

Break detection of annual Swiss temperature series

F. G. Kuglitsch,¹ R. Auchmann,¹ R. Bleisch,² S. Brönnimann,¹ O. Martius,¹
and M. Stewart¹

Received 4 March 2012; revised 25 May 2012; accepted 29 May 2012; published 4 July 2012.

[1] Instrumental temperature series are often affected by artificial breaks (“break points”) due to (e.g.) changes in station location, land-use, or instrumentation. The Swiss climate observation network offers a high number and density of stations, many long and relatively complete daily to sub-daily temperature series, and well-documented station histories (i.e., metadata). However, for many climate observation networks outside of Switzerland, detailed station histories are missing, incomplete, or inaccessible. To correct these records, the use of reliable statistical break detection methods is necessary. Here, we apply three statistical break detection methods to high-quality Swiss temperature series and use the available metadata to assess the methods. Due to the complex terrain in Switzerland, we are able to assess these methods under specific local conditions such as the Foehn or crest situations. We find that the temperature series of all stations are affected by artificial breaks (average = 1 break point / 48 years) with discrepancies in the abilities of the methods to detect breaks. However, by combining the three statistical methods, almost all of the detected break points are confirmed by metadata. In most cases, these break points are ascribed to a combination of factors in the station history.

Citation: 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.

1. Introduction

[2] Instrumental surface temperature series are the foundation for studies of paleo [e.g., Stewart *et al.*, 2011], recent past and present [e.g., Alexandersson and Moberg, 1997; Vincent *et al.*, 2002; Caussinus and Mestre, 2004; Trenberth *et al.*, 2007], and future [Giorgi and Mearns, 1991; Stott and Forest, 2007] climate change. However, most instrumental (“observational”) climate series suffer from non-climatic effects (e.g., the relocation of weather stations, land-use changes, new instruments, and changes in observational hours [Peterson *et al.*, 1998; Aguilar *et al.*, 2003]) which result in inhomogeneities (“break points” or BPs [e.g., Peterson *et al.*, 1998; Szentimrey, 1999; González-Rouco *et al.*, 2001; Cocheo and Camuffo, 2002; Maugeri *et al.*, 2002; Brandsma and Können, 2006; Della-Marta and Wanner, 2006; Kuglitsch *et al.*, 2009, 2010]). Usually, these effects are not documented in metadata. Or, where they are documented, the metadata are in archives which are not readily accessible. Fortunately, statistical methods have been developed to

detect inhomogeneities [e.g., Caussinus and Lyazrhi, 1997; Vincent, 1998; Szentimrey, 1999; Caussinus and Mestre, 2004; Wang *et al.*, 2007; Wang, 2008a, 2008b].

[3] Here, we apply three statistical methods (CAUME based on Caussinus and Mestre [2004]; WANG based on Wang *et al.* [2007], Wang [2008a, 2008b] and X. L. Wang and I. Feng (RHtestV3 User manual, 2010, available at <http://cccma.seos.uvic.ca/ETCCDMI/software.shtml>, 2010); TORETI based on Toreti *et al.* [2012]) to high-quality daily and sub-daily temperature series, integrating reference series from the Swiss, French, and German observation networks. Specifically, we consider more than 60 stations in the Swiss observation network of the Federal Office of Meteorology and Climatology (MeteoSwiss) and explore the potential of different statistical method combinations for detecting major historical station changes and accompanied break points. We chose to use the Swiss observation network as the focus of this study because it has: (1) a high number and density of stations in different environment conditions (e.g., highlands, low-lands, urban and rural areas, and areas affected by local climate effects); (2) many long and relatively complete daily to sub-daily temperature series going back to 1864; and (3) well-documented station histories and metadata that allow a detailed comparison with break points found through statistical methods.

[4] Then, we determine the consistency among statistical break detection methods and metadata. This is fundamental for developing a comprehensively homogenized Swiss station network and for determining the effectiveness of these statistical methods for climate observation networks outside

¹Institute of Geography and Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland.

²Institute of Applied Physics, University of Bern, Bern, Switzerland.

Corresponding author: F. G. Kuglitsch, Institute of Geography, University of Bern, Hallerstrasse 12, CH-3012 Bern, Switzerland. (franz.kuglitsch@gmail.com)

©2012. American Geophysical Union. All Rights Reserved. 0148-0227/12/2012JD017729

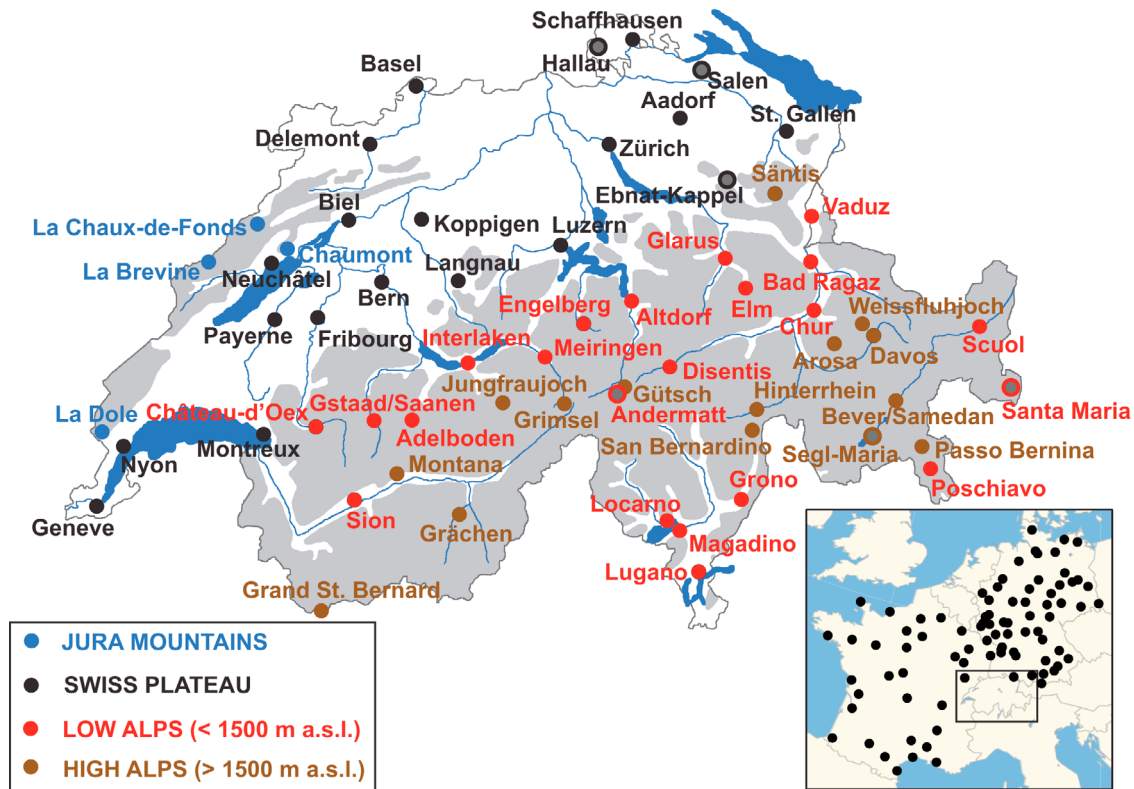


Figure 1. The geographical distribution of (1) the 56 Swiss stations (black, red, blue, and brown dots) used to analyze artificial break points in daily TT, TX, TN, Tmor, Tnoo, Teve series and (2) the 5 stations (gray dots with black, red with brown contours) used to analyze daily Tmor, Tnoo and Teve only. Grey and blue shaded areas indicate altitudes above 1000 m a.s.l. and water bodies, respectively. The inset map shows the geographical distribution of the French and German station network recording TT, TX, and TN, used for reference.

of Switzerland (i.e., regions where detailed station histories may be missing, incomplete, or inaccessible).

2. Data and Methods

[5] The Swiss climate observation network was established in 1864 (cf. Figures 1 and 3–5). Since then, three daily observations and (at most sites) maximum and minimum temperatures have been registered. Around 1980, the network was automated. Currently, these automatic stations are being modernized. A subset of stations, selected by MeteoSwiss as a National Baseline Climatology Network (NBCN), has had historical temperature data digitized in the framework of the DigiHom project [Füllemann *et al.*, 2011]. Previous attempts to homogenize Swiss temperature records have only considered: monthly mean temperatures [Baudenbacher, 1997; Begert *et al.*, 2005], daily maximum temperatures [Della-Marta *et al.*, 2007; Kuglitsch *et al.*, 2009], a small number of stations, and a sole statistical break detection method without the validation or homogeneity assessment of subdaily data.

[6] Here, daily mean (TT), maximum (TX), minimum (TN), morning (Tmor; 06:00 UTC), noon (Tnoo; 12:00 UTC), and evening (Teve; 18:00/20:00 UTC) temperatures, from 61 stations in Switzerland, are screened for artificial break points using statistical methods. These results are compared with metadata. In 1971, the observation times of the evening values changed from 20:00 UTC to 18:00 UTC, affecting evening temperatures and the daily mean temperature. Therefore, the daily mean temperature for non-automatic stations is calculated as:

$$\begin{aligned} \text{until 1970: } TT &= (Tmor + Tnoo + 2 \times Teve)/4 \\ \text{since 1971: } TT &= n - k \times (n - TN) \end{aligned}$$

where $n = (Tmor + Tnoo + Teve)/3$, and $k = \text{factor}$; dependent on location and time [Begert *et al.*, 2003]. For all non-automatic stations, detected breaks in TT must be validated by a simultaneous break in either Tmor, Tnoo and/or Teve. For automatic stations, introduced around 1980 (ANETZ, see auxiliary material, Table S1), daily TT is the mean of

Figure 2. A summary of the detected break points for TT, TX, TN, Tmor, Tnoo, and Teve in (top) Basel, (middle) Altdorf, and (bottom) Grand St. Bernard using metadata and the statistical methods of: (1) CAUME, (2) WANG, and (3) TORETI. The black lines correspond to the length of the daily TT, TX, TN, Tmor, Tnoo, and Teve series. The black crosses indicate the years of detected break points following the respective statistical method. Colored vertical lines represent years when major problems are indicated by metadata. Dotted lines designate longer periods of changes and problems. Grey vertical lines summarize the break points found with statistical methods.

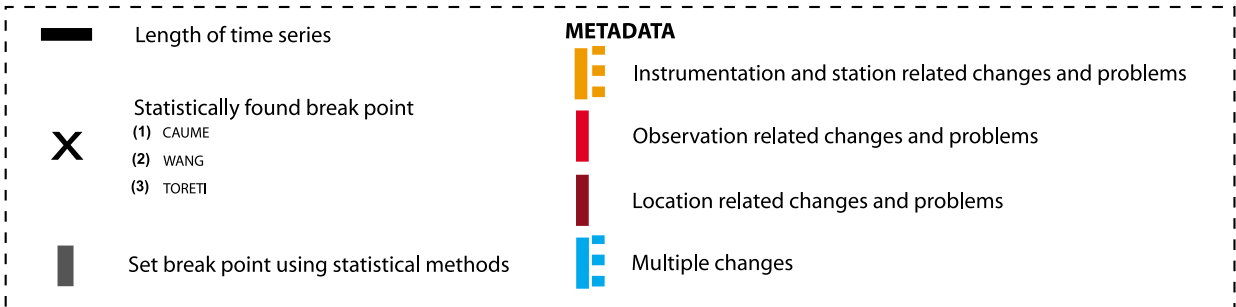
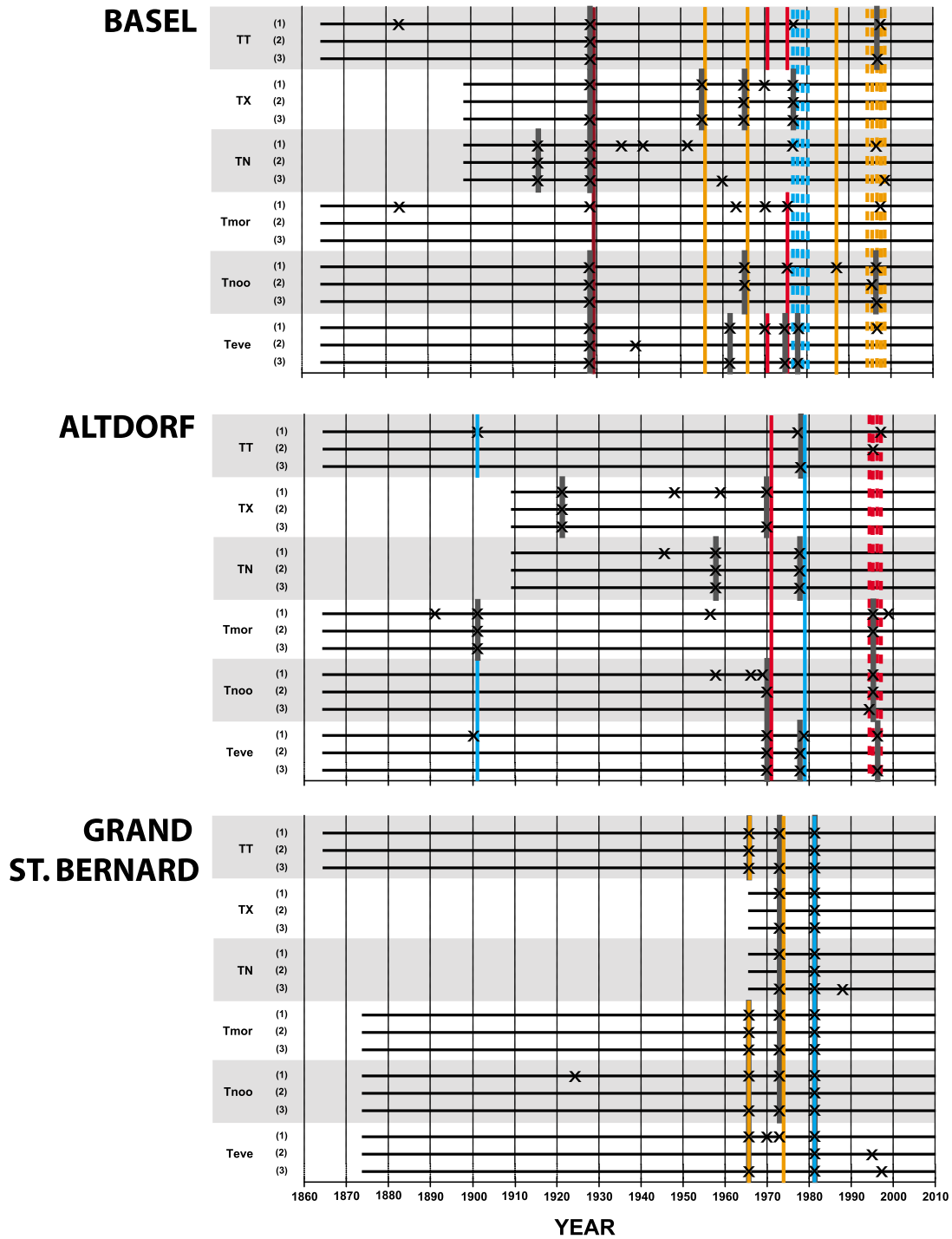
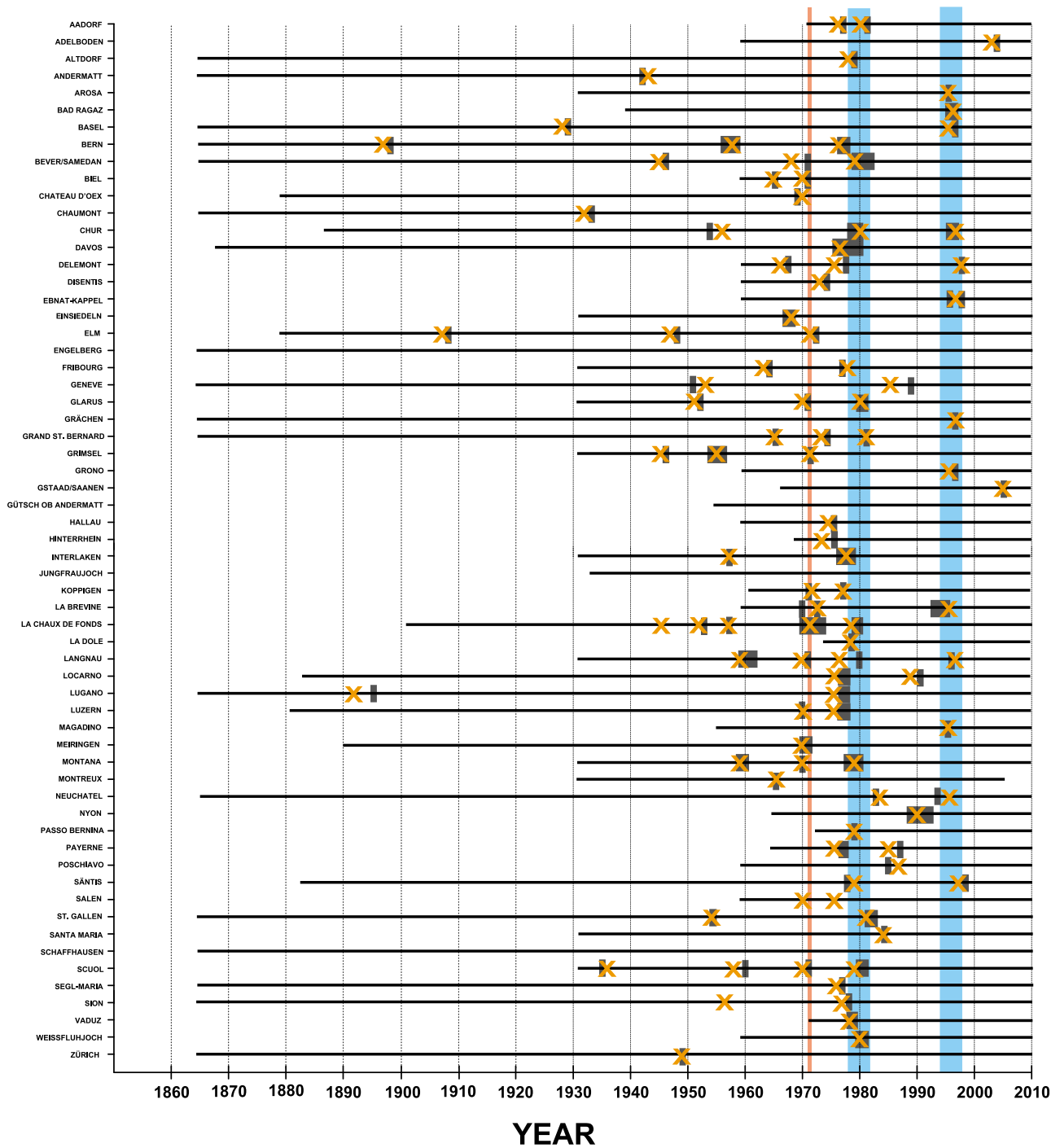


Figure 2



LENGTH OF TIME SERIES	METADATA
TT	using historical records
	All METADATA
SET BREAK POINT	1971: change in observation times
using statistical methods	1978-1982: introduction ANETZ
TT	1994-1998: numerous technical problems and errors

Figure 3

24 hourly values.¹ Therefore, detected breaks in automatic stations cannot be validated using Tmor, Tnoo, and Teve. All the data is publically available at IDAweb, MeteoSwiss, and can be downloaded under http://www.meteoschweiz.admin.ch/web/en/services/data_portal/idaweb.html. The TT, TX, and TN from 32 and 53 stations of the French and German network, respectively, are included in the analysis as reference stations (see section 2.2). These data stem from the European Climate Assessment & Data set (ECA&D (eca.knmi.nl) [Klein Tank et al., 2002]). The geographical distribution of the 61 Swiss stations analyzed, and the German and French reference stations, is presented in Figure 1.

[7] According to environmental conditions and regional climate effects, all stations can be categorized into: (1) JURA MOUNTAINS (4 stations, >1,000 m a.s.l.), (2) SWISS PLATEAU (21 stations, <1,000 m a.s.l.), (3) LOW ALPS (22 stations, <1,500 m a.s.l.), and (4) HIGH ALPS (15 stations, >1,500 m a.s.l.). This enables us to explore the effects of the environmental and regional conditions on the performance of the different break detection methods.

[8] All time series (besides Montreux which stops at 2005) cover the period from 1974 to 2010. Most of the time series go back to 1960, and many extend beyond 100 years. A summary and detailed overview of the time series lengths is provided in Table 1 and Figures 3–5, respectively.

2.1. Data Quality Control

[9] Before assessing homogeneity, the quality of the data must be analyzed in detail. We follow the method of Aguilar et al. [2003] to identify suspicious values (e.g., outliers) and to verify the consistency, tolerance, and temporal coherency of the time series. Missing values replace temperature values exceeding: (1) $\pm 50.0^\circ\text{C}$, (2) ± 4 standard deviations (s) of the full length of the respective station series, and/or (3) a difference of 25°C between consecutive observations. Furthermore, four or more equal consecutive values are removed and we verify that the number of days for each month is represented in the data. Finally, we consider a month to be complete when ≤ 3 days are missing. A year is considered complete when all months are present. A station series is considered complete when no more than three years are missing [Moberg and Jones, 2005].

[10] After applying these criteria to the whole data set, the most complete time period spanned from 1971 to 2000 (for 61 stations; showing a maximum of three missing years). Therefore, this 30-year climatological standard-normal period is the base period for the break detection and calculating correlations [cf. Kuglitsch et al., 2009, 2010].

2.2. Break Detection: Using Statistical Approaches

[11] Under ideal circumstances, statistical break detection should be applied to the mean annual data of multiple

“difference” time series ($Y-R_n$, where Y denotes a so-called *candidate* and R_n various *reference series*) with a highly correlated ($\rho > 0.8$) reference series [e.g., Aguilar et al., 2003; Toreti et al., 2010a]. Break points are more reliably detected in mean annual series because these are less noisy than monthly, daily or sub-daily series [Easterling and Peterson, 1995; Alexandersson and Moberg, 1997; Vincent, 1998; Caussinus and Mestre, 2004]. The use of highly correlated reference series helps to distinguish non-climatic break points from climatic break points [i.e., Kuglitsch et al., 2009; Toreti et al., 2010a]. However, a non-climatic network-wide break point (e.g., the simultaneous introduction of new instruments) can be difficult to detect because it exists in both Y and R_n . To overcome this problem, and to increase the probability of detecting break points affecting an entire network, series from independent meteorological networks (i.e., from different countries) should be included whenever possible.

[12] To detect an unknown number of artificial break points, we apply three methods to identify artificial changes in mean and/or variance: (1) *Caussinus and Mestre* [2004] (hereafter referred to as CAUME), (2) *Wang et al.* [2007], *Wang* [2008a, 2008b], and Wang and Feng (online publication, 2010) (hereafter referred to as WANG), and (3) *Toreti et al.* [2012] (hereafter referred to as TORETI). The method of WANG, also called RHtestsV3, is available online at <http://cccma.seos.uvic.ca/ETCCDMI/software.shtml>. The methods of CAUME and TORETI are programmed using R Statistical Computing. Each of these methods is applied to the mean annual TT, TX, TN, Tmor, Tnoo, and Teve difference series between a candidate station and its ten highest correlated ($\rho > 0.8$) neighboring stations from the Swiss, French and German (French and Swiss only TT, TX, TN) network (Figure 1). For example, for the candidate station “La Chaux-de-Fonds” (Jura, TX) we used the following reference stations: Adelboden 0.94, La Brevine 0.92, Chaumont 0.92, Hinterrhein 0.91, Davos 0.90, Engelberg 0.90, Disentis 0.90, La Dole 0.90, Château-d’Oex 0.89, Gstaad/Saanen 0.89. If one method detects the same break point (i.e., within two consecutive years) in three or more of the ten reference series, this break point is considered to be “true.” A break point year is assumed to be “valid” if either two or three methods show a break point within the same two consecutive years. The year with the greatest number of break points is defined as the year of an inhomogeneity [cf. Kuglitsch et al., 2009].

[13] We highlight the statistical results from Basel (316 m a.s.l., SWISS PLATEAU), Altdorf (440 m a.s.l., LOW ALPS) and Grand St. Bernard (2,472 m a.s.l., HIGH ALPS) and their relationship to their respective metadata. These three stations were selected because they: (1) have a relatively long station history and complete data coverage going back to 1864, (2) represent very different environmental and climatic conditions such as Foehn (Altdorf), a Central European low land situation (Basel), and a high elevation situation on the Alpine divide (Grand St. Bernhard), but at the same time, and (3) are cli-

¹Auxiliary material is available in the HTML. doi:10.1029/2012JD017729.

Figure 3. A summary of the detected break points for the Swiss TT series. The black lines correspond to the length of the daily TT series. The orange crosses indicate the years of detected break points. The gray, red, and blue rectangles indicate periods when metadata suggest major problems and changes in the weather station history.

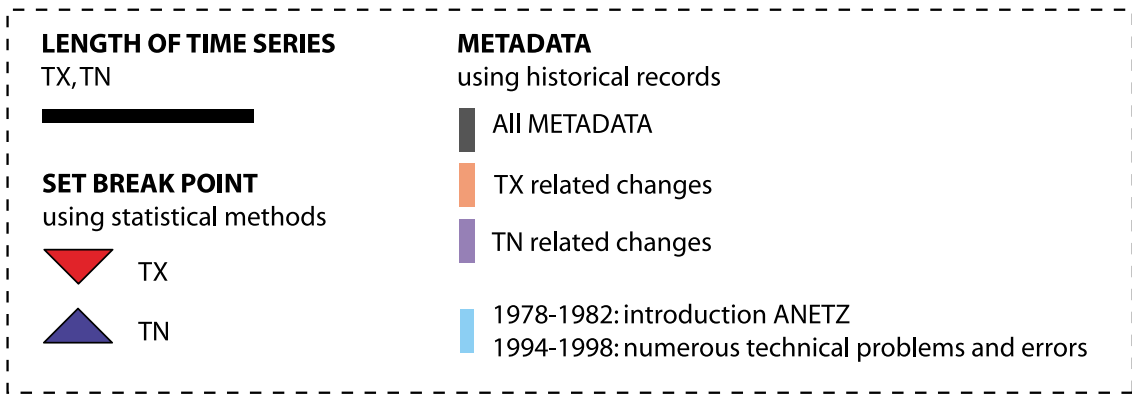
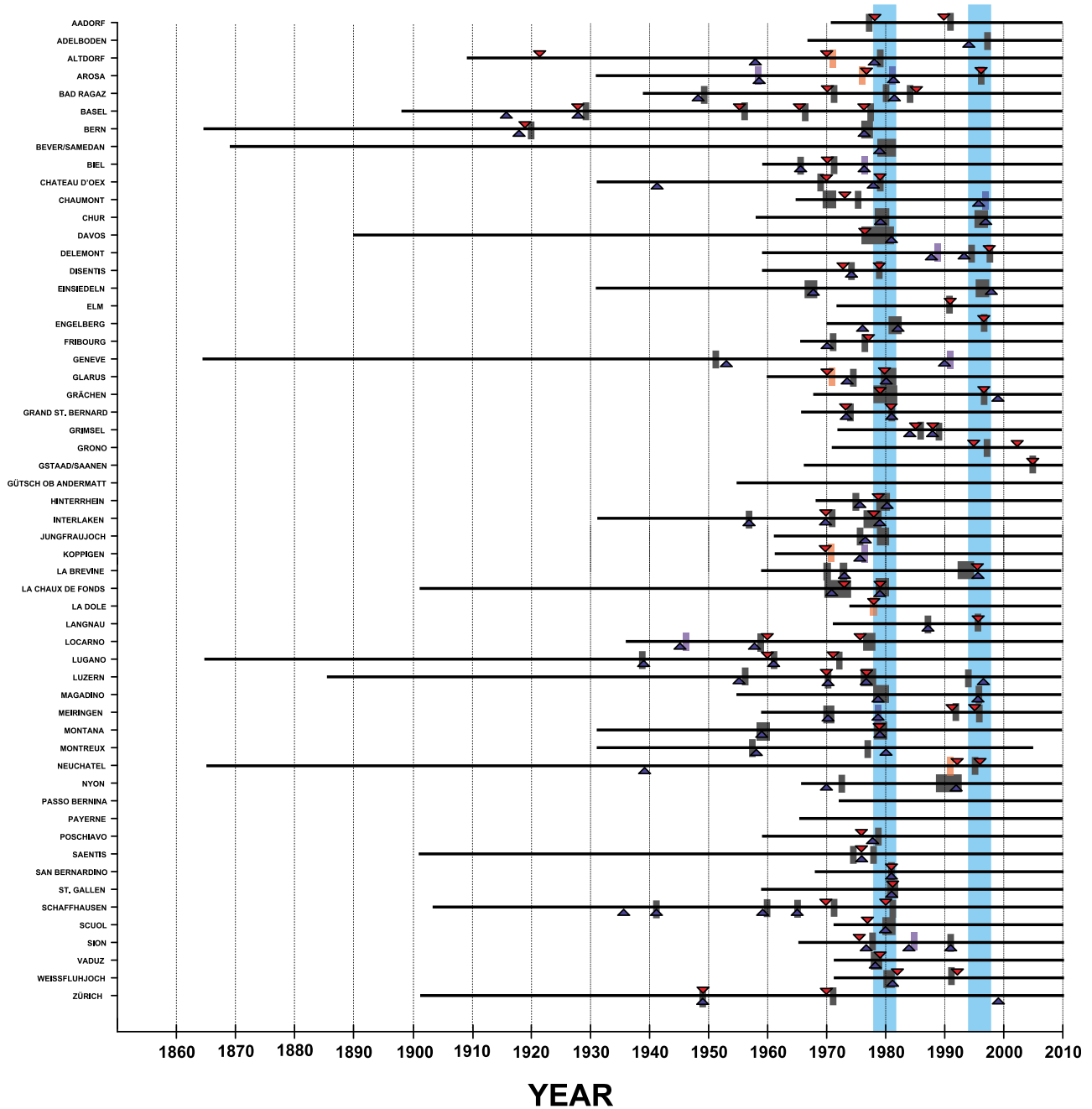


Figure 4

matologically representative for their region (as classified by Figure 1).

2.3. Break Detection: Using Metadata

[14] To test the reliability of the three statistical break detection methods, we compare their results with the metadata provided by MeteoSwiss. These metadata contain information about individual Swiss stations including details about: (1) instrumentation- and station-related changes and problems (e.g., changes in instruments, screens, ventilation, roofs, stairs, type of weather station, station decay, maintenance, renovation, reparations, and paintwork); (2) observation-related changes and problems (e.g., changes of the observer, changes of observational hours, failures and erroneous records, and new instructions to the observer), (3) location-related changes and problems (e.g., station relocation and land-use changes), (4) accidents (e.g., floods), and (5) a combination of the aforementioned causes.

2.4. Combination of Statistical and Metadata Approaches

[15] Identifying a break point using statistical methods can be risky. Statistical methods have different sensitivities to break points [e.g., *Toreti et al.*, 2012; *Venema et al.*, 2012], which introduces the risk of, for example, falsely identifying break points. To minimize this risk, we verify that each break point can be identified using at least two of the three independent statistical methods (which we assume to have comparable detection skills) listed in Section 2.2. This is a similar approach as used in multimodel climate projections [e.g., *Randall et al.*, 2007]. Consistently, in *Toreti et al.* [2012], a combination of statistical break detection methods (evaluated on a simulated data set and instrumental precipitation data) was employed. Then, in a further independent test, we compare these model results with the metadata provided by MeteoSwiss. Specifically, once a break point is deemed “valid” by statistical methods, we investigate whether a corresponding metadata entry (i.e., a discrete date) is found within ± 2 years of the statistically detected year of break point. In case the break point coincides with several (or continuous) dates of metadata entries, the statistically found break point is attributed to the respective year (i.e., a break point attributed to 1980 is between 1980 and 1981). If a statistically “valid” break point is not confirmed by metadata (a rarity; see section 3.1), the break is accepted and attributed to the respective year. Finally, where metadata indicate station changes or problems that are not detected statistically, it is assumed that the time series is free of inhomogeneities.

3. Results and Discussion

3.1. Comparison of Statistical Break Detection Methods

[16] The statistically confirmed (“valid”) break points for Basel, Altdorf, and Grand St. Bernard are presented in Figure 2 according to different parameters, statistical

methods, and metadata. The results for the entire Swiss station network are shown in summarized form in Figures 3–5.

[17] We find strong agreement among the “valid” break points (using the combined application of statistical methods) and metadata, emphasizing the reliability of the proposed approach. However, during the most recent decades, an increasing amount of metadata complicates the confirmation of the detected break points. In particular, the technical development of instruments and the automation of the weather stations often results in complex and time-consuming maintenance works, transmission errors, or other failures spanning over several years (e.g., Basel, 1976–1980, 1994–1999; Altdorf, 1994–1997). Therefore, the associated break points during these periods cannot be reliably detected by metadata and require a statistical evaluation.

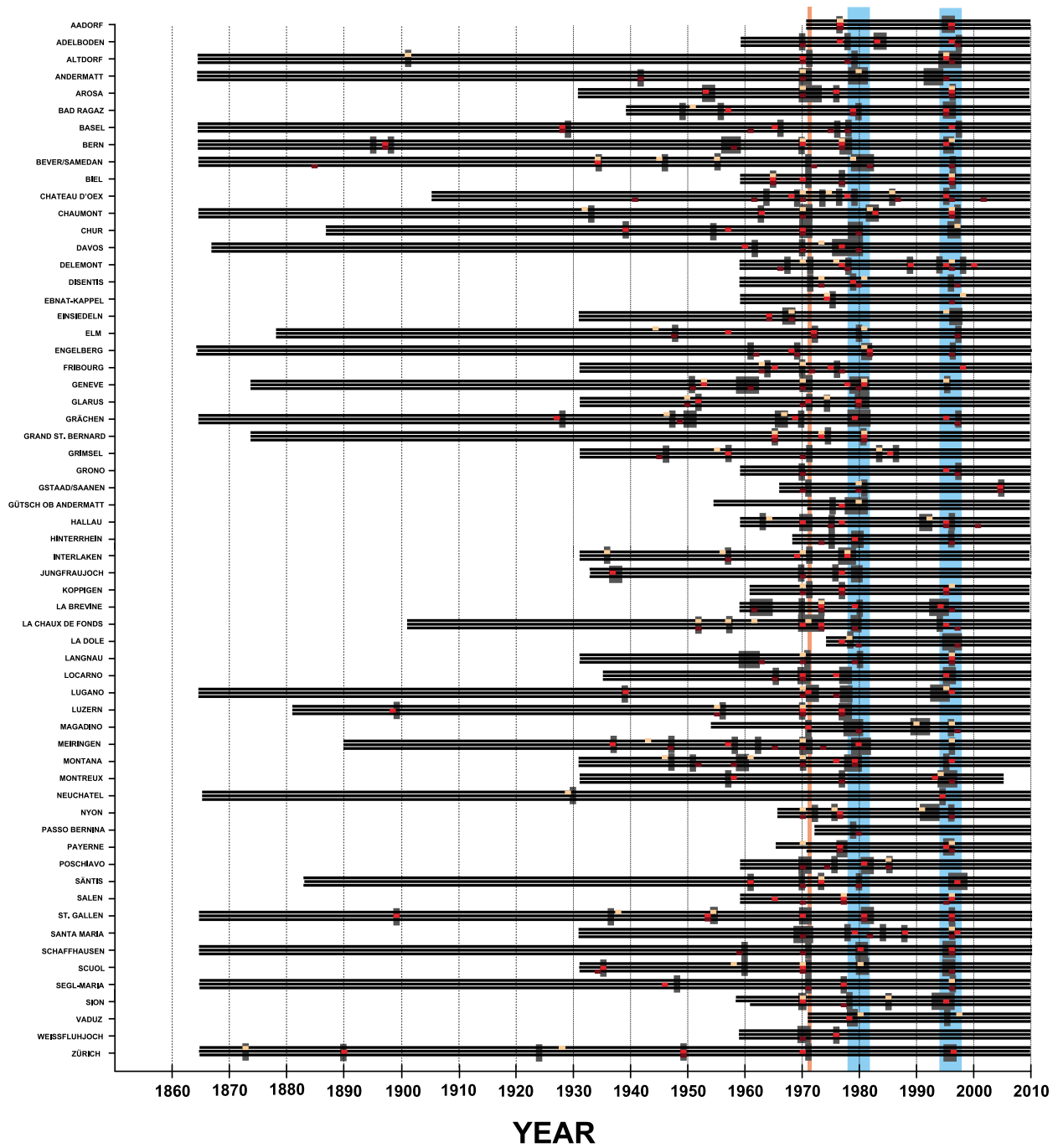
[18] A detailed analysis of the statistically detected break points and metadata identifies inhomogeneities that either: (1) affect only certain parameters, or (2) affect all or most of the parameters (e.g., station relocation in Basel, 1929; change from thermometer to thermograph in Grand St. Bernard, 1965; issues with thermograph in Grand St. Bernard, 1974; introduction of automatic station network ANETZ and station relocation in Grand St. Bernard, 1981). This demonstrates that statistically significant break points can be caused by different or even compound reasons.

[19] In contrast, Basel (Tmor) highlights the importance of combining different break detection methods. Depending on the method, Basel (Tmor) is either homogeneous (WANG, TORETI) or affected by six break points (CAUME). This suggests that the CAUME method tends to overestimate the number of break points and a statistical break detection using only CAUME is not sufficient. Moreover, Basel and Altdorf are affected by statistically significant break points that are not confirmed by metadata (Basel, TN, 1916; Basel, Teve, 1962; Altdorf, TX, 1921; Altdorf, TN, 1958). Local (e.g., spatial exposure of the crest situation at Grand St. Bernard) and often short-time climate effects (e.g., the Foehn in Altdorf) do not affect the reliability of statistical break detection (based on annual means of reference series correlating $\rho > 0.8$) because: (1) the station network in Switzerland is sufficiently dense that multiple stations (≥ 10) are often exposed to the same local and/or short-term climate effects and (2) the statistical break detection methods are based on mean annual temperature data.

[20] To test the potential of the three break detection methods, the break points detected by only one of the three statistical methods are analyzed and compared with the metadata for Basel, Altdorf and Grand St. Bernard (Table 2).

[21] For Basel, six years of break points are found by CAUME (TT, 1883; TN, 1936, 1941, 1952; Tmor, 1883, 1970; Tnoo, 1987). One single year of break point is found by WANG (Teve, 1939) and TORETI (TN, 1960). The metadata indicate that in 1987 the screen was painted and the ventilator was replaced. All other break points are not supported by metadata. This confirms that only break points

Figure 4. A summary of the detected break points for the Swiss TX and TN series. The black lines correspond to the length of the daily TX and TN series. The red and blue triangles indicate the years of detected break points for the TX and TN temperature series, respectively. The red, blue, and gray rectangles indicate periods when metadata suggest major problems and changes in the weather station history.



LENGTH OF TIME SERIES	METADATA
Tmor, Tnoo, Teve	using historical records
	All METADATA
SET BREAK POINT	1971: change in observation times
using statistical methods	1978-1982: introduction ANETZ
Tmor	1994-1998: numerous technical problems and errors
Tnoo	
Teve	

Figure 5

Table 1. A Summary of the Temporal Data Coverage for Each Temperature Parameter

	Temporal Data Coverage in Percent (Number) of Time Series	
	1960–2010	1910–2010
TT	85 (52)	43 (26)
TX	52 (29)	23 (13)
TN	52 (29)	23 (13)
Tmor	85 (52)	39 (24)
Tnoo	85 (52)	39 (24)
Teve	80 (49)	39 (24)

found by two or all three break detection methods are reliable.

[22] In the case of Altdorf, eight years of break points are found by CAUME (TX, 1948, 1959; TN, 1945; Tmor, 1891, 1957, 1999; Tnoo, 1958, 1966). The metadata show: (1) a station relocation in 1892 and (2) a new observer in 1956. For the other years, no metadata is available. WANG and TORETI do not find any single break points.

[23] At Grand St. Bernard, two years of break points are found by CAUME (Tnoo, 1924; Teve, 1970), two single years of break points are found by TORETI (TN, 1988; Teve, 1998), and one year is found by WANG (Teve, 1995). The metadata only confirm: (1) 1971 when the observational times in the evening were changed and (2) 1988 when the weather screen was destroyed by a storm.

[24] For these three locations, perfect consistency was found for TT and the three times daily data (i.e., there was no case where a break was statistically confirmed in TT but not in any of the three sub-daily time series ± 2 years). Overall, the application of three statistical break detection methods and comparison with metadata shows that the CAUME approach detects more breaks than WANG and TORETI, with WANG and TORETI most often in agreement. Break points which were found exclusively by one statistical method and supported by metadata (e.g., CAUME, Basel, Tnoo, 1987) were disregarded. This is because they were sufficiently small to avoid detection by the other statistical methods (WANG and TORETI).

3.2. Results for the Swiss Station Network

[25] When the entire Swiss network is considered, the combined application of the three statistical methods shows that the majority (89%) of the time series is affected by artificial break points. Depending on the temperature parameter, this value can vary from 79% (TX) to 97% (Tnoo), an average of one break point per 48 years (0.021 breaks per station year). However, we find fewer breaks occurring before 1970 (0.010 breaks per station year) and after 1982 (0.013 breaks per station year). Between 1970 and 1982, 68% of the time series are affected by break points

and of all break points, 48% are found between 1970 and 1982 (0.069 breaks per station year). Concurrently, only 23 TX, 23 TN, 23 TT, 24 Tmor, 11 Tnoo, and 8 Teve series, are deemed homogeneous. The high number of break points from 1970 to 1982 can be attributed to the change in the observation times (1st of January, 1971), the introduction of the automatic station network ANETZ, and the frequent instrument changes during the first years of operation [cf. *Begert et al.*, 2005]. Details are given in Table 3.

[26] Statistical break detection aims to find all break points which substantially impact a temperature time series and, therefore, need to be corrected to have a homogeneous temperature time series. Set break points, the break points identified by two or three statistical methods (and therefore needing correction to homogenize a time series), should agree with major changes or problems described in weather station metadata.

[27] The three methods (CAUME, WANG, and TORETI) are well-known for their potential to detect artificial break points in climate series. When all of the temperature parameters for all the stations are considered, CAUME overestimates the number of break points (1,140 break points with only 45% confirmed by metadata), WANG finds fewer break points but with improved accuracy (438 break points with 72% confirmed by metadata), and TORETI isolates 683 break points with 70% confirmed by metadata. When we combine the three methods, we find the best agreement between the station history and the statistics. Specifically, they reveal 649 break points which are “valid” (i.e., have been confirmed by either two or three statistical methods), of which 612 (or 94%) can be confirmed by metadata. A detailed overview of the total number of break points according to different temperature parameters, statistical methods, and relevant metadata is presented in Table 4.

[28] Among the 612 break points of all the temperature parameters which are supported by station metadata, 55.0% can be attributed to a combination of causes, 16.1% are related to changes in location and associated problems, 15.9% are due to changes in observation and related problems, 12.9% are associated with changes in instrumentation or station and accompanying problems, and 0.1% are caused by accidents. A summary of these break points and stations is given in the auxiliary material, Table S2.

[29] The use of statistics to detect break points is particularly valuable because metadata can have several limitations. First, not all of the changes and problems at a weather station can cause break points of a statistically significant size. Second, the volume of metadata spanning over several years makes it difficult to confirm the detected break points. Third, not all changes and problems are recorded in the metadata. Fourth, too many set break points can prevent a reliable correction step in homogenization. Specifically, correction studies based on highly correlated reference series have shown that too many break points (and associated short overlapping periods) render a reliable correction impossible

Figure 5. A summary of the detected break points for the Swiss Tmor, Tnoo, and Teve series. The black lines correspond to the length of the daily Tmor, Tnoo, and Teve series. The pink, red, and burgundy rectangles indicate the years of detected break points for the Tmor, Tnoo, and Teve, respectively. The gray, red, and blue rectangles indicate periods when metadata suggest major problems and changes in the weather station history.

Table 2. A Summary of the Break Points Detected by Only One of the Statistical Methods (1=CAUME, 2=WANG, 3=TORETI) and Related Metadata^a

	Basel			Aldorf			Grand St. Bernard			
	1	2	3	1	2	3	1	2	3	
TT	1883 (NRM)									
TX				1948 (NMA)						
TN	1936, 1941, 1952 (NRM)		1960 (NRM)	1959 (1956, NO)		1945 (NMA)	1988 (weather screen destroyed by storm)			
Tmor	1883, 1970 (NRM)			1891 (1892, station relocation)						
Tnoo	1987 (weather screen newly painted, ventilator replaced)			1957 (1956, NO)		1999 (NMA)	1924 (NRM)			
Teve	1939 (NRM)			1958 (1956, NO)		1966 (NMA)	1970 (1971, NOT)		1995 (NMA)	1998 (NMA)

^aNMA means no metadata available, NRM means no reference in metadata, NO means new observer, and NOT means new observation times.

[e.g., Della-Marta and Wanner, 2006; Toreti et al., 2010b]. However, statistical break detection methods also have strengths and weaknesses (e.g., certain methods might be more or less sensitive and therefore might over- or underestimate the number of breaks). For more details on the accuracy of statistical break detection methods, the reader is advised to consult, for example, Venema et al. [2012].

[30] A detailed overview of all detected break points and associated metadata is presented for the Swiss TT (Figure 3), TX, TN (Figure 4), Tmor, Tnoo, and Teve series in Figure 5. Detailed information about the specific metadata is given in the auxiliary material, Table S1.

[31] To test the consistency of statistically detected break points, results from TT are compared with results from Tmor, Tnoo and Teve. As TT at non-automatic stations is calculated from Tmor, Tnoo and Teve (cf. section 2 for details), breaks in TT must be confirmed by a simultaneous break in Tmor, Tnoo, and/or Teve. For the Swiss station network, problems in consistency have been found for 5 out of 102 set break points in TT (Bever/Samedan in 1968, Langnau in 1976, Lugano in 1892, Elm in 1907, and Montreux in 1965). According to the metadata, Bever/Samedan was affected by new evening observation times in 1971. This is consistent with a break point in Teve in 1972. Langnau was affected by a station relocation in 1980, which caused a break point in Teve in 1979. Lugano and Elm had instrumentation changes in 1895 and 1908. However, this did not cause breaks in Tmor, Tnoo, or Teve. Montreux changed observers in 1965.

4. Conclusion and Outlook

[32] Overall, the application of three statistical break detection methods to high-quality Swiss temperature series and the comparison with the metadata show that the CAUME approach detects more breaks than WANG and TORETI. However, in this study, CAUME also appears to overestimate the number of breaks. Alternatively, WANG and TORETI produce fewer breaks but are more accurate, as confirmed by metadata.

[33] On the other hand, by combining the three statistical approaches, we are able to detect the maximum number of breaks which can be confirmed by metadata. In most instances, these breaks are ascribed to a combination of problems (e.g., instrumentation, station, and observer). These findings confirm that a combined break detection method can provide a wealth of statistical information (even across complex terrain) where metadata are not readily available [cf. also Kuglitsch et al., 2010]. However, for inconsistent breaks (i.e., breaks detected in TT but not in one of the sub-daily series, when TT is based on Tmor, Tnoo, and Teve), metadata may be a valuable tool for interpreting a break point. In these cases, we suggest not to correct break points in TT. For all other cases, if a statistical break point is supported by major station changes ± 2 years in the station history, the set break point for the correction part should be the year confirmed by the station history. Where there is insufficient or too much metadata available, the suggested approach of combining different statistical methods should be effective. When the metadata indicate breaks that are not detected statistically, or are only indicated by one statistical method, the time series can be considered to be free of inhomogeneities.

Table 3. A Summary of the Time Series Affected by Break Points (BP) Categorized by Temperature Parameter

	Time Series Affected by BPs in Percent (Number)	1 BP per Period in Years	1970–1982	
			Time Series Affected by BPs in Percent (Number)	Amount of All BPs in Percent (Number)
TT	93 (57)	57	62 (38)	48 (48)
TX	79 (44)	56	59 (33)	60 (41)
TN	86 (48)	45	59 (33)	50 (42)
Tmor	85 (52)	51	61 (37)	44 (47)
Tnoo	97 (59)	41	82 (50)	44 (60)
Teve	95 (58)	36	87 (53)	50 (75)

Table 4. Detected Break Points According to the Temperature Parameter (listed in left column), the statistical method (CAUME, etc.)^a

Temperature Parameter	Total Number of Break Points			“Valid” Break Points
	CAUME	WANG	TORETI	
TT	165 (75; 45.5%)	65 (50; 76.9%)	101 (75; 74.3%)	102 (98; 96.1%)
TX	112 (56; 50.0%)	52 (37; 71.2%)	76 (54; 71.1%)	68 (66; 97.1%)
TN	154 (64; 41.6%)	54 (39; 72.2%)	94 (61; 64.9%)	84 (77; 91.7%)
Tmor	217 (85; 39.2%)	71 (47; 66.2%)	122 (79; 64.8%)	108 (100; 92.6%)
Tnoo	222 (112; 50.5%)	100 (77; 77.0%)	137 (99; 72.3%)	136 (130; 95.6%)
Teve	270 (123; 45.6%)	96 (67; 69.8%)	153 (112; 73.2%)	151 (141; 93.4%)
ALL	1,140 (515; 45.2%)	438 (317; 72.4%)	683 (480; 70.3%)	649 (612; 94.3%)

^aALL refers to the sum of break points found for all six temperature parameters. “Valid” indicates the set break points (i.e., those which have been confirmed by either two or three statistical methods). The number of break points is presented in bold and the parentheses contain the number and percent of break points confirmed by metadata.

[34] Break detection is the first step in homogenization. For the Swiss network temperature series (TT, Tmax, Tmin, Tmor, Tnoo, and Teve), the results of this first step are presented in this paper. In a next step the set break points will be corrected.

[35] **Acknowledgments.** We are grateful to M. Begert, M. Croci-Maspoli and C. Füllemann from MeteoSwiss and J. Luterbacher, A. Toreti and E. Xoplaki from Justus-Liebig-University Giessen for providing data and discussion. We acknowledge the European COST Action HOME ES0601 (Advances in homogenization methods of climate series: An integrated approach–HOME), the 6th EU Framework program CIRCE (212250) and the NCCR Climate PALVAREX project. The reviewers made useful comments and suggestions and helped to improve the quality of this study.

References

- Aguilar, E., I. Auer, M. Brunet, T. C. Peterson, and J. Wieringa (2003), Guidance on metadata and homogenization, *WMO-TD-No. 1186*, 51 pp., World Meteorol. Org., Geneva, Switzerland.
- Alexandersson, H., and A. Moberg (1997), Homogenization of Swedish temperature data. Part 1: Homogeneity test for linear trends, *Int. J. Climatol.*, *17*, 25–34, doi:10.1002/(SICI)1097-0088(199701)17:1<25::AID-JOC103>3.0.CO;2-J.
- Baudenbacher, M. (1997), *Homogenisierung Langer Klimareihen Dargelegt am Beispiel der Lufttemperatur, Veröffentlichung der Schweiz. Meteorol. Anstalt*, vol. 58, Schweiz. Meteorol. Anstalt, Zürich, Switzerland.
- Begert, M., G. Seiz, T. Schlegel, M. Musa, G. Baudraz, and M. Moesch (2003), *Homogenisierung von Klimamessreihen der Schweiz und Bestimmung der Normwerte 1961–1990: Schlussbericht des Projektes NORM90*, 170 pp., Meteo Swiss, Zurich, Switzerland.
- Begert, M., T. Schlegel, and W. Kirchhofer (2005), Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000, *Int. J. Climatol.*, *25*, 65–80, doi:10.1002/joc.1118.
- Brandsma, T., and G. P. Können (2006), Application of nearest-neighbor resampling for homogenizing temperature records on a daily to sub-daily level, *Int. J. Climatol.*, *26*, 75–89, doi:10.1002/joc.1236.
- Caussinus, H., and F. Lyazrhi (1997), Choosing a linear model with a random number of change-points and outliers, *Ann. Inst. Stat. Math.*, *49*, 761–775, doi:10.1023/A:1003230713770.
- Caussinus, H., and O. Mestre (2004), Detection and correction of artificial shifts in climate series, *J. R. Stat. Soc., Ser. C*, *53*, 405–425, doi:10.1111/j.1467-9876.2004.05155.x.
- Cocheo, C., and D. Camuffo (2002), Corrections of systematic errors and data homogenisation in the daily temperature Padova series (1725–1998), *Clim. Change*, *53*, 77–100, doi:10.1023/A:1014950306015.
- Della-Marta, P. M., and H. Wanner (2006), A method of homogenizing the extremes and mean of daily temperature measurements, *J. Clim.*, *19*, 4179–4197, doi:10.1175/JCLI3855.1.
- Della-Marta, P. M., M. R. Haylock, J. Luterbacher, and H. Wanner (2007), Doubled length of western European summer heat waves since 1880, *J. Geophys. Res.*, *112*, D15103, doi:10.1029/2007JD008510.
- Easterling, D. R., and T. C. Peterson (1995), A new method for detecting undocumented discontinuities in climatological time series, *Int. J. Climatol.*, *15*, 369–377, doi:10.1002/joc.3370150403.
- Füllemann, C., M. Begert, M. Croci-Maspoli, and S. Brönnimann (2011), Digitalisieren und Homogenisieren von historischen Klimadaten des Swiss NBCN – Resultate aus DigiHom, *Arbeitsber. MeteoSchweiz* *236*, 48 pp., Meteo Swiss, Zurich, Switzerland.
- Giorgi, F., and L. O. Mearns (1991), Approaches to the simulation of regional climate change: A review, *Rev. Geophys.*, *29*, 191–216, doi:10.1029/90RG02636.
- González-Rouco, J. F., J. L. Jimenez, V. Quesada, and F. Valero (2001), Quality control and homogeneity of precipitation data in the southwest of Europe, *J. Clim.*, *14*, 964–978, doi:10.1175/1520-0442(2001)014<0964:QCAHOP>2.0.CO;2.
- Klein Tank, A., et al. (2002), Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment, *Int. J. Climatol.*, *22*, 1441–1453, doi:10.1002/joc.773.
- Kuglitsch, F. G., A. Toreti, E. Xoplaki, P. M. Della-Marta, J. Luterbacher, and H. Wanner (2009), Homogenization of daily maximum temperature series in the Mediterranean, *J. Geophys. Res.*, *114*, D15108, doi:10.1029/2008JD011606.
- Kuglitsch, F. G., A. Toreti, E. Xoplaki, P. M. Della-Marta, C. S. Zerefos, M. Türkeş, and J. Luterbacher (2010), Heat wave changes in the eastern Mediterranean since 1960, *Geophys. Res. Lett.*, *37*, L04802, doi:10.1029/2009GL041841.
- Maugeri, M., L. Buffoni, B. Delmonte, and A. Fassina (2002), Daily Milan temperature and pressure series (1763–1998): Completing and homogenising the data, *Clim. Change*, *53*, 119–149, doi:10.1023/A:1014923027396.
- Moberg, A., and P. D. Jones (2005), Trends in indices for extremes in daily temperature and precipitation in central and western Europe, 1901–99, *Int. J. Climatol.*, *25*, 1149–1171, doi:10.1002/joc.1163.
- Peterson, T. C., et al. (1998), Homogeneity adjustments of in situ atmospheric climate data: A review, *Int. J. Climatol.*, *18*, 1493–1517, doi:10.1002/(SICI)1097-0088(199811)18:13<1493::AID-JOC329>3.0.CO;2-T.
- Randall, D. A., et al. (2007), Climate models and their evaluation, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 589–662, Cambridge Univ. Press, Cambridge, U. K.
- Stewart, M. M., I. Larocque-Tobler, and M. Grosjean (2011), Quantitative inter-annual and decadal June–July–August temperature variability ca. 570 BC to AD 120 (Iron Age–Roman Period) reconstructed from the varved sediments of Lake Silvaplana, Switzerland, *J. Quat. Sci.*, *26*, 491–501, doi:10.1002/jqs.1480.
- Stott, P. A., and C. E. Forest (2007), Review. Ensemble climate predictions using climate models and observational constraints, *Philos. Trans. R. Soc. A*, *365*, 2029–2052, doi:10.1098/rsta.2007.2075.
- Szentimrey, T. (1999), Multiple analysis of series for homogenisation (MASH), in *Proceedings of the Second Seminar for Homogenisation of Surface Climatological Data, WMO-TD-No. 962*, pp. 27–46, World Meteorol. Org., Geneva, Switzerland.
- Toreti, A., F. G. Kuglitsch, E. Xoplaki, P. M. Della-Marta, E. Aguilar, M. Prohom, and J. Luterbacher (2010a), A note on the use of the standard normal homogeneity test to detect inhomogeneities in climatic time series, *Int. J. Climatol.*, *31*, 630–632, doi:10.1002/joc.2088.
- Toreti, A., F. G. Kuglitsch, E. Xoplaki, J. Luterbacher, and H. Wanner (2010b), A novel method for the homogenization of daily temperature series and its relevance for climate change analysis, *J. Clim.*, *23*, 5325–5331, doi:10.1175/2010JCLI3499.1.
- 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. Climatol.*, *51*, 317–326, doi:10.1175/JAMC-D-10-05033.1.
- Trenberth, K. E., et al. (2007), Observations: Surface and atmospheric climate change, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the*

- Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 235–336, Cambridge Univ. Press, Cambridge, U. K.
- Venema, V. K. C., et al. (2012), Benchmarking homogenization algorithms for monthly data, *Clim. Past*, 8, 89–115, doi:10.5194/cp-8-89-2012.
- Vincent, L. A. (1998), A technique for the identification of inhomogeneities in Canadian temperature series, *J. Clim.*, 11, 1094–1104, doi:10.1175/1520-0442(1998)011<1094:ATFTIO>2.0.CO;2.
- Vincent, L. A., X. Zhang, B. R. Bonsal, and W. D. Hogg (2002), Homogenization of daily temperatures over Canada, *J. Clim.*, 15, 1322–1334, doi:10.1175/1520-0442(2002)015<1322:HODTOC>2.0.CO;2.
- Wang, X. L. (2008a), Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or F test, *J. Appl. Meteorol. Climatol.*, 47, 2423–2444, doi:10.1175/2008JAMC1741.1.
- Wang, X. L. (2008b), Penalized maximal F test for detecting undocumented mean shift without trend change, *J. Atmos. Oceanic Technol.*, 25, 368–384, doi:10.1175/2007JTECHA982.1.
- Wang, X. L., Q. H. Wen, and Y. Wu (2007), Penalized maximal t test for detecting undocumented mean change in climate data series, *J. Appl. Meteorol. Climatol.*, 46, 916–931, doi:10.1175/JAM2504.1.