Guidelines on Homogenization

2020 edition

WEATHER CLIMATE WATER

WORLD
METEOROLOGICAL
ORGANIZATION

WMO-No. 1245

Guidelines on Homogenization

2020 edition

WORLD
METEOROLOGICAL
ORGANIZATION

WMO-No. 1245

EDITORIAL NOTE

METEOTERM, the WMO terminology database, may be consulted at https://public.wmo.int/en/meteoterm.

Readers who copy hyperlinks by selecting them in the text should be aware that additional spaces may appear immediately following http://, https://, ftp://, mailto:, and after slashes (/), dashes (-), periods (.) and unbroken sequences of characters (letters and numbers). These spaces should be removed from the pasted URL. The correct URL is displayed when hovering over the link or when clicking on the link and then copying it from the browser.

WMO-No. 1245

Chair, Publications Board
World Meteorological Organization (WMO)
7 bis, avenue de la Paix
P.O. Box 2300
CH-1211 Geneva 2, Switzerland

Tel.: +41 (0) 22 730 84 03
Fax: +41 (0) 22 730 81 17
Email: publications@wmo.int

ISBN 978-92-63-11245-3

NOTE

The designations employed in WMO publications and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of WMO concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The mention of specific companies or products does not imply that they are endorsed or recommended by WMO in preference to others of a similar nature which are not mentioned or advertised.

CONTENTS

Page

Page

ACKNOWLEDGEMENTS

This is an updated version of *Guidelines on Climate Metadata and Homogenization* (WMO/TD No. 1186). It was prepared by the Task Team on Homogenization of the Commission for Climatology.

We thank the following people for their outstanding contributions to this publication:

Victor Venema (lead)	Meteorological Institute, University of Bonn, Germany
Blair Trewin	Bureau of Meteorology, Melbourne, Australia
Xiaolan L. Wang	Climate Research Division, Science and Technology Branch, Environment and Climate Change Canada, Toronto, Ontario, Canada
Tamás Szentimrey	Varimax Limited Partnership, Budapest, Hungary
Monika Lakatos	Hungarian Meteorological Service, Climate Division, Hungary
Enric Aguilar	Rovira i Virgili University, Tarragona, Spain
Ingeborg Auer	Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Vienna, Austria
Jose A. Guijarro	State Meteorological Agency (AEMET), Spain
Matthew Menne	National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Information, Dataset Section, Asheville, North Carolina, United States
Clara Oria	Servicio Nacional de Meteorología e Hidrología (SENAMHI) del Perú, Peru
Wilfrid S. R. L. Louamba	Direction de la météorologie, Brazzaville, Congo
Ghulam Rasul	Pakistan Meteorological Department, Islamabad, Pakistan

Further contributions have been made by:

Athanassios Argiriou	Greece
Peer Hechler	WMO
Gregor Vertačnik	Slovenia
Yizhak Yosef	Israel

INTRODUCTION

High-quality, homogeneous time series data are essential to analyse climate variability and change.

Homogenization aims to make data “homogeneous”. This word derives from the ancient Greek and means “of the same nature”. Unfortunately, most long-term raw climate time series do not fulfil this principle and are internally inhomogeneous, that is not of the same nature, and are therefore unsuitable for statistical climate change analysis. Non-climatic factors such as change in the circumstances of the measurements can substantially affect measured values and cause distortions in the statistical behaviour of the time series. The impact of these distortions can be comparable to that of climate change and may lead to erroneous conclusions.

Each data value arriving at the computer of a user is the result of a series of consecutive procedures: a set of instruments is deployed in a location, the value is measured, recorded, undergoes transformations and is included in a database, then it is transmitted to the intermediate and final users. With regard to long-term time series data, some of these procedures may have been altered over the years. In a temperature time series starting in the late nineteenth century, the thermometer has most likely been replaced several times or even substituted with an electronic sensor in recent decades. The shelter has probably changed from an open stand to a Stevenson screen and then possibly to a multi-plate screen if the station has been automated. Around the middle of the twentieth century, many observation stations were moved to airports to service the growing demand for civil aviation. If the station remained at the same location, the surroundings have probably been altered. For example, if the station was deployed in the outskirts of a village 100 years ago, today, it might be surrounded by buildings. Perhaps due to land use changes or other practical reasons, the station has been relocated to a more convenient place in the suburban area. If we combined the observations made in such different periods, they would obviously not be comparable. It is rare, although not unknown, for a climate data record of 100 years or more to be truly homogeneous. For example, in the Australian ACORN-SAT temperature dataset, only 2 out of 112 stations (both starting in the 1940s) were found to be homogeneous for both maximum and minimum temperatures, and in Europe, the period between two detected breaks in temperature observations is thought to be, on average, about 20 years.

The situation described in the previous paragraph implies that we will not be able to make solid inferences about the temporal evolution (for example, compute reliable trends) of climate time series without ensuring that all the observations and the derived time series are comparable. No climate time series should be used without homogenization testing and adjustment, where appropriate, and all National Meteorological and Hydrological Services (NMHSs) and climate data providers that create and deliver climate datasets should routinely conduct homogenization.

There are two fundamental types of homogenization: homogenization of the annual, seasonal or monthly *means*, and homogenization of the *distribution* that also adjusts the variability around the mean as well as higher order statistics of the data. The focus of this guidance will be on the homogenization of the means.

Climatic datasets of any kind typically contain inhomogeneities. This guidance will be limited, however, to the homogenization of instrumental land station data.

Many lessons will be applicable to the homogenization of radiosonde data (Jovanovic et al., 2017, Haimberger et al., 2012) and other types of data. For more information on the homogenization of marine in-situ data, see Kent et al. (2016), Kennedy et al. (2011) and Huang et al. (2015). Some papers on biases and homogenization of satellite data are by Schröder et al. (2016) and Brogniez et al. (2016).

Chapters 1–3 of these guidelines aim at getting new people started with homogenization, and Chapters 4 and 5 discuss more advanced and background topics intended for advanced users and developers of homogenization methods. Parts of this publication that contain more detailed explanations are printed in a smaller font.

At the time of writing, these guidelines are accompanied by a Frequently Asked Questions (FAQs) page, which is available at https://homogenisation.grassroots.is/. These questions can be of a more practical or more transient nature (for example, software bugs and solution for them).

CHAPTER 1. PREREQUISITES

Homogenization is one step in the processing of climate data (see Figure 1). The preceding steps, such as data rescue and quality control, affect the quality of homogenization.

Data rescue is particularly important for data sparse regions and periods, including those cases where we do have data but the station density is not sufficient to reveal important data problems. How well the data could be homogenized should be taken into account in the subsequent climate data analysis.

Before homogenizing a dataset, it is important to know how the variable was measured historically throughout the observing network and what happened with the stations. In the course of the homogenization process, awareness of the amount of missing data is essential. If the volume of missing data crosses a threshold over a period of time, some homogenization approaches may not work as expected. Metadata is also important during homogenization, to validate the results of statistical homogenization and to document what happened in complicated situations.

The final section of this chapter highlights the importance of training and identifies selected recent scientific meetings on homogenization.

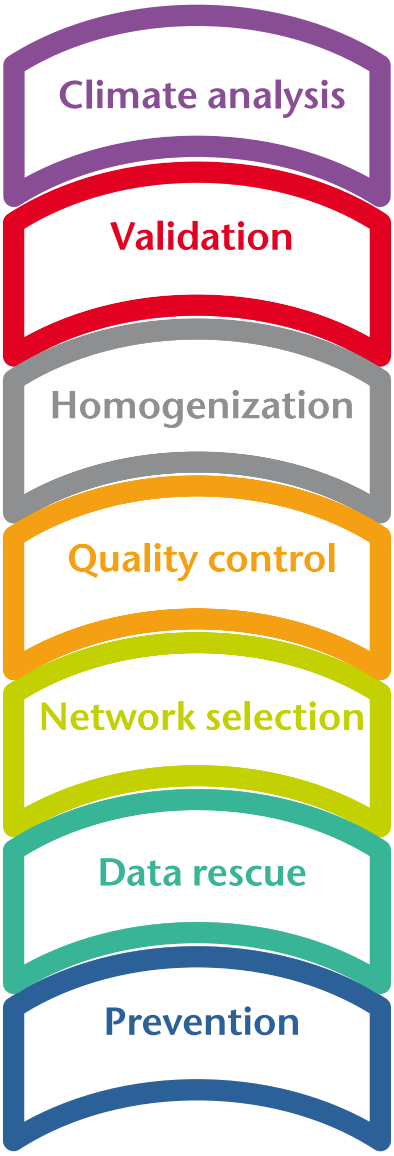

Figure 1. Processing of climate data

1.1 **Engagement with observing station network managers**

The task of developing and maintaining high-quality time series datasets is simpler when there are fewer inhomogeneities that need to be considered. While nothing can be done to prevent inhomogeneities that have already occurred, there is considerable value in managing an observing station network in such a way as to reduce the number of inhomogeneities in current and future data, and/or to facilitate accurate quantification of any inhomogeneities that may occur. This is something that requires engagement between climatologists and station network managers. It can be challenging, especially if different organizations are responsible for managing the observation network and for climate data.

Practices that can help minimize ongoing inhomogeneities include:

– Ensuring that proper field trials are carried out for any new observation system before they are implemented in the full station network. These field trials should preferably be carried out for at least two years (a single year can produce misleading results if that year is climatically unusual) and, especially in large countries, they should take place in climates representative of the full range of climates found within those countries;

– Ensuring that parallel observations take place where any significant change occurs; that such observations start early enough to provide at least two years of overlapping data and that these data are archived and shared (see section 1.2 below);

– Identifying sites that are at risk of closure or significant changes in their environment (for example, through planned building works nearby) at an early stage, to maximize the time available for parallel observations with any new site;

– Where new sites for observing stations are being chosen to replace a current site, the new site should match the topography and local environment of the current site as closely as possible, to mitigate the risk of a climatically significant change;

– If possible, select sites where the surrounding is unlikely to change in the coming decade(s) to centuries;

– Also important is the ongoing collection of metadata. Station identifiers should not be reused, as this can create confusion and may lead to data being merged from two completely unrelated stations. The reuse of station identifiers is a problem particularly for international datasets.

A useful WMO report, *Protocol Measurement Infrastructure Changes* (Brandsma et al., 2019), contains an example of protocol for the management of changes in climate networks.

1.2 **Parallel measurements**

A recommended practice for major changes in observations is to conduct parallel observations between two systems. WMO (2018*a*) suggests that "observations from new instruments should be compared over an extended interval (at least one year; see the *Guide to Climatological Practices* (…) before the old measurement system is taken out of service". The updated *Guide to Climatological Practices* (WMO, 2018*b*) recommends: "Where feasible and practical, both the old and new observing stations and instrumentation should be operated for an overlapping period of at least one year, and preferably two or more years, to determine the effects of changed instruments or sites on the climatological data." This suggested range of overlapping periods reflects the fact that longer is better, but shorter parallel measurements should not be discouraged by specifying long periods, because they are better than none.

These recommended practices are not universally followed (adherence was less common in the past than it is now), and parallel observations are not an option where an inhomogeneity arises because of an unexpected change around the station (for example, in the site environment) rather than at the station itself. Adherence can be improved by making the management of changes part of the operational practice (see section 1.1).

The specific importance of parallel observations for the homogenization process is highlighted in section 2.4.

1.3 Data rescue

Data rescue is the ongoing process of identifying and preserving all data and related metadata, records and climate archives that are at risk of being lost, and of digitizing current and past data into computer-compatible form for easy access. The identification process also entails searching for data that may be held in non-NMHS repositories such as universities, libraries and national archives. In some cases, historical data may be held overseas. Data rescue includes also migration from obsolete or computer corrupted media to modern media and readable formats.

Data rescue plays two important roles in the development of homogenized climate datasets. The first and most obvious is that a dataset will not be homogenized and analysed unless it is in digital form. Less obviously, homogenization is most effective when a candidate station can be compared with a large number of reference stations in the region, something which requires digitized data from those reference stations as well as the candidate station. Reference stations are also useful in quality control, as are sub-daily (for example, hourly) observations at the candidate station.

Chapter 2 expounds the importance of the signal-to-noise ratio (SNR) for homogenization. When the SNR is understood as the variance of the break signal divided by the variance of the noise, it is important that the SNR is above one. Until this level is reached, further data rescue for the period and region should be prioritized, provided further non-digitized data exist. Digitizing clusters close to a candidate station of interest are preferable to a uniform sampling when it comes to data quality control and homogenization. The digitization of short series can also be worthwhile where they can help quality control and homogenization.

A detailed discussion of the practical aspects of data rescue is outside the scope of this publication. For further information, see WMO (2016).

1.4 Quality control

Quality control aims to verify that a reported data value is representative of what was intended to be measured and that it has not been contaminated by unrelated factors (WMO, 2018*b*).

Quality control should be performed before homogenization because large outliers can affect the homogenization process. Sometimes quality control is performed again after homogenization because the higher data quality allows the detection of more subtle erroneous values. Moreover, if the (recent) data was subjected to real-time quality control, an additional quality control of the time-series data must be made in order to achieve the most uniform possible data quality across the full period of record.

Quality control can also be a source of inhomogeneities, especially in daily data when these are analysed for changes in weather variability and extremes. The methods used for quality control of data have often changed over time and this can introduce inhomogeneities if, for example, erroneous data were not detected in the past, but now are being flagged. In many cases, older historical data underwent very limited quality control; in particular, methods that involve spatial intercomparison of data with other stations have only become practical since the introduction of modern computer systems. More recently, the level of manual intervention in quality control is being reduced in many countries. Whereas in the recent past, a data point flagged by an automatic system might be subject to manual review, now more and more quality control is purely automatic, sometimes leading to a greater risk of “false positives” where valid extremes are flagged as suspect.

Derived time series can have quality control problems that were not obvious in single observations. It is desirable to carry out quality control at various timescales to detect the full range of possible error modes.

Discussion of specific quality control tests and methods is outside the scope of this publication. For a fuller discussion of quality control, readers are referred to WMO (1993) and WMO (1986); guidance material was being updated at the time of writing this publication. For a good example of quality control applicable to a global dataset, see Dunn et al. (2012).

1.5 Station history and metadata

The following sub-sections describe the situation of metadata in the past as relevant to current homogenization activities. Modern metadata standards and station identifiers have been defined by the WMO Integrated Global Observing System (WIGOS) (see WMO, 2019).

1.5.1 *What are metadata?*

In the context of homogenization, the term metadata is used to refer to what in some other contexts is called station metadata. Especially important for homogenization is the station history. This includes information about the observation site and instruments, as well as observation and data-processing procedures throughout the station history. Station metadata may include, but is not limited to:

- The location and elevation of a station and dates of changes therein;
- The types and conditions of instruments used at a location;
- Dates of replacements, for example, of instruments and screens;
- Land use and vegetation type in the vicinity of the station;
- Changes in the surrounding area (siting);
- The standard times and frequency at which observations are made;
- The name of the observer(s) (for manual stations);
- Procedures for processing data (for example, the definition of daily mean temperature);
- Dates and results of calibrations or tolerance checks carried out at the station;
- Details of maintenance (scheduled or unscheduled) carried out at the station.

1.5.2 *The value of metadata and their limitations*

Metadata, where they exist, are extremely valuable for data homogenization. While statistical methods for detecting inhomogeneities can provide strong evidence that an inhomogeneity can occur, metadata can indicate the cause of an inhomogeneity and even determine the date (if recorded) of its occurrence with high precision, whereas statistical methods can determine the timing only with limited precision (typically within a few months to a year). Metadata can also provide evidence of inhomogeneities in situations where statistical methods have limited or no effectiveness – for example, where there are no nearby reference stations, or where a change affects a large part of the network at the same time.

Homogenization is normally most effective when it uses a combination of metadata and statistical methods.

There is no consensus on whether metadata should influence the decision to set or not to set a break found by statistical homogenization. Doing so is potentially dangerous because not all breaks are equally well documented. The best-known example would be an urban station where the gradual urbanization is typically not well documented, unlike relocations. Removing only the documented relocations would theoretically make the trend errors larger. On the other hand, with several breaks in the candidate and reference series, homogenization can be a difficult combinatorial problem. In such cases, known likely breaks can help solve this puzzle.

Metadata are most valuable when they are complete. Many National Meteorological Services have gathered good metadata in recent years, but their availability often becomes sparser as one goes further back in history. Moreover, older metadata are more likely to be on paper, which means that they can be difficult to locate or use. As most stations, particularly prior to the 1990s, were installed to support weather forecasting rather than climate applications, recording metadata that was not relevant to a weather forecasting function was often not a high priority. Also, metadata reports sometimes provide a snapshot in time and do not indicate the date of a change (for example, two station inspection reports five years apart may show the station in different places, but the exact date of the move may not be documented).

Some aspects of metadata are also better documented than others. Experience shows that new instruments are often well documented whereas documentation of the site environment, especially outside the immediate vicinity of the instrument enclosure, is often limited or non-existent.

Metadata can take the form of a single point of data (for example, a set of station coordinates) or a more complex piece of information (for example, a photo of the station which provides information on the land surface type and surrounding obstructions). Even the simpler forms of metadata have some level of uncertainty attached. Until recently, station coordinates were rarely recorded with sufficient accuracy to resolve small site moves (in the tens of metres), but such moves may still be climatically significant, especially in complex topography or where they affect the site's exposure to wind. In exposed coastal locations, for example, site moves of less than 50 m have been found to have a 15%–25% impact on observed precipitation. In older datasets, or those exchanged internationally, station coordinates are often only reported to the nearest minute or 0.01°, which only resolves the location of the station to within about 2 km.

Metadata may require considerable interpretation, where multiple pieces of metadata information are used to reach a conclusion. Separating climate relevant from non-relevant metadata documents can be a time-consuming process.

As with data, metadata can sometimes also be erroneous, so it is useful to have multiple lines of evidence for an inhomogeneity where possible.

1.5.3 *Station identifiers*

Determining exactly which station a dataset is associated with is an important part of developing long-term homogeneous datasets. In most countries, a station will have a domestic station identifier, with stations that report internationally also having a WMO station identifier. It may happen that two or more stations share the same WMO station identifier. There may also be other identifiers in use, such as a station name, or an International Civil Aviation Organization (ICAO) code for an airport site.

It is good practice to associate a station identifier with a station and, if that station undergoes a climatically significant change (for example, a significant site move), to cease recording data under that identifier and create a new one for the new location. This maximizes the visibility of the change, while still allowing a long-term dataset to be created by using a composite of multiple identifiers (multiple identifiers are, of course, a necessity if parallel observations are taking place).

It has been, however, common practice to retain a single station identifier even through a significant change. This is particularly true for historical data when awareness of the climatic impact of site changes was lower than it is now.

Note that the WMO station identifier is not a unique station identifier in some national archives. The WMO station identifier was initially for weather forecast purposes and its number of digits is limited. Two or more stations can thus share the same WMO station identifier. In addition, a new station sometimes reuses the identifier of a closed station; there are also cases where national identifiers have been reused.

International datasets pose particular challenges in identifying stations. Typically, such datasets will contain data from multiple sources which may be partial duplicates of each other, with metadata often limited to station coordinates (possibly of limited precision) and a station name (which may have multiple possible spellings, especially when it contains non ASCII characters or if the original name is in a language which does not use the Roman alphabet). Rennie et al. (2014) detail the procedures used to merge data from multiple sources in one major international dataset, the International Surface Temperature Initiative (ISTI) databank.

1.5.4 *Formats and accessibility of metadata*

Metadata can exist in a wide variety of forms. In many National Meteorological Services, some or all recent metadata is held in digital form within a searchable database, but large amounts of historical metadata normally exist only on paper.

Metadata may exist in forms that are specific to an individual station, or in documents that cover a large number of stations (for example, observation procedures often apply to a whole national network and are covered by national-level documents rather than station-specific ones). In many countries, stations are inspected regularly by network management staff, and these inspection reports form a significant part of available metadata. Sometimes observers, their family and newspapers, financial accounting or other external documents can provide additional metadata.

Obtaining access to metadata is often a significant challenge. Substantial quantities of metadata may not be digitized and are only available in paper form. Metadata must be part of data rescue operations since they are as important as the climate records themselves. Most metadata are specific to individual locations and may be locatable in files indexed against that location. Other relevant metadata, especially those dealing with network-wide standards and changes, may be in annual reports or other documents which may be more difficult to trace. To use paper documents, it may be necessary to visit an archive in person.

Finally, metadata are normally available in the local language only.

1.6 **Training**

Training of personnel involved in homogenization is important for the quality of homogenized data. Training has multiple aspects: A good grasp of the general concepts, an understanding of the statistical background of the homogenization process and practical advice for selecting the most appropriate homogenization methods and for handling the selected software. Moreover, the evaluation of the results requires a trained expert.

At the time of writing this publication, relevant training opportunities include:

- The series of seminars “Homogenization and Quality Control in Climatological Databases” held in Budapest, Hungary (HMS, 1996; WMO, 1999; OMSZ, 2001; WMO, 2004; WMO, 2006; WMO, 2010; WMO, 2011; WMO, 2014; OMSZ, 2017). The seminars promote the discussion of homogenization methods, with emphasis on their theoretical aspects, practical applications and evaluation of methods. Most proceedings of these seminars are published through the World Climate Data and Monitoring Programme (WCDMP) series (namely WCDMP-41, WCDMP-56, WCDMP-71, WCDMP-75 and WCDMP-78);

- The annual training session, “Climatology, foundation for climate services”, organized by Météo-France;

- The Data Management Workshops of the European Meteorological Services Network (EUMETNET);

- The annual session on “Climate monitoring: data rescue, management, quality and homogenization” at the annual meeting of the European Meteorological Society;

- The annual session on “Development of climate datasets: homogenization, trends, variability and extremes, including sub-daily timescales” at the General Assembly of the European Geosciences Union;

- The session on “Climate data homogenization and climate trends/variability assessment” at the International Meeting on Statistical Climatology.

CHAPTER 2. HOMOGENIZATION PRACTICE

The purpose of this chapter is to describe general issues related to the development of homogenized datasets. Homogenization is normally best performed with well-tested existing software if available; specific software packages for data homogenization are discussed separately in Chapter 3.

There are many factors that can cause inhomogeneities in a climate record. Some will affect only a few climate elements at any specific location; some only affect a single location, while others may affect an entire observation network or substantial parts of it. This last scenario is important as it (a) can cause large-scale bias and is thus climatologically significant, and (b) is difficult to remove by statistical homogenization if occurring over a short period. Inhomogeneities may have a pronounced seasonal cycle and/or be dependent on weather type (that is, regime-dependent).

Causes of inhomogeneities include:

- **A site relocation**. For example, early stations often started in towns and villages and were later relocated to the outskirts or airports. Early automatic weather stations sometimes had to be placed close to buildings (Menne et al., 2010), while with modern technology it has become easier to place station in pristine locations or move them from the middle of a village to outlying areas (Dienst et al., 2017; Dienst et al., 2019).

- **A change in the local environment**. Examples of changes in micro-siting are growing or cut vegetation around a site, or the construction or removal of a building nearby. Moreover, watering the grass below the instruments may produce an inhomogeneity. Early temperature and precipitation measurements were often performed at a height of several meters, while nowadays 1.5 m to 2 m is standard. Development of the larger environment (urbanization) can cause a gradual local warming leading to a station relocation, which would likely produce a temperature drop. Changes in irrigation practices in the region may lead to artificial cooling (Cook et al., 2014). Moving an anemometer from the roof of an airport to the standard 10 m level often introduces a big drop in surface wind speed (Wan et al., 2010).

- **A change in the instrumentation**. For precipitation, changes in the outside geometry and windshield will affect undercatchment and thus produce inhomogeneities (Leeper et al., 2015). Changes in gauge or instrument type will affect typical measurement errors such as the wetting loss and measuring precision (Wang et al., 2017, 2010). For temperature changes in the screen (see for example, Parker, 1994; Böhm et al., 2010; Buisan et al., 2015) and thus in radiation and wetting, protection of the thermometer is especially important. Even the position of the thermometer within the screen can matter. There was one report of a plastic Stevenson screen letting in the sun on hot days. Mechanical ventilation can reduce radiation errors, but it can also produce stronger wetting errors and may perturb the stable boundary layer. The response time of modern thermometers and their small screens is intrinsically much shorter, which is especially important for the daily minimum and maximum temperature due to turbulent fluctuations. The glass of some early thermometers shrunk in the first years (Winkler, 2009). Calibration errors can also cause inhomogeneities. Mercury thermometers cannot record temperatures below -39 °C, which is why minimum temperature thermometers have typically been replaced with alcohol thermometers. Record temperatures can be outside of the (calibration) range of some automatic weather sensors.

- **A change of observer**. The influence of the observer is especially notable for elements which involve some level of observer judgement, such as cloud or visibility. In voluntary networks, an observer change can also signal a relocation, which might not be documented elsewhere.

- **A change in observation procedures**. For instance, a change in observation time, which often happens simultaneously in the entire network. This is important for fixed-hour measurements, but also for the daily minimum and maximum temperatures (Vincent et al., 2009; Degaetano, 2000; Vose et al., 2003; Karl et al., 1986). Changes in maintenance may also be important, for example, the painting, cleaning and replacement schedule of the screens. With automatic weather stations (AWSs) the loss of ventilation due to icing may not be noticed and damage and soiling may be detected later. Changes in calibration procedures can potentially affect an entire network.

- **A change in data processing**. In much of the world, the daily mean temperature is computed from the daily maximum and minimum temperatures. However, early measurements in Europe were often made at fixed hours and the mean temperature was computed from them, sometimes in combination with the daily minimum or maximum temperature. With AWSs many different definitions of the daily mean have become possible. Quality control and validation procedures, as well as the ability to carry out quality control, have changed over time. For precipitation, another key element is to know how solid precipitation was measured and converted to the equivalent liquid precipitation amount for archiving (Wang et al., 2017).

- **Digitization and database errors**. Digitization may produce inhomogeneities, for example, when a minus sign is forgotten during digitization of a section of data on paper (when the sign is indicated in colour). Data from stations with the same or similar names may be mixed up. In global databases, the data can come from several different sources with unclear provenance. As a result, station data series with similar names may be mixed up, or slightly different series originating from the same station may be kept as different series. Also, metadata such as the station location and units can be wrong. Typical database errors are values that are 10 times too large, especially for precipitation.

The remainder of this chapter consists of 12 sections mainly devoted to the following topics: physical and statistical homogenization, the roles of reference series, how to choose or compose reference series, options if no reference series exist, validation and operational update of homogenization results, and documentation of the homogenization procedure and resulting dataset.

2.1 Physical and statistical homogenization

Whether a homogenization method is called physical or statistical generally depends on how the corrections are computed. Sometimes the main evidence of inhomogeneity in a candidate station is statistical in nature: a large jump in one time series or a candidate station that behaves clearly differently from its neighbours. In such cases, the corrections need to be computed statistically, hence the method used is called statistical homogenization.

Sometimes the reasons for the inhomogeneities are known and this can help make more accurate adjustments. Physical homogenization refers to this case where the adjustments are being estimated using a physical relationship between different variables. For example, a logarithmic wind profile, which represents the relationship between surface wind speed and anemometer height and surface roughness length, was used to adjust surface wind speed data for anemometer height changes (Wan et al., 2010); and the hydrostatic model, which represents the relationship between station and sea level pressures and dry-bulb temperatures, was used to correct errors in both station and mean sea level pressure data due to errors in station elevation (Wan et al., 2007).

Another example would be when the time of observation changes and observations made at both reading times can be used to correct this problem (Vose et al., 2003; Vincent et al., 2009). This is also referred to as physical homogenization, although the method used to estimate the adjustment is not strictly physical.

When there is a documented station relocation, we know what happened physically, but the size of the jump needs to be determined statistically, thus this is still considered as a statistical homogenization. The use of parallel data to estimate the size of the jump (for a relocation) is considered statistical homogenization using a very good reference station, which can be expected to produce more reliable/robust results.

Physical homogenization can also include statistical estimates. For example, for changes in the time of observation in the United States, the National Oceanic and Atmospheric Administration (NOAA) developed a correction method based on hourly observations (Vose et al., 2003). To be able to correct data from stations that only have daily observations, multiple linear regression was used to compute monthly time of observation bias corrections from stations with hourly data. Predictors were the station coordinates (time zone, latitude and longitude), observation hour, average diurnal temperature range and average day-to-day temperature difference (Karl et al., 1986). Similar methods were developed for Canada to account for a change in observation time from 00:00 to 06:00 UTC in 1961 (Vincent et al., 2009), whilst more recent 1-minute resolution data was used to assess the expected impact of an observation time change from 00:00 to 09:00 local time in 1964 at some Australian sites (Trewin, 2012).

It is best to use a physical relationship to estimate the adjustments in cases where known physical relationships between variables are sufficiently robust that the related adjustments are likely to give better results than statistical homogenization, such as in the cases of Wan et al. (2010) and Wan et al. (2007) mentioned above. However, adjustments estimated on the basis of a physical relationship need accurate metadata (such as anemometer heights, station elevations) and may need also other data (such as dry-bulb temperature and surface roughness in the examples above). These metadata or data are often not available. Physical homogenization can be done only when the cause of inhomogeneities is known and the related metadata/data is available. Thus, physical homogenization is normally additional work in the sense that it can only be applied to a part of the inhomogeneities and statistical homogenization should thus always be also applied to correct likely remaining ones.

When both physical and statistical methods are possible, it is recommended to compute both adjustments and compare them with each other. Comparing both types of adjustment can help identify problems. Whether physical or statistical homogenization is preferred for the actual corrections depends on the accuracy of the corrections. In most cases the physical adjustments are more accurate, but not always. First applying physical adjustments may make the remaining inhomogeneity too small to be detected with statistical homogenization, but still large enough to be climatologically problematic on the large scale. Thus, choosing the most accurate method may give better results than applying both methods one after another.

2.2 Selecting the data to be homogenized

2.2.1 *Which stations should be selected?*

The guidance in this subsection applies to the homogenization of datasets rather than a single time series.

There are two broad approaches that can be used for selecting the data to be homogenized in a national or regional dataset.

The first is to include all stations that meet the preset criteria for factors such as length of record and data completeness, while the second is to consider only the stations, chosen for the best observational quality (for example, best site standards, fewest documented moves) or geographic representativeness.

Both approaches have been used for major national datasets; for example, national homogenized datasets from the United States, Canada and Spain (Menne et al., 2009; Vincent et al. (2012); Guijarro, 2013) include all available stations with a sufficiently long record, whereas the Swiss, another Spanish and Australian national datasets (Begert et al., 2005; Brunet et al., 2006; Trewin, 2013) select stations assessed to be of the best quality and representativeness out of a broader national network, which is several times larger.

A strong selection reduces the amount of work, which means that more attention can be given to single stations. The goal should be to obtain a SNR, defined as the square root of the variance of the break signal divided by the variance of the noise signal (Lindau and Venema, 2018*a*)) greater than one. Using several stations has the advantage that regional climates can also be studied, and maps can be produced.

Methods to compute the SNR of a difference time series are introduced in Lindau and Venema (2018*a*, 2019).

2.2.2 *How to prepare data for homogenization*

Once a set of stations has been selected for homogenization, the next step is to determine which data from those stations will be used in the change point detection process. Decisions made here are important because some inhomogeneities may have a seasonal cycle – for example, a change in the wind shielding of a precipitation gauge at a cold-climate site may have little impact on summer precipitation, but a large impact on precipitation in winter when much of it falls in the form of snow.

Three possible options are:

(a) To use only annual data (annual sum or mean, sometimes combined with magnitude of seasonal cycle);

(b) To use monthly or seasonal time series in parallel with each other (and/or with annual data);

(c) To use monthly or seasonal data (either in their original form or as anomalies) as a single time series (also referred to as serial or consecutive monthly or seasonal time series).

The use of annual data typically has the most favourable SNR for inhomogeneities that affect all, or a substantial part, of the year. Compared to serial monthly or daily data, the lower SNR is, however, exactly compensated by the larger number of values (in the case of white noise and as long as the breaks are on the 1st of January; Lindau and Venema, 2018*a*). On the other hand, using annual data may result in failure to detect inhomogeneities that only affect part of the year, or inhomogeneities that have opposite impacts in different seasons and cancel each other out in an annual mean (for example, where a site is moved from an exposed coastal location to one further inland, it is likely to be warmer in summer but cooler in winter). In such cases using the seasonal cycle or the option (b) above can improve the results.

Using monthly or seasonal time series in parallel (that is, carrying out tests separately on time series for each month/season with one data point per year) allows the detection of seasonally varying inhomogeneities. In some cases, a station which shows no significant inhomogeneity at an annual timescale may show significant seasonal signals, for example, where opposite signals in summer and winter cancel each other out in an annual mean. This method requires the consolidation of information from the various monthly/seasonal tests with each other, and with annual values, to determine a final set of potential inhomogeneities. However, these individual parallel monthly (or seasonal) series are the same length as the corresponding annual data series, yet they are noisier. Consecutive monthly (or seasonal) series form longer time series, a fact that partly compensates for the noise. A test on parallel monthly data helps only in case of dense networks.

Some methods use monthly or seasonal data as a single time series, also called serial monthly/seasonal or consecutive monthly/seasonal series/methods. This increases the temporal resolution in detecting inhomogeneities and, like parallel monthly methods, avoids assigning the entire year with the break to the homogeneous subperiod either before or after the break. However, serial monthly/seasonal methods may also miss a seasonally variable signal. Such methods may also be complicated by the fact that, in some climates, not only does the mean value of a time series vary seasonally, but its variability may also have a seasonal cycle (for example, in most mid- and high-latitude locations in the northern hemisphere, temperature variability is substantially greater in winter than in summer). Also, autocorrelation in consecutive monthly or seasonal series is usually higher than in annual series, hence non-negligible, and should be taken into account in the statistical test applied.

Methods using annual data cannot determine the timing of an inhomogeneity, without reference to metadata, to a precision greater than one year. However, if the SNR of monthly or seasonal data is not large enough, the change date will have a clear uncertainty (Lindau and Venema, 2016). Hence, the greater apparent precision which might be obtained from a monthly method can be partly illusory. The greater temporal precision may also sometimes be helpful in focussing the search for metadata.

The most widely used methods are option (a) above, for automatic homogenization methods, and option (b) with annual data, for manual homogenization. Option (b) is likely to detect a broader range of inhomogeneities if applied properly. It is also possible to use two or the three options combined (Xu et al., 2017, 2013).

2.3 Statistical detection or determination of change points

Metadata should never be assumed to be complete and statistical determination of breakpoints should always be applied. For a homogeneity test, the null hypothesis is that the candidate series is homogeneous, and the alternative hypothesis is that the candidate series is not homogeneous (has one or more change points). In general, a statistical test works by comparing the value of a test statistic to its critical value corresponding to the chosen level of significance α (usually 5%). The null hypothesis is rejected when the test statistic value exceeds the critical value. The significance level α is the likelihood of the null hypothesis being rejected wrongly.

Even for a documented change point (for example, when the time of change is documented in metadata and known to the person analysing the data), one still needs to use a statistical test to determine whether or not the documented non-climatic change is statistically different from zero at a chosen significance level. Furthermore, even when an inhomogeneity is statistically significant, there may still be a large uncertainty in determining the size of the adjustment.

Inhomogeneities are mostly abrupt but can also be gradual (for example, in cases of growing vegetation or urbanization). In statistical homogenization, we typically model inhomogeneities as a step function. This also works well for gradual inhomogeneities because there are often additional jumps during the gradual inhomogeneity, and also because fitting linear trends to model gradual inhomogeneities is often inaccurate, as their behaviour in time can be non-linear. Thus, in practice, homogenizing gradual inhomogeneities with multiple breaks works well (Venema et al., 2012).

Most of the homogeneity tests are developed to detect mean shifts and thus cannot detect any variance shift or probability distribution change that is not accompanied by a mean shift. A variant of the Kolmogorov–Smirnov (K–S) test developed by Dai et al. (2011) can be used to detect unknown changes in the probability distribution (including variance shift) of the data. Unfortunately, this method has not been included in any data homogenization software. Szentimrey (2018) is working on a method for the detection and correction of breaks in the mean and standard deviation for normally distributed data.

After a list of change points has been determined to consist of significant non-climatic change points, the adjustments needed to homogenize the candidate series must be estimated; this is discussed in section 2.6. Detection and correction are mostly performed by comparison with neighbouring stations; more information on reference series can be found in section 2.4. Readers are referred to Chapter 3 for detailed descriptions of the software packages available for detecting and testing change points and estimating adjustments used in the homogenization of climate data.

2.3.1 *Incorporating metadata in statistical tests*

A statistical break test provides an estimate of change point position. Its accuracy depends on the SNR; if the SNR is well above one, the break position can be estimated accurately (Lindau and Venema, 2016).

Finding documentation for a break can be time consuming. The statistical evidence for a break can reduce the period over which the metadata needs to be investigated. The SNR indicates how broad this period would need to be. For smaller SNRs, errors of several data-points before or after the true change point are possible, as shown in Figure 2.

When the SNR is very low, a random segmentation of the time series can even explain as much of the actual break signal as the segmentation estimated by a homogenization method (see Figure 3). In other words, the test rightly detects that the series has inhomogeneities, but their estimated positions may be determined more by noise than by the break signal.

When combining statistically estimated breakpoint locations with documented breaks, it should be borne in mind that the metadata may be wrong. The likelihood of metadata being wrong and the uncertainty in the data depend on the situation and are subjective. Typically, if the metadata provide a specific date, the metadata is more accurate, but if the statistical evidence is strong, it can take precedence. If a statistically detected break is within a short period (that is, within the uncertainty of the break detection) of a change documented without a specific date, one can attribute the break to the documented change.

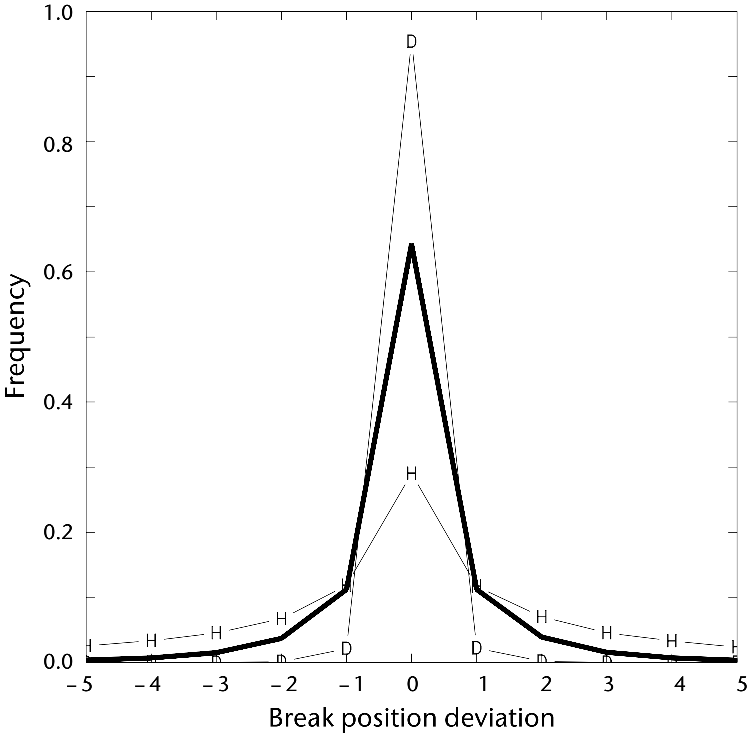

Figure 2. Distribution of the break position deviation for three different signal-to-noise ratios (SNRs). The thick line is for an SNR of 1, that is, where the break variance is as large as the noise variance. The line marked D represents the double SNR and the line marked H half the SNR (Figure from Lindau and Venema (2016)).

The statistical tests for detecting unknown change points are different from those for determining the statistical significance of documented change points. In case of one unknown breakpoint, every position is tested (multiple testing problem) and the maximum expected difference under the null hypothesis is thus larger. The commonly used tests for detecting single or multiple unknown change points (mean shifts) are maximal *t* tests — for example, the standardized normal homogeneity test of Alexandersson (1986) and the penalized maximal *t* test of Wang et al. (2007) – or maximal *F* tests, such as the penalized maximal *F* test of Wang (2008*b*), and the tests of Lund and Reeves (2002) based on the two-phase regression model. Modern multiple-breakpoint methods, such as PRODIGE, MASH, ACMANT and HOMER, effectively test all break combinations and thus also take multiple testing into account; see Chapter 3 for details on the methods mentioned in this paragraph.

The commonly used tests for determining the significance of documented change points are the regular Student *t* test, when a reference series is used, and the regular *F* test, when the series is tested without using a reference series or when a trend difference between the candidate and reference series is suspected (Lund and Reeves 2002; Wang 2003; Wang 2008*a*). As explained above, the critical values for the regular *t* or *F* test are much smaller than those for the corresponding maximal *t* or maximal *F* test (see, for example, Lund and Reeves (2002) and Wang (2003)). Thus, the use of the regular *t* or *F* test for detecting unknown change points would lead to too many false alarms (declare many change points that are actually insignificant).

When metadata are available, one can first use a statistical test to identify significant unknown change points; then, one can add all additional documented change points and test their statistical significance. This is necessary because some documented changes in observing site, procedure or instrumentation might not induce any significant change in the candidate data series, while the statistically identified change points should be taken into account when testing the documented breakpoints using the homogeneous subperiod before and after the documented break.

The test for documented breakpoints should only be used for breakpoints of a limited number of clearly documented, highly likely causes, such as relocations and screen-type changes. A test for a documented breakpoint should not be used for regularly occurring events such as maintenance or calibration. Note that including too many documented breaks for testing will reduce the sample size (that is, the segments are shorter) and thus increase the uncertainty of the results. Tests for unknown breakpoints are also commonly applied, which is equivalent to notably raising the significance level used (Lund and Reeves 2002, Wang 2003).

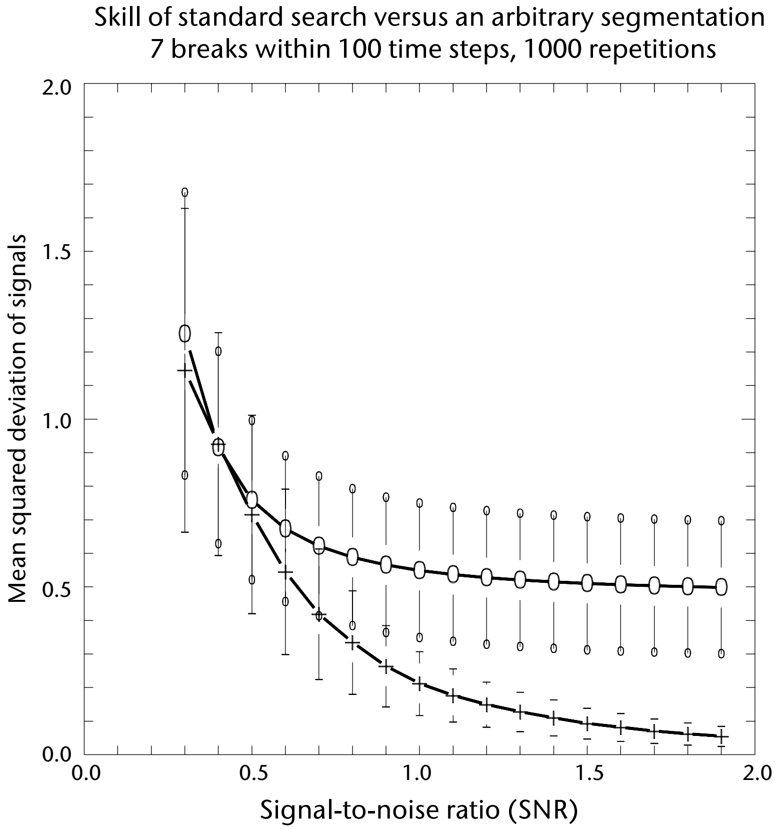

Figure 3. Accuracy of the estimation of the break signal as a function of the signal-to-noise ratio (SNR). The curve with circles shows the results of a breakpoint detection method while the other curve shows the results of random segmentations. When the SNR is high (right side of graph) the break signal is accurately estimated by the break detection methods (plusses): the squared deviation between detected and inserted signal is small. A random segmentation, however, can also explain half of the variance of the break signal. When the SNR is 0.5 or below, the segmentation of the homogenization method is no better than random segmentation. Because the detected breaks explain both noise and break variance, the variance is larger than expected in case of noise, and the detection is statistically significant. The problem is that the algorithm correctly detects that the observations have inhomogeneities but is unable to determine their positions (see Lindau and Venema (2018*a*) for details).

2.4 Reference series

The task of detecting, and adjusting for, an inhomogeneity in a climate time series is made more challenging by the fact that any climate time series will contain substantial noise, which arises from natural climate and weather variability and measurement errors. This can make inhomogeneities difficult to find. For example, it will be difficult or impossible for any statistical method to detect a 0.4 °C inhomogeneity in a time series with a standard deviation of 1 °C.

A common method of addressing this problem is comparing the candidate station's time series to a (neighbouring) reference station's time series. The most typical example is to compute a difference time series between the candidate and reference series; then detect and adjust the series based on this difference series. The steps to follow in case no reference stations are available are discussed below in section 2.5.

The natural variability that exists in a candidate series will also exist in the reference series, hence creating a series that compares the candidate and reference series will remove much of the influence of natural climate variability while retaining the effect of the inhomogeneity at the candidate site.

Another advantage of using a reference series is that no assumptions about the statistical nature of the climatic variability are necessary, since the use of a representative reference series removes that necessity. The difference time series can be assumed to contain white noise (or auto-correlated noise) and inhomogeneities, which greatly simplifies the statistical problem. Typically, a reference series will comprise data from one or more locations near the candidate station.

The best possible set of candidate and reference series is a parallel observation of the old and new situation at the same site, for example, when a new instrument system is introduced, while the old system continues to be in operation for a period of time.

Parallel observations are most valuable if the "old" part of the parallel observation system is representative of conditions before the start of the parallel observation period. However, a common scenario is that the old site environment was changed during the period of parallel observations, which makes the parallel observations unrepresentative of the old site environment. Comparison with neighbouring station data could help identify this problem and is highly recommended.

If there are no parallel measurements available, one can consider making them with the equipment used before and after the break or, if the equipment is not available, with replicas created for the experiment. This is especially recommended for studying the influence of a historical transition that affected a large part or the entire network (Brunet et al., 2011; Mekis and Vincent, 2011; Quayle et al., 1991).

A reference data series can be a composite computed from several neighbouring stations (composite reference) or a single neighbouring station (pairwise homogenization). In the latter case, inhomogeneities in the comparison series can belong to the candidate as well as the reference series and multiple pairs must be investigated to determine which station the break belongs to.

There are four considerations for the selection or weighing of reference series:

(a) The set of references need to overlap over the full period of the candidate series;

(b) The weights should reduce the noise of the difference series;

(c) The influence of inhomogeneities in the reference should be reduced;

(d) The similarity of the regional climate signal in the candidate and reference series should be enhanced.

These four considerations conflict with each other and the optimal solution is unclear. Consequently, many different methods are used to select reference stations and to assign weights to neighbouring stations when computing a composite reference from them. Common weighting methods use weights based on correlations, the correlations of the first difference series, kriging (optimal interpolation) weights, the inverse distance from the candidate and height difference. Also, giving all reference stations the same weight is used to avoid the excessive influence of a very inhomogeneous, but very close or very well correlated, neighbouring station.

Stations might be excluded when the distance or the height difference is too large or the correlation too low to select references with a similar climate.

In many cases, using distance is similar to using correlations and the correlation between two stations will decay reasonably monotonically with the distance between them, but this may, for instance, not be the case in mountains, where altitude is also important. The correlation matrix may also be anisotropic, for example, temperatures at an exposed coastal station are likely to be better correlated with another similarly exposed station 100 km further along the coast than with a station 50 km inland.

Regarding the correlation between two stations, let $B(t)$ and $R(t)$ denote the candidate series and the potential reference series, respectively. The series $\Delta_B(t) = [B(t) - B(t–1)]$ and $\Delta_R(t) = [R(t) - R(t–1)]$ are called the first difference series. The correlation between the first difference series $\Delta_B(t)$ and $\Delta_R(t)$ is often used to select a reference series, because this correlation value will be much less affected by any inhomogeneities that might exist in the candidate and/or reference series (an inhomogeneity will only generate one bad value in the first difference series, but a segment of bad values in the candidate or reference series).

2.4.1 *Overlap*

Data from earlier periods are typically sparser and it is hard to find well-correlated reference stations. In order to have reference stations for earlier periods, references with poorer correlations may also have to be part of the set of reference stations.

The theoretical minimum number of stations needed for statistical homogenization is three. In practice, five stations (four references) are required in order to obtain good results in more complex situations. This requirement typically determines the start year of a homogenized dataset.

Usually networks have fewer data in the beginning (and sometimes near the end) of their operation. Having overlapping stations for the early period normally means that stations with lower correlations need to be selected. When adding shorter series to a composite reference, the beginning or the end of the time series can introduce an inhomogeneity. Similar problems can arise from missing data periods.

When selecting or weighting references based on their correlations, one should take into account that the computed correlations have considerable uncertainties, especially in the case of short series. Thus, a high correlation of a short series may be a coincidence and a longer series with a lower correlation may be more reliable.

Short series of a few years are also problematic because the detectability of a break depends on being able to detect a statistically significant difference in the mean value before and after a break. Thus, besides the size of the inhomogeneity, the number of samples, that is, the length of the homogeneous subperiods, is also important.

2.4.2 *Noise reduction*

Kriging provides an optimal estimate of observations at the candidate station, given the observations of the references. A composite reference computed as a weighted average of the references using kriging weights will reduce the noise of the difference time series optimally. The kriging weights are computed using the cross-correlation matrix.

Because of the other three considerations mentioned above and for ease of use, the correlations themselves, or inverse distances, are also often used to compute the weights for a composite reference series.

The detectability of an inhomogeneity is mainly a function of the SNR. In the case of one breakpoint, the SNR is often defined as the ratio of the size of that inhomogeneity to the standard deviation of the time series under consideration. Reducing the amount of noise in a time series by using a reference will increase the likelihood of detecting an inhomogeneity of a given size.

As shown in Figure 4, in the case of one single change point in a data time series of length N=600, the chance for reasonably accurate detection of inhomogeneities increases from around 53% when the ratio $\Delta/SD=0.5$ to around 91% (99%) when $\Delta/SD=1.0$ (1.5); it is ≥ 99.99% when $\Delta/SD \geq 2.0$. A lower SNR is also associated with a larger uncertainty range of detection power. See section 2.3.1 for a similar argument in relation to the multiple breakpoint case. Trying to obtain a SNR higher than 1 is thus a high priority. For higher SNR values, the other three considerations (under 2.4) become important.

A pair of stations in pairwise homogenization will have twice as many breaks as a difference time series based on a composite reference, because the composite offers a better estimate of the regional climate signal than a single station. Thus, if the SNR is low, the use of composite references may be preferable.

2.4.3 *Inhomogeneities in the reference*

A reference series should be homogeneous or at least be homogeneous around the time the candidate series is inhomogeneous. Inhomogeneities in reference series can easily be mistaken as inhomogeneities in the candidate series. Such mistakes can be greatly reduced by visualizing

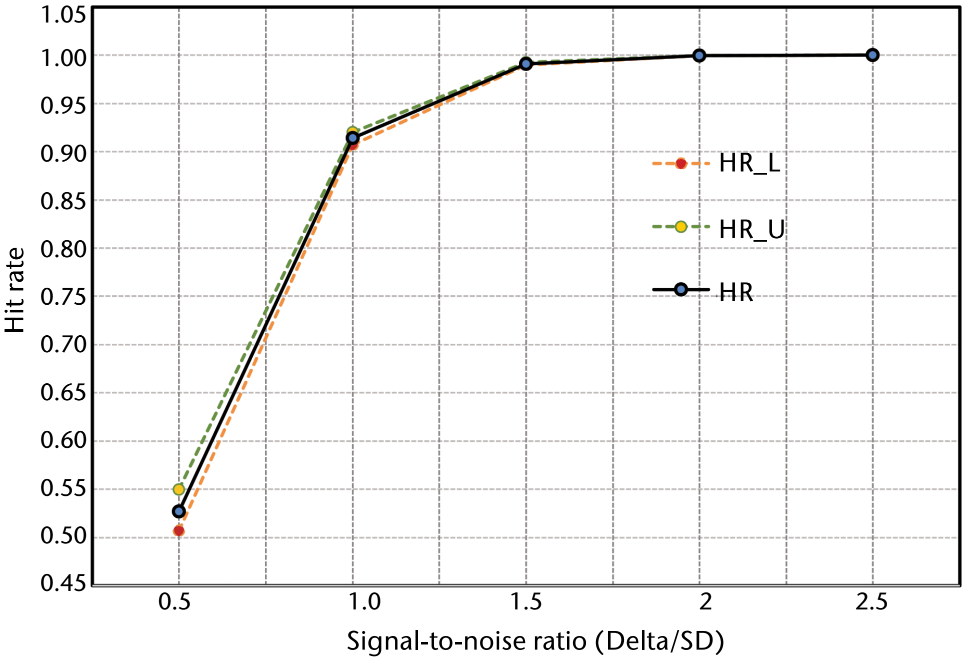

Figure 4. Power of detection, that is, hit rates (HR, vertical axis) as a function of the ratio Δ/SD (horizontal axis) of the size Δ of an inhomogeneity to the standard deviation (SD) of the candidate series (based on Table 1e of Wang (2008*a*)). The lower (HR L) and upper (HR U) bounds represent the 95% confidence interval (uncertainty range) of the hit rates.

the difference series or the regression fits. To reduce the risk that a composite reference may have significant inhomogeneities of its own around the time of interest, a sufficient number of reference stations needs to have significant weights.

In case of widespread inhomogeneities occurring in many or all stations over a short period, one needs to be very careful when computing a composite reference. The composite may not contain visible jumps, but if this widespread inhomogeneity has a bias this will also be present in the composite reference. This is especially problematic for computing corrections.

Carefully designed iterative procedures that remove reference stations with breaks from the computation of the composite reference or correct the breaks in the reference stations reduce this problem. Pairwise homogenization methods are best suited for dealing with such difficult cases, but other methods that run through iterative processes of break correction also yield good results.

2.4.4 *References with similar climate signals*

The selection of references with similar climate signals is best performed by an expert on the basis of an understanding of the local climate. In automatic methods, similarity is often estimated by the cross-correlations, together with thresholds for the maximum distance and height difference and minimum correlation.

However, a high correlation does not guarantee that the reference stations are from the same climatological region although they share a similar signal. Additional objective criteria could be the Köppen climate classes, the size of the seasonal cycle, the daily cycle, the exposure, the soil moisture and vegetation.

In case of low-density networks, the number of reference stations may need to be limited to ensure they all belong to the same climate region.

In detection methods that involve counting the number of detected breaks, it is especially important that all stations have a similar climate signal. Such a step is sometimes used to collate the results of several break detection tests or in the attribution of pairwise homogenization methods, where the station with the break is determined by the number of pairs that have a break. If not all pairs belong to the same climate region, it is possible that many remote reference stations may falsely suggest a break, whereas a smaller number of stations from the same climate region would not. In such cases, it is appropriate to select fewer reference stations or to apply higher weights to nearby stations.

2.4.5 ***Using references***

References are mostly used as follows: (a) the difference between candidate and reference is used for normally distributed variables; (b) the ratio of candidate to reference is used for approximately log-normally distributed variables.

In the difference series approach, the test is applied to, and the adjustments are estimated from, the difference series $D(t) = [X(t) - Y(t)]$ where X denotes the candidate station and Y denotes the reference time series (or station). The ratio series would be computed as $R(t) = [X(t) / Y(t)]$. Temperature and most other variables stem from additive processes and have an approximately normal distribution. Monthly precipitation and wind speed are multiplicative processes, which produce approximately a log-normal distribution. In some climates, even the monthly averages of these variables may contain zeros. In such cases, one can transform the data to an approximately normal distribution and use a difference time series.

The covariate approach is less common; it refers to the use of the reference series as a covariate in a regression-based test such as the multiple-regression-based test of Vincent (1998). In this case, the test is applied to the residual series of the regression of the candidate on the reference series, for example, residuals $\varepsilon(t) = [B(t) - \hat{B}(t)]$ of the fitted regression $\hat{B}(t) = [\hat{a} + \hat{b}R(t)]$.

2.5 **Options if no useful reference station exists**

For some observational series, such as those in northern Canada (most of the sites are 400 km–800 km apart) on remote islands or in the Antarctic, there is no useful reference available. What matters for homogenization is the SNR, not distance; for wind, precipitation or humidity, the SNR will be small at much shorter distances than for temperature and especially for sea-level pressure. These remote stations are important and have a large weight in regional or global mean data series because these sites represent a large area and often provide scientifically interesting data from early periods. These situations require expert judgement and are best dealt with by retrieving and studying metadata. The following approaches can be used to homogenize data from such remote sites.

First, one can look for other related variables or data sources to use as reference. For example, for coastal and island stations, sea-surface temperatures may be another possible reference series (Cowtan et al., 2018). However, air and sea-surface temperature do have different trends and variability. Cloud cover has been used to homogenize sunshine duration observations, and sunshine duration data have been used to homogenize surface radiation data (Yang et al., 2018). Care must be taken that this does not remove variability in sunshine duration due to variations in cloud microphysics and aerosols. Surface-air temperatures from a reanalysis dataset can be used as a reference to test surface-air temperature recorded at a remote site, keeping in mind that reanalysis data have inhomogeneities of their own.

Second, there are homogenization methods that can be used without using a reference series (absolute statistical homogenization). However, homogeneity testing without a reference series is much less reliable and should never be done with a fully automatic procedure. One should visually inspect the original and deseasonalized candidate series and should study all available metadata to determine the final list of change points to be adjusted. Estimation of the size of adjustment is also more uncertain in this case. The higher uncertainty of such homogenized datasets should be quantified and clearly communicated.

An important reason why absolute homogenization is less accurate is that the series is “noisier”. Furthermore, this “noise” is partially due to long-term variations in the climate system and is harder to distinguish from inhomogeneities than uncorrelated noise.

The problems can be reduced by identifying and modelling the low-frequency variations within a homogenization procedure, for example, using the method developed by Wen et al. (2011). An example of application of this method to the Fort Nelson (Canada) cloudiness time series is shown in Figure 5: it identified correctly two shifts, along with a 12.5-year cycle, an annual cycle, and a negative trend; the combination of these components is shown as the blue curve in the graph. If the 12.5-year cycle in this time series were ignored, that is, if the PMFred algorithm were applied to this time series directly, it would identify three false change points and fail to identify the true change point. Such noise reduction techniques work especially well for absolute homogenization but can also be explored for relative homogenization.

As mentioned above, reference series may be derived from other variables. For example, Wan et al. (2010) and Minola et al. (2016) used geostrophic wind speeds derived from sea-level pressure gradients over a triangular area (formed by three stations for surface pressure observations) as a reference series to homogenize surface-wind speed data from stations within that area. However, geostrophic wind is probably not a good reference for surface winds over tropical/subtropical regions and regions of complex topography. Dai et al. (2011) used empirical relationships between the anomalies of air temperature and vapor pressure derived from recent observations, when dewpoint depression (DPD) reports were available under those conditions, to adjust artificial sampling effects by estimating missing DPD reports for cold (T < 30 °C) and dry (DPD artificially set to 30 °C) conditions. For coastal and island stations, sea-surface temperatures may be another possible reference series.

Reanalysis is independent of surface data for most variables and thus has potential as a reference series where no other suitable reference series exist. This approach has been used for the homogenization of wind on the Iberian Peninsula and in Australia (Azorin-Molina et al., 2014 and 2019). For upper-air data, reanalysis data are used in the RAOBCORE and RICH datasets (Haimberger et al., 2012).

Because upper-air temperatures (on which reanalyses are based) generally have longer decorrelation length scales than surface temperatures, reanalyses may be of value as a reference series where no good surface neighbours exist. For the same reason, it is expected that inhomogeneities in the reanalysis would be visible in the difference times series with many stations, which may allow the operator to determine whether the inhomogeneity is in the station data or the reanalysis. Large-scale break inhomogeneities are present in reanalysis data when satellite datasets are introduced or changed. Atmospheric models often have (regional) differences in observations. When more observations become available, the reanalysis can thus (gradually) shift from the model base state to observed climatology.

If the data contain a clear break whose size cannot be estimated with sufficient accuracy, absolute homogenization can also be used to determine the date of this break and to disregard the data before this cut-off date. However, in the case of data sparse regions, such as the Arctic, not using such data can lead to bias. In the case of data from early periods, digitization of more data may help resolve the conflict.

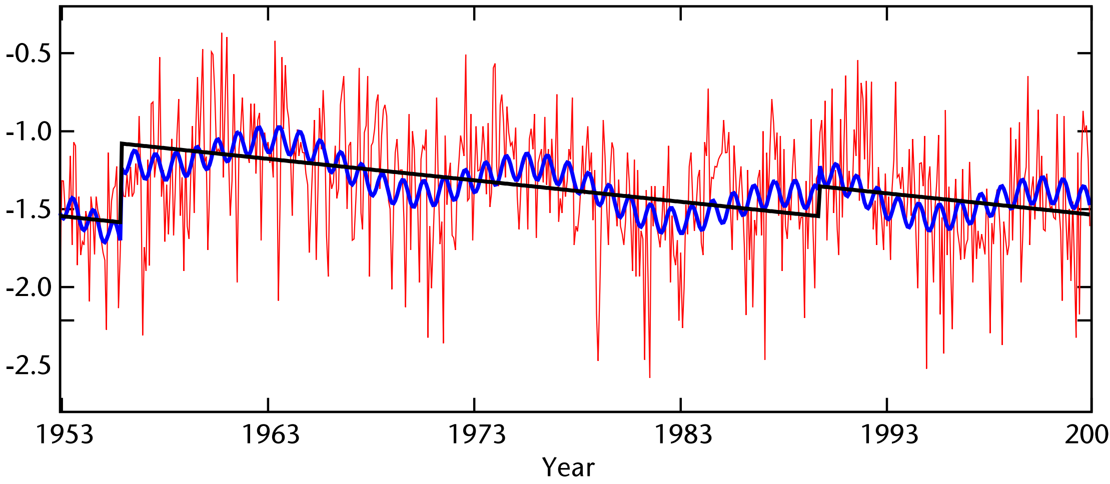

Figure 5. The Fort Nelson (Canada) cloudiness time series (red) and the climate signal and shifts (blue curve) identified by the method of Wen et al. (2011). The thick black line shows the trend and shifts, excluding the periodic components (this is part of Figure 8 of Wen et al. (2011)).

2.6 Statistical adjustments

The most accurate way to compute corrections is to consider corrections for all breaks in a (regional) network simultaneously; this principle is called joint correction. Such a method was introduced by Caussinus and Mestre (2004): it decomposes the raw data into a regional climate signal for all stations and a step function to model the breaks and noise for every station, which is minimized. This method helped to improve the corrections for nearly all contributions to the HOME[1] benchmarking study (Venema et al., 2012) that were not yet using this method (Domonkos et al., 2013). When all breaks are detected, this method on average removes large-scale trend bias perfectly, while its uncertainty is determined by the noise of the difference time series; errors in set of detected breaks lead to undercorrection of any trend errors (Lindau and Venema, 2018*b*).

Many homogenization methods do not set two breakpoints close to each other. It is common to treat two nearby breaks as a single break in correction, using data before the first break and after the second break as the basis for longer-term adjustments, and treat data between the two breaks separately. As an example, in the Australian ACORN-SAT methodology, breaks are only treated separately if at least four years apart (Trewin, 2018), and breaks detected using the RHtests package with (or without) a reference series are at least five (or 10) data points apart.

These correction methods can be applied to annual, seasonal and monthly data. Many inhomogeneities have a seasonal cycle and would not be corrected by computing corrections at the annual scale and applying them as fixed corrections for every month. Numerical experiments with the PRODIGE homogenization method on the HOME benchmark dataset showed that temperature monthly corrections for each calendar month separately were most accurate. Because of the seasonal cycle of inhomogeneities, annual corrections performed less well. For precipitation, annual corrections were most accurate, although the inhomogeneities did have a seasonal cycle. However, the uncertainty of the estimated monthly corrections for each calendar month separately was likely too large for precipitation data, leading to less accurate homogenized data. Correcting temperature monthly/seasonally and precipitation annually is likely a good rule of thumb, although in sparser networks than the typical European ones, as studied in HOME, temperature corrections could behave similarly to precipitation corrections. The SNR is likely what matters most, rather than the meteorological element.

Correction methods for the distribution of daily data (Trewin and Trevitt, 1996; Della-Marta and Wanner, 2006; Mestre et al., 2011; Wang et al., 2010; Wang et al., 2013; Trewin, 2013) can also be used to correct data with a seasonal cycle, which typically dominate the total variance. For example, when a Canadian station near the St Lawrence River

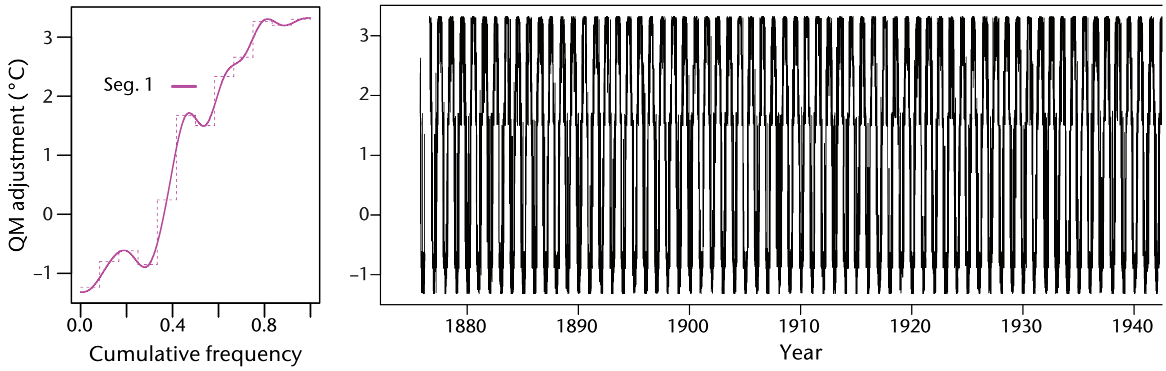

Figure 6. Effect of relocating a station from a near-river site to an inland site on daily maximum surface-air temperatures. Shown here is the distribution over quantiles (left panel) and time series (right panel) of the quantile matching adjustments estimated from a difference (candidate-minus-reference) series, which are needed to adjust the near-river site data (this is part of Figure 1 of Wang et al. (2013)).

[1] European Cooperation in Science and Technology (COST) Action on Homogenization

was moved to an inland site, the inhomogeneity in the daily maximum surface-air temperature series showed a clear seasonal cycle (see Figure 6, right panel). The inland site is much warmer in summer, and a little colder in winter than the near-river site.

One practical issue is whether the whole homogeneous subperiod of data record should be used to estimate the adjustments. By default, the whole period is used to make full use of the limited amount of data to estimate the corrections, but sometimes it may make sense to deviate from this. When the homogeneous subperiod is long, the added benefit of more data diminishes, while the risk of remaining inhomogeneities or noticeable differences in climate change between candidate and reference increases. This is especially true when the reference is not optimal. When absolute homogenization is applied, it is common to limit the periods to 10 years before and after the break. It is best not to use (adjusted) data beyond the adjacent homogeneous subperiods.

An additional issue, as discussed earlier, is where the period immediately before or after an inhomogeneity is not representative of the broader long-term behaviour of the station. This can occur, for example, where a station is moved after a sudden deterioration in its exposure; in such cases, it would be appropriate to exclude data between the change in exposure and the site move when making longer-term adjustments (see figure 7). In practice, issues of this type are often difficult to detect in statistical testing. Also, where the date of the break is uncertain (has an SNR below one; Lindau and Venema, 2016), or the statistical breaks and the metadata suggest different dates, it can be justified to exclude some data around the break when estimating corrections.

The use of composites without removing series with breaks is not recommended for computing corrections. The breaks that bias the network average changes the most are those that occur in all stations. When this happens over a short period, the reference would have a similar bias as the candidate, and the large-scale bias would largely remain after correction.

It is usual practice to correct the data to match the conditions of its most recent homogeneous section. By doing so, incoming future data will still be homogeneous unless further changes occur at the station.

2.7 Data review and multiple rounds of homogenization

The final step is the validation of the homogenized data. No matter how well the data are homogenized, perfection will not be achieved, and some residual inhomogeneity will remain in the adjusted series. It is necessary to critically evaluate the work and to review the homogenized data.

This review should look at individual series and assess whether the new values make sense: is the seasonal cycle preserved? Are the values in the expected range for the station? Do the adjustments differ radically between adjacent months, etc. If a full dataset has been homogenized, it is extremely useful to look at the regional coherence of the temporal evolution of the series, as well as to compare the adjustment series with the known changes in the network and compute the adjustments made for specific inhomogeneity types. This can involve carrying out homogeneity tests on the homogenized data – for example, by comparing them with those from other homogenized stations in the region, or testing for anomalous trends at individual homogenized stations. If the network includes multiple station types, it can be useful to compare them. One may be able to compare the results with previous homogenization exercises and neighbouring countries. The consistency among climate elements should also be investigated, for example, the mean, maximum and minimum temperature. In addition to quantifying the results, visual inspection of the adjustments, homogenized data and difference series can also help to identify problems.

In the case of manual homogenization with methods using a composite reference, it is normal for an initial homogenization process to fail to fully address all homogeneity issues with the dataset under consideration. There are a number of reasons why this can occur; the most common ones include:

- Undetected inhomogeneities in one or more reference series or in a parallel observation pair at one of the stations during the period of parallel observations;

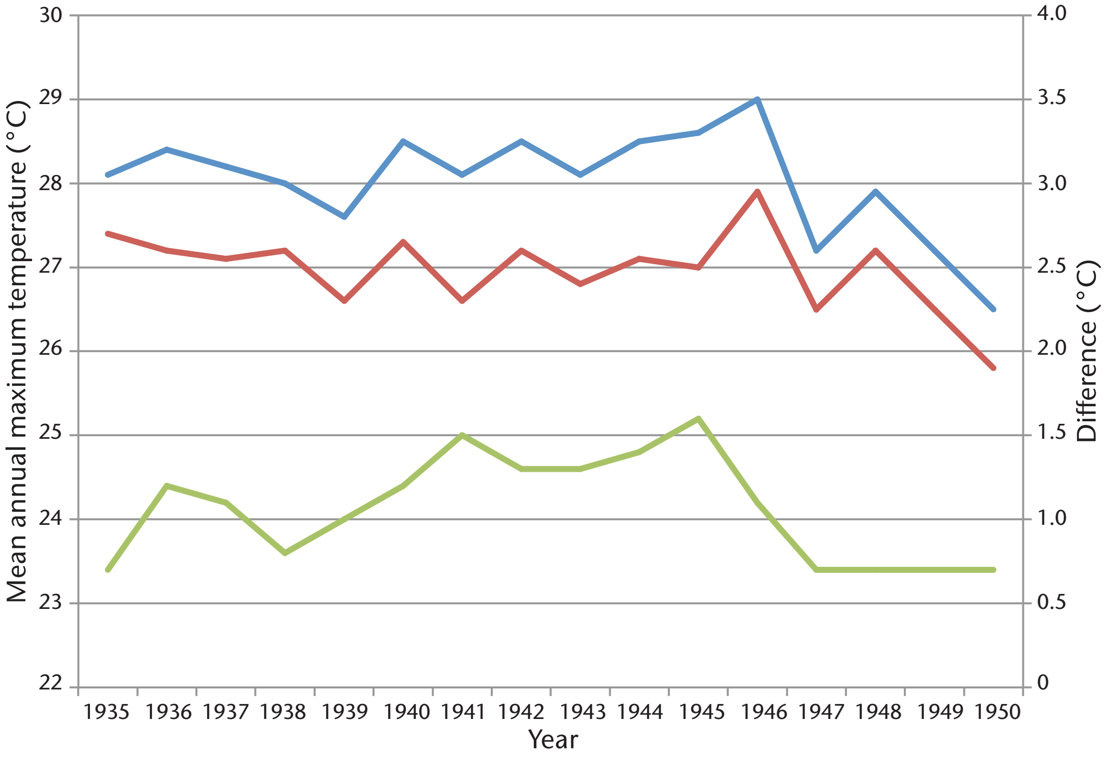

Figure 7. An example of unrepresentative data before a change. At Gayndah, Australia (blue line; left axis), the screen deteriorated progressively after 1940, before it was replaced in October 1945. The maximum temperature difference (green line; right axis) between Gayndah and the mean of three reference sites, Dalby, Brisbane and Emerald (red line; left axis), increased from 1.0 °C to about 1.5 °C in the years before the screen change, before dropping to 0.7 °C after the change.

- Anomalous climatic conditions around the time of an inhomogeneity, resulting in an unrepresentative adjustment (for example, a particularly wet or dry period immediately before or after an inhomogeneity);

- Conditions at a site, shortly before or after an adjustment, that are unrepresentative of the longer-term record. One common scenario here is where a site move takes place because of recent construction work; data from the old site after the work started may not be representative of the old site prior to the construction work and should, therefore, not be used in determining the required adjustment for a long-term dataset.

Anomalies can also occur for other reasons; for example, a data quality problem at a station in an individual month during a period of parallel observations may, depending on the method used, affect adjustments for that month, but not for any other.

The most effective way of dealing with these issues is to carry out a second-round homogenization process in conjunction with a data review.

Where issues are identified through the second-round process, depending on the nature of the issue, options for addressing them include:

- Repeating the homogenization using a different set of reference stations, if possible. A useful approach, if a number of reference series are being used, is to estimate the size of the adjustment that would occur when using a single reference station, for each of the stations individually, and remove any reference station that generates results that are excessively anomalous relative to other reference stations;

- Using a different time period as the basis for adjustment, for example, if a station moved in 1951 but a new building was built near the old station in 1949, the period ending in 1948 (rather than the period ending in 1950 or 1951) should be used as the basis for long-term adjustment.

When making multiple rounds of homogenization, one should never assume that the data of a previous round were homogeneous but should compute all corrections anew. Otherwise the solution may drift away from the truth because of repeated homogenization.

Sometimes, even after a second (or third) round of homogenization, some stations will still show anomalous trends relative to other stations. This may occur because of changes that gradually affect the local climate over an extended period (for example, the site being encroached upon by an expanding urban area, or increasing levels of irrigated agriculture in a district), or because of natural local effects (for example, a coastal site being cooled by increased levels of coastal upwelling in the nearby ocean). If it can be established with a reasonable level of confidence that such anomalous trends have a specific non-climatic cause (such as urbanization), one option is to remove them from the dataset or to exclude them from some products based on the dataset (for example, not including stations influenced by urbanization in assessments of long-term climate change).

2.8 Documentation

When an adjusted dataset is produced, the adjustments should be properly documented and published. Such documentation should include the dates covered by the adjustment, whether the inhomogeneity was identified through statistical methods or metadata and, if possible, what the likely cause of the adjustment was. Summary statistics of the influence of adjustments (all and per metadata category) can help the user assess the quality of the data. A best practice is to share both the raw and homogeneous data as well as the metadata on the identified breaks.

The methods used in the development of any homogenized dataset should be properly documented in an accessible form. This optimally includes an Open Access paper published in the peer-reviewed scientific literature. The paper should contain at least a detailed description of the methods used, available in the same location as the homogenized data themselves. Best practice is to write a clear well-documented code, bearing in mind that it will be published alongside the data.

2.9 Operational maintenance of a homogenized dataset

The initial development of a homogenized dataset is a substantial undertaking. Normally, one of the main purposes of homogenized datasets is to serve as the underlying dataset for products (for example, a national or global temperature anomaly), which means that for those products to continue to be updated, the underlying dataset needs to continue to be updated too. An advantage of automatic homogenization methods is that they can be easily applied when new data comes in.

Homogenized datasets tend to be constructed in such a way that the most recent data are unadjusted. This allows the dataset to be updated by appending new data without further adjustment. An exception to this may be when the older data are considered to be a more reliable long-term reference; for example, in a precipitation network, if the bulk of the network is manual but a small number of automatic stations are being introduced, it may be more appropriate to adjust the automatic stations so that their data are equivalent to the earlier manual data for better spatial consistency of the network as a whole. Another reason to adjust to earlier manual observations is that they are often of higher quality and more accurate than automated observations.

Over time, a homogenized dataset will become out of date. There are two major factors that contribute to this. Firstly, some stations which are part of the original dataset will close over time (sometimes replaced in the network by new stations nearby, which can be used as the basis for a composite). Secondly, new inhomogeneities may occur at stations that remain in the dataset.

It is recommended that a reassessment be undertaken of any homogenized dataset at least every five years. This reassessment should include:

- A check of the status of all stations in the existing dataset and, if they are closed or no longer reporting reliably, whether they can be replaced with an alternative station that can become part of a composite record;

- A search of recent metadata (covering the period since the last update) for all stations in the dataset;

- Incorporation of any relevant historical data that have become available (for example, data recently digitized as part of data rescue activities);

- In the case of manual methods, at least a statistical testing for inhomogeneities in the most recent part of the record should be made. This includes a reassessment of the last few years of the previous version of the dataset, as inhomogeneities in the last (or first) few years of a time series are difficult to detect and quantify and the additional new observations may allow more reliable assessments to be made. Especially with automatic methods a full homogenization of the full dataset is recommended.

2.10 Network-wide issues and options for dealing with them

On occasions, changes will occur that affect all stations in a national network or a substantial proportion of those stations, at the same time or over a period of years. Examples of such changes include:

- A change in observation time either explicit (such as the change in the observing period for daily data from 00.00–00.00 UTC to 06.00–06.00 UTC in Canada in 1961) or implicit (for example, Australian stations continued to observe at the same local clock time when daylight saving time was introduced in the early 1970s, introducing an effective one-hour shift in standard observation time during the summer);

- A major change in instrument type such as a change in standard thermometer screens (for example, in the transition to AWSs), or the introduction of a new type of radiosonde (upper-air observations are particularly susceptible to this type of change, as they do not involve much fixed infrastructure, hence it is possible for any change to be implemented rather quickly);

- A change in observation procedures or definitions such as a change in the unit of cloud amount measurement from 1/10 to 1/8, or a change of units (for example, from Fahrenheit to Celsius);

- A change in algorithms used for data analysis – such as a change in the definition of daily mean temperature (for example, from the mean of eight 3-hourly observations to the mean of the daily maximum and minimum temperature).

Network-wide changes can be particularly challenging to deal with in a homogenization process. Since they normally apply to most or all stations in a particular region, the use of reference stations from the same network will be of limited use both in detecting the inhomogeneity and in determining its likely impact. In addition, a change whose impact may not be significant or detectable at an individual station (for example, a 0.2 °C temperature inhomogeneity) may be significant in a national mean, if it affects all stations in a network, or in the global mean, if it represents a widespread technological or organizational change.

Some possible strategies for addressing such changes include:

- Changes affecting the entire network are not always mentioned in metadata databases, which typically document known changes to individual stations. On the other hand, because they are important national events, they are often mentioned in annual reports;

- If a change affects most, but not all, of the stations in a network (for example, it affects all automatic stations, but not manual ones), compare the affected and unaffected stations;

- Compare with observations near the border in neighbouring countries that are unaffected by the change (this is only effective where there are such stations; it normally requires a land border and access to observations from other countries, which may not always be easy to obtain);

- Compare with another data type that was not affected by the change. For example, compare temperatures at the surface with radiosonde temperatures at 850 hPa (or reanalysis fields based on those), or compare measured wind speeds with geostrophic winds derived from mean sea-level pressure fields;

- Use alternative data to indicate the possible impact of the change. For example, for a historical change in observation time, even though high-resolution sub-daily data from the period when the change took place may not be available, it may be possible to use high-resolution data from recent years to estimate what the impact of a past observation time change may have been. For example, Vincent et al. (2009 and 2012) used hourly temperature data to correct a bias in daily minimum temperatures caused by a change in observing time across Canada. In Austria (Böhm et al., 2010) and Spain (Brunet et al., 2006) parallel experiments were performed to quantify the changes due to the transition to Stevenson screens.

Such methods can produce reasonably coarse results, and it may only be possible to quantify the impact of such inhomogeneities at a national or regional level (or to determine that a change had no significant impact at that level), without fully accounting for different impacts that any inhomogeneity may have had at individual stations.

Network-wide inhomogeneities can be solved in different ways depending on the resources available. The first version of the US Historical Climate Network dataset contained a physical adjustment for the transition from Stevenson screens (Cotton Region shelters) to Maximum-Minimum Temperature System (MMTS) AWS systems. This transition happened in many stations over a short period of time and data were consequently difficult to homogenize using statistical homogenization, comparing a candidate station with a composite reference (the average signal over several neighbours), because neighbouring stations were often also affected. The adjustments were based on estimates of those candidates that had neighbouring stations without the transition. After designing a new homogenization method, making pairwise comparisons, it became possible to deal with this difficult situation and the physical adjustments were no longer applied.

2.11 Specific challenges for multinational datasets

Combining data from different networks and countries has its advantages as it reduces problems with network-wide inhomogeneities. However, developing a homogeneous dataset at a global or regional (multinational) scale also presents certain challenges, including:

- **Large dataset size**. Such datasets will often contain information from hundreds or thousands of stations. These will normally be beyond the practical size limit for manual homogenization methods, requiring the use of automated methods or automation-assisted manual methods;

- **Limited access to metadata**. As metadata are normally archived at the national level, with only the most basic metadata exchanged internationally, there is often limited or no capacity to incorporate metadata into the homogenization process for multinational datasets (even if metadata can be accessed, their use will often require interpretation of documents in the local language). Sometimes, even determining exactly which station the data come from can be challenging, with a major task in global dataset development being identifying and consolidating duplicate datasets from different sources (for example, Rennie et al., 2014). The sharing of metadata will hopefully improve with the WMO Observing Systems Capability Analysis and Review (OSCAR) Tool.

– **Limited access to potential reference series and other relevant data**. Multinational datasets usually consist of data from selected climate stations only. In contrast, in national datasets, it is often possible to draw on additional data (for example, shorter-term stations, which have too little data to be considered for a long-term dataset but can still be useful as reference stations in specific subperiods), as well as data from extra meteorological elements and sub-daily data, while global datasets are typically single-element datasets. The European Centre for Medium-range Weather Forecasts (ECMWF) is working on a multi-element database where observations of several meteorological elements from one station are kept together (Dunn and Thorne, 2017).

Consequently, automated homogenization methods are used for such datasets, and correlations between candidate and reference series will be lower, making it more difficult to detect smaller inhomogeneities and increasing the uncertainty of the adjustments that are made.

One option is to draw on national-level information to the extent possible. In the HadCRUT global temperature dataset, maintained by the University of East Anglia and the UK Met Office, national-level homogenized datasets are used where they are available (Jones et al., 2012). Xu et al. (2017) built on and expanded several homogenized national datasets to develop a homogenized global dataset.

2.12 Conclusion: Good practices for homogenization

Chapter 2 outlined a range of issues and considerations in the development of long-term homogeneous datasets. The extent to which these can be implemented will vary widely, depending on the access dataset developers have to relevant data and metadata, the tools and computer systems available to them, the support they can get, the density of the underlying observation network and the size of the dataset under consideration.

There are a number of principles which can be considered as good practice for data homogenization, including:

1. Data homogenization is applied most effectively through a combination of statistical methods and metadata (for example, Yosef et al., 2018). If this is not possible (for example, because metadata are not available or because a lack of reference series makes statistical homogenization difficult), homogenization is likely to be less effective.

2. Statistical homogenization should always be applied; it cannot be assumed that metadata are perfect.

3. If there are known issues with a dataset (for example, a network-wide change of observation time), these should be dealt with before more station-specific inhomogeneities are considered. Usually, the use of reference series directly or in pairwise comparison will not identify network-wide changes.

4. Reference series should be used in statistical homogenization methods if at all possible.

5. It is important to try to obtain an SNR higher than one and it is worthwhile to maximize the SNR further.

6. Once a draft version of a homogenized dataset has been prepared, methods that assume the composite reference to be homogeneous need a second round of homogenization, as discussed in section 2.7 above.

7. Homogenized datasets should be fully updated at least every five years.

8. Where major changes are anticipated, parallel observations should be set up and performed for at least two years. A useful document on managing network changes can be found in Brandsma et al. (2019).

9. When an adjusted dataset is produced, the adjustments should be properly documented and published.

10. The methods used in the development of any homogenized dataset should be properly documented in an accessible form.

While no method can guarantee a fully homogeneous dataset, and any homogenized dataset will have some level of uncertainty associated with the adjustments involved in creating it, following these principles should maximize the likelihood that a dataset will be sufficiently homogeneous to be effectively used for the development of long-term climate data products.

CHAPTER 3. SELECTING STATISTICAL HOMOGENIZATION SOFTWARE

This chapter aims to guide the reader through the numerical methods and software packages that can be used for various homogenization tasks.

3.1 Statistical homogenization packages

The list and table below describe the publicly available homogenization packages (in alphabetical order) used in climatology at the time of writing this publication. New methods are regularly being developed and the list below should not be considered exhaustive. Other methods are described in the scientific literature but are not included here as their software has not yet been released in a form that is usable by the broader community. At the time of writing this publication, the table was kept up to date at http://www.climatol.eu/tt-hom/.

Many of the descriptions below refer to a benchmarking study of homogenization methods that was part of the HOME project (Venema et al., 2012). Section 3.2.2 contains further discussion of this study.

ACMANT is a homogenization package for temperature and precipitation data. It is one of the most accurate automatic methods for homogenizing temperature networks without metadata.

AnClim implements all common detection and correction methods with one graphical user interface. The AnClim contribution to the HOME benchmark, using an ensemble approach that includes many methods and settings, was not outstanding. The package, however, offers access to many methods that can also be used by themselves in the more standard way. AnClim, together with ProClimDB (not free), helps automatize many database-related tasks.

Bayesian MDL is a multiple-breakpoint method that is freely available but still at the stage of research. It is mentioned for completeness by developers of homogenization methods (Li et al., ArXiv 2017).

Berkeley Earth is a homogenization and interpolation method used for global temperature datasets. The homogenization corrections are computed in the interpolation part. Thus, it does not return station data but a field or an estimate of the regional climate at the location of the station.

Climatol applies the Standard Normal Homogeneity Test (SNHT) to split the series into homogeneous subperiods. It computes a full series for every homogeneous subperiod. In its final stage, all missing data are estimated from other subperiods of the same station (when available) or from neighbouring stations. Climatol is one of the packages that is most tolerant of data gaps and one which can make use of available metadata.

GAHMDI solves the multiple-breakpoint problem with a global search algorithm (genetic algorithm). It is packaged together with HOMAD, a method to correct the distribution of daily data. No intercomparison or documentation is available beyond two published articles (Toreti et al., 2010 and 2012).

GSIMCLI uses geostatistical methods to compute the null hypothesis numerically using the Monte Carlo approach. It is a new method and there is no information from intercomparison studies on its performance. It provides a graphical user interface and automation for networks.

HOMER was designed as part of the COST[1] Action HOME, but its performance was not benchmarked. It implements several multiple-breakpoint methods. Using the pairwise option, it is the successor of the multiple-breakpoint method PRODIGE and thus expected to be one

[1] European Cooperation in Science and Technology (COST). A funding programme to stimulate collaboration in Europe.

of the best manual methods. It includes the ANOVA correction method, which is likely the most accurate correction method available. The joint detection option (R package ‘cghseg’ implemented in HOMER) is best not used alone (see Gubler et al., 2017).

iCraddock implements the Craddock test in a pairwise fashion. This manual method is subjective but performs well for small networks and is recommended by HOME. It can also be used for daily data (Brugnara et al., 2012).

MASH is an automatic homogenization algorithm based on hypothesis testing that is designed to work with inhomogeneous references and uses a multiple-breakpoint approach. It is excellent.

PHA, or pairwise homogenization algorithm, is used by NOAA to homogenize its national (U.S. Historical Climatology Network (USHCN)) and global (Global Historical Climatology Network (GHCN)) temperature datasets. It is highly robust and recommended for large datasets. It can use metadata.

ReDistribution Test is a single-breakpoint test for vector wind.

RHtests implements several break detection tests taking into account autocorrelations and the distance from the edges. RHtests and Anclim are the only methods in this list that have the option of homogenizing without a reference. The reference series, where used, must be computed by the operator. It includes a test for documented breakpoints. Note that these tests are not designed for use with a fully automatic procedure; an analysis of the automated detection results is required to finalize the results.

SNHT refers to an R package in the Comprehensive R Archive Network (CRAN), which implements the well-known Standard Normal Homogeneity Test in a way modified by Haimberger (2007) and in the pairwise approach used by Menne and Williams (2009).

Table 1. Overview of the characteristics of homogenization packages

Package	*Resolution detection*[a]	*Detection method*	*Reference use*[b]	*Resolution correction*[c]	*Correction method*	*Primary operation*	*Metadata use*	*Variable*[d]	*Documentation*[e]	*Reference*
ACMANT[f]	Year, month	Multiple breakpoints	Composite	Year, month, [day]	Joint (ANOVA)	Automatic	No	Any	User guide	Domonkos and Coll (2017)
AnClim	Year, month	Many	Composite, pairwise	Year, month, day	Several	Interactive, automatic	Yes	Any	Manuals	Štěpánek et al. (2009)
Berkeley Earth	Month	Splitting	Composite	n/a	n/a	Automatic	Yes	T	Article	Rohde et al. (2013)
Climatol	Month (serial), day	Splitting	Composite	Month (serial), day	Missing data filling	Automatic	Yes	Any	Manual and user guide	Guijarro (2018)
GAHMDI HOMAD	Month (serial), day	Multiple breakpoints	Selection	Day	Higher-order moment method	Automatic	Yes	T	None	Toreti et al. (2010, 2012)
GSIMCLI	Year, month	Multiple breakpoints	Composite	See footnote[g]	See footnote[g]	Automatic and interactive	No	T, p	Manuals	Ribeiro et al. (2017), Costa and Soares (2009)
HOMER	Year, season, month	Multiple breakpoints	Pairwise, joint	Year, month	Joint (ANOVA)	Interactive	Yes	Any	Basic user guide+courses	Mestre et al. (2013)
iCraddock	Year, season, month	Splitting	Pairwise	Year, season, month, day	Daily: smoothed monthly corrections	Interactive	Yes	Any	None	Craddock (1979), Brunetti et al. (2006)
MASH	Year, season, month	Multiple breakpoints	Composite	Month, [day]	Multiple comparisons	Automatic and interactive	Yes	Any	User guide	Szentimrey (2008, 2014)
ReDistribution Test	Readings	Single breakpoints	No reference	n/a	n/a	Interactive	No (but it is interactive)	Wind	None	Petrovic (2004)
RHtests	Year, month, day	Splitting	Selection or no reference	Year, month, day	Multi-phase regression	Interactive	Yes	Any	User guide+courses	Wang (2008a and b), Wang and Feng (2013)

Package	*Resolution detection*[a]	*Detection method*	*Reference use*[b]	*Resolution correction*[c]	*Correction method*	*Primary operation*	*Metadata use*	*Variable*[d]	*Documentation*[e]	*Reference*
R package SNHT	Year, month	Splitting	Composite and pairwise	Month	Composite and multiple comparisons	Automatic	No	T	Help files	Haimberger (2007), Menne and Williams, (2009)
PHA	Year	Splitting	Pairwise	Year, [month]	Multiple comparisons	Automatic	Yes	T	Plain text notes	Menne et al. (2009)

[a] If not noted otherwise in the column Resolution, "month" means detection is on multiple monthly series in parallel;

[b] Options are: Operator reference **selection**, averaging (**composite** reference), removing references with breaks from composite (**no reference**), **pairwise**, and **joint** detection;

[c] Square brackets mean that the resolution is supported by the software, but the corrections are not computed at that resolution;

[d] Options are: T = temperature; p = precipitation; any = Gaussian and log-normal distribution or additive and multiplicative models;

[e] A user guide is limited to a few pages and shorter than a manual;

[f] ACMANT can detect breaks in both annual averages and seasonal cycle in parallel;

[g] The corrections are computed at the resolution of the data (annual or monthly series). Corrections are applied to a metric computed by GSIMCLI (user-defined: percentile, mean or median) of the probability density function (pdf) of the candidate station, which is estimated using composite references.

3.2 Performance of statistical homogenization methods

There are two indications as to whether homogenization has improved a dataset or not. First, the breaks found should fit the breaks known from the station history. Second, no more breaks would be detected in the dataset if it were homogenized again, and the results should be regionally coherent, physically consistent and climatologically plausible (see section 2.7 on data review). However, these assessments do not prevent setting too many breaks and removing real regional climate variability (over-homogenization). In addition, these indications are not accurate enough to help select the best statistical homogenization method. Therefore, we mostly rely on what is known about the performance of these methods from the scientific literature for general cases to select appropriate homogenization methods.

There are two lines of evidence on the general performance of statistical homogenization methods: theoretical principles (section 3.2.1) and numerical studies (section 3.2.2). They strengthen each other and both are needed to gain confidence. The design principle of a homogenization method may be theoretically sound but implementation details matter, and the method may still perform poorly in a numerical comparison. Conversely, numerical studies only test specific scenarios, which may not be realistic for the task at hand, and our understanding helps us to figure out what is important and realistic.

3.2.1 *Theoretical principles*

If a normally distributed uncorrelated difference (candidate minus reference) series contains one break at a known date, the appropriate test is a simple t-test for the difference in the mean before and after the break. If the same series is known to contain only one break at an unknown position, multiple testing needs to be considered, and the appropriate test is the Standard Normal Homogeneity Test (SNHT) (Alexandersson, 1986) or the Penalized Maximal *t* Test (PMT) (Wang et al., 2007).

However, climate series typically contain more than one break, and the reference series may also contain breaks, which should not be falsely attributed to the candidate series. How statistical homogenization methods solve these two problems seems to be the main determinant of the performance of statistical homogenization methods.

There are three ways to detect multiple breakpoints in one series:

1. Sometimes, single-breakpoint tests are performed over moving windows. However, to reduce the probability that there are not multiple breakpoints in the window its length needs to be short. This makes the method less sensitive and this approach is not much used.[2]

2. Traditionally, single-breakpoint methods are used, and the series is split at the most significant breakpoint, after which the two new series are tested again (hierarchical splitting and variants thereof).

3. Modern multiple-breakpoint methods effectively test all possible multiple break combinations.

Theoretically, multiple-breakpoint methods are the most accurate way to solve this problem.

[2] The moving-window detection method can be useful when you want to remove only clear breaks, not gradual inhomogeneities, for example, to study the gradual (nonlinear) warming due to urbanization one may want to remove only the effect of relocations (Zhang et al., 2014).

There are also several approaches to handling inhomogeneities in the reference series:

1. Averaging over a large number of reference stations. This removes large obvious jumps in the composite reference series, but it often happens that a large part of a network experiences a similar transition over a few years or decades. The bias due to this transition would also largely be in such a composite reference and may reduce detection power.

2. Selecting reference stations without a break around the breakpoint in the candidate station or, alternatively, correcting the breaks in the references before using them. These approaches must be used iteratively, as the breaks in the references must be known. Using previously homogenized data is potentially dangerous. Therefore, such approaches must be validated with particular care in case they remove the large-scale biases.

3. Detecting breaks on pairs of stations. Here the reference station is not assumed to be homogeneous, with breaks in both stations being detected as breaks in the difference series between them. A second "attribution" step is necessary to determine which of the breaks detected in the pairs belong to which station.

4. Joint detection of all breaks in a network of multiple stations simultaneously. This is a complicated combinatorial puzzle and computationally more demanding than the other methods.

Joint detection is theoretically the optimal solution. However, at the moment, this method is used only in the homogenization package HOMER, which does not work well. So, in practice, methods 2 and 3 are preferred.

A modern joint correction method (often called ANOVA) has been developed by Caussinus and Mestre (2004). This method disaggregates a network of stations into one regional climate signal, a step function for every station to model the inhomogeneities and noise. Corrections are computed by minimizing the noise. The corrections of this methods are unbiased if all breaks are correctly identified (Lindau and Venema, 2018*b*). Thus, all breaks should be included in the adjustment process, including those that are close to each other. This method has been shown to produce more accurate results than traditional methods for the dense European networks simulated in HOME (Domonkos et al., 2013).

3.2.2 ***Numerical studies***

Traditionally, validation studies have focussed on break detection scores of the detection methods. This can help understand how the algorithm works, but it is not clear what the optimal compromise for climatological analysis of the homogenized data between the hit rate and the false alarm rate is. More recent work has included error measures, such as the root mean square error and the remaining uncertainty in the trend after homogenization, which assess the performance of entire homogenization methods and are of relevance to climatological users of the data (Domonkos, 2011; Venema et al., 2012; Williams et al., 2012).

To some extent, the results of validation studies will depend on the metric(s) used for evaluation. The way in which homogenization is carried out may also influence the results (for example, whether metadata were used, how the algorithms were operated and how well-trained the operator was). In the case of manual and semi-automatic methods, clear differences were found among operators (Venema et al., 2012), so a clear distinction should be made between the homogenization method/package and the validated homogenized datasets (contributions). Especially for RHtests, which implements several detection tests and correction methods and which can be used in many ways, the results of a validation study may not be representative.

Williams et al. (2012) only studied the pairwise homogenization algorithm, Domonkos (2011) compared a large number of automatic homogenization algorithms, while the COST Action HOME (Venema et al., 2012) included nearly all state-of-the-art and most-used methods, including several manual ones. It should be borne in mind that these three studies were carried out for dense networks, so the performance of the methods will be lower for sparser networks. Moreover, the ranking of the methods could be different for other networks.

These numerical results support the idea that algorithms designed to solve problems related to multiple breakpoints and inhomogeneous references can obtain accuracies that are clearly higher than traditional methods. HOME recommended the following algorithms: ACMANT, iCraddock, MASH and PRODIGE and, for large networks, PHA. Again, it should be noted that the recommendation/conclusion could have been different if the study had been done in a different way with a different benchmarking dataset. In particular, the HOME results are not representative of the whole RHtestsV3 package. Actually, no benchmark study that applies methods in a fully automatic procedure can evaluate the whole RHtests because this package was designed for interactive operation, including manual or automation-assisted human intervention.

The above validation studies involved the generation of an artificial station network where the inhomogeneities were known. The HOME benchmark dataset aimed to model networks for temperature and precipitation for Europe. Its validation data are quite realistic, but the break variance is about two times too high. The study did not include explicit large-scale trend biases; they were thus small and difficult to correct. Furthermore, its high station density is not representative of the early instrumental period or sparse networks such as those of northern European or developing countries. An upcoming benchmarking study of the International Surface Temperature Initiative (ISTI; Thorne et al., 2011) aims to resolve these issues (Willett et al., 2014). New validation results are being produced in the framework of the MULTITEST project (http://www.climatol.eu/MULTITEST/) at the time of writing these guidelines.

In the HOME benchmark, all contributions made the temperature data more homogeneous, except for the contribution using absolute homogenization. This illustrates how dangerous absolute methods can be, especially when they are used with a fully automatic procedure as in the benchmark study (see section 2.5). But it should be noted that to make the benchmark blind, the regional climate signal used in the HOME dataset was more variable than usual and that this dataset was thus more difficult than a real observing dataset case for absolute homogenization. Nevertheless, absolute homogenization is generally less reliable and should be used with extra caution and never automatically. For precipitation data, only the best methods were able to improve the homogeneity.

Validation studies can also be based on high-quality homogenized data. For example, in Gubler et al. (2017) information on inhomogeneities based on data homogenized with high station density was used to study the performance of homogenization methods applied to a thinned sparser network. The advantage of this procedure is that the inhomogeneities are by definition realistic. The problem is that even with a high station density homogenization will not be perfect. These kinds of study provide synergy with those using simulated data. Gubler et al. (2017) studied four different ways to operate HOMER and found that HOMER using joint detection should not be used by itself. The way in which metadata was used in this study did not improve the results. The breaks due to the transition to AWSs were often not detected in the sparse network, while such historical transitions are supposed to be the strength of such pairwise methods.

Kuglitsch et al. (2012) also used homogenized Swiss data to validate the PRODIGE and Toreti homogenization methods and the FindU.wRef function of RHtestsV3. The results indicate that PRODIGE detects more breaks that can be confirmed by metadata, but that it also has a much higher false alarm rate than RHtestsV3 or the Toreti method: PRODIGE found 1140 breaks (of which 515 are confirmed by metadata, 45.2%), while the RHtestsV3 and Toreti method found respectively 438 and 683 breaks, 72.4% and 70.3% of which are confirmed by metadata. However, only break detection accuracy was studied; the study did not evaluate the adjustment methods, nor the entire homogenization procedure (that is, detection and adjustment combined). Thus, the effect of overestimating the number of breaks or failure to detect real breaks on the homogenization results is unclear.

3.3 Automated and manual methods

Homogenization methods may be fully automated, and referred to as automatic methods (that is, no human intervention is required beyond the selection of the dataset), or they may involve some level of manual intervention, and be referred to as manual methods. Areas that may involve human judgement in manual methods include:

- The selection of reference stations to be used (whereas an automatic method may use, for example, a purely distance-based or correlation-based criterion);
- Merging information from statistical methods and metadata;
- Determining which inhomogeneities identified by a statistical method should be retained;
- Determining which period to use for a comparison (for example, not using a climatically anomalous year or a less reliable data period in comparing two sites).

Both automatic and manual methods have been used successfully in many countries. Manual methods do have the advantage of allowing the introduction of information about a station that may not be easily quantifiable (for example, known gradual changes in the local site environment), and also of allowing readier identification of anomalous results at individual stations (although the risk of anomalous results is reduced if some of the practices described in section 2.7 on data review are implemented). Many manual methods can be made fully automated by automation of the related human intervention procedure.

However, manual methods do have the disadvantage of being labour-intensive and requiring expert judgement, which may not be available in some cases. Furthermore, automatic methods can be better validated because it is easy to compute many cases and settings. This promotes faster improvements in the capabilities of automatic methods. This also means that we have more reliable estimates of the uncertainties. Additionally, removing uncertainties due to human factors increases the likelihood of achieving the accuracy expected from validation studies.

Automated or semi-automated methods are recommended in cases that do not necessarily require human judgement (see the four bullet points above) and for operators with limited experience in homogenization. They are also the only practical option for very large datasets, especially global or regional datasets.

3.4 Use cases

The performance of the homogenization methods mentioned in the previous section is an important consideration, which will clearly influence the quality of your homogenized data, although most of the methods listed in this chapter will at least improve temperature data under most conditions. This section presents a range of use cases to illustrate how to weight various criteria and select a package that fits a given task. A further consideration is whether the package can handle large amounts of metadata. The size of the network is important: the larger it is the more automatic methods are preferred. The station density is also an important factor: it is easier to carry out homogenization for high-density networks than for medium- to low-density networks.

Moreover, the availability of local expertise or training opportunities are a reason to prefer a specific method; for the HOME benchmark, clear differences were found among contributions from different operators using the same method, which may be related to experience with the method. Automatic methods are less influenced by the availability of such expertise and are recommended for less experienced users. But absolute homogenization methods should never be applied automatically.

When the significance of documented breaks (breaks at known times) has be determined, the RHtests package can be used: it is the only package that can test both documented and undocumented breaks in tandem. If a single breakpoint in a difference time series at a known time has to be checked, the t-test implemented in many computational science packages can be used.

When the network is too sparse and there are effectively no references that can be used (including the availability of reanalysis or proxy series), the only option is to use absolute homogenization. Such methods are implemented in RHttests and AnClim software.

For a small network of fewer than 50–100 stations with few data gaps, some knowledge of the station history and sufficient time to devote to obtaining a good dataset, both the pairwise method in HOMER and iCraddock are good options. Advantages of HOMER over iCraddock are that the positions of the breaks in the pairs are determined objectively, which also speeds up the task, and HOMER natively supports joint correction. If the network is sparse and the references may have slightly different climate signals, iCraddock is a good option because the operator can assess graphically whether any differences are climatologically to be expected or give reason to suspect an inhomogeneity. When interactive options are not used, ACMANT and MASH can also work with small networks. In such cases the operator carefully selects the series to be used for homogenization. For very large networks (hundreds or thousands of stations) this becomes cumbersome. The size of the network is a relatively flexible consideration: the developer of iCraddock homogenized a dataset with about 700 series on the basis of his experience.

For midsized networks (more than 100 stations), the automatic methods ACMANT, Climatol and MASH are attractive options. MASH can handle up to 500 stations, ACMANT v.4 up to 5 000, and Climatol's network size is limited only by the available memory in the computer (some thousands of stations in practice). Climatol and MASH (if the volume of missing data is not too big) would be preferred for making use of available metadata.

The Pairwise Homogenization Algorithm was made for continental and global datasets and can also handle messy datasets with short series and missing data. Such would be the main selection considerations for this method. With some automated human intervention, the RHtests is the only other package that has been used to homogenize a global dataset (Xu et al., 2017).

CHAPTER 4. HISTORY OF HOMOGENIZATION

The implications of inhomogeneities for data analyses have long been recognized and homogenization has a long history. In September 1873, at the International Meteorological Congress in Vienna, Carl Jelinek requested information on national multi-annual data series (k.k. (kaiserlich-königlich) Hof- und Staatsdruckerei, 1873). Decades later, at the international conference for directors of meteorological services in 1905, G. Hellmann (k.k. Zentralanstalt für Meteorologie und Geodynamik, 1906) still regretted the absence of homogeneous climatological time series due to changes in the surrounding of stations and to new instruments, and pleaded for stations with a long record, *Säkularstationen* (centennial observing stations), to be kept as homogeneous as possible. Although this Conference recommended maintaining a sufficient number of stations under unchanged conditions, today these basic inhomogeneity problems still exist.

4.1 Detection and adjustment

In early days, documented change points were removed with the help of parallel measurements. Differing observing times at the astronomical observatory of the k.k. University of Vienna (Austria) were adjusted by using multi-annual 24-hour measurements of the astronomical observatory of the k.k. University of Prague (today Czech Republic). Measurements of Milano (Italy), between 1763 and 1834, were adjusted to 24-hour means by using measurements of Padova (Kreil, 1854*a* and *b*).

In the early twentieth century, Conrad (1925) applied and evaluated the Heidke criterion (Heidke, 1923) using ratios of two precipitation series. As a consequence, he recommended the use of additional criteria to test the homogeneity of series dealing with the succession and alternation of algebraic signs: the Helmert criterion (Helmert, 1907) and the painstaking Abbe criterion (Conrad and Schreier, 1927). The Helmert criterion for pairs of stations and Abbe criterion were still described as appropriate tools in the 1940s (Conrad 1944). Some years later, the double-mass principle was popularized for break detection (Kohler, 1949).

4.2 Reference series

Julius Hann (1880) studied the variability of absolute precipitation amounts and ratios between stations. He used these ratios for quality control. This inspired Brückner (1890) to check precipitation data for inhomogeneities by comparison with neighbouring stations; he did not use any statistics.

In their book, *Methods in Climatology*, Conrad and Pollak (1950) formalized this relative homogenization approach, which is now the dominant method for detecting and removing the effects of artificial changes. The building of reference series, by averaging the data from many stations in a relatively small geographical area, was subsequently recommended by the WMO Working Group on Climatic Fluctuations (WMO, 1966).

Papers by Alexandersson (1986) and Alexandersson and Moberg (1997) made the Standard Normal Homogeneity Test (SNHT) popular. The broad adoption of this test was accompanied by clear guidance on how to use the test together with references to homogenize station data.

4.3 Modern developments

The Standard Normal Homogeneity Test is a single-breakpoint method, but climate series typically contain more than one break. Thus, a major step forward was the design of methods specifically conceived to detect and correct multiple change-points and work with inhomogeneous references (Szentimrey, 1999; Mestre, 1999; Caussinus and Mestre, 2004). These types of method were shown to be more accurate by the benchmarking study of the EU COST Action HOME (Venema et al., 2012).

A paper by Caussinus and Mestre (2004) also provided the first description of a method that corrects all series of a network simultaneously. This joint correction method improved the accuracy of all but one contribution to the HOME benchmark which was not yet using this approach (Domonkos et al., 2013).

The ongoing work to create appropriate datasets for climate variability and change studies promoted the continuous development of better methods for change-point detection and correction. To enhance this process the Hungarian Meteorological Service started a series of seminars on homogenization in 1996 (Hungarian Meteorological Service, 1996; WMO, 1999; OMSZ, 2001; WMO, 2004; WMO, 2006; WMO, 2010; WMO, 2011; WMO, 2014; OMSZ, 2017).

CHAPTER 5. THEORETICAL BACKGROUND OF HOMOGENIZATION

This chapter considers some theoretical aspects of monthly time series homogenization. It revisits many homogenization problems mentioned in earlier chapters be. This chapter is intended for users that are interested in a better analytical understanding and for scientists who would like to develop their own methods.

In practice, monthly series are homogenized mostly as means. The aim of these homogenization procedures is to detect inhomogeneities of the mean and to adjust the series.

For the detection of inhomogeneities of monthly data and their adjustment, solutions are needed for the following mathematical problems:

- Statistical spatio-temporal modelling of the series;
- Methodology for comparison of candidate and reference series;
- Breakpoint (change point) and outlier detection;
- Methodology for adjustment of series.

This chapter will, furthermore, discuss the following topics:

- Quality control procedures;
- Missing data completion;
- Usage of metadata;
- Manual versus automatic methods;
- Evaluation of methods (theoretical, benchmark).

5.1 General structure of the additive spatio-temporal models

The statistical spatio-temporal modelling of a series is a fundamental question. Adequate comparison, breakpoint detection and adjustment procedures depend on the statistical model chosen. If the data series are normally distributed (for example, temperature), the additive spatio-temporal model can be used. The general form of this additive model for the monthly series of several stations in a small climate region can be written as follows,

$$X_{j,m}(t) = \mu_m(t) + S_{j,m} + IH_{j,m}(t) + \varepsilon_{j,m}(t), \quad (j = 1,2,\ldots,N;\ m = 1,2,\ldots,12;\ t = 1,2,\ldots,n), \qquad (1)$$

where

$j = 1,2,\ldots,N$	station index
$m = 1,2,\ldots,12$	month index
$t = 1,2,\ldots,n$	year
$\mu_m(t)$	common and unknown climate change signal or temporal expected values or temporal trend of the stations
$S_{j,m}$	spatial expected values or spatial trend of the stations
$IH_{j,m}(t)$	inhomogeneity signals with type of 'step-like function', generally assumed with unknown breakpoints T and shifts $IH_{j,m}(T) - IH_{j,m}(T+1) \neq 0$, and $IH_{j,m}(n) = 0$.

Consequently, the expected values or means are:

$$E(X_{j,m}(t)) = \mu_m(t) + S_{j,m} + IH_{j,m}(t), \quad (j = 1,2,\ ,N;\ m = 1,2,\ldots,12;\ t = 1,2,\ldots,n),$$

and superimposed on these means are normal noise series, written in vector form:

$$\boldsymbol{\varepsilon}_m(t) = [\varepsilon_{1,m}(t),\ldots..,\ \varepsilon_{N,m}(t)]^{\mathrm{T}}\ \ N(\mathbf{0},\boldsymbol{C}_m), \quad (m = 1,2,\ldots,12;\ t = 1,2,\ldots,n).$$

The matrices $\boldsymbol{C}_m(m = 1,2,...,12)$ include the spatial covariances between the stations and they are assumed to be without any climate change or inhomogeneity over the years. This is because the existing methods developed for homogenization of monthly series assume that there is no climate change or inhomogeneity in the higher order moments.

In general, the vector noise terms $\boldsymbol{\varepsilon}_m(t)$ also have some temporal autocorrelations, this issue is detailed below in Remark 1 (items 1.1 and 1.2).

The aim of homogenization is to detect the inhomogeneity signal $IH_{j,m}(t)$ and to adjust the original raw monthly series $X_{j,m}(t)$, that is:

$$X_{H,j,m}(t) = X_{j,m}(t) - I\hat{H}_{j,m}(t) \quad (j = 1,2,...,N;\ m = 1,2,...,12;\ t = 1,2,...,n),$$

where $I\hat{H}_{j,m}(t)$ is the estimated inhomogeneity signal.

Remark 1

In practice there are some differences between absolute and relative methods.

Absolute homogenization: only one station data series is used: $N = 1$. The main problem of absolute methods is that it is essentially impossible to separate the unknown climate change and variability signal $\mu_m(t)$ from the inhomogeneity $IH_{j,m}(t)$ without additional information on the variability of the climate signal, for which homogeneous data would be needed.

Relative homogenization: data series from more than one station are used and compared to each other, that is, $N > 1$. The comparison makes it possible to filter out the common unknown climate change and variability signal $\mu_m(t)$.

The two basic strategies for solving the problem of homogenization are:

1.1 The monthly series (Equation 1 above) are examined serially as one time series in chronological order. The specific problems raised by this type of examination are:

- The annual cycle or seasonality of $\mu_m(t)$, $E_{j,m}$, $IH_{j,m}(t)$, $\boldsymbol{C}_m(m = 1,2,...,12)$ the last covariances implicitly include the standard deviations and spatial correlations;
- The temporal autocorrelation between the elements of adjacent months.

1.2 The monthly, seasonal and annual series may be examined independently in parallel. Then the series examined may be:

- The derived monthly series in each calendar month $(m = 1,2,...,12)$ separately, $X_{j,m}(t) = \mu_m(t) + S_{j,m}+ IH_{j,m}(t) + \varepsilon_{j,m}(t) \quad (j = 1,2,...,N;\ t = 1,2,...,n)$
- The derived seasonal series in each calendar season $(s = 1,2,3,4)$ separately, $X_{j,s}(t) = \mu_s(t) + S_{j,s}+ IH_{j,s}(t) + \varepsilon_{j,s}(t) \quad (j = 1,2,...,N;\ t = 1,2,...,n)$ and
- The derived annual series, $X_{j,y}(t) = \mu_y(t) + S_{j,y}+ IH_{j,y}(t) + \varepsilon_{j,y}(t) \quad (j = 1,2,...,N;\ t = 1,2,...,n)$.

For this type of examination, there is no need to consider the annual cycle or the seasonality and the temporal autocorrelation. In the case of a specific month, season or year, it can be accepted that the elements of vector series $\boldsymbol{\varepsilon}_m(t) \ (t = 1,...,n)$, $\boldsymbol{\varepsilon}_s(t) \ (t = 1,...,n)$ or $\boldsymbol{\varepsilon}_y(t) \ (t = 1,...,n)$ are totally independent in time. However, after the parallel examinations, a synthesis is necessary for the estimated monthly inhomogeneity signals $I\hat{H}_{j,m}(t)$ $(m = 1,2,...,12)$. If the data series are quasi lognormal distributed (for example, precipitation), a multiplicative model can be used that can be transformed into an additive one by applying a logarithmic transformation procedure.

5.2 Methodology for comparison of series in the case of an additive model

In the relative homogenization approach, a chosen candidate station series is compared to the other stations as reference series. That removes the common climate change and variability signal, which is unknown.

Using the notation of Equation 1, all the examined station series $X_{j,m}(t)$ $(j = 1,...N)$ must be taken as candidate and reference series alike. Furthermore, the reference series are not assumed to be homogeneous since, in general, there is no information about this. The main aim of the comparison is to remove the unknown climate change signal $\mu_m(t)$.

The problems related to series comparison include:

- Pairwise comparison;
- Creation of composite reference series;
- Constitution of difference series;
- Multiple comparisons of series.

These topics are very important for detection and adjustment of inhomogeneities, because efficient series comparison can increase both detection power and correction accuracy. The development of efficient comparison methods can be based on the examination of the spatial covariance structure of data series.

The difference series between pairs are $Z_{j,m}(t) = X_{j,m}(t) - X_{j,m}(t)$. However, the constitution of difference series can be formulated in a more general way as well.

Assuming that $X_{j,m}(t)$ is the candidate series and the others are the reference series, then the difference series belonging to the candidate series can be constituted as,

$$Z_{j,m}(t) = X_{j,m}(t) - \sum_{i\neq j} \lambda_{ji,m} X_{i,m}(t) = IH_{j,m}(t) - \sum_{i\neq j} \lambda_{ji,m}\, IH_{i,m}(t) + \varepsilon_{Z_{j,m}}(t), \tag{2}$$

where $\sum_{i\neq j} \lambda_{ji,m} X_{i,m}(t)$ is a composite reference series with condition of $\sum_{i\neq j} \lambda_{ji,m} = 1$ for the weighting factors. As a result of the last condition, the unknown climate change and variability signal $\mu_m(t)$ has been removed. Consequently, the inhomogeneity can be detected by examining the above difference series. Since the difference series (2) generally includes inhomogeneities from both candidate and reference, it may be useful to select better quality reference series and to create several composite reference series for a candidate series. Multiple comparisons or examination of multiple difference series can facilitate the attribution of the detected inhomogeneity for the candidate series.

There are three considerations for the selection or weighting of reference series:

(1) To reduce the noise of the difference series;
(2) To reduce the influence of inhomogeneities in the reference;
(3) To make sure that the reference series has a regional climate signal similar to that of the candidate.

To increase the signal-to-noise ratio (SNR) in order to increase the power of detection, we must decrease the variance of the noise term $\varepsilon_{Z_{j,m}}(t)$. The optimal weighting factors $\lambda_{ji,m}(t)$ that minimize the variance are determined by the spatial covariance matrices C_m uniquely (optimal interpolation or kriging weights). To reduce the influence of inhomogeneities in nearby stations, which get large kriging weights, it may be desirable to average over several stations or give far away stations a relatively stronger weight. On the other hand, to make sure that all stations experience the same regional climate signal, it may be desirable to limit the number of stations in sparse networks.

It can be proven mathematically that using the maximum likelihood principle for joint detection and/or joint correction (assuming normal distribution) the difference series are examined indirectly, and the weighting factors of the reference series are determined by the spatial covariance matrices. We will return to this question in section 5.3.1 on the model selection approach using penalized likelihood methods.

5.3 Methodology for breakpoint (changepoint) detection

One of the basic tasks of homogenization is the examination of difference series (2) in order to detect the breakpoints and to attribute the appropriate ones to the candidate series.

The scheme of the breakpoint detection is as follows: let $Z(t)$ be a difference series according to the formula (2), that is,

$$Z(t) = IH_z(t) + \varepsilon_Z(t) \quad (t = 1,\ldots,n), \tag{3}$$

where $IH_z(t)$ is a mixed inhomogeneity of difference series $Z(t)$ with breakpoints from candidate and references. In general, the number of breakpoints, their positions and sizes are unknown, and we assume $\varepsilon_Z(t)$ is a normal noise series. In the case of parallel homogenization (Remark 1, 1.2) $\varepsilon_Z(t)$ is a normal white (uncorrelated) noise series since its elements are assumed to be independent in time.

Remark 2
Outlier detection is the main quality control procedure for the monthly data in this process. If an outlier is not removed, it may produce two neighbouring breakpoints whose sizes are the same in absolute value, but with opposite signs.

Returning to the detection procedures, the basic types are hierarchical splitting (stepwise detection) and multiple breakpoint detection. The hierarchical splitting procedure is a repeated single-breakpoint detection. The multiple breakpoint detection procedures were developed for the estimation of all breakpoints in a candidate series.

For the detection of breaks, two classical methods from mathematical statistics are used: maximum likelihood estimation and hypothesis testing.

5.3.1 *Breakpoint detection based on maximum likelihood estimation*

Penalized likelihood methods are based on model selection (segmentation). In such methods, joint maximum likelihood estimation is given for the breakpoints assuming normal distribution of the difference series and using some penalty term. The reason for the penalty term is that the number of breakpoints is unknown. The methods may be different in the penalty terms or criteria, for example, the Akaike criterion (AIC), the Schwarz or Bayesian information criterion (BIC) or the Caussinus-Lyazrhi criterion. The penalty terms depend on some a priori probability of break at each time.

Theoretically, this methodology could also be applied to the joint detection of the breakpoints of all the series examined. However, the joint likelihood function, assuming normal distribution, depends on the inverse of the spatial covariance matrix, which can cause complicated technical problems in case of larger networks. However, a non-trivial statistical model problem to be solved is that the climate signals may be different in larger networks.

5.3.2 *Breakpoint detection based on hypothesis testing*

Hypothesis testing is another method for the detection of breakpoints in difference series. The null hypothesis is that the series under test is homogeneous and that it is Gaussian white noise. Such methods also assume normal distribution, therefore the test statistics are derived from the t-type statistics in general. If no reference is used (absolute homogenization) F-type statistics are derived because of the regression of the unknown temporal trend. The significance and the power of such procedures can be defined according to the probabilities of two types of error: type one errors detect false breakpoints, while type two errors miss real breakpoints. A compromise between these two types of error must be found. The test statistics can be compared to the critical value that depends on the given significance level. In the case of multiple test statistics, the critical values can be calculated by Monte Carlo methods.

Most of these methods are stepwise, iterative single-breakpoint detection methods. However, multiple-breakpoint detection procedures can also be developed. The essence of this procedure is that between neighbouring detected breakpoints the homogeneity can be accepted, and between non-neighbouring detected breakpoints the homogeneity cannot be accepted. In addition, confidence intervals can also be given for the breakpoints that make the automatic use of metadata possible.

An advantage of these methods is that the results can be evaluated and validated by comparison with the test statistics before and after homogenization.

5.3.3 *Attribution of the detected breakpoints for the candidate series*

During the breakpoint detection procedures, the difference series are examined with mixed inhomogeneities (Equation 2). There are two basic ways to tackle this problem:

1. If only one difference series is examined for a candidate series, all the detected breakpoints are attributed to the candidate series. In this case, possible inhomogeneities in the references are a very serious problem. Therefore, it is necessary to select more, better-quality reference series for a composite reference series to be created preferably without breakpoints (it is recommended to use at least four reference stations).

2. If several difference series are examined for a candidate series, the mixed inhomogeneity is less of a problem, but the attribution of the breakpoints to the candidate series is not a trivial task. A synthesis is, therefore, necessary and the key task of the homogenization software is to apply automatic procedures for the attribution.

5.4 Methodology for adjustment of series

Besides detection, another basic task is the adjustment of series. Calculation of the adjustment factors can be based on the examination of difference series for estimation of shifts at the detected breakpoints. In general, point estimation is used for shifts at the detected breakpoints.

There are methods that use the standard least squares technique after breakpoint detection for joint estimation of the shifts of all the series examined. The generalized least squares estimation technique based on spatial covariance matrix could be the most efficient in such cases, and it would be equivalent to maximum likelihood estimation for the shifts in the normal distribution case. However, a non-trivial statistical model problem to be solved is that the climate signals may be different in larger networks.

Another method consists in calculating the adjustment factors on the basis of some confidence intervals given for the shifts at the detected breakpoints. Such confidence intervals also allow for the automatic use of metadata.

Remark 3

Completing missing data or filling gaps is essentially an interpolation problem. Use of the spatial covariances for the calculation of the weighting factors of the predictors is strongly recommended, in order to decrease the interpolation error.

5.5 Possibilities for evaluation and validation of methods

For a real understanding of the available methods, a theoretical evaluation of their mathematical basis is indispensable.

Another possibility is a blind comparison and validation study of the homogenization methods. In this case, the methods are tested on a realistic benchmark dataset. The benchmark contains simulated data with inserted inhomogeneity. Testing the methods on a generated benchmark dataset seems to be an objective validation procedure, however, there are limits to such types of examination.

The interpretation of benchmark results depends on different factors, such as:

- Tested methods (quality, manual or automatic);
- Testing benchmark dataset (quality, adequacy);
- Operators (skilled or unskilled);
- Methodology of evaluation (validation statistics).

The creation of an adequate benchmark dataset and the development of appropriate validation statistics are critical points. They require strong theoretical mathematical background, for example, to understand which statistical characteristics of a benchmark are important and they need to be modelled realistically, possibly with more detailed studies to find out which range is realistic in real networks (Lindau and Venema, 2019).

One of the advantages of automatic homogenization methods is that they lend themselves to objective validation for two reasons: (a) objective validation needs relatively large datasets to perform well and automatic homogenization methods require less effort to homogenize these large datasets; (b) in the case of manual methods, not only the method but also the operator is assessed, which may bias the results in either direction.

GLOSSARY

AIC. Akaike Information Criterion, a penalty function.

Analysis of variance (ANOVA). A joint correction method computing all corrections of a network simultaneously assuming that all have the same regional climate signal and that breaks are a step function. This method minimizes the noise using a least square method, so the equations are the same as those of the statistical test for differences in means.

Annual homogenization. Homogenization using annual data (averages or sums). It may also include the magnitude of the seasonal cycle.

Benchmarking. Performance testing of homogenization methods using realistic open data, which typically involves multiple methods, contributions or operators. Benchmarking is a community effort and, as such, it is more than just a test for validation of a method.

BIC. Bayesian Information Criterion, a penalty function.

Candidate/base station. the station to be homogenized.

Correction. An adjustment made to raw observations aiming to make them more homogeneous.

Daily homogenization. Homogenization using daily data (averages or sums). In the case of daily homogenization, the default is to consider all days serially in one long time series.

HadCRUT. Global temperature dataset compiled by the Hadley Centre of the UK Met Office and the Climatic Research Unit (CRU) of the University of East Anglia.

Homogeneous sub-period (HSP). A segment between adjacent changepoints for which the data could be considered homogeneous (between the inhomogeneities).

iCOADS. International Comprehensive Ocean-Atmosphere Data Set

Joint methods. Homogenization methods that jointly detect or correct all inhomogeneities in multiple stations in the same step.

Monthly homogenization. Homogenization using monthly data (averages or sums). This can be done with 12 monthly time series in parallel (default) or with all months considered as one long time series. Often in combination with annual homogenization.

Multiple-breakpoint method. Homogenization method that detects or corrects multiple inhomogeneities in one station/time series in the same step.

PMT. Penalized Maximal t Test

PRODIGE. A homogenization method. A multiple-breakpoint homogenization method based on the Caussinus-Lyazrhi criterion.

Reference time series. An independent time series used to assess the homogeneity of a candidate station, usually a neighbouring station or derived from several neighbouring stations.

Seasonal homogenization. Homogenization using seasonal data (averages or sums). This can be done with four seasonal time series in parallel (default) or with all seasons considered as one long time series. Often in combination with annual homogenization.

Shelter/screen. Enclosure to shield meteorological instruments to adequately record atmospheric conditions in accordance with WMO standards. There are many different types of screen.

UTC. Universal Time Coordinated

REFERENCES

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *Journal of Climatology*, 6, pp. 661–675, https://doi.org/10.1002/joc.3370060607.

Azorin-Molina, C. et al., 2014: Homogenization and assessment of observed near-surface wind speed trends over Spain and Portugal, 1961–2011. *Journal of Climate*, 27, pp. 3692–3712, https://doi.org/10.1175/JCLI-D-13-00652.1.

Azorin-Molina, C. et al., 2019: An approach to homogenize daily peak wind gusts: An application to the Australian series. *International Journal of Climatology*, 18 pp., https://doi.org/10.1002/JOC.5949.

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

Böhm, R. et al., 2010: The early instrumental warm-bias: a solution for long central European temperature series 1760–2007. *Climatic Change*, 101, pp. 41–67, https://doi.org/10.1007/s10584-009-9649-4.

Brandsma, T., J. Van der Meulen and V.K.C. Venema, 2019: Protocol Measurement Infrastructure Changes (PMIC). Soon to be published as a WMO report.

Brogniez, H., et al., 2016: A review of sources of systematic errors and uncertainties in observations and simulations at 183 GHz. *Atmospheric Measurement Techniques*, 9, pp. 2207–2221, https://doi.org/10.5194/amt-9-2207-2016.

Brückner, E., 1890: *Klimaschwankungen seit 1700 nebst Bemerkungen über Klimaschwankungen der Diluvialzeit.* E.D. Hölzel, Wien and Olnütz.

Brugnara Y. et al., 2012: High-resolution analysis of daily precipitation trends in the central Alps over the last century. *International Journal of Climatology*, 32, pp. 1406–1422, https://doi.org/10.1002/joc.2363.

Brunet, M. et al., 2006: The development of a new dataset of Spanish daily adjusted temperature series (SDATS) (1850–2003). *International Journal of Climatology,* 26, pp. 1777–1802, https://doi.org/10.1002/joc.1338

Brunet, M. et al., 2011: The minimization of the screen bias from ancient Western Mediterranean air temperature records: an exploratory statistical analysis. *International Journal Climatology*, 31, pp. 1879–1895, https://doi.org/10.1002/joc.2192.

Brunetti M. et al., 2006: Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series. *International Journal of Climatology*, 26, pp. 345–381, https://doi.org/10.1002/joc.1251.

Buisan, S.T., C. Azorin-Molina and Y. Jimenez, 2015: Impact of two different sized Stevenson screens on air temperature measurements. *International Journal of Climatology,* 35(14), 4408–4416, https://doi.org/10.1002/joc.4287.

Caussinus, H. and O. Mestre, 2004: Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society, Series C (Applied Statistics),* 53, pp. 405–425, https://doi.org/10.1111/j.1467-9876.2004.05155.x.

Conrad, V., 1925: Homogenitätsbestimmung meteorologischer Beobachtungsreihen. *Meteorologische Zeitschrift*, 482–485.

Conrad V. and O. Schreier, 1927: Die Anwendung des Abbe'schen Kriteriums auf physikalische Beobachtungsreihen. *Gerland's Beiträge zur Geophysik*, XVII, 372.

Conrad V., 1944: *Methods in Climatology.* Harvard University Press, 228 p.

Conrad, V. and C. Pollak, 1950: *Methods in Climatology*. Harvard University Press, Cambridge, MA, 459 p.

Costa, A.C. and A. Soares, 2009: Homogenization of climate data: Review and new perspectives using Geostatistics. *Mathematical Geosciences*, 41, pp. 291–305, https://doi.org/10.1007/s11004-008-9203-3.

Cowtan, K., R. Rohde and Z. Hausfather, 2018: Evaluating biases in sea surface temperature records using coastal weather stations. *Quarterly Journal of the Royal Meteorological Society*, https://doi.org/10.1002/qj.3235.

Cook, B.I. et al., 2014: Irrigation as an historical climate forcing. *Climate Dynamics*, 44: 1715, https://doi.org/10.1007/s00382-014-2204-7.

Craddock, J.M., 1979: Methods of comparing annual rainfall records for climatic purposes. *Weather*, 34, pp. 332–346, https://doi.org/10.1002/j.1477-8696.1979.tb03465.x.

Dai, A. et al, 2011: A new approach to homogenize daily radiosonde humidity data. *Journal of Climate*, 24, 965991, https://doi.org/10.1175/2010JCLI3816.1.

Degaetano, A.T., 2000: A serially complete simulated observation time metadata file for U.S. daily Historical Climatology Network stations. *Bulletin of the American Meteorological Society*, 81: 1, pp. 49–68, https://doi.org/10.1175/1520-0477(2000)081<0049:ASCSOT>2.3.CO;2.

Della-Marta, P.M. and H. Wanner, 2006: A method of homogenizing the extremes and mean of daily temperature measurements. *Journal of Climate*, 19, pp. 4179–4197. https://doi.org/10.1175/JCLI3855.1.

Dienst, M. et al., 2017: Removing the relocation bias from the 155-year Haparanda temperature record in Northern Europe. *International Journal of Climatology*, 37: pp. 4015–4026, https://doi.org/10.1002/joc.4981.

Dienst, M. et al., 2019: Detection and elimination of UHI effects in long temperature records from villages – A case study from Tivissa, Spain. *Urban Climate*, 27, pp. 372–383, https://doi.org/10.1016/j.uclim.2018.12.012.

Domonkos, P., 2011: Efficiency evaluation for detecting inhomogeneities by objective homogenization methods. *Theoretical and Applied Climatology*, 105, pp. 455–467. https://doi.org/10.1007/s00704-011-0399-7.

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

Domonkos, P., V. Venema and O. Mestre, 2013: Efficiencies of homogenisation methods: our present knowledge and its limitation. In *Seventh Seminar for Homogenization and Quality Control in Climatological Databases Jointly Organized with the COST ES0601 (HOME) Action MC Meeting*, Budapest, Hungary, 24–27 October 2011. *Climate data and monitoring* (WCDMP-No. 78), Geneva, WMO, pp. 11–24.

Dunn, R. and P. Thorne, 2017: Towards an integrated set of surface meteorological observations for climate science and applications. Proceedings of the 19th European Geosciences Union General Assembly, Vienna, Austria.

Dunn, R.J.H.et al., 2012: HadISD: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, 8, pp. 1649–1679, https://doi.org/10.5194/cp-8-1649-2012.

Gubler, S. et al., 2017: The influence of station density on climate data homogenization. *International Journal of Climatology*, 37, pp. 4670–4683, https://doi.org/10.1002/joc.5114.

Guijarro, J.A., 2013: Temperature trends. In *Adverse weather in Spain* (C. García-Legaz C and F. Valero, eds.). Madrid, AMV Ediciones, pp. 297–306.

Guijarro, J., 2018: Climatol, Version 3.1.2, https://CRAN.R-project.org/package=climatol and http://www.climatol.eu/.

Haimberger, L., 2007: Homogenization of radiosonde temperature time series using innovation statistics. *Journal of Climate*, 20(7): pp. 1377–1403, https://doi.org/10.1175/JCLI4050.1.

Haimberger, L., C. Tavolato and S. Sperka, 2012: Homogenization of the global radiosonde dataset through combined comparison with reanalysis background series and neighboring stations. *Journal of Climate*, 25, pp. 8108–8131, https://doi.org/10.1175/JCLI-D-11-00668.1.

Hann, J., 1880: *Untersuchungen über die Regenverhältnisse von Österreich-Ungarn. II. Veränderlichkeit der Monats- und Jahresmengen.* S.-B. Akad. Wiss. Wien.

Heidke, P., 1923: Quantitative Begriffsbestimmung homogener Temperatur- und Niederschlagsreihen. *Meteorologische Zeitschrift*, pp. 114–115.

Helmert, F.R., 1907: *Die Ausgleichrechnung nach der Methode der kleinsten Quadrate*. 2. Auflage, Teubner Verlag.

Huang, B. et al, 2015: Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4). Part I: Upgrades and Intercomparisons. *Journal of Climate*, 28, pp. 911–930, https://doi.org/10.1175/JCLI-D-14-00006.1.

Hungarian Meteorological Service (OMSZ), 1996: *Proceedings of the First Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, 6–12 October 1996, 44 p.

Hungarian Meteorological Service (OMSZ), 2001: *Third Seminar for Homogenization and Quality Control in Climatological Databases*. Budapest.

Hungarian Meteorological Service (OMSZ), 2017: *9th Seminar for Homogenization and Quality Control in Climatological Databases and 4th Conference on Spatial Interpolation Techniques in Climatology and Meteorology,* Budapest, Hungary, 3–7 April 2017 (T. Szentimrey, L. Hoffmann and M. Lakatos, eds.). Budapest, https://doi.org/10.21404/9.SemHQC4.ConfSI.2017.

Jones, P.D. et al., 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research*, 117: D05127, https://doi.org/10.1029/2011JD017139.

Jovanovic, B. et al., 2017: Homogenized monthly upper-air temperature dataset for Australia. *International Journal of Climatology*, 37, pp. 3209–3222, https://doi.org/10.1002/joc.4909.

Karl, T.R. et al., 1986: A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States, *Journal of Applied Meteorology and Climatology*, 25:2, pp. 145–160, https://doi.org/10.1175/1520-0450(1986)025<0145:AMTETT>2.0.CO;2.

Kennedy, J.J. et al., 2011: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *Journal of Geophysical Research – Atmospheres*, 116:D14103, https://doi.org/10.1029/2010JD015218.

Kent, E. et al., 2016: A call for new approaches to quantifying biases in observations of sea-surface temperature. *Bulletin of the American Meteorological Society*, 98, pp. 1601–1616, https://doi.org/10.1175/BAMS-D-15-00251.1.

k.k. Hof- und Staatsdruckerei, 1873: *Bericht über die Verhandlungen des internationalen Meteorologen-Congresses zu Wien*, 2–10. September 1873, Protokolle und Beilagen.

k.k. Zentralanstalt für Meteorologie und Geodynamik, 1906: *Bericht über die internationale meteorologische Direktorenkonferenz in Innsbruck*, September 1905. Anhang zum Jahrbuch 1905. k.k. Hof-und Staatsdruckerei.

Kohler M.A., 1949: On the use of double-mass analysis for testing the consistency of records and for making required adjustments. *Bulletin of the American Meteorological Society*, 30, pp. 188–189, https://doi.org/10.1175/1520-0477-30.5.188.

Kuglitsch, F.G, 2012: Break detection of annual Swiss temperature series. *Journal of Geophysical Research – Atmospheres*, 117:D13105, https://doi.org/10.1029/2012JD017729.

Kreil, K., 1854*a*: Mehrjährige Beobachtungen in Wien vom Jahre 1775 bis 1850. *Jahrbücher der k.k. Central-Anstalt für Meteorologie und Erdmagnetismus.* I. Band – Jg 1848 und 1849, pp. 35–74.

Kreil, K., 1854*b*: Mehrjährige Beobachtungen in Mailand vom Jahre 1763 bis 1850. *Jahrbücher der k.k. Central-Anstalt für Meteorologie und Erdmagnetismus*. I. Band – Jg 1848 und 1849, pp. 75–114.

Leeper, R.D., J. Rennie and M.A. Palecki, 2015: Observational perspectives from U.S. Climate Reference Network (USCRN) and Cooperative Observer Program (COOP) network: Temperature and precipitation comparison. *Journal of Atmospheric and Oceanic Technology*, 32, pp. 703–721, https://doi.org/10.1175/JTECH-D-14-00172.1.

Li, Y.,R. Lund and A. Hewaarachchi, 2017: Multiple changepoint detection with partial information on changepoint times. *ArXiv:1511.07238* (manuscript), https://arxiv.org/abs/1511.07238.

Lindau, R. and V.K.C. Venema, 2013: On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records. *Idojaras,* 117(1):1, pp. 1–34.

Lindau, R. and V. Venema, 2016: The uncertainty of break positions detected by homogenization algorithms in climate records. *International Journal of Climatology*, 36(2), pp. 576–589, https://doi.org/10.1002/joc.4366.

Lindau, R. and V.K.C. Venema, 2018*a*: The joint influence of break and noise variance on the break detection capability in time series homogenization. *Advances in Statistical Climatology, Meteorology and Oceanography*, 4, pp. 1–18. https://doi.org/10.5194/ascmo-4-1-2018.

Lindau, R. and V.K.C. Venema, 2018*b*: On the reduction of trend errors by the ANOVA joint correction scheme used in homogenization of climate station records. *International Journal of Climatology*, 38(14), pp.5255–5271, https://doi.org/10.1002/joc.5728.

Lindau, R. and V.K.C. Venema, 2019: A new method to study inhomogeneities in climate records: Brownian motion or random deviations? *International Journal of Climatology*,. 39(12), https://doi.org/10.1002/joc.6105.

Lu, Q.Q., R. Lund and T.C.M. Lee, 2010: An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1), 299–319, https://doi.org/10.1214/09-AOAS289.

Lund, R. and J. Reeves, 2002: Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate,* 15, pp. 2547–2554, https://doi.org/10.1175/1520-0442(2002)015%3C2547:DOUCAR%3E2.0.CO;2.

Mekis, É. and L.A. Vincent , 2011: An overview of the second generation adjusted daily precipitation dataset for trend analysis in Canada. *Atmosphere-Ocean*, 49:2, pp. 163–177, https://doi.org/10.1080/07055900.2011.583910.

Menne, M.J., C.N. Williams Jr. and M.A. Palecki, 2010: On the reliability of the U.S. surface temperature record. *Journal of Geophysical Research – Atmospheres*, 115:D11, https://doi.org/10.1029/2009JD013094.

Menne, M.J. and C.N. Williams Jr., 2009: Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22:7, pp. 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

Menne, M.J., C.N. Williams Jr. and R.S. Vose, 2009: The U.S. Historical Climatology Network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, 90, pp. 993–1007, https://doi.org/10.1175/2008BAMS2613.1.

Mestre O., 1999: Step-by-step procedures for choosing a model with change-points. In *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*, (9–13 November 1998, Budapest, Hungary) (WCDMP-No.41, WMO/TD-No. 962). Geneva, WMO, pp. 15–26.

Mestre, O. et al., 2011: SPLIDHOM: A method for homogenization of daily temperature observations. *Journal of Applied Meteorology and Climatology,* 50, pp. 2343–2358, https://doi.org/10.1175/2011JAMC2641.1.

Mestre, O. et al., 2013: HOMER: A homogenization software – Methods and applications. *Idojaras*, 117, 47–67.

Minola, L., C. Azorin-Molina and D. Chen, 2016: Homogenization and assessment of observed near-surface wind speed trends across Sweden, 1956–2013. *Journal of Climate,* 29, pp. 7397–7415, https://doi.org/10.1175/JCLI-D-15-0636.1.

Parker, D.E., 1994: Effects of changing exposure of thermometers at land stations. *International Journal Climatology*, 14, pp. 1–31, https://doi.org/10.1002/joc.3370140102.

Petrovic, P., 2004: Detecting of inhomogeneities in time series using Real Precision Method. In: *Fourth Seminar for Homogenization and Quality Control in Climatological Databases* (WCDMP-No. 56, WMO/TD-No. 1236). Geneva, WMO.

Quayle, R.G.et al., 1991: Effects of recent thermometer changes in the cooperative station network. *Bulletin of the American Meteorological Society*, 72, pp. 1718–1723. https://doi.org/10.1175/1520-0477(1991)072%3C1718:EORTCI%3E2.0.CO;2.

Rennie, J.J., 2014: The international surface temperature initiative global land surface databank: Monthly temperature data release description and methods. *Geoscience Data Journal*, 1:2, pp. 75–102, https://doi.org/10.1002/gdj3.8.

Ribeiro, S. et al., 2017: GSIMCLI: A geostatistical procedure for the homogenization of climatic time series. *International Journal of Climatology*, 37:8, pp. 3452–3467, https://doi.org/10.1002/joc.4929.

Rohde, R. et al., 2013: Berkeley Earth Temperature Averaging Process. *Geoinformatics & Geostatistics: An Overview,* 1:2, https://doi.org/10.4172/2327-4581.1000103.

Schröder, M. et al., 2016: The GEWEX water vapor assessment: Results from intercomparison, trend, and homogeneity analysis of total column water vapor. *Journal of Applied Meteorology and Climatology*, 55, pp. 1633–1649, https://doi.org/10.1175/JAMC-D-15-0304.1.

Štěpánek, P., P. Zahradníček and P. Skalák 2009: Data quality control and homogenization of the air temperature and precipitation series in the Czech Republic in the period 1961–2007. *Advances in Science and Research*, 3, pp. 23–26, https://doi.org/10.5194/asr-3-23-2009.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). In *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, 9–13 November 1998 (WCDMP-No. 41, WMO/TD-No. 962). Geneva, WMO, pp. 27–46.

Szentimrey, T., 2008: Development of MASH homogenization procedure for daily data. In *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases*, Budapest, Hungary, 29 May–2 June 2006 (WCDMP-No. 71, WMO/TD-No. 1493). Geneva, WMO, pp. 123–130.

Szentimrey, T., 2014: *Manual of homogenization software MASHv3.03*. Hungarian Meteorological Service, 71 p.

Szentimrey, T., 2018: New version MASHv4.01 for joint homogenization of mean and standard deviation (EMS Annual Meeting: European Conference for Applied Meteorology and Climatology 2018, Budapest, Hungary, 3–7 September 2018). *EMS Annual Meeting Abstracts,* Vol. 15, EMS2018-331.

Thorne, P.W.et al, 2011: Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bulletin of the American Meteorological Society*, 92, ES40–ES47, https://doi.org/10.1175/2011BAMS3124.1.

Toreti A. et al., 2010: A novel method for the homogenization of daily temperature series and its relevance for climate change analysis. *Journal of Climate*, 23, pp. 5325–5331, https://doi.org/10.1175/2010JCLI3499.1.

Toreti, A. et al., 2012: A novel approach for the detection of inhomogeneities affecting climate time series. *Journal of Applied Meteorology and Climatology*, 51, 317–326, https://doi.org/10.1175/JAMC-D-10-05033.1.

Trewin, B.C. and A.C.F. Trevitt, 1996: The development of composite temperature records. *International Journal of Climatology*, 16, pp. 1227–1242, https://doi.org/10.1002/(SICI)1097-0088(199611)16:11%3C1227::AID-JOC82%3E3.0.CO;2-P.

Trewin, B.C., 2012: *Techniques involved in developing the Australian Climate Observations Reference Network - Surface Air Temperature (ACORN-SAT) dataset.* CAWCR Technical Report 49. Melbourne, Centre for Australian Weather and Climate Research, http://cawcr.gov.au/technical-reports/CTR_049.pdf.

Trewin, B., 2013. A daily homogenized temperature dataset for Australia. *International Journal of Climatology*, 33:6, pp. 1510–1529, https://doi.org/10.1002/joc.3530.

Trewin, B.C., 2018: *The Australian Climate Observations Reference Network – Surface Air Temperature* (ACORN-SAT*) version 2.* Bureau Research Report – BRR032. Melbourne, Bureau of Meteorology.

Venema, V.K.C. et al., 2012: Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8, pp. 89–115, https://doi.org/10.5194/cp-8-89-2012.

Vincent, L.A., 1998: A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate,* 11, pp. 1094–1104, https://doi.org/10.1175/1520-0442(1998)011<1094:ATFTIO>2.0.CO;2.

Vincent, L.A. et al., 2002: Homogenization of daily temperatures over Canada. *Journal of Climate*, 15, 1322–1334, https://doi.org/10.1175/1520-0442(2002)015%3C1322:HODTOC%3E2.0.CO;2.

Vincent, L.A. et al., 2009: Bias in minimum temperature introduced by a redefinition of the climatological day at the Canadian Synoptic stations. *Journal of Applied Meteorology and Climatology*, 48, pp. 2160–2168. https://doi.org/10.1175/2009JAMC2191.1.

Vincent, L.A. et al., 2012: A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. *Journal of Geophysical Research – Atmospheres*, 117:D18110, https://doi.org/10.1029/2012JD017859.

Vose, R.S. et al., 2003: An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network. *Geophysical Research Letters*, 30:20, 2046, https://doi.org/10.1029/2003GL018111.

Wan, H., X.L. Wang and V.R. Swail, 2007: A quality assurance system for Canadian hourly pressure data. *Journal of Applied Meteorology and Climatology*, 46, 1804–1817, https://doi.org/10.1175/2007JAMC1484.1.

Wan, H., X.L. Wang and V.R. Swail, 2010: Homogenization and trend analysis of Canadian near-surface wind speeds. *Journal of Climate*, 23, pp. 1209–1225, https://doi.org/10.1175/2009JCLI3200.1.

Wang, X.L., 2003: Comments on "Detection of undocumented changepoints: A revision of the two-phase regression model". *Journal of Climate*, 16, pp. 3383–3385, 10.1175/1520-0442(2003)016<3383:CODOUC>2.0.CO;2.

Wang, X.L., Q.H. Wen and Y. Wu, 2007: Penalized maximal *t* test for detecting undocumented mean change in climate data series. *Journal of Applied Meteorology and Climatology*, 46, pp. 916–931, https://doi.org/10.1175/JAM2504.1.

Wang, X.L., 2008*a*: Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal *t* or *F* test. *Journal of Applied Meteorology and Climatology*, 47, pp. 2423–2444, https://doi.org/10.1175/2008JAMC1741.1.

Wang, X.L., 2008*b*: Penalized maximal *F* test for detecting undocumented mean shift without trend change. *Journal of Atmospheric and Oceanic Technology*, 25, pp. 368–384, https://doi.org/10.1175/2007JTECHA982.1.

Wang, X. et al., 2010: New techniques for the detection and adjustment of shifts in daily precipitation data series. *Journal of Applied Meteorology and Climatology*, 49, pp. 2416–2436, https://doi.org/10.1175/2010JAMC2376.1.

Wang, X.L. and Y. Feng, 2013: *RHtestsV4 User Manual.* Climate Data and Analysis Section - Environment and Climate Change Canada. Published online July 2013, https://github.com/ECCC-CDAS.

Wang, X.L., Y. Feng and L.A. Vincent, 2013: Observed changes in one-in-20 year extremes of Canadian surface air temperatures. *Atmosphere-Ocean*, 52, pp. 222–231, https://doi.org/10.1080/07055900.2013.818526.

Wang, X.L. et al., 2017: Adjusted daily rainfall and snowfall data for Canada. *Atmosphere-Ocean*, 55:3, pp. 155–168, https://doi.org/10.1080/07055900.2017.1342163.

Wen, Q. H., X.L. Wang and A. Wong, 2011: A hybrid-domain approach to modeling climate data time series. *Journal of Geophysical Research - Atmospheres,* 116, D18112, https://doi.org/10.1029/2011JD015850.

Willett, K.M. et al., 2014: A framework for benchmarking of homogenization algorithm performance on the global scale. *Geoscientific Instrumentation, Methods and Data Systems*, 3, pp. 187–200, https://doi.org/10.5194/gi-3-187-2014.

Winkler, P., 2009: Revision and necessary correction of the long-term temperature series of Hohenpeissenberg, 1781–2006. *Theoretical and Applied Climatology*, 98, 259–268, https://doi.org/10.1007/s00704-009-0108-y.

World Meteorological Organization (WMO), 1966: *Climatic Change: Report of a Working Group of the Commission for Climatology* (WMO–No. 195). Geneva.

———, 1999: *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, 9–13 November 1998 (WMO/TD-No. 962; WCDMP-No. 41). Geneva.

———, 2004: *Fourth Seminar for Homogenization and Quality Control in Climatological Databases*, Budapest, Hungary, 6–10 October 2003 (WMO/TD-No. 1236, WCDMP-No. 56). Geneva.

———, 2006: *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases*, Budapest, Hungary, 29 May–2 June 2006 (WMO/TD-No. 1493, WCDMP-No 71). Geneva.

———, 2010: *Proceedings of the Sixth Seminar for Homogenization and Quality Control in Climatological Databases*, Budapest, Hungary, 26–30 May 2008 (WMO/TD-No. 1576, WCDMP-No. 76). Geneva.

———, 2011: *Seventh Seminar for Homogenization and Quality Control in Climatological Databases jointly organized with the COST E2S0601 (HOME) Action MC Meeting*, Budapest, Hungary, 24–27 October 2011 (M. Lakatos, T. Szentimrey and E. Vincze, eds.) (WCDMP-No. 78). Geneva.

———, 2014: *Eighth Seminar for Homogenization and Quality Control in Climatological Databases and Third Conference on Spatial Interpolation Techniques in Climatology and Meteorology,* Budapest, Hungary, 12–16 May 2014 (M. Lakatos, T. Szentimrey and A. Marton, eds.) (WCDMP-No. 84). Geneva.

———, 2019: *WIGOS Metadata Standard* (WMO-No 1192). Geneva.

———, 2018*a*: *Guide to Instruments and Methods of Observation* (WMO-No. 8). Geneva.

———, 2018*b*: *Guide to Climatological Practices* (WMO-No. 100). Geneva.

Xu, W., Q. Li, P. Jones, X. L. Wang, B. Trewin, S. Yang, C. Zhu, G. Ren, P. Zhai, J. Wang, L. Vincent, A. Dai, Y. Gao, Y. Ding, 2017: A new integrated and homogenized global monthly land surface air temperature dataset for the period since 1900. Climate Dynamics, 50, 2513–2536. https://doi.org/10.1007/s00382-017-3755-1.

Xu W. et al, 2013: Homogenization of Chinese daily surface air temperatures and analysis of trends in the extreme temperature indices. *Journal of Geophysical Research – Atmospheres*, 118(17):9708–9720, https://doi.org/10.1002/jgrd.50791.

Yang, S., X.L. Wang and M. Wild, 2018: Homogenization and trend analysis of the 1958-2016 in-situ surface solar radiation records in China. *Journal of Climate*, 31, 4529–4541, https://doi.org/10.1175/JCLI-D-17-0891.1.

Yosef, Y., E. Aguilar and P. Alpert, 2018. Detecting and adjusting artificial biases of long-term temperature records in Israel. *International Journal of Climatology,* 38(8), 3273–3289, https://doi.org/10.1002/joc.5500.

Zhang, L. et al., 2014: Effect of data homogenization on estimate of temperature trend: a case of Huairou station in Beijing Municipality. *Theoretical and Applied Climatology*, 115, pp. 365–373, https://doi.org/10.1007/s00704-013-0894-0.

SUGGESTED READING

Alexandersson, H. and A. Moberg, 1997: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *International Journal of Climatology*, 17, pp. 25–34, https://doi.org/10.1002/(SICI)1097-0088(199701)17:1<25::AID-JOC103>3.0.CO;2-J.

Auchmann, R. and S. Brönnimann, 2012: A physics-based correction model for homogenizing sub-daily temperature series. *Journal of Geophysical Research – Atmospheres,* 117: D17119, https://doi.org/10.1029/2012JD018067.

Brohan, P. et al., 2006: Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850. *Journal of Geophysical Research – Atmospheres*, 111, D12106, https://doi.org/10.1029/2005JD006548.

Domonkos, P., 2011: Adapted Caussinus-Mestre algorithm for networks of temperature series (ACMANT). *International Journal of Geosciences,* 2, pp. 293–309. https://doi.org/10.4236/ijg.2011.23032.

Hawkins, D.M., 1972: On the choice of segments in piecewise approximation. *Journal of the Institute of Mathematics and its Applications*, 9, pp. 250–256.

Killick, R., 2016: *Benchmarking the Performance of Homogenisation Algorithms on Daily Temperature Data.* PhD thesis, Department of Mathematics, University of Exeter, UK. http://hdl.handle.net/10871/23095.

Maraun, D. et al., 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics,* 48, RG3003, https://doi.org/10.1029/2009RG000314.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate* 26:6, pp. 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1.

Morozova, A.L. and M.A. Valente, 2012: Homogenization of Portuguese long-term temperature data series: Lisbon, Coimbra and Porto. *Earth System Science Data*, 4, 187–213, https://doi.org/10.5194/essd-4-187-2012.

Rienzner, M. and C. Gandolfi, 2013: A procedure for the detection of undocumented multiple abrupt changes in the mean value of daily temperature time series of a regional network. *International Journal of Climatology*, 33, pp. 1107–1120, https://doi.org/10.1002/joc.3496.

Rustemeier, E. et al., 2017: HOMPRA Europe - A gridded precipitation dataset from European homogenized time series. In *9th Seminar for Homogenization and Quality Control in Climatological Databases and 4th Conference on Spatial Interpolation Techniques in Climatology and Meteorology*, Budapest, Hungary, 3–7 April 2017 (T. Szentimrey, L. Hoffmann and M. Lakatos, eds.). Budapest, Hungarian Meteorological Service (OMSZ).

Von Storch, H., 1999: On the use of "inflation" in statistical downscaling. *Journal of. Climate*, 12, 3505–3506, https://doi.org/10.1175/1520-0442(1999)012<3505:OTUOII>2.0.CO;2.

World Meteorological Organization (WMO), 1986: *Guidelines on the quality control of surface climatological data* (WMO/TD-No. 111, WCP-85). Geneva.

———, 1993: *Guide on the Global Data-processing System* (WMO-No. 305). Geneva.

———, 2003: *Guidelines on climate metadata and homogenization* (WMO/TD-No. 1186, WCDMP No. 53). Geneva.

———, 2004: *Guidelines on climate data rescue* (WMO/TD-No. 1210, WCDMP-No. 55). Geneva.

———, 2016: *Guidelines on Best Practices for Climate Data Rescue* (WMO-No. 1182). Geneva.

Yosef, Y., I. Osetinsky-Tzidaki and A. Furshpan, 2015: Homogenization of monthly temperature series in Israel – An integrated approach for optimal break-points detection. *Eighth Seminar for Homogenization and Quality Control in Climatological Databases and Third Conference on Spatial Interpolation Techniques in Climatology and Meteorology*, Budapest, Hungary, 12–16 May 2014 (WCDMP-No. 84). Geneva, WMO.

For more information, please contact:

World Meteorological Organization

7 bis, avenue de la Paix – P.O. Box 2300 – CH 1211 Geneva 2 – Switzerland

Strategic Communications Office

Tel.: +41 (0) 22 730 87 40/83 14 – Fax: +41 (0) 22 730 80 27

Email: communications@wmo.int

public.wmo.int

JN 2069