. Maria is the station with most relative flags.

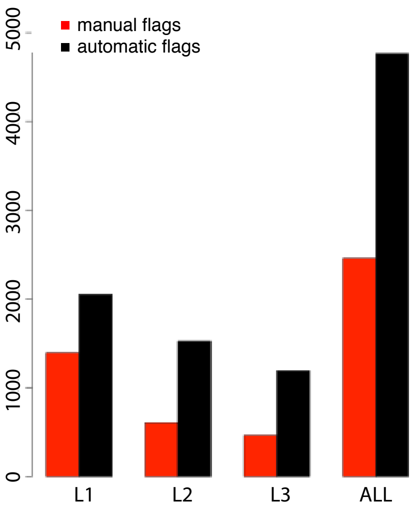


Figure 7: Comparison of automatic flags (black) and final flags (red) by QC Levels (e.g. Level 1 = "L1") and for all flags("ALL", right bars).

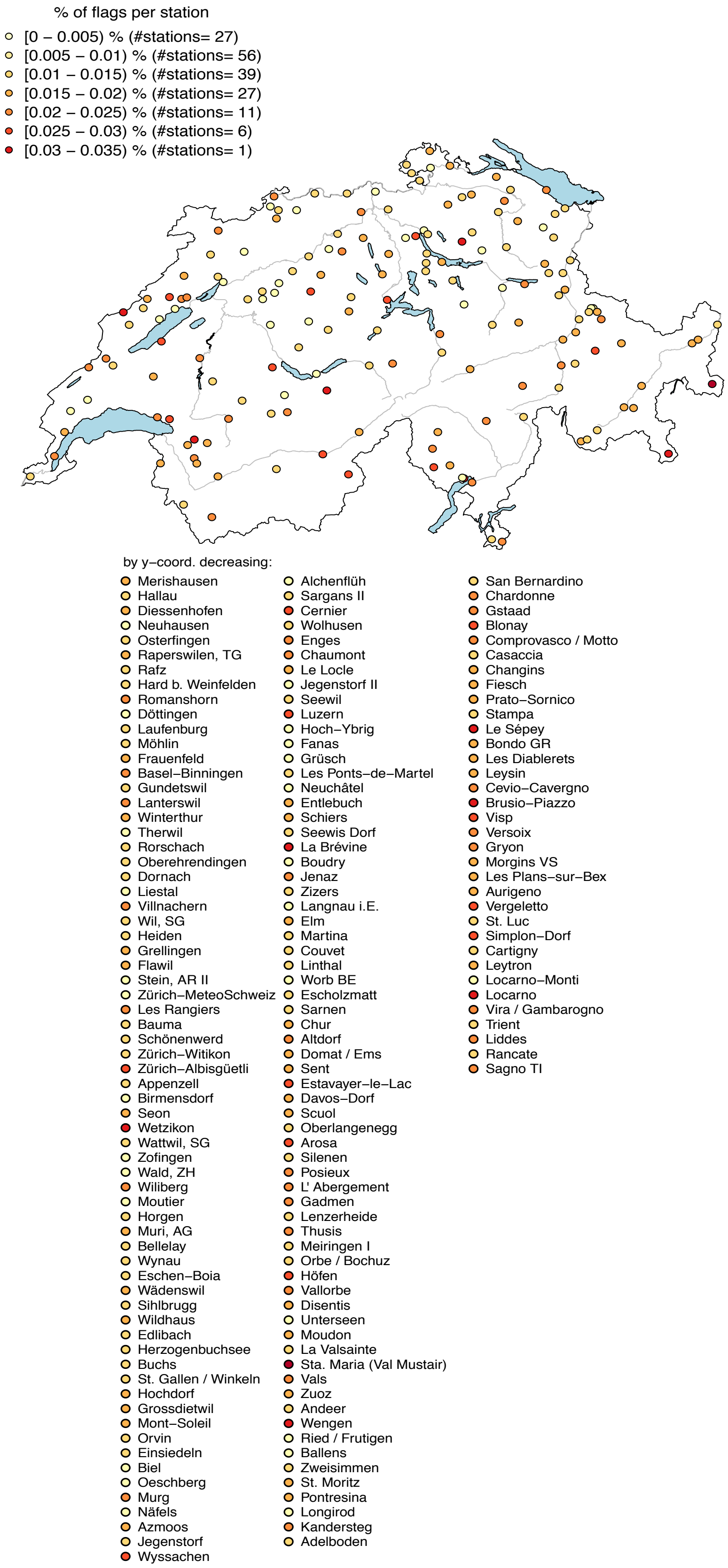


Figure 8: Relative number of flags per station.

3.2 Homogeneity Assessment

The breakpoint detection algorithm described above was applied to 2ˈ925 of a total of 9ˈ455 phenological series, resulting in 156 significant breakpoints. 153 series were found inhomogeneous (i.e. they have at least one significant breakpoint). The observations that have been flagged during the quality control were excluded from the breakpoint detection. By ignoring the quality control, experiment 5REF_NOQC, we obtained 8% less breakpoints, i.e. the quality control had a noteworthy impact on the performance of the breakpoint detection. Figure 9 shows an example of a record that has been found inhomogeneous by the algorithm.

The example is for the full flowering of the horse chestnut (*Aesculus hippocastanum*) in Altdorf. In this example we have high correlations (up to 0.8) with the reference series and all of them have all the 35 years covered by the candidate. One significant breakpoint is detected in 1995, related to a change of observer. This breakpoint was detected by two of the three tests (SNHT and Pettitt) by three reference series each. Hence, the breakpoint is barely significant. The second panel in Figure 11, giving the standardized differences, shows that until 1995 the flowering in Altdorf was usually among the latest of all sites, while after 1995 Altdorfis often the earliest (except for the last few years). A similar breakpoint in 1995 (not shown) corresponding to the same change of observer was detected at the same station for the full flowering of the European elder (*Sambucus nigra*) and of the field daisy (*Leucanthemum vulgare*), the latter with the highest possible significance (all reference series in all tests saw the breakpoint); this adds confidence that a change of observer did cause inhomogeneities in the Altdorf series in 1995.

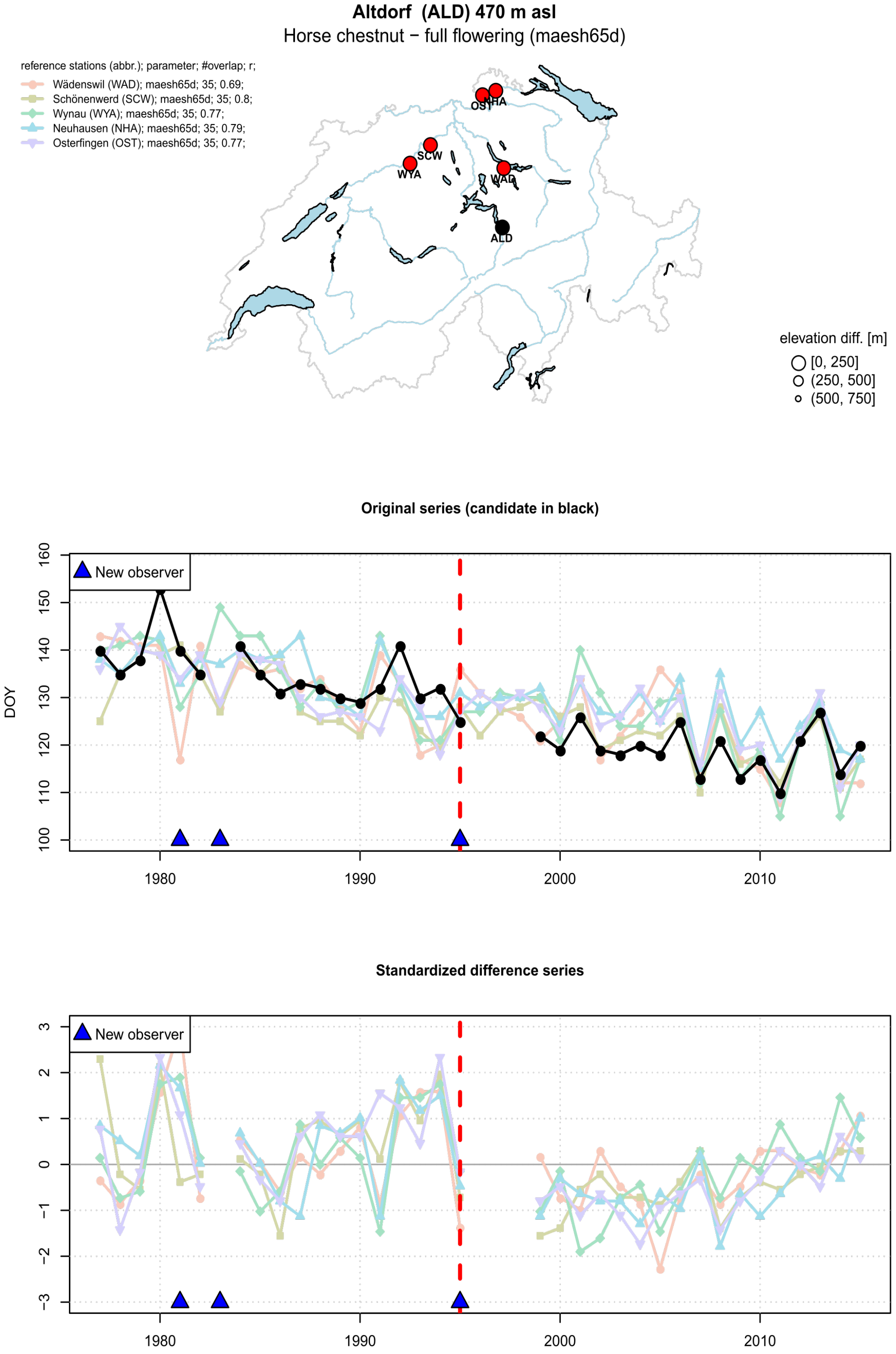


Figure 9: Breakpoint detection summary plot for the full flowering of the horse chestnut in Altdorf. The map shows the position of the candidate (black point) and reference series (red points), the vertical dashed line in the time series indicates the position of the breakpoint.

Figure 10 shows the occurrence of breakpoints for each year, as well as the occurrence of changes of observer. Aside from the 1950s (where the low number of stations inflates the frequency of observer changes), the number of observer changes is particularly high at the end of the 1980s, in the mid-1990s and in the late 2000s. As one would expect, the main peaks in the occurrence of breakpoints are close to those in the changes of observer. The 1950s are again a special case: here no breakpoints were detected, because the quantity of data is too small and often not enough suitable reference series can be found. Similarly, in the 2000s the breakpoint detection works less well, because the sample after the breakpoint is too small.

The largest number of breakpoints is detected in 1987, a year with a not particularly high number of changes of observer. The 14 breakpoints detected in this year affect only 9 stations: two of them (Wildhaus and Zweisimmen) have 2 breakpoints, while the station of Murg is accountable for 4 breakpoints caused by a new observer, representing nearly 30% of the breakpoints in that year. A similar coincidence of multiple breakpoints can be found for 1995. Therefore, caution is required in the interpretation of isolated peaks in the breakpoint frequency, which are strongly affected by random noise due to the limited number of total breakpoints. It is also important to remark that breakpoints in the 1980s and 1990s are more often detected also because those years are usually in the middle of long series (1986 and 1993 are the two most common middle years).

The period 1986-1989 shows 3 to 4 times more breakpoints than surrounding 4-year periods. This is a period characterized by a rapid temperature increase in Switzerland and similarly rapid changes in phenological variables (Schleip et al. 2008, Reid et al. 2016), but unfortunately it also coincide with a particularly large number of new observers (about 20% of stations affected). If there was an increase of false detections caused by the rapid climate change, we would expect the fraction of breakpoints related to changes of observer to decrease. However, for 1986-1989 this fraction is 57%, even larger than the overall average of 54%.

We conclude that the large number of breakpoints detected at the end of the 1980s and in the mid-1990s are mainly related to more frequent new observers and better operating conditions for the homogeneity tests.

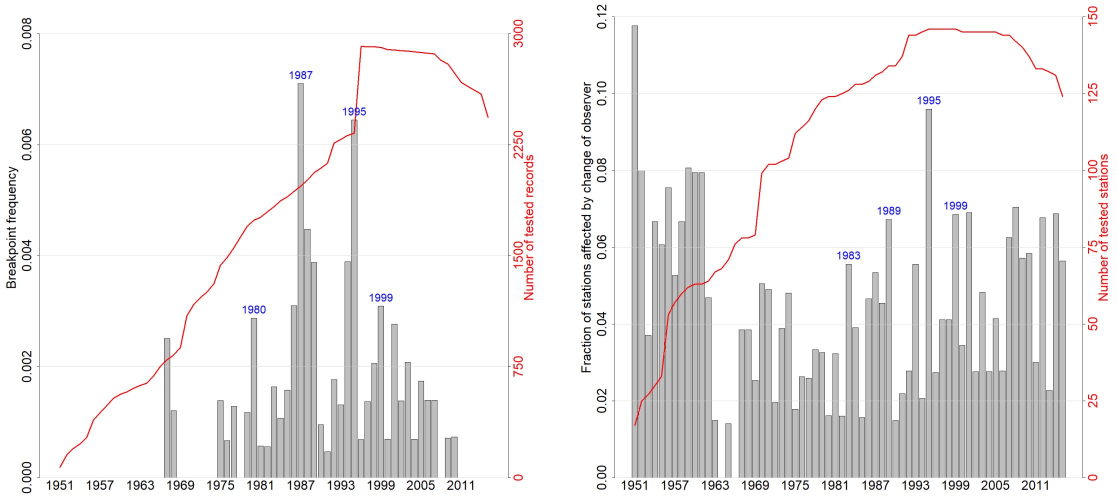


Figure 10: Left: Number of significant breakpoints detected relative to the number of tested series. Right: Number of changes of observer relative to the number of tested stations. The red lines depict the number of tested series/stations.

3.2.1 Feasibility of the Breakpoint Detection

The breakpoint detection was applied to 73.9% (2ˈ925) of the phenological series with at least 20 years of observations (thereof, 2ˈ082 entire series where subjected to the break detection, for 843 series only segments could be used). For the remaining 26.1% with at least 20 observations (1ˈ035) it was not possible to find enough suitable reference series.

In general, late phases (fruit maturity, leaf colouring, leaf drop) have lower spatial correlation than spring phases, because they are less strongly driven by temperature, therefore it is harder to find suitable reference series for them.

Figure 11 shows a map of the breakpoint detection feasibility. Here we clearly see that the mountainous regions (Jura and Alps) are those where finding reference series is more difficult. Even on the plateau, though, some stations have one quarter of the parameters with insufficient reference series.

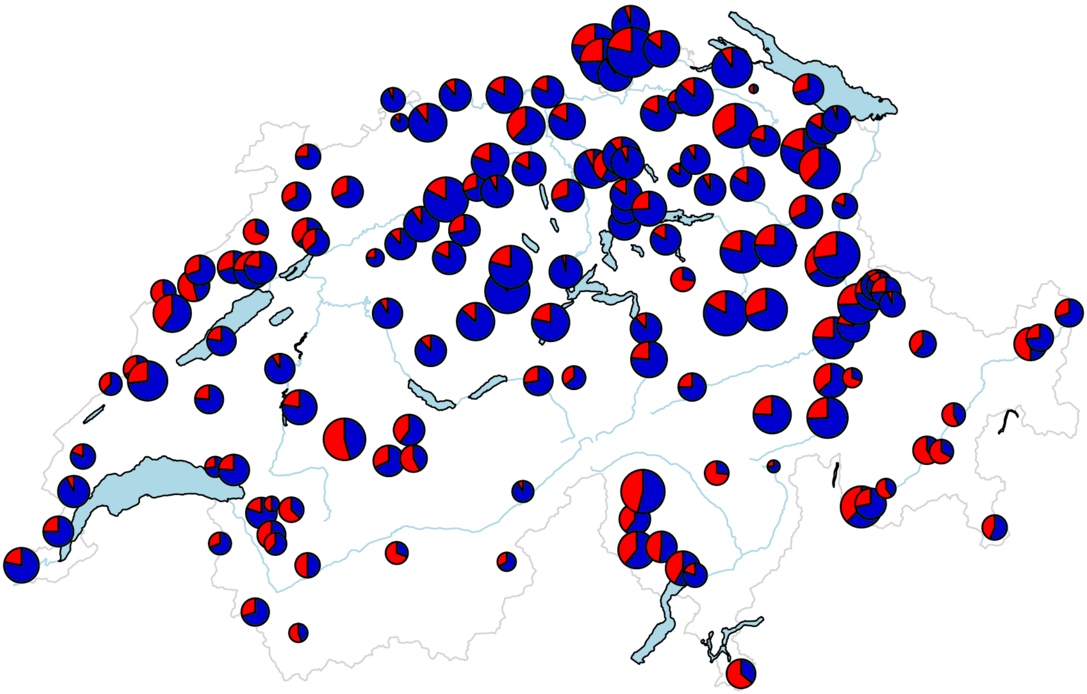


Figure 11: Breakpoint detection feasibility for each station. The red fraction of the pies is the fraction of series with at least 20 observations that did not have enough reference series. The area of the pies is proportional to the number of parameters.

3.2.2 Elevation Difference

Figure 12 shows a statistics of correlation as a function of elevation difference between candidate and reference series. As shown by the red line, more than half of all reference series used in the whole data set were drawn from stations no more than 125 meters higher or lower than the candidate station. The impact of the elevation difference on correlation almost disappears already above 250 meters. Also the differences between phases fade quickly, whereas until 375 meters the phases related to flowering have on average larger correlations.

There are no appreciable differences among species in the correlation changes with the elevation (not shown). Even the mean absolute values of the correlations for a given elevation difference range do not differ significantly.

These results support our choice of using reference stations that lie up to 750 metres higher or lower than the candidate.

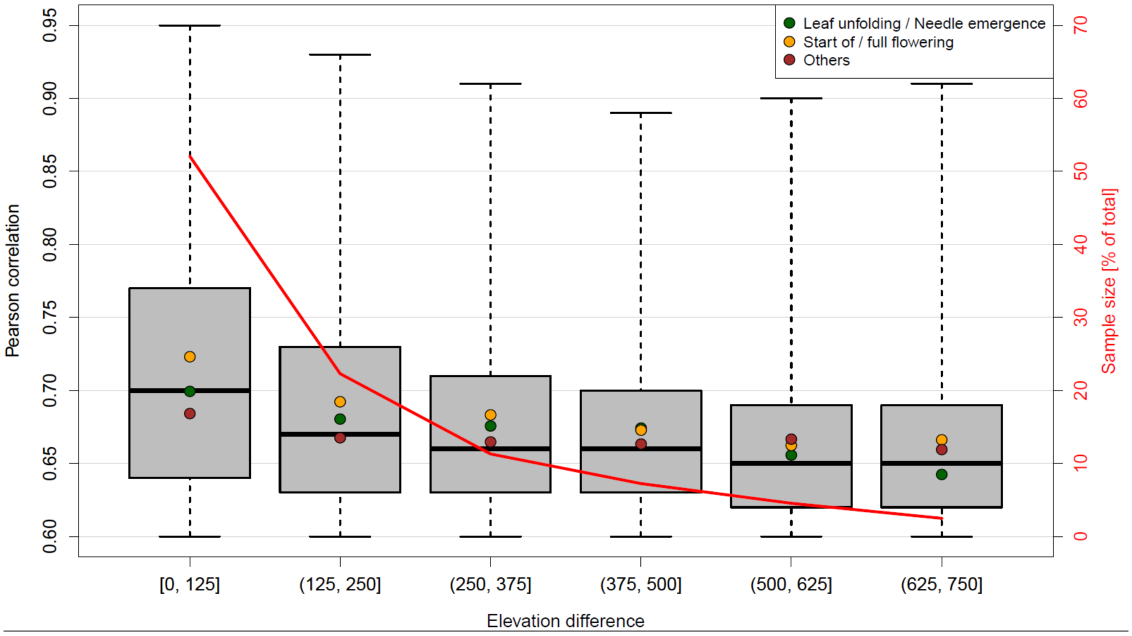


Figure 12: Boxplots of the correlations of the reference series separated by elevation differences. The points show the averages of different phases, the red lines show how many reference series contributed to each boxplot.

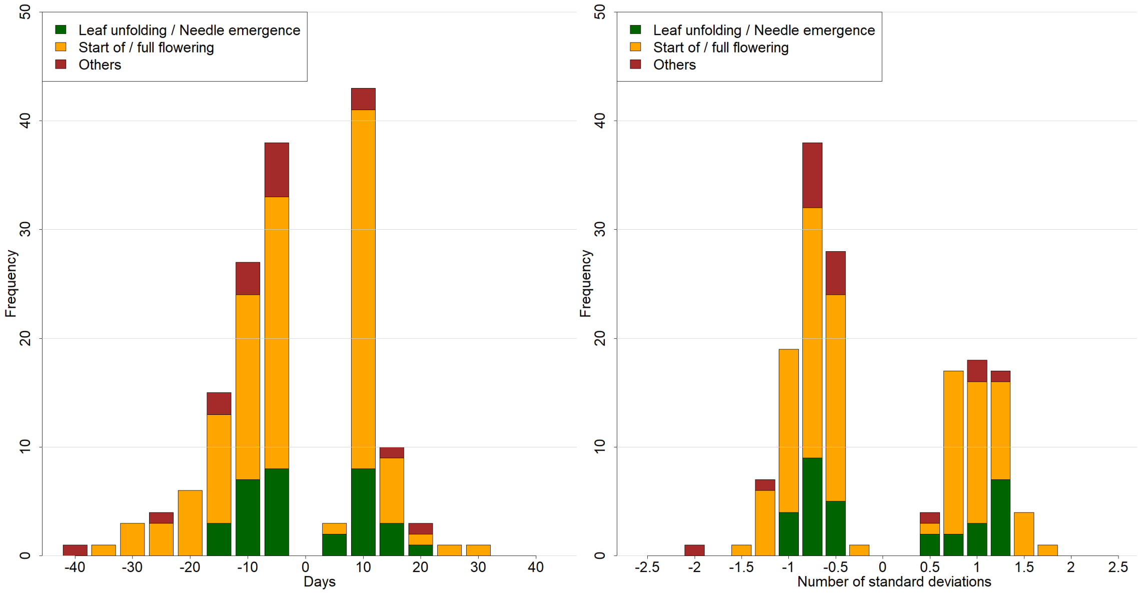


Figure 13: Histograms of the absolute (left) and standardized (right) size of the detected inhomogeneities.

3.2.3 Size of the Inhomogeneities

We estimated the size of each inhomogeneity in a series from the same five reference series used in the breakpoint detection for that particular series. Figure 13 shows the distribution of the absolute (left panel) and standardized (right panel) sizes of all breakpoints. The distribution is bimodal because breakpoints with a size close to zero are too small to be detected. Moreover, the distribution is not symmetric: there are significantly more negative changes (i.e. anticipation of the phenological phase after the breakpoint; 61%) than positive (39%). This asymmetry is found across all phases, although the leaf unfolding is slightly less affected (56% vs. 44%). The reason of the asymmetry is unknown but it is likely related to the changes of observer, which show a higher incidence of negative changes (66%) than the other breakpoints (55%). We did not detect any significant variability of the asymmetry over time.

The estimation of the size of the inhomogeneities is itself difficult and not always reliable. In the example shown in Figure 14, one breakpoint (1998) reaches the significance threshold, but then is barely significant (three reference series in two tests show a break). A second possible breakpoint in correspondence of a second change of observer in 2001 is only detected by one test and is therefore not significant. Judging from the bottom plot in Figure 14, the size of the two breakpoints is similar and it is of about one standard deviation; however, since the breakpoint in 2001 was not significant, the whole period 1999-2015 is used to calculate the size of the first breakpoint and this results in an estimated size of only 0.4 standard deviations (i.e. 5 days). There would possibly be a third breakpoint around 1965, but the first 14 years of the record were ignored by the detection algorithm (yellow shading) since not enough reference series were available in that period.

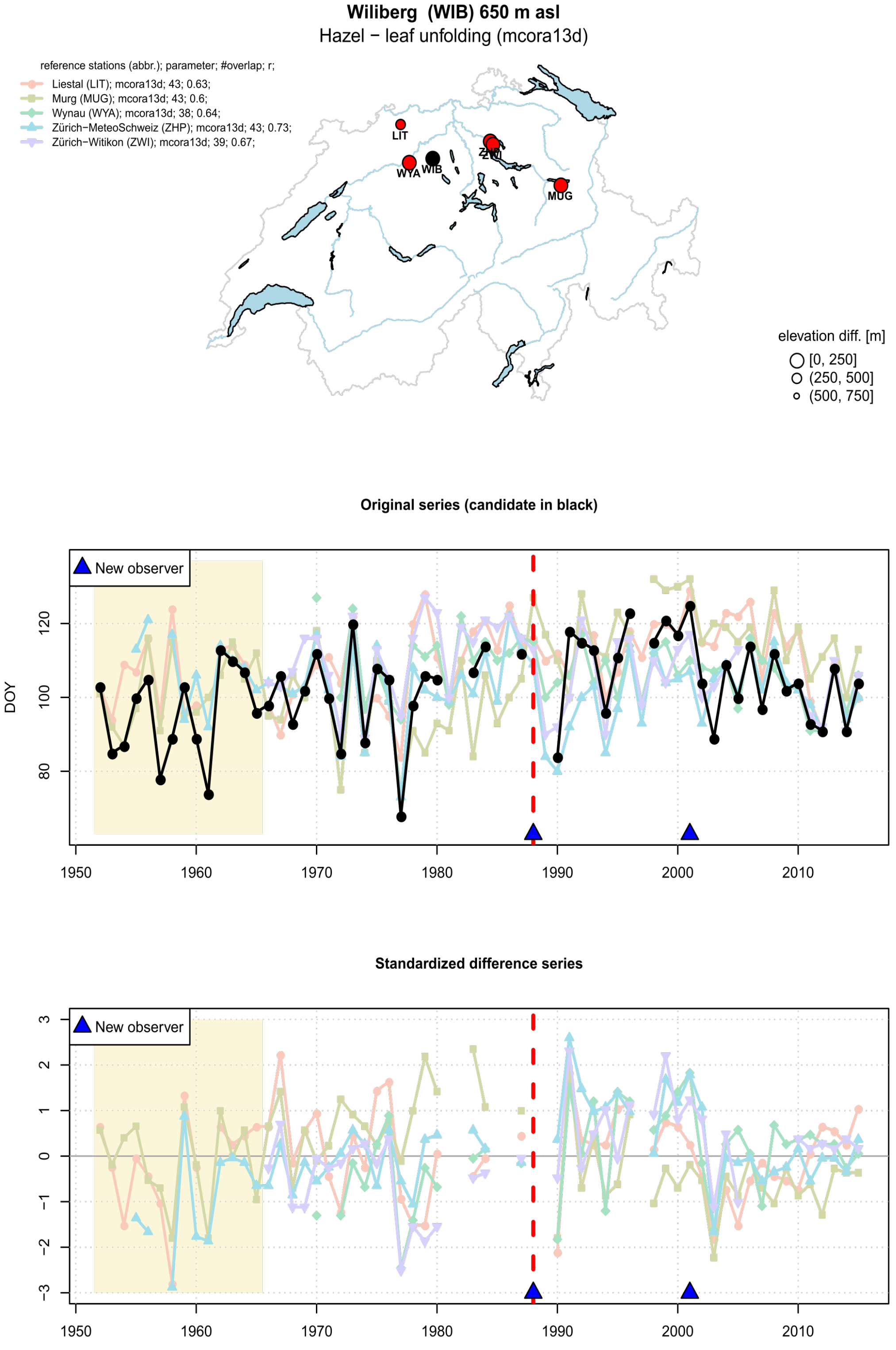


Figure 14: Example with likely undetected breakpoint. The yellow shading indicates the sub-period when not enough reference series were available.

3.2.4 Evaluation

A proper evaluation would require a benchmark consisting of surrogate series. Such benchmarks have only recently been developed in the climate sciences (e.g. Venema et al. 2012) and require detailed knowledge and physics of the causes of the breakpoints and of their statistical properties. Breakpoints in phenological series are arguably rather different from those affecting temperature series. For instance, events such as parasites attacking a plant do not have a correspondence in climate data.

Therefore, we can only provide a subjective validation, based on the visual analysis of a randomly selected sub-sample of series by three experts: Y. Brugnara, T. Rutishauer, R. Auchmann. Each expert analysed all series that were found inhomogeneous by the algorithm, plus 100 randomly selected homogeneous series, comparing the target series with up to ten reference series (including the five used in the breakpoint detection). Sub-periods that did not undergo breakpoint detection were ignored.

None of the detected breakpoints was found implausible. This finding is based on subjective criteria and does not guarantee that all breakpoints are true. More than half (54%) of the breakpoints are associated to changes of the observer. In very few cases (3%), the breakpoint was judged to be possibly misplaced by two years or more, or to represent rather a trend (i.e. an inhomogeneity developing gradually over several years). An example of the latter is shown in Figure 15. In 16% of the inhomogeneous series with a single breakpoint we found that multiple breakpoints could be likely (such as in Fig. 14).

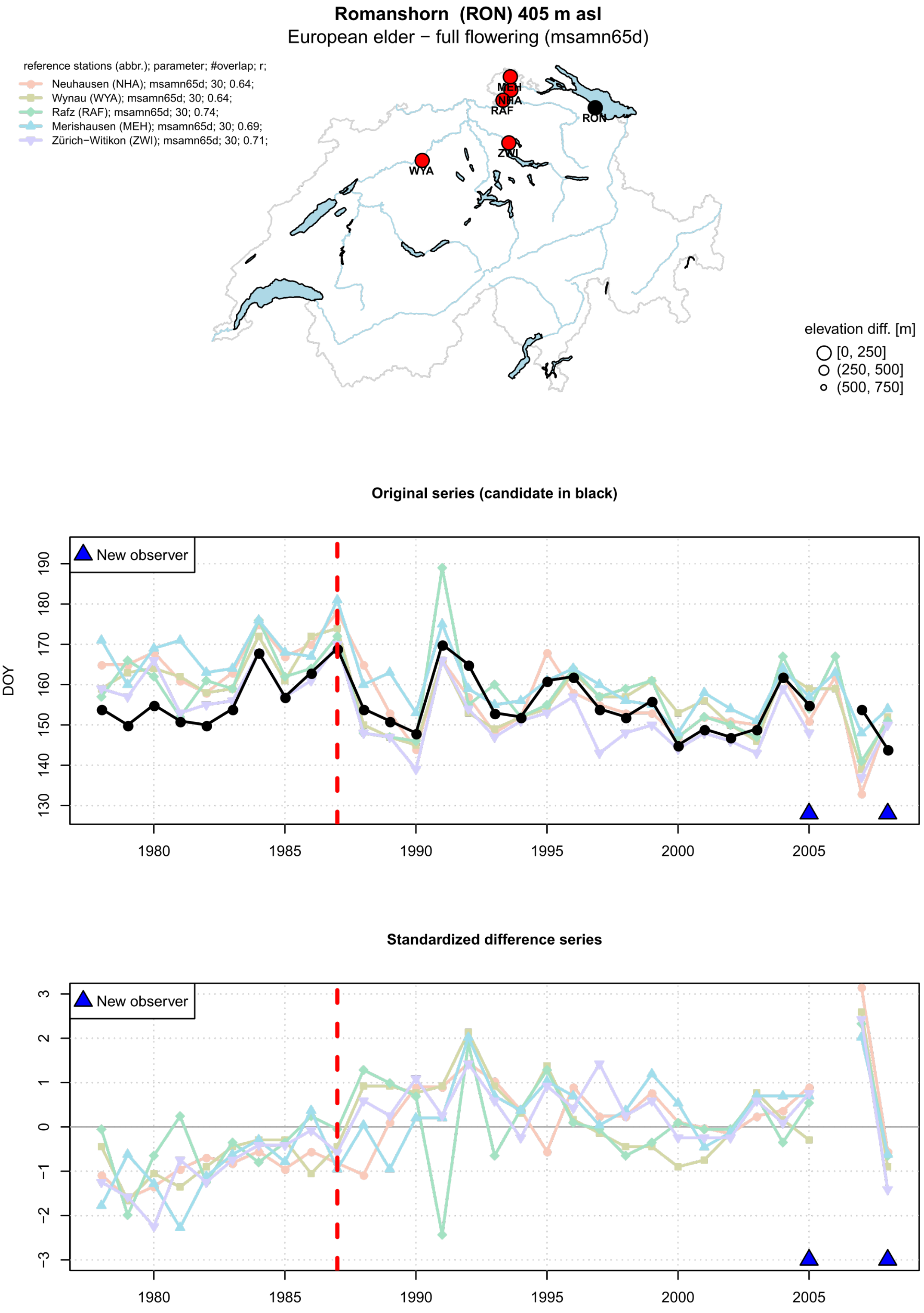


Figure 15: Example of a trend-like inhomogeneity.

In 10% of the 100 randomly selected homogeneous series (no breakpoint found by the automatic algorithm) at least one of the experts disagreed, suggesting at least one probable undetected breakpoint. If we assume that this percentage applies to the whole dataset, we can estimate that for 277 series where no breakpoint was detected by the algorithm, the experts judgment would disagree. Note that any break detection approach has to find a balance between undetected breaks and false detections, the approach used here is clearly more focused to limit false detections. We conclude that large breaks (breakpoints with a large signal-to-noise ratio, relative to the noise of the record) could be identified by the method, while false detections are low, at the price of undetected small-to-medium size breaks.

Changing the parameters of the detection and the rules for the selection of the reference series can significantly improve the power of detection, at the price of more false detections. To test this, an additional experiment (not shown in Table 2) was performed: the setup of experiment 5REF was applied using 8 reference series instead of five. This enhanced the chances that at least three references will detect a breakpoint. We obtained a number of inhomogeneous series corresponding to 80% of the expected inhomogeneous series (i.e., 153+277). Similarly, if we relax the biological restrictions for the selection of the reference series and use eight references (Experiments ALL1 and ALL2, see Sect. 2.3), we detect 72% of the expected inhomogeneities. These percentages are unrealistically high, implying that among the detected breakpoints there are many false detections.

To put these numbers in context, it is worth mentioning that the best hit rate (i.e. ratio between detected and total breakpoints) of automated breakpoint detection algorithms for temperature datasets does not exceed 40%, when a maximum false detection rate of 5% is allowed (i.e.; one false detection every 20 years of data); for precipitation, the best reported hit rate is 26% (Venema et al. 2012). Considering that we estimated about 20% of the inhomogeneous series to have multiple breakpoints, an estimate power of detection of our algorithm is around 30% (assuming that all multiple breakpoints are double breakpoints, which is not true). This does not take into account the breakpoints in those periods and series that did not have enough reference series, and we are still assuming that there are no false detections. Therefore, a realistic estimation is that the hit rate for phenological data is similar to what we would obtain for precipitation.

We summarize that breakpoint detection in combination with a thorough analysis of metadata such as observer changes and other causes of inhomogeneities may contribute to a complementary, more robust estimation of breaks and shifts in phenological series. In the view of a complete quality control of the dataset, the analyses presented above aim at contributing additional information on series and station quality in terms of homogeneity.

3.3 Classification Scheme

3.3.1 Quality Classes

The right panel of Figure 16 shows the distribution (density) of scores for all series while the left panel shows the unweighted score. Seventy-two series (0.8% of all series) reached the maximum score of 1. Those series are assigned to the highest class (dark green) for each criterion and arguably represent the longest, high quality, homogeneous series of the SPN. In Class 2 we find 634 series (6.7%) and in Class 3 1'305 (13.8%). Classes 4 and 5 comprise 11.3% and 27.2% of all series. The largest number of series (3'053, i.e. 43.3%) are in Class 6, i.e, "low valuable series", 8% of all series are of very low quality (Class 7).

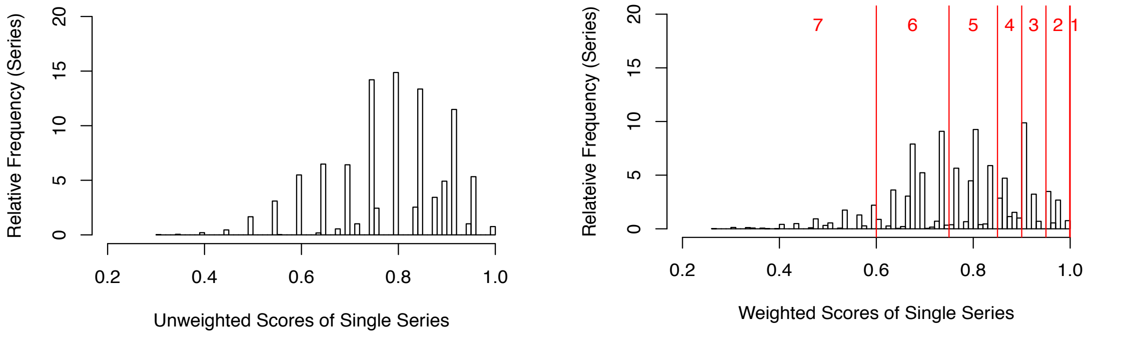


Figure 16: Density of unweighted (left) and weighted scores (right) of all series, with boundaries of quality classes (red lines) and class number (red numbers).

3.3.2 Examples

Figure 17 and Figure 18 show the resulting plots with traffic lights for a Quality Class 2 (highly valuable) and a Quality Class 5 (medium valuable) series, respectively.

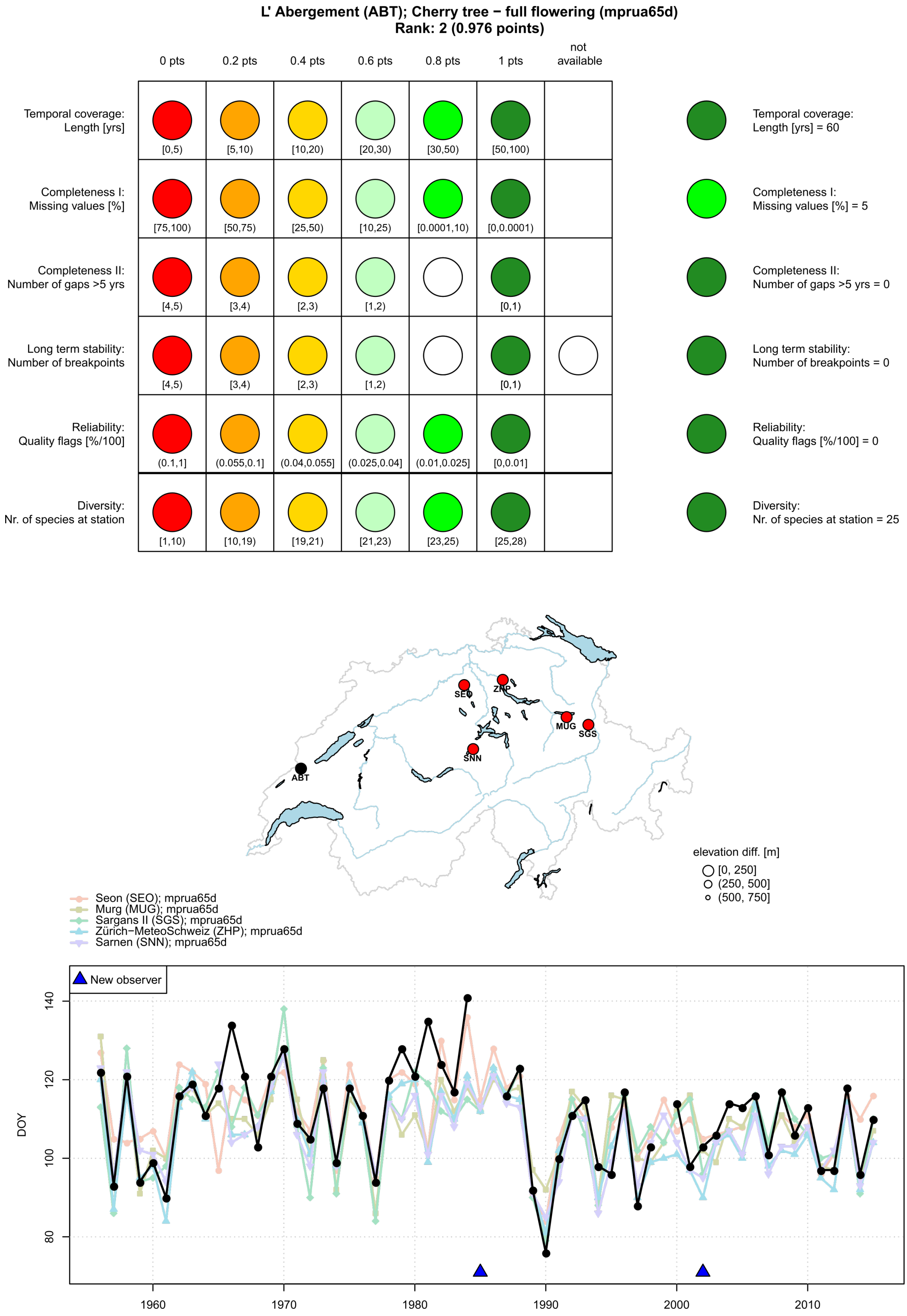
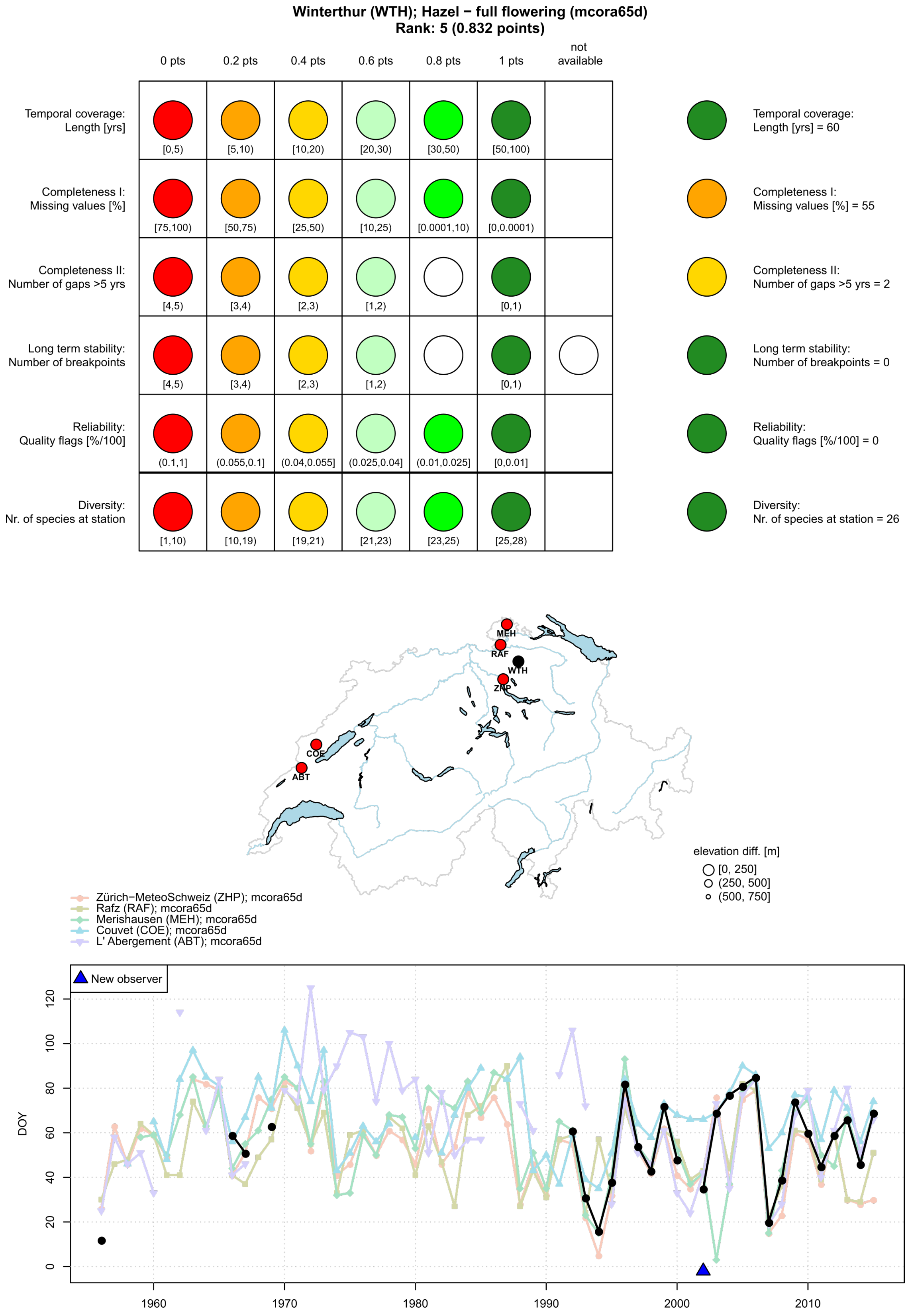


Figure 17: Results from example: L'Abergement (ZWS), Cherry tree - full flowering, "Class 2" series (highly valuable series).



Winterthur (WTH); Hazel − full flowering (mcora65d)
Rank: 5 (0.832 points)

	0 pts	0.2 pts	0.4 pts	0.6 pts	0.8 pts	1 pts	not available	
Temporal coverage: Length [yrs]	[0,5)	[5,10)	[10,20)	[20,30)	[30,50)	[50,100)		Temporal coverage: Length [yrs] = 60
Completeness I: Missing values [%]	[75,100)	[50,75)	[25,50)	[10,25)	[0.0001,10)	[0,0.0001)		Completeness I: Missing values [%] = 55
Completeness II: Number of gaps >5 yrs	[4,5)	[3,4)	[2,3)	[1,2)	○	[0,1)		Completeness II: Number of gaps >5 yrs = 2
Long term stability: Number of breakpoints	[4,5)	[3,4)	[2,3)	[1,2)	○	[0,1)	○	Long term stability: Number of breakpoints = 0
Reliability: Quality flags [%/100]	(0.1,1]	(0.055,0.1]	(0.04,0.055]	(0.025,0.04]	(0.01,0.025]	[0,0.01]		Reliability: Quality flags [%/100] = 0
Diversity: Nr. of species at station	[1,10)	[10,19)	[19,21)	[21,23)	[23,25)	[25,28)		Diversity: Nr. of species at station = 26

Figure 18: Results from example: Winterthur (WTH), Hazel - full flowering, "Class 5" series (medium valuable series).

3.3.3 Sensitivity of Classes

The goal of this work was to use one classification system that can be applied across as many series as possible, despite the fact that some may not have scores for all indicators. It is therefore important to assess the sensitivity of the system towards missing or wrong information (e.g. if no breakpoint detection could be performed due to e.g. the series not meeting length requirements or not finding enough reference series, the criteria “homogeneity” is not considered and only the weights of the remaining four criteria are considered). For 26.1% of data series (1ˈ035) with at least 20 observations no breakpoint detection could be made, for 20,3% of the series (843) only segments could be used. In total, the break detection was applied to 2ˈ925 of a total of 9ˈ455 phenological series, for 153 series at least one significant breakpoint was found.

If a series underwent the breakpoint detection and a break was detected, the criteria “homogeneity” is assigned 0.6 (out of 1) points. If no break could be detected the series is assigned the full 1 point. Assuming a series underwent a breakpoint detection but an existing break in reality could not be detected by the test, it depends much on the levels of the other criteria how “wrongly” the series was ranked (Table 6).

Table 6 shows the sensitivity of the classification to the outcome of the breakpoint detection. The colours represent classes (see legend on the top right). The points of all other factors except homogeneity are shown in the first column, the points when considering the break detection are shown in the remaining columns for all possible break detection outcomes. If no breakpoint detection could be performed (second column), the score is between those for zero (third column) and one break (fourth column; closer to the score of zero breaks if all other factors have high scores, see bottom rows of Table 6). Having a falsely detected break considerably decreases the rank, particularly when moving from zero to one breaks. Being concerned about false detection, our approach will arguably be overly optimistic.

Table 6: Impact of breakpoint detection on the classification. The first columns shows the score when not considering the breakpoint detection, while the other columns show the score for different results of the breakpoint detection. Colors indicate the class related to each score.

Other Factors	no Breakpoint Detection	0 Breaks	1 Break	2 Breaks	3 Breaks	4 Breaks	
0.85	0.85	0.895	0.775	0.715	0.655	0.595	Class
0.855	0.855	0.899	0.779	0.719	0.659	0.599	1
0.86	0.86	0.902	0.782	0.722	0.662	0.602	2
0.865	0.865	0.906	0.786	0.726	0.666	0.606	3
0.87	0.87	0.909	0.789	0.729	0.669	0.609	4
0.875	0.875	0.913	0.793	0.733	0.673	0.613	5
0.88	0.88	0.916	0.796	0.736	0.676	0.616	6
0.885	0.885	0.92	0.8	0.74	0.68	0.62	7
0.89	0.89	0.923	0.803	0.743	0.683	0.623	
0.895	0.895	0.927	0.807	0.747	0.687	0.627	
0.9	0.9	0.93	0.81	0.75	0.69	0.63	
0.905	0.905	0.934	0.814	0.754	0.694	0.634	
0.91	0.91	0.937	0.817	0.757	0.697	0.637	
0.915	0.915	0.941	0.821	0.761	0.701	0.641	
0.92	0.92	0.944	0.824	0.764	0.704	0.644	
0.925	0.925	0.948	0.828	0.768	0.708	0.648	
0.93	0.93	0.951	0.831	0.771	0.711	0.651	
0.935	0.935	0.955	0.835	0.775	0.715	0.655	
0.94	0.94	0.958	0.838	0.778	0.718	0.658	
0.945	0.945	0.962	0.842	0.782	0.722	0.662	
0.95	0.95	0.965	0.845	0.785	0.725	0.665	
0.955	0.955	0.969	0.849	0.789	0.729	0.669	
0.96	0.96	0.972	0.852	0.792	0.732	0.672	
0.965	0.965	0.976	0.856	0.796	0.736	0.676	
0.97	0.97	0.979	0.859	0.799	0.739	0.679	
0.975	0.975	0.983	0.863	0.803	0.743	0.683	
0.98	0.98	0.986	0.866	0.806	0.746	0.686	
0.985	0.985	0.99	0.87	0.81	0.75	0.69	
0.99	0.99	0.993	0.873	0.813	0.753	0.693	
0.995	0.995	0.997	0.877	0.817	0.757	0.697	
1	1	1	0.88	0.82	0.76	0.7	

Example 1: If a series is in Class 1 (i.e. all criteria have full points) but the break detection mistakenly found no break (which in reality exists) the series is classified to Class 1 with 1 average point. However, if this series would have received only 0.6 points for the homogeneity criteria (i.e. for 1 breakpoint) the series would have yielded 0.88 average points (Table 6, last row, second last column) and the “true” class would be Class 4 with 0.12 points “lost” due to break detection. On the other hand, if this Class 1 series could not have undergone a break detection because of not enough suitable reference series, the series would still have 1 average point and have Class 1.

Example 2: L’Abergement – Cherry tree full flowering with 0.976 points in Class 2 (Fig. 17). This series underwent a break detection with no break found. However, assuming a break is apparent but hasn’t been detected, the series would have 0.856 points and be in Class 4 (minus 0.12 points). If this series could not undergo a breakpoint detection, the series would have 0.966 points and still be in the same class, Class 2.

3.3.4 Points by Parameters

We calculated the average of points for each series by parameter (Fig. 19). The figure shows error bars for the mean (+/- 2 standard errors (SE)) as well as on their most left panel the series frequency for each parameter. To partly account for the varying number of observations due to the different start of observing a specific parameter Figure 19 shows only observations from the early parameters that started being observed in 1951, parameters that started in 1996 are shown in Figure 20.

The list can be used to select the parameters with the best observation quality which are best suited for long term studies. The parameters with highest points and therefore with the most valuable data series are the full flowering of dandelion, the needle emergence of European larch and the full flowering of cherry tree as the top three (Fig. 19).

As stated above, the shorter parameters that have only being observed since1996 have in general less points (Fig. 20). The most valuable parameters with the highest points are the start of flowering of apple and cherry tree and the leaf unfolding of sycamore maple as top three.

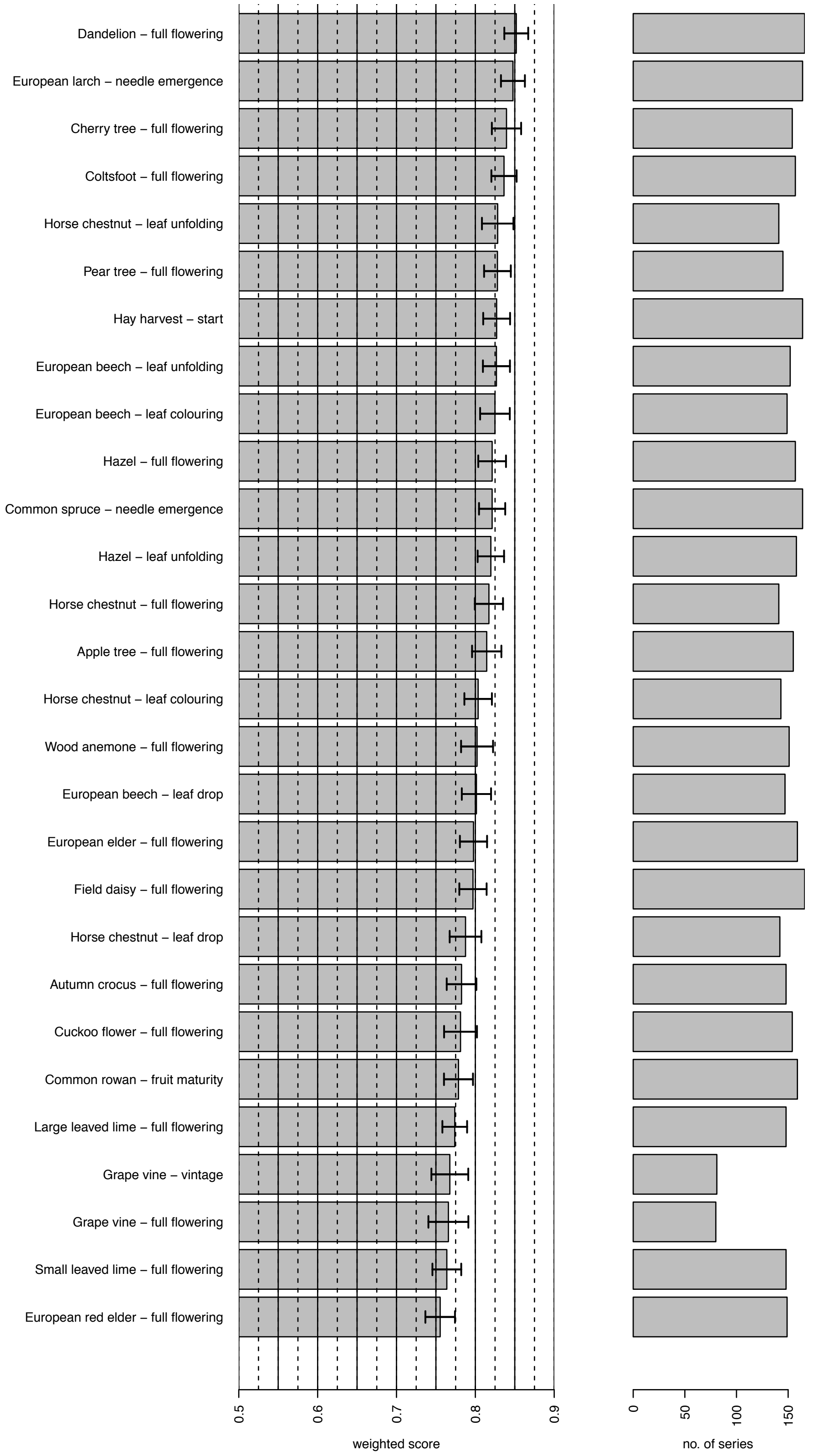


Figure 19: Left: Barplot of weighted scores for each parameter starting in 1951 with +/- 2 SE. Right: Number of series per parameter.

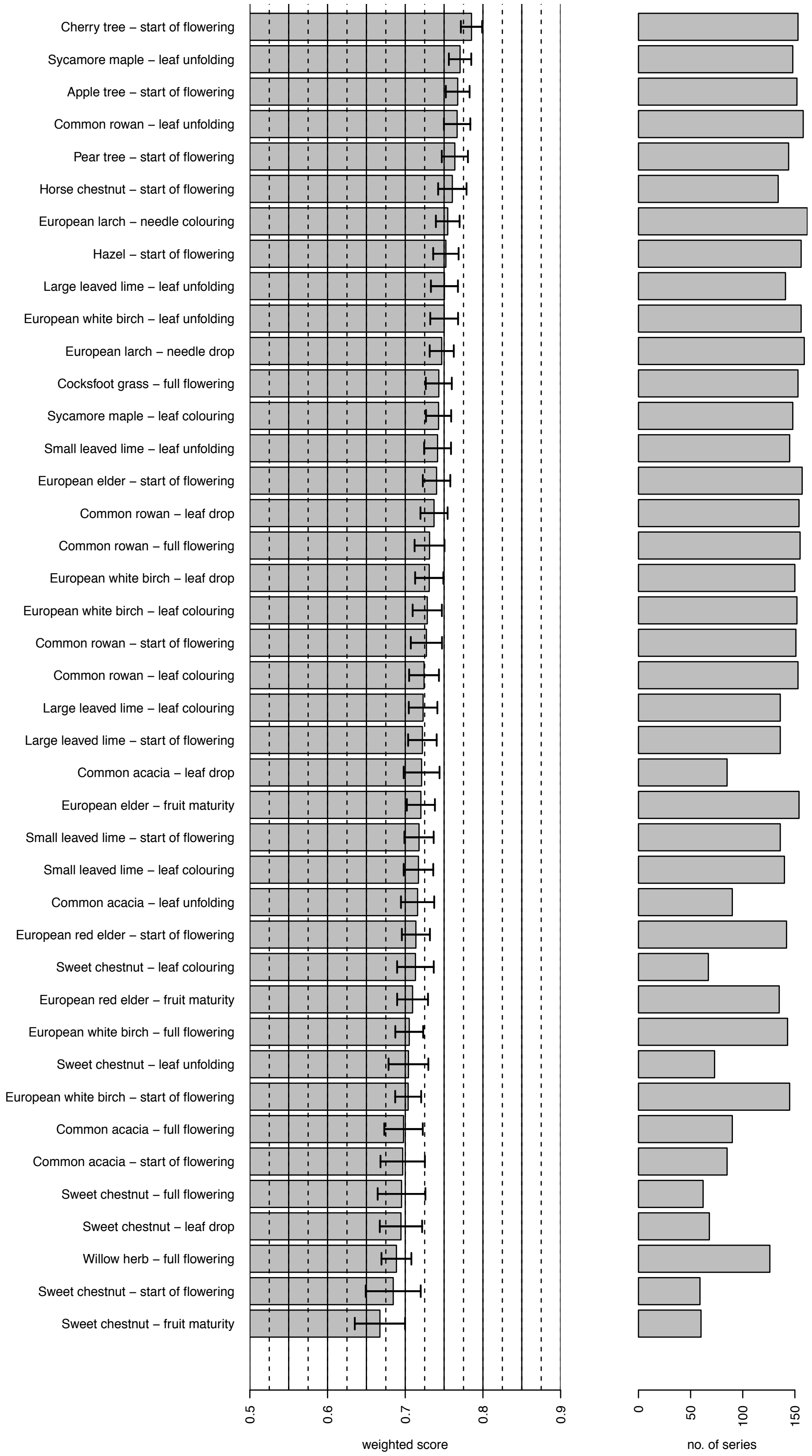


Figure 20: Left: Barplot of weighted scores for each parameter starting in 1996 with +/- 2 SE. Right: Number of series per parameter.

3.3.5 Station Classes

The station scores without applying the stability criterion at any station and without applying the diversity criterion are shown in Figure 21. The station scores with and without applying the diversity criterion (both including the stability criterion) are shown in Figures 23 and 22, respectively. Green circles denote stations with a high score (>0.8).

For 40 stations the class changes between Figures 21 and 22, 40 stations move up one class when the stability criterion is accounted for. The difference between the classes in Figures 22 and 21 (with and without stability) results from series that could undergo a break detection and no breakpoint was found, or when a breakpoint was detected. Hence, the 157 breaks that were detected do not notably influence the station class. However, where no break was detected, those series do influence the station class of 40 stations. Class 1 series are not affected because all other criteria have full scores. When adding stability, ten stations move up to Class 2 (from Class 3), 12 stations move up to Class 3, 11 to Class 4 and seven move up to Class 5.

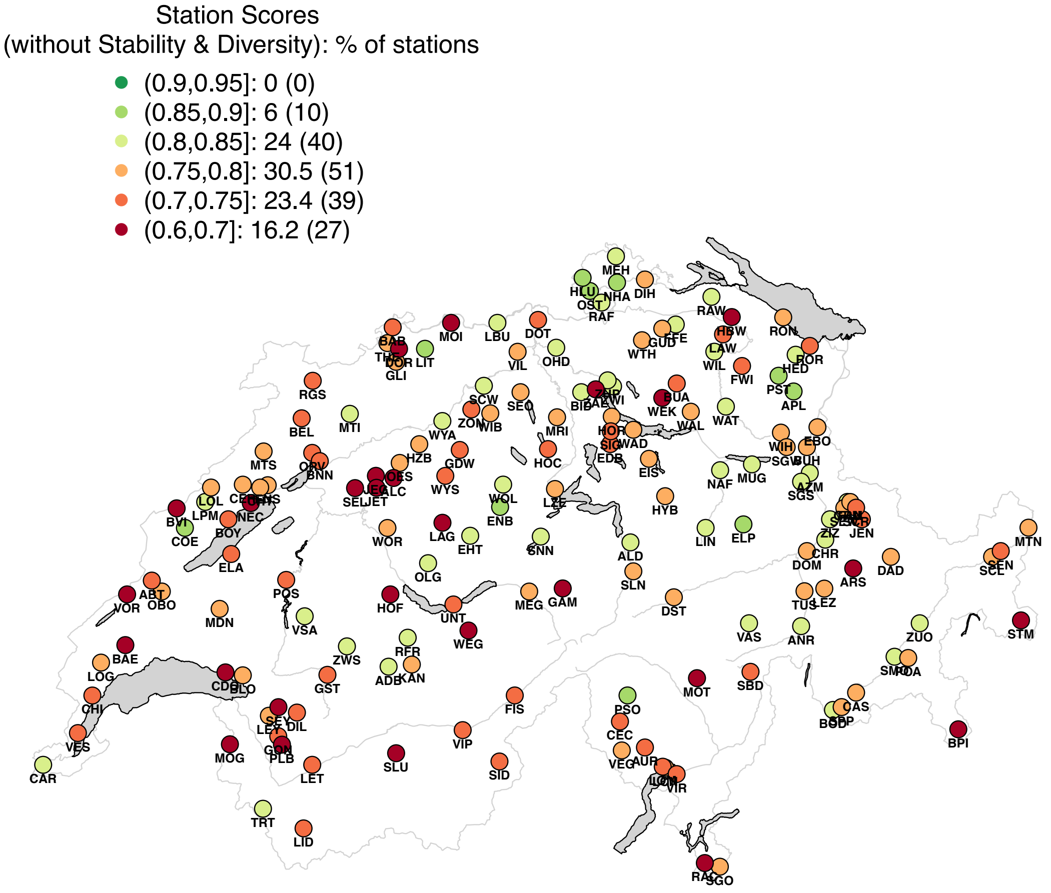


Figure 21: Map of stations with station scores leaving out the stability and diversity criteria.

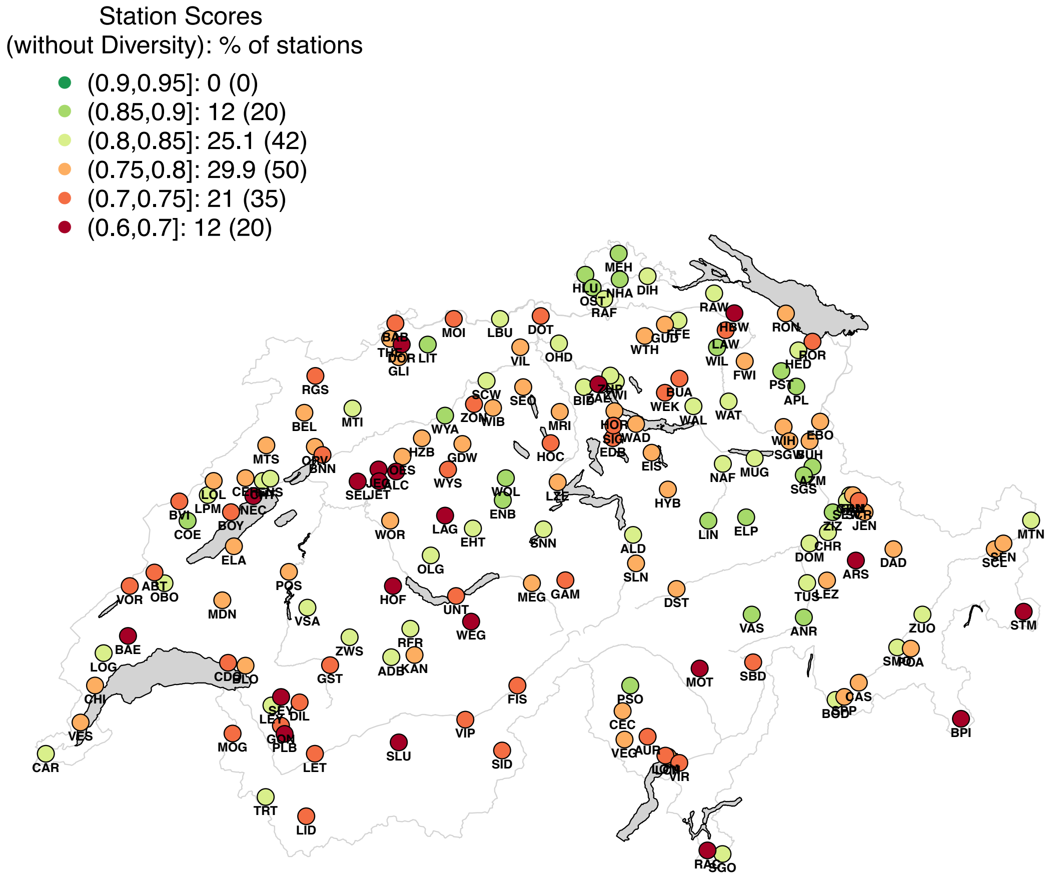


Figure 22: Map of stations with station scores leaving out the diversity criterion.

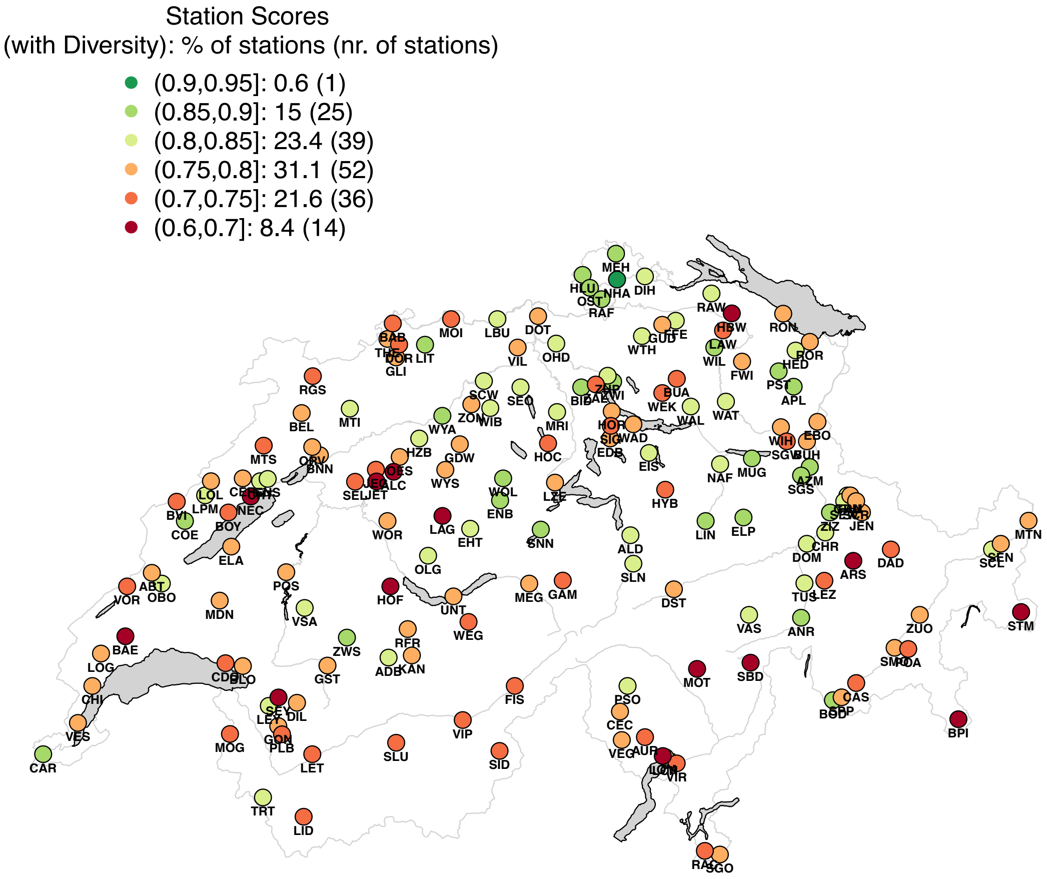


Figure 23: Map of stations with station means using all criteria.

When applying the diversity criterion 68.7% of the stations do not change their class. 6.5% of the stations lose one class, 24.8% of the stations gain one class. No station changes for more than one class.

Figure 24 shows the stations where more than one breakpoint was detected. The station most affected by inhomogeneities was that of Horgen (HOR), where 6 series out of 22 are affected by breakpoints. However, Horgen also has six class 2 series, a visual inspection of which shows that only one (European elder – full flowering) shows a probable missed break. Domat/Ems (DOM) has 4 Class 2 series, none of them shows visually a missed break. Osterfingen (OST) has 12 Class 2 series, none of them shows visually a missed break.

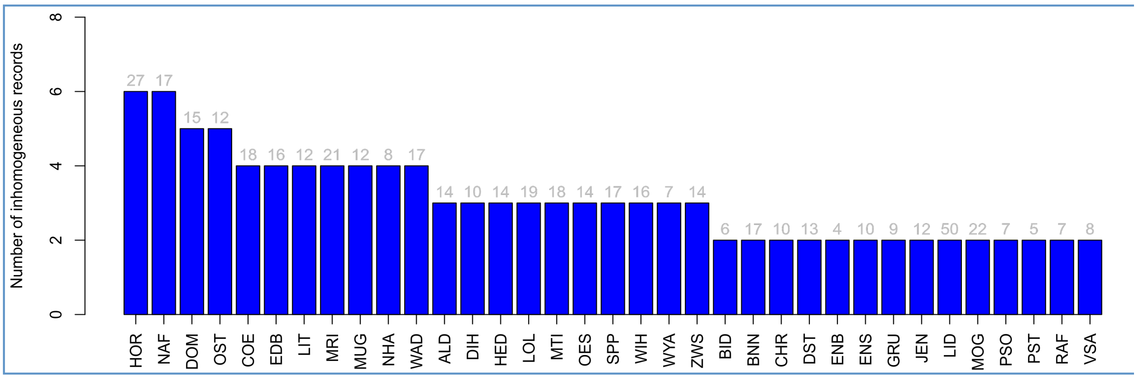


Figure 24: Number of inhomogeneous series found at each station (only stations with at least two inhomogeneous series are shown). The grey number on top of each bar represents the percentage of inhomogeneous series relative to the number of analysed series.

Figure 25 shows histograms of the weighted station mean which includes diversity (left panel) and just the station mean without diversity (mean of all series points, right panel). The top panels show all data, the bottom panels show station means from series that started being observed in 1951 only.

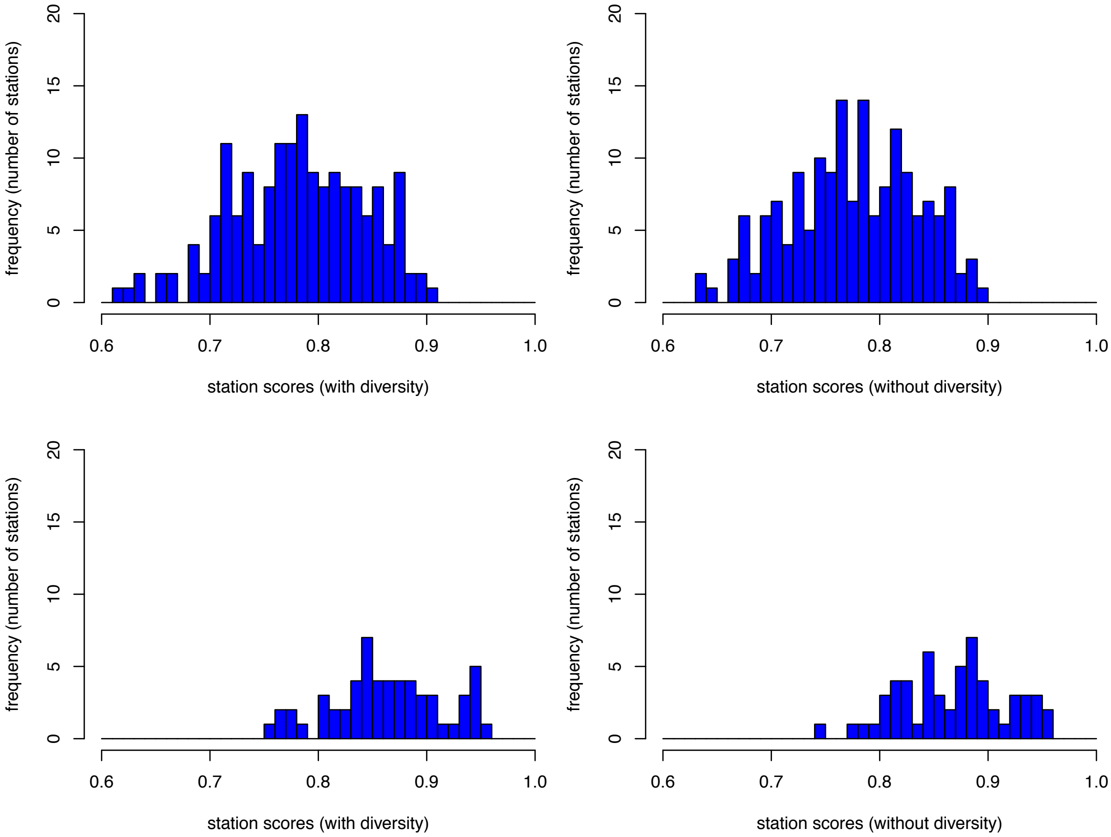


Figure 25: Top panel: Frequency station scores including diversity (left) and simple station scores without diversity (right). Bottom row: Same as top row, but station scores calculation only with series starting in 1951.

As for the series scores, also the station scores are influenced largely by the length of the series observed at a station. Figure 25 shows in the bottom panels the station scores using only parameters that started in 1951. Clearly the distribution shifted to the right, meaning that higher station scores could be achieved (Fig. 25, left with diversity, right without diversity). Again, applying the diversity criteria to the stations does not change the station means much (only 31% of the stations move one class up or down) because of the low weight for the diversity criterion.

3.4 Characterization of Classes

We analyzed the most frequent types of series in each class in terms of their characteristics. Figures 26-29 show for each class the different types of series in the class. For example, Figure 26 left panel shows that Class 1 has only two different types of series, represented by two rows. The first row shows the first type which is characterized by one point in each category (dark green dots). 67 (93.06%) of the series of Class 1 are of this type. The second type in Class 1 has also one point in each category except category "Long-term Stability" where no points could be assigned because those series could not undergo a break detection.

Class 2 (Fig. 26, right panel) comprises five different record types. More than three quarters of all Class 2 series belong to two types: One type has full points in all but one category ("Completeness I") and has a few missing values, the other type has full points in each but one category ("Temporal Coverage") where the series are only 30-50 years long but otherwise complete, homogeneous and have no quality flag.

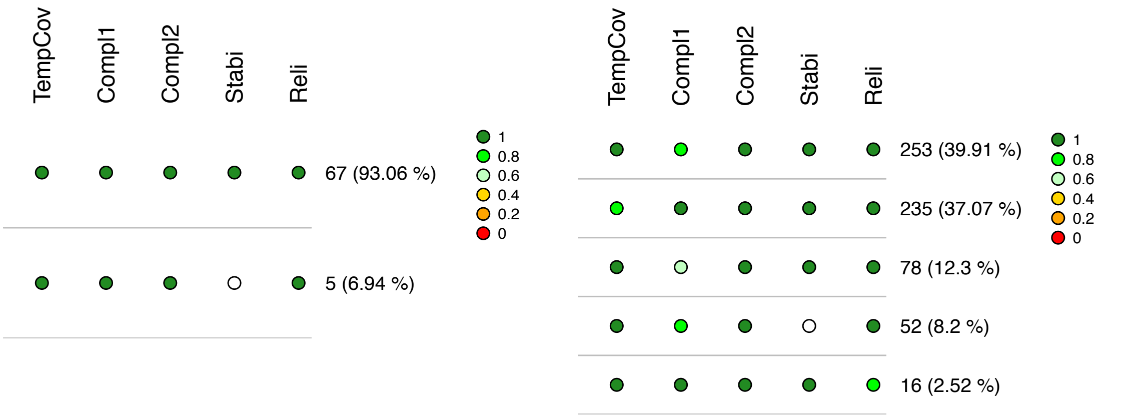


Figure 26: Characterization of Class1 (left) and Class 2 (right). The columns represent the classification criteria described in Section 2.4.

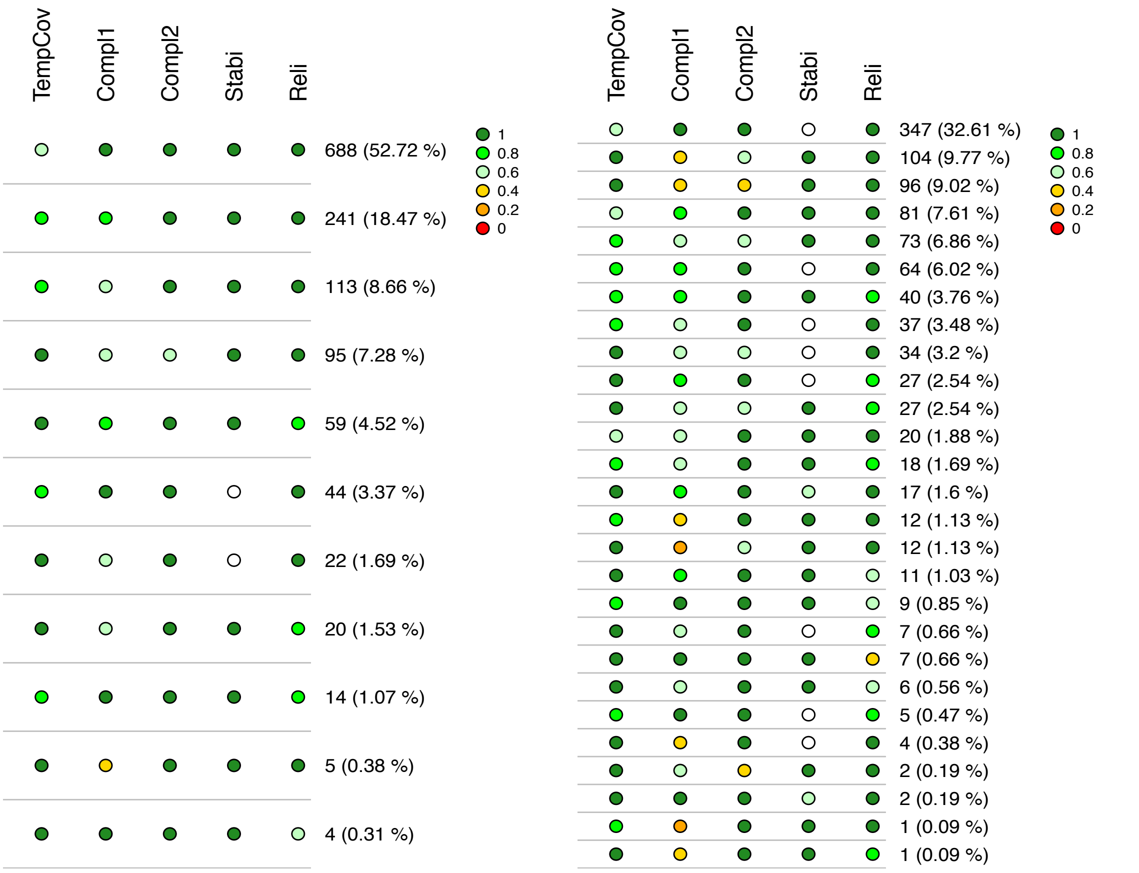


Figure 27: Same as Fig. 26 for Class 3 (left) and Class 4 (right).

In Class 3 (Fig. 27, left panel) 70% of the series are mostly represented by two different types. One type (52.72% of the Class 2 series) are complete, homogeneous series, with no quality flag and no missing values, but the series are only between 20-30 years long. The second type (18.47% of the Class 2 series) is characterized by series that are 30-50 years long, have a few single missing values but no gaps, and are otherwise homogeneous and have no quality flag.

Class 4 (Fig. 27, right panel) comprises a variety of types, where the first type comprises one third of the series. Series of the first type (Fig. 27, first row) are 20-30 years long and could not undergo a break detection, but are complete and have no quality flag. The second and third type (each around 9% of the series) is characterized by series which are more than 50 years long but have 25-50% single missing values including data gaps >5 years.

TempCov	Compl1	Compl2	Stabi	Reli	1 0.8 0.6 0.4 0.2 0
0.4	1	1	○	1	816 (31.78 %)
0.6	0.8	1	○	1	378 (14.72 %)
0.6	0.6	1	○	1	324 (12.62 %)
0.4	0.8	1	○	1	322 (12.54 %)
0.6	0.4	1	○	1	106 (4.13 %)
0.8	0.4	0.6	1	1	44 (1.71 %)
1	0.4	0.6	○	1	43 (1.67 %)
0.8	0.4	0.6	○	1	32 (1.25 %)
1	0.8	1	1	0.4	26 (1.01 %)
1	0.4	0.6	1	0.8	24 (0.93 %)
1	0.2	0.4	○	1	20 (0.78 %)
0.8	1	1	0.6	1	20 (0.78 %)
1	0.2	0.6	○	1	19 (0.74 %)
1	0.4	0.4	○	1	19 (0.74 %)
0.8	0.8	1	0.6	1	19 (0.74 %)
1	0.2	0.4	1	1	16 (0.62 %)
0.8	1	1	1	0.2	16 (0.62 %)
0.8	0.6	0.6	○	1	14 (0.55 %)
0.6	1	1	0.6	1	14 (0.55 %)
0.6	1	1	1	0.4	14 (0.55 %)
0.8	0.4	1	○	1	13 (0.51 %)
0.8	0.8	1	○	0.8	13 (0.51 %)
0.8	0.6	0.6	1	0.8	13 (0.51 %)
0.8	0.6	1	0.6	1	12 (0.47 %)
0.8	0.4	0.6	1	0.8	12 (0.47 %)
1	0.6	0.6	1	0.4	11 (0.43 %)
1	0.8	1	1	0.2	11 (0.43 %)
1	0.4	0.4	1	0.8	10 (0.39 %)
0.8	0.8	1	1	0.4	10 (0.39 %)
1	0.6	0.6	○	0.8	9 (0.35 %)
0.8	0.6	1	1	0.6	9 (0.35 %)
1	0.2	0.2	1	1	8 (0.31 %)
0.8	0.6	1	○	0.8	8 (0.31 %)
1	0.8	1	0.6	0.8	8 (0.31 %)
0.8	0.8	1	1	0.6	8 (0.31 %)
1	0	0.6	○	1	7 (0.27 %)
1	0.4	0.6	○	0.8	7 (0.27 %)
1	0.2	0.4	1	0.8	6 (0.23 %)
0.8	0.8	1	○	0.6	6 (0.23 %)
0.8	0.6	1	1	0.4	6 (0.23 %)

TempCov	Compl1	Compl2	Stabi	Reli	1 0.8 0.6 0.4 0.2 0
0	1	1	○	1	637 (20.86 %)
0.4	0.6	1	○	1	428 (14.02 %)
0.2	1	1	○	1	391 (12.81 %)
0.4	0.4	1	○	1	261 (8.55 %)
0.6	0.4	0.6	○	1	171 (5.6 %)
0.2	0.6	1	○	1	156 (5.11 %)
0.2	0.4	1	○	1	140 (4.59 %)
0.4	0.4	0.6	○	1	134 (4.39 %)
0.6	1	1	○	0.4	93 (3.05 %)
0.6	0.2	0.6	○	1	80 (2.62 %)
0.2	0.2	1	○	1	73 (2.39 %)
0.6	0.6	1	○	0.4	56 (1.83 %)
0.6	0.8	1	○	0.4	55 (1.8 %)
0.4	0.2	1	○	1	33 (1.08 %)
0.8	0.2	0.6	○	1	26 (0.85 %)
0.8	0.2	0.4	○	1	16 (0.52 %)
0.4	1	1	○	0.4	16 (0.52 %)
0.8	0.8	1	1	0.2	15 (0.49 %)
1	0	0.4	○	1	14 (0.46 %)
0.8	0	0.6	○	1	12 (0.39 %)
0.6	0	0.6	○	1	10 (0.33 %)
0.6	1	1	○	0.2	10 (0.33 %)
0.6	0.2	1	○	1	9 (0.29 %)
0.8	0.4	0.4	○	1	9 (0.29 %)
0.8	0	0.4	○	1	9 (0.29 %)
0.8	0.6	0.6	0.6	1	8 (0.26 %)
0.8	0.4	0.6	○	0.8	8 (0.26 %)
1	0.4	0.6	1	0.4	7 (0.23 %)
0.8	0.4	0.6	0.6	1	6 (0.2 %)
1	0.2	0.2	○	0.8	6 (0.2 %)
0.8	0.8	1	○	0.2	6 (0.2 %)

Figure 28: Same as Fig. 26 for Class 5 (left) and Class 6 (right).

For Class 5 (Fig. 28, left panel) the first four types comprise around 70% of the series. The first type (32% of all Class 5 series) is characterized by 10-20 year long series which could not undergo a break detection but are otherwise complete and have no quality flag. The second, third and fourth type (14.7%, 12.6%, 12.5%, respectively) are also characterized by shorter series (10-30 years long), that could not undergo a break detection, but which also have a few single missing values, in contrast to the first type.

In Class 6 (Fig. 28, right panel) the first five types account for around 50% of all series. The remaining series are distributed into many different types. In general the first eight types are characterized by reliable series (no quality flags), but short series (less than 20 years) with more or less missing values. The first type (~21%) comprises shorter than 5 years long, complete series with no quality flags. The second and fourth type are series with 10-20 years of observations but 10-25% (second type) and 25-50% (fourth type) single missing values. The third type is similar to the first type but with a little longer series (5-10 years). Almost all of the Class 6 series could not undergo a break detection.

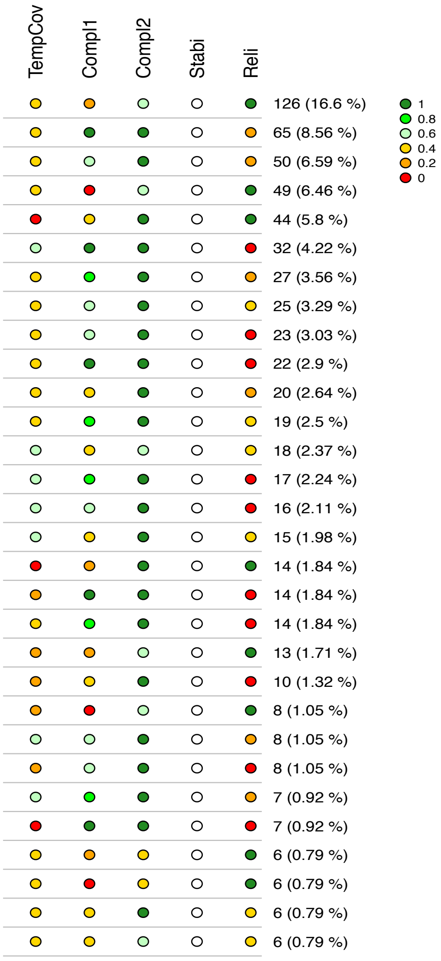


Figure 29: Same as Fig. 26 for Class 7. For display reasons types that only have less than six representative series are not shown.

Class 7 (Fig. 29) series comprise various types of series. The first four types (together ~40%) comprise 10-20 year long series with either many missing values (and for type one and four data gaps > 5 years) or quality issues (type three has a few missing values and quality issues). In contrast to the higher class most of the Class 7 series have quality issues. None of the series could undergo a break detection.

3.5 List of Most Valuable Series and Stations

Figure 30 shows the locations of the 17 stations with at least one series with a maximum score of 1. With 12 and 11 series, respectively, reaching the maximum score, Merishausen (MEH) and Liestal (LIT) are the two stations with the highest number of top-scoring series (Appendix C, first row, "highly valuable series"). Appendix C contains the complete table of the series with maximum scores.

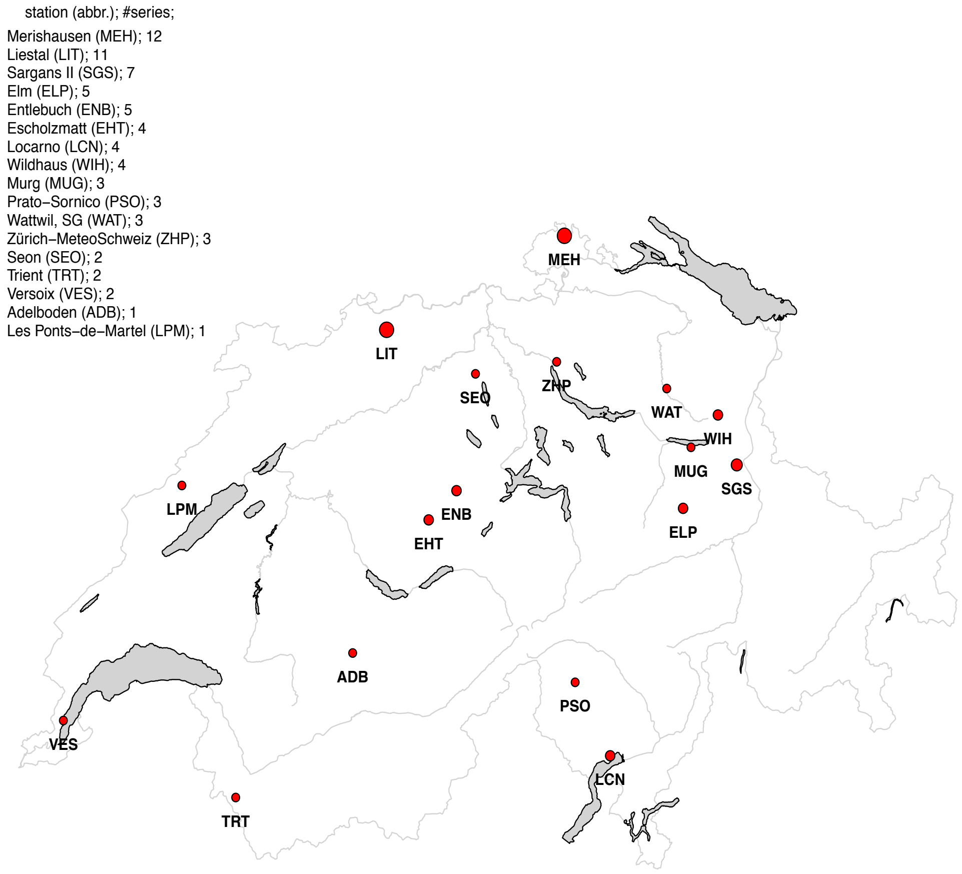


Figure 30: Map of stations with number of series with the maximum score. The circle size of the stations increases with number of series.

3.6 Recommendations for GCOS Switzerland Stations

Very valuable stations can either be defined by the presence of very long data series with high quality or by a high station score of all series. We selected additionally the longest data series of the SPN with more than 60 years of observations. According to the GCOS Climate Monitoring Principles, ensuring high data quality and homogeneity of series is of outmost importance for long term climate monitoring (WMO 2016). For climate analysis the use of high quality, long, and homogeneous series is crucial in order to perform reliable analysis for, e.g. climate change applications.

There are currently (including observations until 2015) 850 series that have at least 60 years of observations. Thereof, 284 series have less or equal to three missing values. From these 284 series, 189 have no quality flag. From the 189 series, 21 could not undergo a break detection, the other 168 have been subjected to a break detection. The break detection showed one break for ten series and two breaks for one record. For 157 series no break could be detected by the break detection algorithm. To ensure the reliability of the 157 homogeneous series and the 21 series that could not undergo a break detection, those series had been inspected visually for possible missed breakpoints. The following code was applied to each of those series:

- 1: series has undergone break detection, no break detected, also visually no break found
- 2: series has undergone break detection, no break detected, visually not clear if break/breaks
- 3: series has undergone break detection, no break detected, visually a clear break found
- 4: series has not undergone break detection, visually no break found
- 5: series has not undergone break detection, visually maybe a break/breaks found
- 6: series has not undergone break detection, visually a clear break found
- 7: series has not undergone break detection, no information available from neighbouring stations to judge series

Out of the 178 (157 homogeneous plus 21 not tested) series almost 80% have no visual break (i.e. 73.6% have code 1 plus 6.2% with code 4). All results of the visual inspection are summarized in Table 7. Here only very long series underwent an expert inspection, hence more (~20%) likely undetected breakpoints could be found than in the independent sample inspected in Section 3.2.5 (~10%).

Table 7: Results of visual inspection of homogeneity of the 178 stations > 60 years long with less than 3 missing values and no quality flag.

Visual Inspection Code Number	Short Code Description	Absolute Number	Relative Number [%]
1	tested, no visual break	131	73.6
2	tested, maybe visual break	21	11.8
3	tested, clear visual break	6	3.4
4	not tested, no visual break	11	6.2
5	not tested, maybe visual break	4	2.3
6	not tested, clear visual break	0	0
7	not tested, no reference information	5	2.8

All 178 that underwent the visual inspection for homogeneity are listed in Appendix D with their subsequent visual code, length characteristics, class, and information on whether the station is currently a GCOS Switzerland station. The series are ordered by class and station, starting with Class 1 series (note that stations can repeat if they have Class 1 and Class 2 series).

In total, 27 of the 29 listed stations in Appendix D have at least one series that has a visual code 1 or 4 (i.e. no visual break found). Nine of those 27 stations are current GCOS Switzerland stations: Liestal, Enges, Murg, Rafz, Sarnen, Trient, Valsainte, Versoix and Wildhaus. Davos, St. Moritz and Prato-Sornico, which are also GCOS Switzerland stations, are not among the 27 listed stations and hence not in the list in the Table 6. The importance of the stations of Davos and St. Moritz for GCOS Switzerland is that they represent alpine regions and Prato-Sornico is the most valuable station (highest station scores) in Ticino. The station score for Prato-Sornico is 0.87 (the earliest series in Prato-Sornico start in 1957), for Davos 0.80 (from series starting in 1951: 0.84) and for St. Moritz 0.84 (for series starting in 1951 0.83). St. Moritz has the highest station score of all stations above 1500 m.a.s.l. Other stations above this elevation with high scores are Zuoz with a score of 0.83, Pontresina with a score of 0.80 (longest series of both stations are 46 years long) as well as Davos with 0.80.

The 27 station with their number of series that have a code 1 or a code 4 from the visual inspection are listed below in Table 8.

Table 8: Characteristics of the 27 stations with a length of more than 60 years that have at least one series with visual code 1 or 4, i.e., without a break visually detected by the expert.

Station Name	Number of Parameters with Visual Code 1 or 4	Current GCOS Station yes/no	Station score	Station score (only series starting in 1951)
Liestal (LIT)	13	yes	0.87	0.94
Sargans II (SGS)	12	no	0.86	0.95
Sarnen (SNN)	12	yes	0.84	0.94
Murg (MUG)	11	yes	0.84	0.93
Couvet (COE)	9	no	0.86	0.94
Escholzmatt (EHT)	9	no	0.84	0.95
Elm (ELP)	8	no	0.86	0.94
Rafz (RAF)	7	yes	0.84	0.94
Seon (SEO)	6	no	0.79	0.9
Wattwil, SG (WAT)	6	no	0.83	0.92
Wildhaus (WIH)	6	yes	0.8	0.9
Zürich-MeteoSchweiz (ZHP)	6	no	0.85	0.92
La Valsainte (VSA)	5	yes	0.82	0.9
Les Ponts-de-Martel (LPM)	5	no	0.8	0.92
Cartigny (CAR)	4	no	0.84	0.94
Versoix (VES)	4	yes	0.75	0.88
Kandersteg (KAN)	3	no	0.77	0.87
Wiliberg (WIB)	3	no	0.78	0.9
Appenzell (APL)	2	no	0.87	0.89
Le Locle (LOL)	2	no	0.79	0.88
Trient (TRT)	2	yes	0.82	0.88
Disentis (DST)	1	no	0.78	0.89
Enges (ENS)	1	yes	0.8	0.9
Gryon (GON)	1	no	0.75	0.87
Simplon-Dorf (SID)	1	no	0.72	0.89
Thusis (TUS)	1	no	0.81	0.89
Vals (VAS)	1	no	0.86	0.88

Figure 31 shows a map with the location of the 27 stations. Nine red stations are current GCOS Switzerland stations, 18 blue stations are not. Additionally the three GCOS Switzerland stations that are not among the 27 stations are also shown in white. The blue stations in Figure 33 (currently no GCOS Switzerland stations) could be recommended as possible future GCOS Switzerland stations.

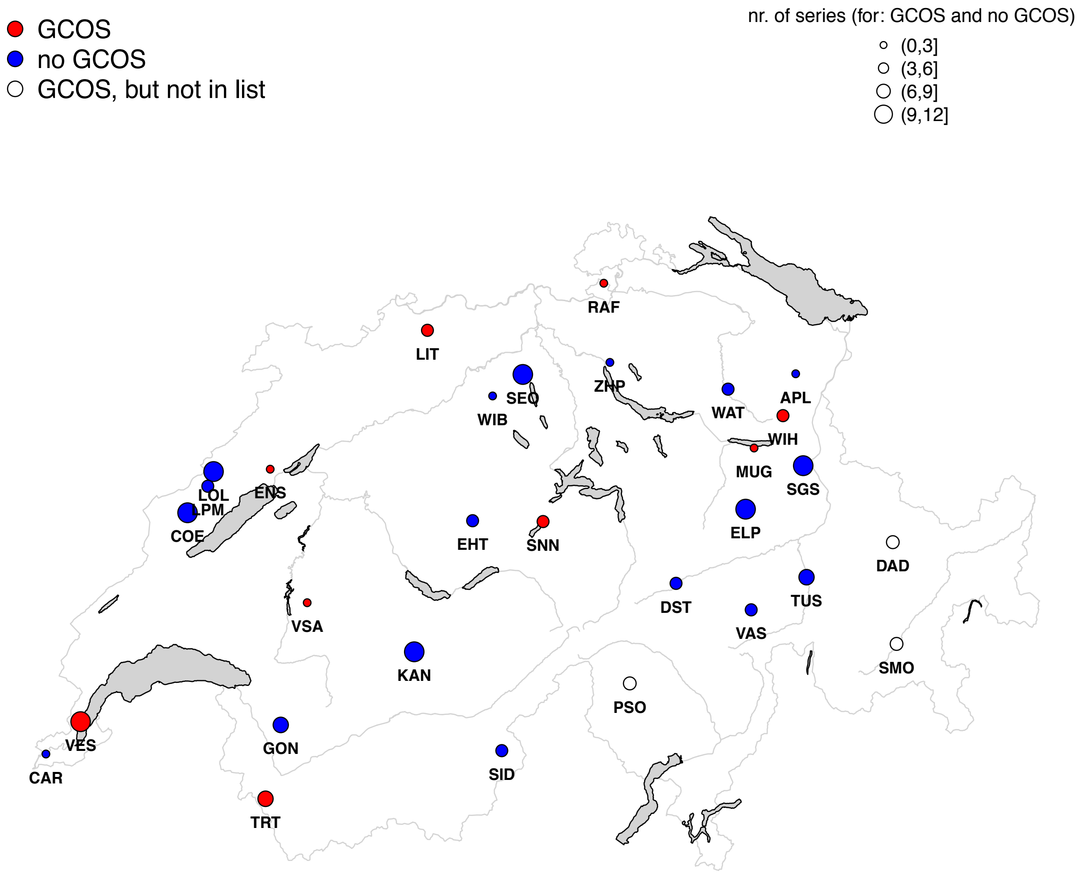


Figure 31: Map of 27 stations that have more than 60 years of observations, not more than 3 missing values, have no quality flag and are homogeneous. Current GCOS Switzerland stations that fulfill these requirements are shown in red (9 stations), others in blue (18 stations), their point size represents the number of series per station (see legend top right). Current GCOS Switzerland stations that do not fulfill those requirements are shown in white (3 stations).

Figure 32 (top) shows the 14 stations with the highest (>0.9) station scores (calculated from series starting in 1951). We find all 14 stations also in Figure 31, which shows that the station scores well represent highly valuable stations. Scores of all stations (calculated from series starting in 1951 are shown in Fig. 32, bottom).

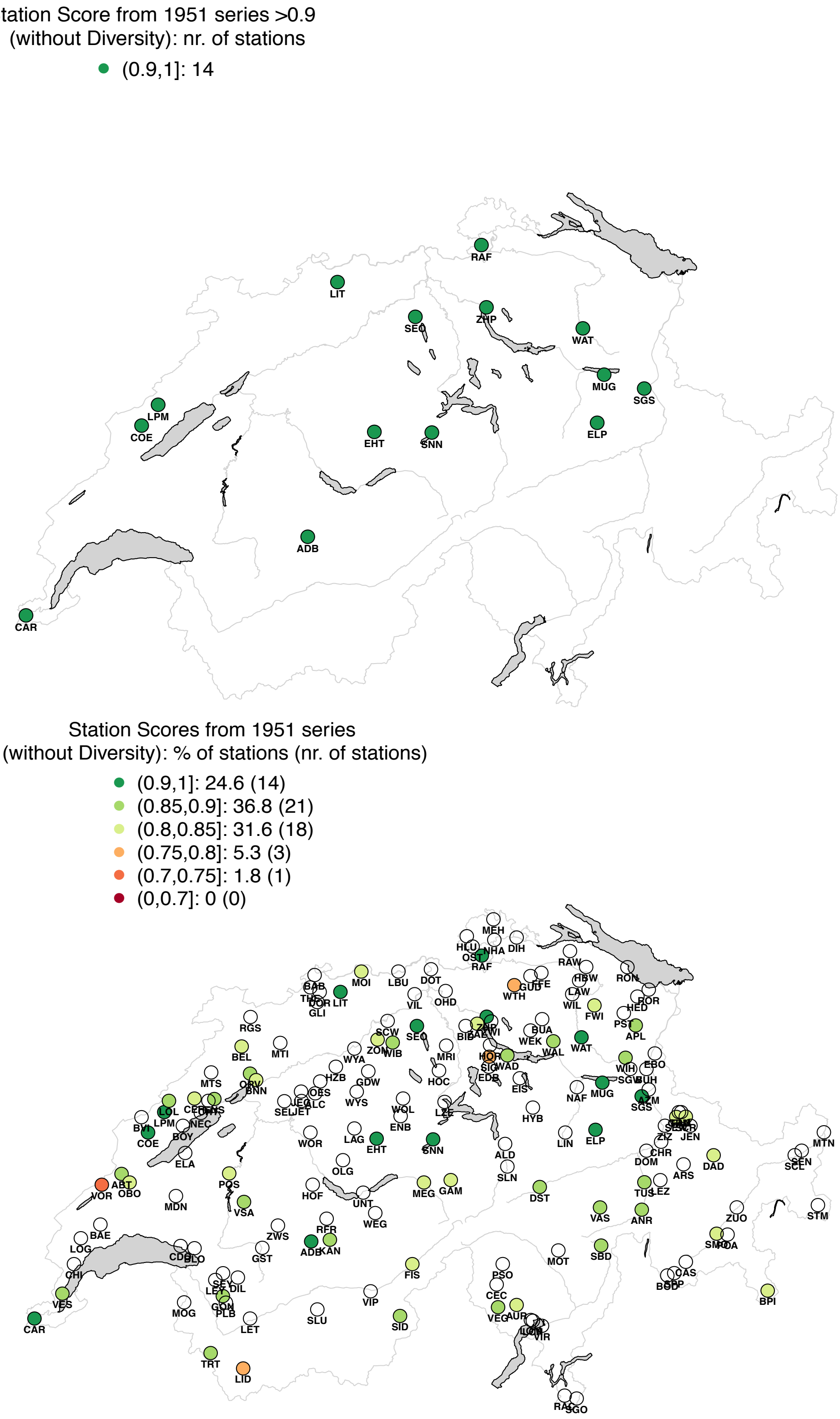


Figure 32: Top: Map of stations with a high station score over 0.9 (calculated from series starting in 1951). Bottom: Map of all station scores (calculated from series starting in 1951). White circles denote stations that have no series starting in 1951.

Potential future GCOS Switzerland stations comprise stations that have series that meet several GCOS requirements such as high data quality and homogeneity but are not yet 60 years long, however have the best prerequisites to become GCOS Switzerland stations in 20 to 30 years. Series that have no quality issues, are almost complete (maximum of 3 missing values), homogeneous (or could not yet undergo a breakpoint detection) and are between 30 and 50 years long are shown in Figure 33 (Note that series of Class 1, which are all longer than 50 years, are shown in Fig. 32). Those 585 potential GCOS Switzerland series are Class 2 (300), 3 (236) and 4 (49) series from 78 different stations (Fig. 35). Figure 33 also shows the number of potential future GCOS Switzerland series per station (circle size). Especially additional alpine stations as well as stations in Ticino and the South-East of Switzerland could serve as potential future GCOS Switzerland stations.

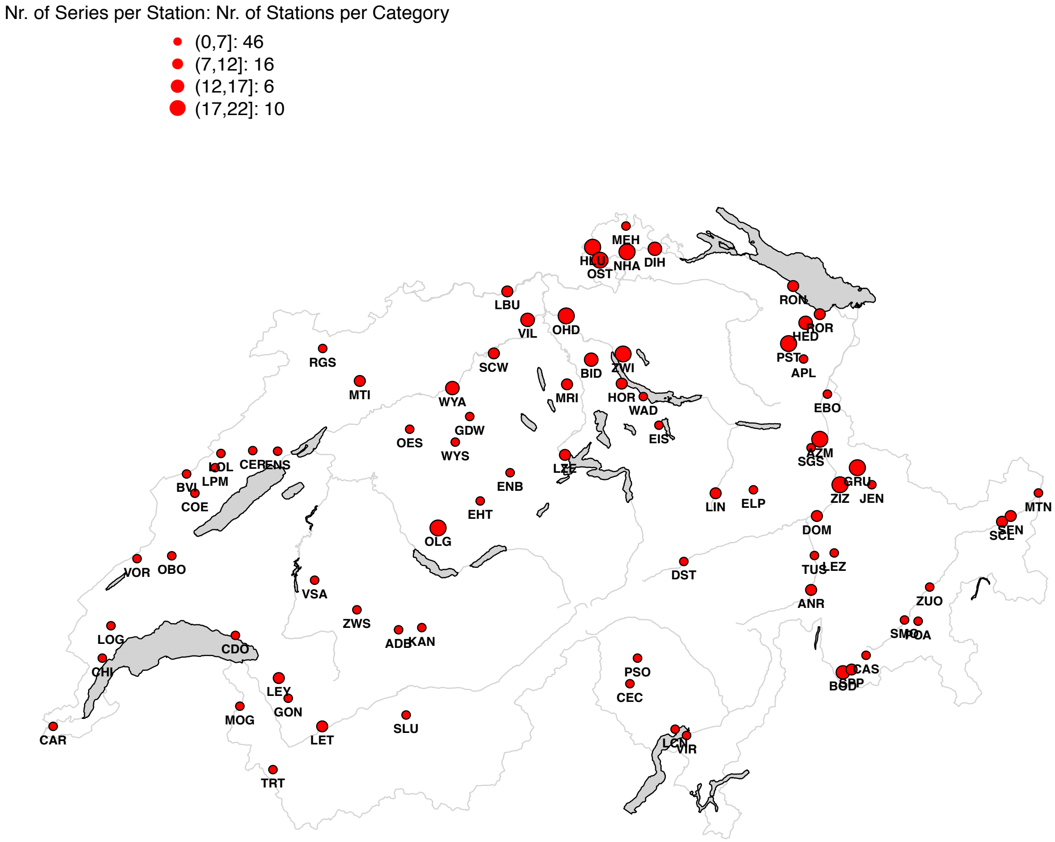


Figure 33: Map of 78 stations that have between 30 and 50 years of observations, not more than 3 missing values, have no quality flag and have no detected break (or could not undergo a breakpoint detection). The point size represents the number of series per station (see legend top left).

4 Summary and Outlook

The goal of the project PhenoClass was the thorough assessment and subsequent classification of all data series and stations of the SPN. We defined relevant criteria for the classification system of phenological data series and stations of the SPN. Methods for determining all criteria were developed, their subsequent indicator levels/states quantified for each record, and criteria had been weighted and translated into the classification scheme using a point system. We assessed the data quality and homogeneity of the series of the SPN as basic criteria for the classification.

In terms of data quality we found that the automatic QC yielded 4019 observations with at least one flag, this is 1.96% of the observations. Around 49.8% of the 4019 automatic flags were confirmed by two individual experts, 319 additional flags were introduced (when experts found a quality issue during the inspection that was not captured by the automatic QC). A combination of the automatic and experts control yielded so called final flags, which affect 2319 observations (= ~1.13% of all observations were attributed a final flag). Hence, the combination enabled a reduction of flags by 49%.

In terms of homogeneity assessment, breakpoint detection could be performed for 73.9% of all series with at least 20 observation years (though for 20.3% of the series, not the entire series could be tested). Many records were however, shorter than 20 years. In total, the break detection was applied to 2925 of a total of 9455 phenological series, for 153 series at least one significant breakpoint was found.

The novel classification scheme for all phenological series of the SPN was developed using the criteria data quality, homogeneity, completeness, and length. Subjecting all series of the SPN to the procedure resulted in a hierarchical point-based ranking of all series and the identification of the most valuable Swiss phenological series (series with the highest score). Furthermore, series that meet the GCOS requirements (i.e. high data quality, longer than 60 years, homogeneous) were assessed as well as future potential GCOS Switzerland stations (which not yet meet the GCOS requirements due to shortness of series) where added.

Each year new phenological observations are added to the dataset. The algorithm described here has been developed especially for the historical dataset. However, the classification procedure can also be applied in the future to a growing dataset (Note that for recent observations a QC procedure was developed for the SPN data and is applied after 2015; Pietragalla et al. 2016). The dataset has now been analysed until 2015. The analyses should however be continued with new observations to identify new stations that then meet the GCOS requirements (i.e. high data quality, longer than 60 years, homogeneous). The station Merishausen could potentially meet the requirements by the end of 2018.

Especially the break detection will profit from a larger dataset as series are becoming longer (more series can be tested). This may lead to more reliable break detection results.

In the PhenoClass project the QC served the purpose of identifying reliable stations and records. However, from the expert QC control, lessons can be learned that could be implemented (e.g. in a probabilistic way) into an automatic QC. For instance, 75% of absolute value flags were confirmed by the visual control but only 30-40% of all temperatuere-related flags were confirmed. Note that this can also result from the thresholds set for each method and should be tested before any implementation.

Further future tasks concerning the QC could be to compare flagged observations with the original data sheets and correct possible digitization errors. Also for Level 3 of the QC (comparison with temperature) only the temperature dataset until 2011 was used. The procedure could be updated with the now available updated temperature dataset from 2011.

For the experts control more metadata on the siting of the observed plant or special weather conditions could be helpful. Some outliers may be caused by e.g. a wind storm (e.g. sudden leaf drop). Additional information on special occasions can support the expert control. Furthermore, information on the orientation or siting (such as e.g. northern slope) could be helpful to identify series that represent smaller scale features. Hence a systematic collection of additional metadata (siting, orientation, special weather conditions) by the observers would not only support the QC but also support the break detection.

Additionally, for the break detection the systematic collection of changes that may lead to an inhomogeneity would be helpful in the evaluation of breaks. Such metadata comprise the change in observers and the change of observed plants (or notable illness of plants which may lead to gradual inhomogeneities). Our results suggest that the changes of observer can introduce systematic negative trends on a network scale. More detailed metadata would help understand the causes and allow to develop specific guidelines for the new observers in order to reduce the problem.

For the break detection, it is important to use quality controlled data (to reduce the signal-to-noise). In our project, the final QC flags were set only if the experts agreed on keeping a flag. More rigorous data flagging would reduce the signal-to-noise ratio at the price of wrongly flagged observations.

Further improvements of the break detection procedure can also comprise the definition of reference series e.g. changing the biological restriction, which was much discussed within this project. Adjustments could comprise, e.g., considering all parameters within a time window of ± 20–30 days or restricting the time window to the same season.

Further adjustments concerning the break detection methods could comprise tests on varying correlation thresholds. However, correlation alone is not sufficient to guarantee that a series is a suitable reference and more information on further causes of inhomogeneities might again be helpful, e.g. what causes gradual inhomogeneities and the collection of such information. In contrast to relative homogenization, absolute homogeneity tests (candidate series are tested without the use of correlated reference series) would not need reference series, but in the case of phenological series they are not recommended, because sudden changes in the mean state of a parameter do not imply in general an artificial inhomogeneity. Besides, issues such as consecutive breakpoints in one series or breakpoints at the beginning or end of the series are hard to overcome by automatic procedures. A visual inspection of homogeneity can better detect this kind of inhomogeneities; however, it requires some degree of expertise and familiarity with the data (again, metadata on changes of plants, observers and information on the siting can be helpful here). Still only relatively large breakpoints can be found and only when highly correlating reference series are available, at the price of high working time investments.

Geographical distance and even the alpine divide do not seem to influence much the correlation between at least some phenological series. An alternative way to improve the break detection could be then to use records from other networks, such as the BernClim network or data from nearby countries, although they are not likely to have a large impact on the whole data set due to the limited quantity of common observed parameters and consistency issues in the observation procedures. In terms of the classification scheme, the weights of the criteria can be adjusted and additional tests on various threshold of the point system performed. The system is flexible enough to simply add or remove criteria for future applications or adjust weights. Another way of applying the system is to develop a system just for subsamples of data series (e.g. using long series only, or only series where a break detection could be performed). In this way a more equal treatment of these subsamples can be ensured. However, especially the applicability of a classification system on a complex, diverse dataset such as the SPN is probably the biggest advantage of the classification system as used in this project.

References

Alexandersson, H. and A. Moberg, 1997: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *Int. J. Climatol*. 17: 25-34.

Anderson, D.M., E.M. Mauk, E.R. Wahl, C. Morrill, A.J. Wagner, D. Easterling, and T. Rutishauser, 2013: Global warming in an independent record of the past 130 years.
Geophys. Res. Lett. 40, 189–193. doi:10.1029/2012GL054271.

Defila, C. and B. Clot, 2001: Phytophenological trends in Switzerland. *Int. J. Biometeorol*. 45, 203–207.

Frei, C. 2014: Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. *Int. J. Climatol.* 34: 1585–1605. doi:10.1002/joc.3786.

Fu et al. 2015: Declining global warming effects on the phenology of spring leaf unfolding. *Nature*, 526, 104–107. doi:10.1038/nature15402.

Ge Q., H. Wang, J. Zheng, T. Rutishauser, and J. Dai, 2014: A 170 year spring phenology index of plants in eastern China. *J. Geophys. Res.* 119, 301-311. doi:10.1002/2013JG002565**.**

Güsewell, S, 2014: Phenological responses to changing temperatures: representativeness and precision of results from the Swiss Phenological Network. *Master Thesis in Biostatistics (STA495)*. University of Zurich, P. 1–144.

IPCC, 2007: Summary for Policymakers. In: Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds. Cambridge University Press, Cambridge, UK, 7-22..

Körner C. and D. Basler, 2009: Phenology under global warming. *Science*, 327, 1461-1462. doi:10.1126/science.1186473

Kuglitsch F. G., R. Auchmann, R. Bleisch, S. Brönnimann, O. Martius, and M. Stewart, 2012: Break detection of annual Swiss temperature series. *J. Geophys. Res. 117*, D13105. doi:10.1029/2012JD017729.

Menzel, A., T.H. Sparks, N. Estrella, et al. 2006: European phenological response to climate change matches the warming pattern. *Global Change Biol*. 12, 1969–1976. doi:10.1111/j.1365-2486.2006.01193.x.

MeteoSwiss 2018: National Climate Observing System (GCOS Switzerland). Update 2018. Available online at www.gcos.ch/inventory.

Pettitt, A.N. 1979: A non-parametric approach to the change-point detection. *Appl. Statist.* 28(2), 126-135.

Pietragalla, B., V. Knechtl, C. Sigg, and B. Clot, 2016: New automated quality control of phenological data in Switzerland. *EGU General Assembly Conference Abstracts* EPSC2016-6278.

Reid, P.C., R.E. Hari, G. Beaugrand, et. al, 2016: Global impacts of the 1980s regime shift. *Glob. Chang. Biol.* 22: 682–703. doi: 10.1111/gcb.13106.

Rutishauser, T., J. Luterbacher, C. Defila, D. Frank, and H. Wanner, 2008: Swiss Spring Plant Phenology 2007: Extremes, a multi-century perspective and changes in temperature sensitivity. *Geophys. Res. Lett.* 35, L05703. doi: 10.1029/2007GL032545..

Rutishauser, T., R. Stöckli, J. Harte, and L. Kueppers, 2012: Climate change: Flowering in the greenhouse. *Nature*, 485:448–449, doi :10.1038/485448a.

Schleip, C., J. Luterbacher, and A. Menzel, 2008: Time series modeling and central European temperature impact assessment of phenological series over the last 250 years. *J. Geophys. Res. 113*, G04026. doi :10.1029/2007JG000646.

Seiz, G. and N. Foppa, 2007: Nationales Klima-Beobachtungssystem (GCOS Schweiz). Publikation von MeteoSchweiz und ProClim, 92 S.

Stöckli, R., T. Rutishauser, I. Baker, M. Liniger, and S. Denning, 2011: Global Reanalysis of Vegetation Phenology, *J. Geophys. Res*. 116, G03020, doi:10.1029/2010JG001545.

Studer, S., C. Appenzeller, and C. Defila, 2005: Inter-annual variability and decadal trends in Alpine spring phenology: A multivariate analysis approach. *Clim. Change,* 73: 395–414. doi:10.1007/s10584-005-6886-z.

Toreti, A., F.G. Kuglitsch, E. Xoplaki, and J. Luterbacher, 2012: A novel approach for the detection of inhomogeneities affecting climate time series. *J. Appl. Meteorol. Clim.*, 51: 317–326. doi: 10.1175/JAMC-D-10-05033.1.

Venema, V., O. Mestre, E. Aguilar, et al. 2012: Benchmarking homogenization algorithms for monthly data. *Clim. Past,* 8, 89-115. doi:10.5194/cp-8-89-2012.

Wang, X.L. 2008: Penalized maximal F test for detecting undocumented mean shift without trend change, *J. Atmos. Oceanic Technol*. 25, 368–384, doi:10.1175/2007JTECHA982.1.

WMO, 2016: The global observing system for climate: implementation needs. GCOS-200. World Meteorological Organization, 315 pp.

Wolkovich, E.M., B.I. Cook, J.M. Allen, et al. 2012: Warming experiments underpredict plant phenological responses to climate change. *Nature,* 485, 494–497. doi:10.1038/nature11014.

Acknowledgements

PhenoClass was financed by MeteoSwiss in the framework of GCOS Switzerland. We would like to thank Yang Feng for providing the last version of the RHtests software.

Abbreviations

DWD	Deutscher Wetterdienst (German Weather Service)
GCOS	Global Climate Observing System
PEP	Pan European Phenology
QC	Quality Control
SNHT	Standard Normal Homogeneity Test
SPN	Swiss Phenology Network

Appendix A

Table A: Biological rules to define inconsistent observations.

Nr.	Parameter 1	Parameter 1 Name	Test	Parameter 2	Parameter 2 Name
1	mmald65d	Apfelbaum - Allgemeine Blüte	<	mmald60d	Apfelbaum - Beginn der Blüte
2	macep94d	Bergahorn - Allgemeine Blattverfärbung	<=	macep13d	Bergahorn - Allgemeine Blattentfaltung
3	mpyrc65d	Birnbaum - Allgemeine Blüte	<	mpyrc60d	Birnbaum - Beginn der Blüte
4	mfags94d	Buche - Allgemeine Blattverfärbung	<=	mfags13d	Buche - Allgemeine Blattentfaltung
5	mfags95d	Buche - Allgemeiner Blattfall	<	mfags94d	Buche - Allgemeine Blattverfärbung
6	mcass60d	Edelkastanie - Beginn der Blüte	<=	mcass13d	Edelkastanie - Allgemeine Blattentfaltung
7	mcass65d	Edelkastanie - Allgemeine Blüte	<	mcass60d	Edelkastanie - Beginn der Blüte
8	mcass87d	Edelkastanie - Allgemeine Fruchtreife	<=	mcass65d	Edelkastanie - Allgemeine Blüte
9	mcass95d	Edelkastanie - Allgemeiner Blattfall	<=	mcass87d	Edelkastanie - Allgemeine Fruchtreife
10	mcass94d	Edelkastanie - Allgemeine Blattverfärbung	<=	mcass65d	Edelkastanie - Allgemeine Blüte
11	mcass95d	Edelkastanie - Allgemeiner Blattfall	<	mcass94d	Edelkastanie - Allgemeine Blattverfärbung
12	mbetp65d	Hängebirke - Allgemeine Blüte	<	mbetp60d	Hängebirke - Beginn der Blüte
13	mbetp94d	Hängebirke - Allgemeine Blattverfärbung	<=	mbetp65d	Hängebirke - Allgemeine Blüte
14	mbetp95d	Hängebirke - Allgemeiner Blattfall	<	mbetp94d	Hängebirke - Allgemeine Blattverfärbung
15	mcora65d	Haselstrauch - Allgemeine Blüte	<	mcora60d	Haselstrauch - Beginn der Blüte
16	mprua65d	Kirschbaum - Allgemeine Blüte	<	mprua60d	Kirschbaum - Beginn der Blüte
17	mlard94d	Lärche - Allgemeine Nadelverfärbung	<=	mlard13d	Lärche - Allgemeiner Nadelaustrieb
18	mlard95d	Lärche - Allgemeiner Nadelfall	<	mlard94d	Lärche - Allgemeine Nadelverfärbung
19	mrobp60d	Robinie - Beginn der Blüte	<=	mrobp13d	Robinie - Allgemeine Blattentfaltung
20	mrobp65d	Robinie - Allgemeine Blüte	<	mrobp60d	Robinie - Beginn der Blüte
21	mrobp95d	Robinie - Allgemeiner Blattfall	<=	mrobp65d	Robinie - Allgemeine Blüte
22	maesh65d	Rosskastanie - Allgemeine Blüte	<	maesh60d	Rosskastanie - Beginn der Blüte

23	maesh95d	Rosskastanie - Allgemeiner Blattfall	<	maesh94d	Rosskastanie - Allgemeine Blattverfärbung
24	msamr65d	Roter Holunder - Allgemeine Blüte	<	msamr60d	Roter Holunder - Beginn der Blüte
25	msamr87d	Roter Holunder - Allgemeine Fruchtreife	<=	msamr65d	Roter Holunder - Allgemeine Blüte
26	msamn65d	Schwarzer Holunder - Allgemeine Blüte	<	msamn60d	Schwarzer Holunder - Beginn der Blüte
27	msamn87d	Schwarzer Holunder - Allgemeine Fruchtreife	<=	msamn65d	Schwarzer Holunder - Allgemeine Blüte
28	mtilp60d	Sommerlinde - Beginn der Blüte	<=	mtilp13d	Sommerlinde - Allgemeine Blattentfaltung
29	mtilp65d	Sommerlinde - Allgemeine Blüte	<	mtilp60d	Sommerlinde - Beginn der Blüte
30	mtilp94d	Sommerlinde - Allgemeine Blattverfärbung	<=	mtilp65d	Sommerlinde - Allgemeine Blüte
31	msora60d	Vogelbeere - Beginn der Blüte	<=	msora13d	Vogelbeere - Allgemeine Blattentfaltung
32	msora65d	Vogelbeere - Allgemeine Blüte	<	msora60d	Vogelbeere - Beginn der Blüte
33	msora87d	Vogelbeere - Allgemeine Fruchtreife	<=	msora65d	Vogelbeere - Allgemeine Blüte
34	msora94d	Vogelbeere - Allgemeine Blattverfärbung	<=	msora87d	Vogelbeere - Allgemeine Fruchtreife
35	msora95d	Vogelbeere - Allgemeiner Blattfall	<	msora94d	Vogelbeere - Allgemeine Blattverfärbung
36	mvitv89d	Weinrebe - Weinlese	<=	mvitv65d	Weinrebe - Allgemeine Blüte
37	mtilc60d	Winterlinde - Beginn der Blüte	<=	mtilc13d	Winterlinde - Allgemeine Blattentfaltung
38	mtilc65d	Winterlinde - Allgemeine Blüte	<	mtilc60d	Winterlinde - Beginn der Blüte
39	mtilc94d	Winterlinde - Allgemeine Blattverfärbung	<=	mtilc65d	Winterlinde - Allgemeine Blüte
40	mcora13d	Haselstrauch - Allgemeine Blattentfaltung	<=	mcora65d	Haselstrauch - Allgemeine Blüte

Appendix B

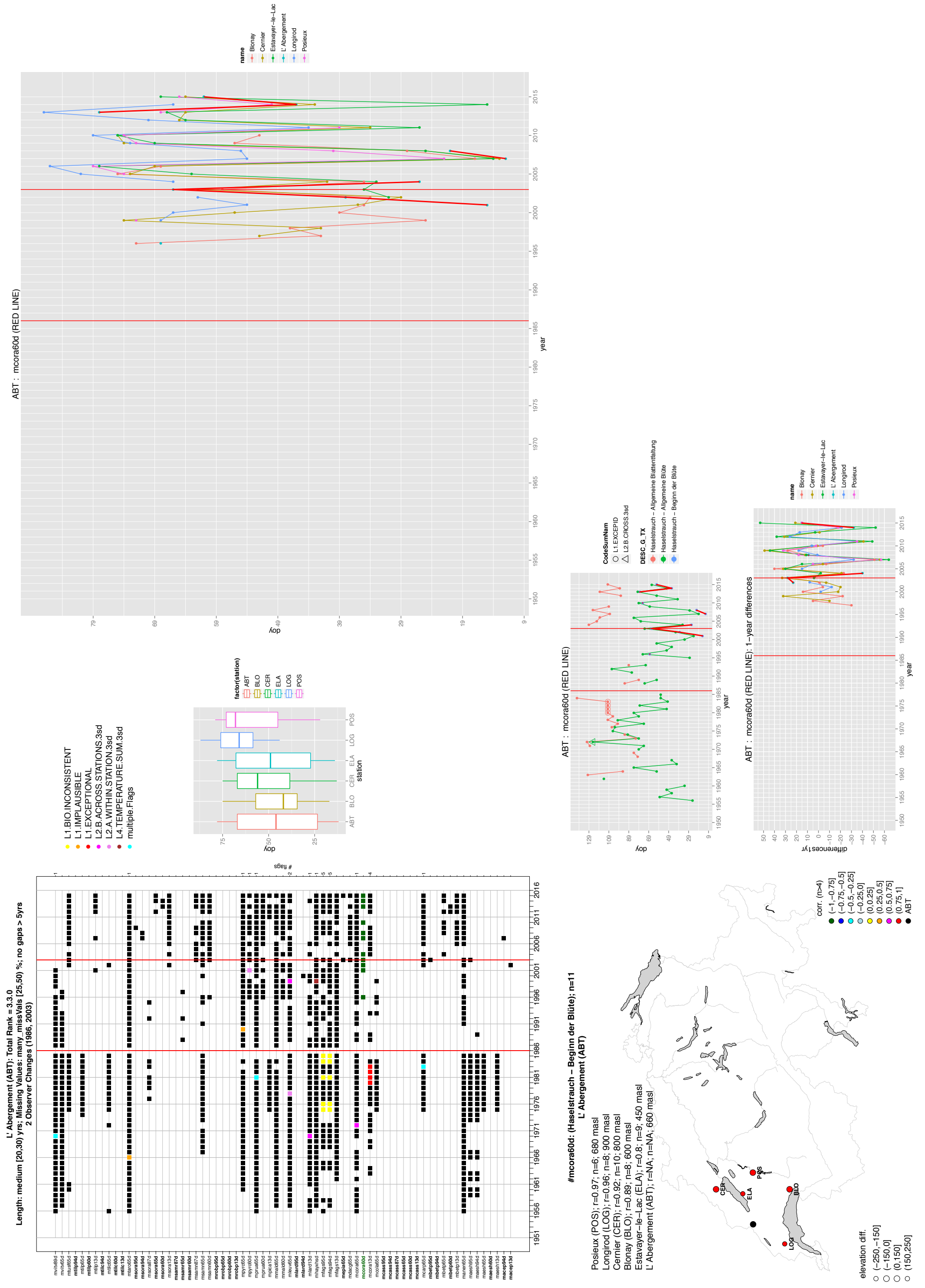


Figure B: Example of an inspection sheet for L’Abergement.

Appendix C

Table C: List of station with highest score (1) series.

Stat.	Station Name	Parameter	Parameter Name	Start	End
ADB	Adelboden	manen65d	Wood anemone - full flowering	1965	2015
EHT	Escholzmatt	mfags13d	European beech - leaf unfolding	1956	2015
		mfags94d	European beech - leaf colouring	1956	2015
		mleuv65d	Field daisy - full flowering	1956	2015
		mhayxhsd	Hay harvest - start	1956	2015
ELP	Elm	maesh13d	Horse chestnut - leaf unfolding	1956	2015
		msamr65d	European red elder - full flowering	1956	2015
		mlard13d	European larch - needle emergence	1956	2015
		mcarp65d	Cuckoo flower - full flowering	1956	2015
		mleuv65d	Field daisy - full flowering	1956	2015
ENB	Entlebuch	maesh13d	Horse chestnut - leaf unfolding	1958	2015
		manen65d	Wood anemone - full flowering	1958	2015
		mtaro65d	Dandelion - full flowering	1958	2015
		mcarp65d	Cuckoo flower - full flowering	1958	2015
		mpyrc65d	Pear tree - full flowering	1958	2015
LCN	Locarno	maesh13d	Horse chestnut - leaf unfolding	1966	2015
		mfags13d	European beech - leaf unfolding	1966	2015
		mlard13d	European larch - needle emergence	1966	2015
		mprua65d	Cherry tree - full flowering	1966	2015
LIT	Liestal	maesh13d	Horse chestnut - leaf unfolding	1951	2015
		maesh65d	Horse chestnut - full flowering	1951	2015
		mfags13d	European beech - leaf unfolding	1951	2015
		mfags94d	European beech - leaf colouring	1951	2015
		mfags95d	European beech - leaf drop	1951	2014
		msamn65d	European elder - full flowering	1951	2015
		mpica13d	Common spruce - needle emergence	1951	2015
		mtusf65d	Coltsfoot - full flowering	1951	2015
		manen65d	Wood anemone - full flowering	1951	2015
		mtaro65d	Dandelion - full flowering	1951	2015
		mcarp65d	Cuckoo flower - full flowering	1951	2015
LPM	Les Ponts-de-Martel	mfags13d	European beech - leaf unfolding	1951	2015
MEH	Merishausen	maesh13d	Horse chestnut - leaf unfolding	1964	2015
		mfags94d	European beech - leaf colouring	1959	2015

		mfags95d	European beech - leaf drop	1959	2015
		mcora13d	Hazel - leaf unfolding	1959	2015
		mcora65d	Hazel - full flowering	1959	2015
		msamn65d	European elder - full flowering	1959	2015
		mlard13d	European larch - needle emergence	1959	2015
		mtusf65d	Coltsfoot - full flowering	1959	2015
		mcarp65d	Cuckoo flower - full flowering	1959	2015
		mcola65d	Autumn crocus - full flowering	1959	2015
		mprua65d	Cherry tree - full flowering	1959	2015
		mpyrc65d	Pear tree - full flowering	1959	2015
MUG	Murg	maesh13d	Horse chestnut - leaf unfolding	1952	2015
		mfags94d	European beech - leaf colouring	1951	2015
		mprua65d	Cherry tree - full flowering	1952	2015
PSO	Prato-Sornico	manen65d	Wood anemone - full flowering	1957	2015
		mpyrc65d	Pear tree - full flowering	1957	2015
		mhayxhsd	Hay harvest - start	1957	2015
SEO	Seon	mprua65d	Cherry tree - full flowering	1952	2015
		mpyrc65d	Pear tree - full flowering	1953	2015
SGS	Sargans II	maesh13d	Horse chestnut - leaf unfolding	1956	2015
		mfags13d	European beech - leaf unfolding	1956	2015
		mtaro65d	Dandelion - full flowering	1956	2015
		mprua65d	Cherry tree - full flowering	1956	2015
		mmald65d	Apple tree - full flowering	1956	2015
		mvitv65d	Grape vine - full flowering	1956	2015
		mhayxhsd	Hay harvest - start	1956	2015
TRT	Trient	mtaro65d	Dandelion - full flowering	1951	2015
		mprua65d	Cherry tree - full flowering	1951	2015
VES	Versoix	maesh13d	Horse chestnut - leaf unfolding	1952	2014
		mtaro65d	Dandelion - full flowering	1952	2014
WAT	Wattwil, SG	maesh13d	Horse chestnut - leaf unfolding	1951	2015
		mlard13d	European larch - needle emergence	1952	2015
		mprua65d	Cherry tree - full flowering	1952	2015
WIH	Wildhaus	mcora13d	Hazel - leaf unfolding	1951	2015
		mlard13d	European larch - needle emergence	1951	2015
		mtaro65d	Dandelion - full flowering	1951	2015
		mleuv65d	Field daisy - full flowering	1951	2014
ZHP	Zürich-MeteoSchweiz	maesh13d	Horse chestnut - leaf unfolding	1955	2015

	mtaro65d	Dandelion - full flowering	1955	2015

Appendix D

Table D: List of 178 over 60 years long series without a quality flag, no automatically detected break and less than 3 missing values. Visual Code: 1: series has undergone break detection, no break detected, also visually no break found; 2: series has undergone break detection, no break detected, visually not clear if break/breaks; 3: series has undergone break detection, no break detected, visually a clear break found; 4: series has not undergone break detection, visually no break found; 5: series has not undergone break detection, visually maybe a break/breaks found; 6: series has not undergone break detection, visually a clear break found; 7: series has not undergone break detection, no information available from neighbouring stations to judge series.

Stat.	Station Name	Parameter	Parameter Name	Class	Length	Miss Vals	Homogeneous (1) / not tested(NA)	Visual Code	GCOS y/n
EHT	Escholzmatt	mfags13d	European beech - leaf unfolding	1	60	0	1	1	n
		mfags94d	European beech - leaf colouring	1	60	0	1	2	n
		mleuv65d	Field daisy - full flowering	1	60	0	1	1	n
		mhayxhsd	Hay harvest - start	1	60	0	1	1	n
ELP	Elm	maesh13d	Horse chestnut - leaf unfolding	1	60	0	1	1	n
		msamr65d	European red elder - full flowering	1	60	0	1	1	n
		mlard13d	European larch - needle emergence	1	60	0	1	1	n
		mcarp65d	Cuckoo flower - full flowering	1	60	0	1	1	n
		mleuv65d	Field daisy - full flowering	1	60	0	1	1	n
LIT	Liestal	maesh13d	Horse chestnut - leaf unfolding	1	65	0	1	1	y
		maesh65d	Horse chestnut - full flowering	1	65	0	1	1	y
		mfags13d	European beech - leaf unfolding	1	65	0	1	1	y
		mfags94d	European beech - leaf colouring	1	65	0	1	1	y
		mfags95d	European beech - leaf drop	1	64	0	1	1	y

		msamn65d	European elder - full flowering	1	65	0	1	1	y
		mpica13d	Common spruce - needle emergence	1	65	0	NA	5	y
		mtusf65d	Coltsfoot - full flowering	1	65	0	1	1	y
		manen65d	Wood anemone - full flowering	1	65	0	1	1	y
		mtaro65d	Dandelion - full flowering	1	65	0	1	1	y
		mcarp65d	Cuckoo flower - full flowering	1	65	0	1	1	y
LPM	Les Ponts-de-Martel	mfags13d	European beech - leaf unfolding	1	65	0	1	1	n
MUG	Murg	maesh13d	Horse chestnut - leaf unfolding	1	64	0	1	3	y
		mfags94d	European beech - leaf colouring	1	65	0	1	1	y
		mprua65d	Cherry tree - full flowering	1	64	0	1	1	y
SEO	Seon	mprua65d	Cherry tree - full flowering	1	64	0	1	1	n
		mpyrc65d	Pear tree - full flowering	1	63	0	1	1	n
SGS	Sargans II	maesh13d	Horse chestnut - leaf unfolding	1	60	0	1	2	n
		mfags13d	European beech - leaf unfolding	1	60	0	1	1	n
		mtaro65d	Dandelion - full flowering	1	60	0	1	1	n
		mprua65d	Cherry tree - full flowering	1	60	0	1	1	n
		mmald65d	Apple tree - full flowering	1	60	0	1	1	n
		mvitv65d	Grape vine - full flowering	1	60	0	1	1	n
		mhayxhsd	Hay harvest - start	1	60	0	NA	4	n
TRT	Trient	mtaro65d	Dandelion - full flowering	1	65	0	1	1	y
		mprua65d	Cherry tree - full flowering	1	65	0	1	1	y
VES	Versoix	maesh13d	Horse chestnut - leaf unfolding	1	63	0	1	1	y
		mtaro65d	Dandelion - full	1	63	0	1	1	y

			flowering						
WAT	Wattwil, SG	maesh13d	Horse chestnut - leaf unfolding	1	65	0	1	1	n
		mlard13d	European larch - needle emergence	1	64	0	1	1	n
		mprua65d	Cherry tree - full flowering	1	64	0	1	1	n
WIH	Wildhaus	mcora13d	Hazel - leaf unfolding	1	65	0	1	1	y
		mlard13d	European larch - needle emergence	1	65	0	1	1	y
		mtaro65d	Dandelion - full flowering	1	65	0	1	1	y
		mleuv65d	Field daisy - full flowering	1	64	0	1	1	y
ZHP	Zürich-MeteoSchweiz	maesh13d	Horse chestnut - leaf unfolding	1	61	0	1	1	n
		mtaro65d	Dandelion - full flowering	1	61	0	1	1	n
		mcarp65d	Cuckoo flower - full flowering	1	61	0	1	1	n
ABT	L' Abergement	mprua65d	Cherry tree - full flowering	2	60	3	1	2	n
APL	Appenzell	mlard13d	European larch - needle emergence	2	60	1	1	1	n
		mtaro65d	Dandelion - full flowering	2	60	1	1	1	n
CAR	Cartigny	maesh13d	Horse chestnut - leaf unfolding	2	62	3	1	2	n
		maesh65d	Horse chestnut - full flowering	2	62	3	1	2	n
		mcora13d	Hazel - leaf unfolding	2	62	3	1	1	n
		mlard13d	European larch - needle emergence	2	62	3	1	1	n
		mtaro65d	Dandelion - full flowering	2	62	3	1	1	n
		mprua65d	Cherry tree - full flowering	2	62	3	1	1	n
COE	Couvet	maesh13d	Horse chestnut - leaf unfolding	2	62	2	1	1	n

		maesh94d	Horse chestnut - leaf colouring	2	62	1	NA	5	n
		mfags94d	European beech - leaf colouring	2	62	1	1	1	n
		mfags95d	European beech - leaf drop	2	62	1	1	1	n
		mpica13d	Common spruce - needle emergence	2	62	2	1	1	n
		mtusf65d	Coltsfoot - full flowering	2	62	1	NA	4	n
		manen65d	Wood anemone - full flowering	2	62	3	NA	5	n
		mtaro65d	Dandelion - full flowering	2	62	1	1	1	n
		mleuv65d	Field daisy - full flowering	2	62	1	1	1	n
		mprua65d	Cherry tree - full flowering	2	62	3	1	1	n
		mpyrc65d	Pear tree - full flowering	2	62	3	1	1	n
		mhayxhsd	Hay harvest - start	2	62	1	NA	7	n
DST	Disentis	mtusf65d	Coltsfoot - full flowering	2	60	2	1	1	n
EHT	Escholzmatt	mfags95d	European beech - leaf drop	2	60	1	1	1	n
		mcora13d	Hazel - leaf unfolding	2	60	3	1	2	n
		msamn65d	European elder - full flowering	2	60	1	1	1	n
		mtaro65d	Dandelion - full flowering	2	60	1	1	1	n
		mprua65d	Cherry tree - full flowering	2	60	1	1	1	n
		mpyrc65d	Pear tree - full flowering	2	60	2	1	1	n
		mmald65d	Apple tree - full flowering	2	60	2	1	1	n
ELP	Elm	maesh95d	Horse chestnut - leaf drop	2	60	3	NA	4	n
		mcora13d	Hazel - leaf unfolding	2	60	1	1	1	n
		mpica13d	Common spruce - needle emergence	2	60	2	1	1	n

		mmald65d	Apple tree - full flowering	2	60	3	1	1	n
		mhayxhsd	Hay harvest - start	2	60	1	NA	7	n
ENS	Enges	mfags94d	European beech - leaf colouring	2	65	3	1	2	y
		mprua65d	Cherry tree - full flowering	2	65	3	1	1	y
GON	Gryon	mtaro65d	Dandelion - full flowering	2	60	1	1	1	n
KAN	Kandersteg	mlard13d	European larch - needle emergence	2	60	1	NA	1	n
		manen65d	Wood anemone - full flowering	2	60	1	NA	7	n
		mtaro65d	Dandelion - full flowering	2	60	1	1	1	n
		mcola65d	Autumn crocus - full flowering	2	60	1	NA	4	n
LIT	Liestal	mcora65d	Hazel - full flowering	2	62	2	1	1	y
		mtilc65d	Small leaved lime - full flowering	2	60	1	1	1	y
		mlard13d	European larch - needle emergence	2	64	1	1	1	y
LOL	Le Locle	mcora13d	Hazel - leaf unfolding	2	60	3	NA	4	n
		mtaro65d	Dandelion - full flowering	2	60	1	1	1	n
LPM	Les Ponts-de-Martel	maesh13d	Horse chestnut - leaf unfolding	2	65	1	1	1	n
		maesh65d	Horse chestnut - full flowering	2	65	2	1	1	n
		mtusf65d	Coltsfoot - full flowering	2	65	2	1	2	n
		mtaro65d	Dandelion - full flowering	2	65	1	1	1	n
		mhayxhsd	Hay harvest - start	2	65	3	NA	4	n
MUG	Murg	maesh94d	Horse chestnut - leaf colouring	2	65	2	NA	4	y
		mfags13d	European beech - leaf unfolding	2	64	1	1	2	y
		mcora13d	Hazel - leaf	2	64	1	1	1	y

			unfolding						
		msamn65d	European elder - full flowering	2	65	2	1	1	y
		mlard13d	European larch - needle emergence	2	64	1	1	1	y
		mpica13d	Common spruce - needle emergence	2	65	3	1	2	y
		mtusf65d	Coltsfoot - full flowering	2	63	3	1	1	y
		manen65d	Wood anemone - full flowering	2	64	1	1	1	y
		mcarp65d	Cuckoo flower - full flowering	2	64	1	1	1	y
		mleuv65d	Field daisy - full flowering	2	65	1	1	2	y
		mmald65d	Apple tree - full flowering	2	65	3	1	1	y
		mvitv89d	Grape vine - vintage	2	65	1	1	1	y
RAF	Rafz	mcora13d	Hazel - leaf unfolding	2	63	2	1	3	y
		mcora65d	Hazel - full flowering	2	64	1	1	1	y
		mpica13d	Common spruce - needle emergence	2	64	2	1	1	y
		mtusf65d	Coltsfoot - full flowering	2	64	1	1	1	y
		manen65d	Wood anemone - full flowering	2	64	1	1	2	y
		mtaro65d	Dandelion - full flowering	2	64	1	1	1	y
		mcarp65d	Cuckoo flower - full flowering	2	64	1	1	1	y
		mleuv65d	Field daisy - full flowering	2	64	1	1	2	y
		mprua65d	Cherry tree - full flowering	2	64	2	1	1	y
		mpyrc65d	Pear tree - full flowering	2	64	2	1	1	y
SEO	Seon	maesh13d	Horse chestnut - leaf unfolding	2	64	2	1	1	n
		mfags13d	European beech - leaf unfolding	2	64	1	1	1	n

		mfags94d	European beech - leaf colouring	2	63	2	1	2	n
		mtusf65d	Coltsfoot - full flowering	2	62	3	1	1	n
		mtaro65d	Dandelion - full flowering	2	64	2	1	2	n
		mhayxhsd	Hay harvest - start	2	64	3	1	1	n
SGS	Sargans II	maesh94d	Horse chestnut - leaf colouring	2	60	3	1	1	n
		maesh95d	Horse chestnut - leaf drop	2	60	2	1	1	n
		mfags94d	European beech - leaf colouring	2	60	3	NA	4	n
		mfags95d	European beech - leaf drop	2	60	2	NA	4	n
		msamn65d	European elder - full flowering	2	60	1	1	1	n
		mlard13d	European larch - needle emergence	2	60	1	1	2	n
		mpyrc65d	Pear tree - full flowering	2	60	1	1	1	n
SID	Simplon-Dorf	mtaro65d	Dandelion - full flowering	2	64	3	1	1	n
SNN	Sarnen	maesh13d	Horse chestnut - leaf unfolding	2	62	1	1	1	y
		maesh65d	Horse chestnut - full flowering	2	62	1	1	1	y
		maesh94d	Horse chestnut - leaf colouring	2	62	2	NA	7	y
		maesh95d	Horse chestnut - leaf drop	2	62	3	NA	5	y
		mfags13d	European beech - leaf unfolding	2	62	1	1	1	y
		mcora13d	Hazel - leaf unfolding	2	62	1	1	1	y
		mtusf65d	Coltsfoot - full flowering	2	62	1	1	1	y
		manen65d	Wood anemone - full flowering	2	62	1	1	1	y
		mcarp65d	Cuckoo flower - full flowering	2	62	1	1	1	y
		mleuv65d	Field daisy - full flowering	2	62	1	1	1	y
		mprua65d	Cherry tree - full	2	62	1	1	1	y

			flowering						
		mpyrc65d	Pear tree - full flowering	2	62	1	1	1	y
		mmald65d	Apple tree - full flowering	2	62	1	1	1	y
		mhayxhsd	Hay harvest - start	2	62	2	NA	4	y
TRT	Trient	mmald65d	Apple tree - full flowering	2	65	1	1	2	y
		mhayxhsd	Hay harvest - start	2	65	1	NA	7	y
TUS	Thusis	mtaro65d	Dandelion - full flowering	2	60	1	1	1	n
VAS	Vals	maesh13d	Horse chestnut - leaf unfolding	2	60	3	1	1	n
VEG	Vergeletto	mfags94d	European beech - leaf colouring	2	60	1	1	3	n
		mlard13d	European larch - needle emergence	2	60	2	1	3	n
VES	Versoix	maesh65d	Horse chestnut - full flowering	2	63	2	1	1	y
		mtilp65d	Large leaved lime - full flowering	2	63	3	1	2	y
		mtusf65d	Coltsfoot - full flowering	2	63	1	1	2	y
		mcarp65d	Cuckoo flower - full flowering	2	62	1	1	1	y
VSA	La Valsainte	mfags13d	European beech - leaf unfolding	2	60	1	1	1	y
		mlard13d	European larch - needle emergence	2	60	1	1	1	y
		manen65d	Wood anemone - full flowering	2	60	2	1	1	y
		mtaro65d	Dandelion - full flowering	2	60	1	1	1	y
		mhayxhsd	Hay harvest - start	2	60	2	NA	4	y
WAT	Wattwil, SG	mfags13d	European beech - leaf unfolding	2	65	1	1	1	n
		manen65d	Wood anemone - full flowering	2	64	1	1	2	n
		mtaro65d	Dandelion - full flowering	2	64	1	1	3	n

		mpyrc65d	Pear tree - full flowering	2	64	1	1	1	n
		mmald65d	Apple tree - full flowering	2	64	1	1	1	n
WIB	Wiliberg	mfags94d	European beech - leaf colouring	2	64	3	1	3	n
		manen65d	Wood anemone - full flowering	2	64	2	1	1	n
		mcarp65d	Cuckoo flower - full flowering	2	64	1	1	1	n
		mprua65d	Cherry tree - full flowering	2	64	2	1	1	n
WIH	Wildhaus	mfags94d	European beech - leaf colouring	2	65	2	1	2	y
		mfags95d	European beech - leaf drop	2	65	2	1	1	y
		mhayxhsd	Hay harvest - start	2	64	2	1	1	y
ZHP	Zürich-MeteoSchweiz	maesh65d	Horse chestnut - full flowering	2	60	1	1	1	n
		mcora65d	Hazel - full flowering	2	61	1	1	1	n
		mprua65d	Cherry tree - full flowering	2	60	1	1	1	n

MeteoSchweiz
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

MeteoSvizzera
Via ai Monti 146
CH-6605 Locarno Monti
T +41 58 460 97 77
www.meteosvizzera.ch

MétéoSuisse
7bis, av. de la Paix
CH-1211 Genève 2
T +41 58 460 98 88
www.meteosuisse.ch

MétéoSuisse
Chemin de l'Aérologie
CH-1530 Payerne
T +41 58 460 94 44
www.meteosuisse.ch