DATA ANALYSIS

DETECTING AND REPAIRING INHOMOGENEITIES IN DATASETS

Assessing Current Capabilities

To study climate variability, the original surface observations are indispensable, but these have to be used with care. Long observational records always contain changes due to nonclimatic factors. Such inhomogeneities can be either sudden jumps (breaks) or gradual trends in one station.

Most surface stations are not operated for climatic purposes, but rather to meet the needs of weather forecasting, agriculture, and hydrology. Consequently, the average period between detected inhomogeneities, or breaks, in Western instrumental records is only 15–20 years. The typical size of the breaks is of the same order as the climate-change signal during the twentieth century. Specific inhomogeneities are typical for certain periods and common to many stations; these can collectively lead to artificial biases in climate trends across large regions. Inhomogeneities are thus a significant source of uncertainty in the estimation of secular trends and decadal-scale variability.

To the general public, the best-known inhomogeneity is probably the urban heat-island effect. The temperature in cities can be warmer than in the surrounding countryside, especially during calm nights. As cities have grown, they have encroached on many weather stations, raising the ambient temperature. Worldwide, the advent of aviation led to relocation of stations from cities to nearby, typically cooler airports. In general, relocations are an important cause of inhomogeneities.

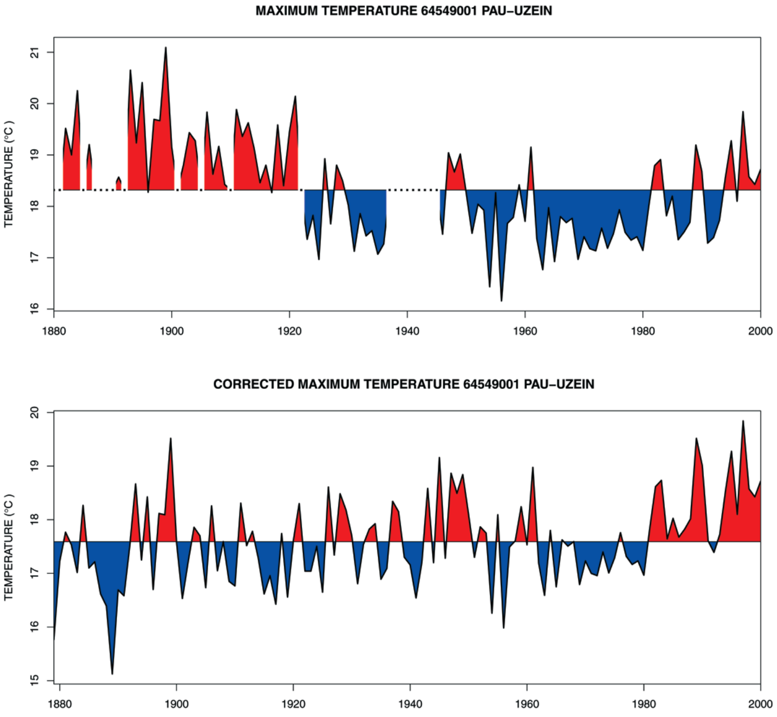
Inhomogeneities caused by changes in the screens that protect the instruments from radiation and wetting are also common. In nineteenth-century Europe, it was common to install the instruments in a metal screen near a window on a north-facing wall. However, the building may warm the screen, leading to higher temperature measurements. When this problem was realized, screens such as the cotton region shelter were introduced. Other typical causes of inhomogeneities are changes in the surrounding environment (e.g., land-use change and building activity). An important recent inhomogeneity is the changeover to automatic weather stations.

Ideally, the date of a change of instruments, location, or observing practices would be recorded as "metadata," and parallel measurements made with the original and the new setup for several years, allowing reliable estimation of the inhomogeneity. By making parallel measurements with replicas of historical instruments, screens, etc., the influence of some historical inhomogeneities can still be studied today.

However, metadata are often incomplete or lacking, so statistical homogenization is necessary as well. The most commonly used principle to detect and remove the artificial changes is relative homogenization. This assumes that nearby stations are exposed to almost the same climate signal, but not any nonclimatic changes. By looking at the difference between nearby stations, the year-to-year variability of the climate is removed, as well as the regional climatic trend. In such a difference time series, a clear and persistent jump can easily be detected and can only be due to changes in the measurement conditions.

A jump (break) in a difference time series of a pair of stations does not pinpoint the responsible station. Furthermore, time series typically have more than just one jump. These two features make statistical homogenization a challenging and beautiful statistical problem. Homogenization algorithms typically differ in how they solve these two fundamental problems.

Indeed, many statistical procedures have been developed to detect and correct inhomogeneities. As



Raw (top) annual averages of maximum temperatures in Pau, a city on the northern edge of the Pyrenees. The data were homogenized (bottom) using a pairwise comparison with surrounding stations, which were also used to fill gaps in the Pau record. The change in 1921 (–1.4°C), which is clearly visible in the uncorrected station data shown here, is due to a relocation from the primary school in Pau-Lescar to the military airport and a shelter change. However, several other changes had significant influence on this series: a shelter move in 1932 (–0.67°C), the relocation to Pau-Uzein civil airport in 1946 (+0.53°C), and installation of a Mistral automatic weather station in 1985 (–0.46°C).

a result, a coordinated European initiative, Advances in Homogenization Methods of Climate Series: An Integrated Approach (HOME), was established in order to facilitate comparisons of methods, to produce standard methods, and to promote the most efficient methods of homogenization.

As part of this initiative, we created a benchmark temperature and precipitation dataset with inhomogeneities inserted to make it a realistic test of homogenization methods (see paper published in *Climates of the Past*, 10 January 2012). We applied all the most common and most developed algorithms for homogenization to this made-up dataset. The main novelty of this experiment was that it was a blind test—the benchmark was generated, and the analysis of results performed, by independent researchers who did not homogenize the data themselves.

THE BENCHMARKING TEST. The benchmark dataset created for our test mimics station networks and their data problems with unprecedented realism. Homogeneous surrogate data were generated reproducing the cross- and autocorrelation functions, as well as the non-Gaussian distribution of climate observations. Added to these data were random break-type inhomogeneities, as well as breaks occurring simultaneously in multiple stations. Furthermore, local trends were inserted, either continuing at the end (to model, for instance, the urban heat-island effect) or reverting to baseline (to model growing vegetation that is subsequently cut back). The sizes of the breaks and local trends follow a normal distribution with a width of 0.8°C. Finally, a stochastic nonlinear network-wide trend was added. Everyone was invited to homogenize the data; 25 homogenized blind contributions were returned.

We tested three types of homogenization methods: absolute, relative, and direct algorithms. In absolute homogenization, only the station time series itself is used. With this approach, it is difficult to distinguish small inhomogeneities from climate variability. In traditional relative homogenization, a candidate series is compared with a composite reference time series computed from its neighboring stations. This composite reference is assumed to be homogeneous due to averaging, which is only approximately true. The main research impetus for the last two decades has been the development of a new approach to relative homogenization—the so-called direct homogenization algorithms that also function with an inhomogeneous reference time series.

The first main conclusion is that relative homogenization improves the temperature data; it reduces the root-mean-square error of the data and its linear trend coefficients and does not cause artificial climate trends. This conclusion can be stated with confidence because the test was blind and because of the realism of the data. The exceptions, where relative homogenization made the data more inhomogeneous, could mostly be explained by inexperienced users or be traced back to algorithms (or parts thereof) newly written for this exercise. This shows an important disadvantage of blind studies: mistakes discovered after the results are shared with participants cannot be corrected. The results also demonstrate that statistical absolute homogenization can make climate data more inhomogeneous. In contrast to the results for temperature, the results for precipitation are more mixed; still, all but one relative method did improve the station trends.

The second main conclusion is that direct homogenization algorithms are clearly better than traditional ones. A realistic benchmark dataset was needed to see this difference with such clarity. With mathematical argumentation, climatological reasoning, and the benchmark metrics all pointing in the same direction, we thus strongly recommend the use of direct homogenization algorithms.

The performance ranking of the homogenization methods depends on the error metric considered; whether the root-mean-square error is computed on the monthly, yearly, or decadal data; and whether it is computed on the station data or on the network mean climate signal. These rankings also do not correlate strongly with the error in the linear trend estimates (or break detection scores). In other words, it is difficult to compute one error metric that would signify the remaining error after homogenization for all climatic purposes. The computation and communication of the remaining uncertainties of homogenized data should be one of the research priorities for the coming years.

We feel that benchmarking has helped the homogenization community to mature. The discussions about properties of benchmarks, the nature of inhomogeneities, and homogenization methods, as well as the joint work on the same dataset, helped to bring scientists closer together in a way that writing individual papers cannot. The International Surface Temperature Initiative has started a follow-up benchmarking program for homogenization algorithms. This benchmark will be global and even more realistic, especially due to the inclusion of metadata, biased inhomogeneities, and random missing data.

Everyone is invited to download and analyze the benchmark dataset. The homogeneous, inhomogeneous, and homogenized datasets are published on the Internet. Another offspring of the European initiative is a package with homogenization methods written in the statistical programming language R (known as HOMER). This open-source, state-of-the-art package is based on the best homogenization methods and also performs basic quality control. Furthermore, a mailing list for researchers working on homogenization has been started. All these resources can be accessed via the initiative website, www.homogenisation.org, which will be kept running for the coming years and which contains an extensive bibliography.

With advanced and well-validated statistical methods, the homogenization of annual and monthly station data is a mature field. The homogenization of daily data, however, is still in its infancy. Daily data are essential for studying extremes of weather and climate, and therefore are the basis for important political decisions with huge socioeconomic consequences. For such studies, the complete distribution needs to be homogenized. Looking at the physical causes of inhomogeneities, one would expect that many of them especially affect the tails of the distribution of the daily data. The IPCC AR4 report warns that changes in extremes are often more sensitive to inhomogeneous climate-monitoring practices than changes in the mean. This is of concern, given that homogenization methods for daily data are often limited to adjustments on the mean of the distribution. Some correction algorithms for the distribution do exist, but these only reliably correct the first three moments, have currently only been applied to some networks, and require highly correlated neighboring stations. A better understanding of the nature of daily inhomogeneities and better tools to correct them will be the main challenges for the coming years.

—Victor Venema
University of Bonn

FOR FURTHER READING

Venema, V., and Coauthors, 2012: Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, **8**, 89–115.