Creating surface temperature datasets to meet 21st Century challenges

Met Office Hadley Centre, Exeter, UK

7th-9th September 2010

White papers background

Each white paper has been prepared in a matter of a few weeks by a small set of experts who were pre-defined by the International Organising Committee to represent a broad range of expert backgrounds and perspectives. We are very grateful to these authors for giving their time so willingly to this task at such short notice. They are not intended to constitute publication quality pieces – a process that would naturally take somewhat longer to achieve.

The white papers have been written to raise the big ticket items that require further consideration for the successful implementation of a holistic project that encompasses all aspects from data recovery through analysis and delivery to end users. They provide a framework for undertaking the breakout and plenary discussions at the workshop. The IOC felt strongly that starting from a blank sheet of paper would not be conducive to agreement in a relatively short meeting.

It is important to stress that the white papers are very definitely not meant to be interpreted as providing a definitive plan. There are two stages of review that will inform the finally agreed meeting outcome:

1. The white papers have been made publicly available for a comment period through a moderated blog.
2. At the meeting the approx. 75 experts in attendance will discuss and finesse plans both in breakout groups and in plenary. Stringent efforts will be made to ensure that public comments are taken into account to the extent possible.

Creation of quality controlled homogenised datasets from the databank

Blair Trewin, Xiaolan L. Wang, Phil Jones, Matthew Menne, Robert Lund, Graham Machin

Remit

- Whether different methods will be required for different timescales (monthly, daily, sub-daily).
- Whether some common metrics would be appropriate.
- Whether efforts should be focussed on a particular area or timescale first, and if so, why.
- The role of national and regional assessments.
- How to engender multiple efforts.
- Potential for novel approaches (e.g. an open source community effort, the use of reanalyses output instead of neighbours as the expectation field, Bayesian approaches).
- The importance of a consistent approach to assessing uncertainty.

Introduction and terminology

Inhomogeneities in temperature records can arise for a wide variety of reasons (for a broad review of these see Trewin (2010)). The process of creating a homogenised dataset from the databank involves two principal stages: the **detection** of inhomogeneities (also known as **changepoints or shifts**) in the data, and making **adjustments** to remove those inhomogeneities and create a homogeneous dataset. For some applications only detection is required, with users making their own judgements about adjustments (if any). The extent to which users wish to remove inhomogeneities may also vary depending on the application: for example, in detection of global climate change signals, it is desirable to remove any inhomogeneities arising from urbanisation, but some users may be explicitly interested in anomalous local trends arising from urban growth.

There are multiple approaches in the literature to both the detection and adjustment problem. In order to preserve the real climate signal (trend and low-frequency variations) in the series under homogeneity study (the candidate series), most approaches compare the **candidate series** to one or more **reference series**. The reference series should contain the same regional climate signal as the candidate series and be highly positively correlated with the candidate series. The reference series most commonly used in currently published methods for surface temperature data are based on data from a neighbouring station, or a combination of neighbouring stations (more in the section on Reference series)).

Quality control refers to the identification of errors in an individual observation, or small number of observations (such as might arise from, for example, a clerical or observation error, or a short-term instrument fault). These are sometimes characterised as random (as opposed to systematic) errors. This has often been a neglected area in the development of homogenised data sets. Whilst most national databanks now have quality control procedures implemented, this has not necessarily been the case historically, and an important part of the development of a homogenised

dataset is applying, as far as possible, a level of quality control to the existing data which is comparable to that applied to current data. Particularly important in the context of extremes is not automatically rejecting observations purely because they are outliers in a statistical distribution. Large random data errors may also complicate detection and adjustment of inhomogeneities. The boundary between data quality issues and inhomogeneities may sometimes be blurred (e.g. in a case where an instrument is out of calibration over a period of several months) and such cases require decisions as to whether the data can be adjusted or should be rejected altogether.

Detection and adjustment of inhomogeneities are complex problems. Metadata are often incomplete or non-existent, and the majority of long-term time series used in climate change analyses will have multiple inhomogeneities, which substantially increases the statistical challenges of detection. Creating a 'perfect' data set from historical data is a practical impossibility, as there is a lower bound below which inhomogeneities, and errors in individual observations, are effectively undetectable. Quantifying this lower bound gives important information as to one source of systematic (type B) uncertainty in the data. Where inhomogeneities are identified, there is also an uncertainty in determining their size and hence any necessary adjustments.

Methods of detection of inhomogeneities

Many changepoint detection methods have been developed and applied to time series of climate data. Most commonly used are likelihood ratio tests, such as the Standard Normal Homogeneity (SNH) test and its variants, two-phase regression (TPR) based tests, Potter's method, etc. There are also tests that use penalized likelihood criteria (e.g., Akaike's information criteria), as well as CUSUM and nonparametric tests, and Bayesian approaches. Reeves et al. (2007) and Peterson et al. (1998) give relatively comprehensive reviews on this topic although there are more recent developments. The regression based tests are most powerful when the assumed mean structure (e.g., no trend in the SNH test, a constant trend in a TPR test…) and normality of errors hold. One should not expect a single method to perform optimally over all forms of mean structures. This stresses the need for using multiple methods. The chosen method(s) should be able to address the most common types of inhomogeneities expected to exist in climate data sets.

Many of these methods are developed for time series containing "at most one changepoint (AMOC)". However, a long term climate data time series often contains multiple changepoints. A commonly used method to deal with multiple changepoints is a sort of stepwise testing algorithm, in which the time series being tested is divided into segments by a number of most probable changepoints and then an AMOC sub-series containing two neighbouring segments is tested to determine the most probable position and significance of the changepoint within this sub-series (e.g., Wang 2008). Segmentation methods usually perform reasonably well, especially when the mean shifts in a series all are in the same direction (either increasing or decreasing), but they could become more delicate when the shifts take opposite signs or occur close together. A direct approach to the problem of multiple changepoints is a penalized log-likelihood procedure developed by Caussinus and Mestre (2004).

There exist two types of changepoints, depending on whether the time of change is known or unknown. As pointed out by Lund and Reeves (2002), test statistics for detecting an unknown changepoint are different from those for assessing statistical significance of a known changepoint. This is because detection of an unknown changepoint involves a search for the most probable time of change by maximizing the related statistic and thus the test statistic has an extreme type of distribution (much higher percentiles), while such a search is not needed when the time of change is known (documented in metadata). However, apart from the RHtestsV3 package (Wang and Feng 2010), few other packages include tests for both known and unknown changepoints. Hierarchical Bayesian methods with a prior chosen to depend on a metadata record have the potential to consolidate both known and unknown changepoints, but have not been extensively developed to date.

Such techniques are generally designed to detect changepoints at a specific point in time. Most are less suited to detecting cases where an anomalous local trend may develop over a period of time, for example as a result of urbanisation (although the extent to which urbanisation would be manifested as an anomalous trend, as opposed to one or more step changes as conditions change in the vicinity of the observation site, is not fully resolved).

Reference series and network-wide inhomogeneities

Most published homogenisation methods for temperature have used one or more reference series based on a number of neighbouring stations. This may involve, for example, the estimation of a background field at the location of the candidate station from a distance-weighted mean of neighbouring stations, or the pairwise comparison of the candidate station with neighbours. The best choice of methods is an area of active discussion.

A good reference series should contain the same regional climate signal as the candidate series and be homogeneous, or at least be homogeneous in a sub-period in which the candidate series is likely inhomogeneous when a pair-wise comparison method is used. The uncertainty in the extent to which the reference series represents the regional climate signal in the candidate series is a source of uncertainty in the homogenization process. The accuracy of individual values in the reference series will also affect the accuracy of the homogenized data when the reference series is involved in deriving the adjustments. In most cases (especially for temperature), it is reasonable to assume that the candidate and reference series have the same regional climate signal; homogeneity can also be assumed, especially over sub-periods of the record.

Station-based reference series are, however, not appropriate for a situation where a change is implemented across a network at the same time (e.g. a change in instrument type, or a change in observation time). Furthermore, a network-wide change may induce an inhomogeneity which is undetectable at an individual station but may be highly significant in a larger dataset (for example, a hypothetical change in instrument type that caused an inhomogeneity of +0.2°C would have a major impact on a global mean). For known network-wide changes, an option is, where possible, to establish an experimental comparison (e.g. between an old and new instrument type), or a physics-based model linking old and new methods. The former approach has been used in a number of studies to assess the impact of changing from manual thermometers to

automated temperature sensors. Comparing instruments with a traceable standard is significant in this context (although in theory all instruments should be traceably calibrated to reliable national standards for the appropriate quantities, in practice this is unlikely to have occurred in all countries or throughout the period of historical record). There have also been attempts to recreate historical thermometer exposures to compare with modern standards (e.g. Böhm et al., 2010). Physics-based modelling has been used to assess the impact of changes in sea-surface temperature measurement methodology.

While reanalyses have inhomogeneities of their own, some of them are independent of surface data (especially surface air temperature data) and thus may have potential as a reference series for network-wide changes where no other suitable reference series exist. Because upper-air temperatures (on which reanalyses are based) generally have longer decorrelation length scales than surface temperatures, reanalyses may also be of value as a reference series where no good surface neighbours exist (e.g. sparsely populated regions and remote islands). For coastal and island stations, sea surface temperatures may be another reference series possibility. As yet no major published data set has used either method.

Methods of adjustment to remove inhomogeneities

Many existing homogenised datasets at the national and international level take one of two approaches: either the exclusion of stations found to be inhomogeneous, or the application of a single adjustment for each inhomogeneity, applied uniformly across the year. More recently, a wider range of adjustment techniques have been put forward. These include:

(a) Calculation of adjustments separately for each month, or season.
(b) Calculation of adjustments calculated from monthly data but smoothed across the annual cycle with a different adjustment for each date (widely described in the literature (e.g. Brunet et al, 2007) as ‘daily adjustment’). A number of national and regional datasets have used this method.
(c) Calculation of different adjustments for different parts of the daily temperature frequency distribution, or for different weather types. Such methods have been applied to a number of test sets of stations, but Australia and Canada are the only countries known to have produced a national dataset adjusted in this way.

Historically, adjustment methods have received less attention in the literature than detection methods, but are currently an active area of research, in particular through the European HOME project (www.homogenisation.org).

Appropriate timescales for homogenisation

Temperature data exist on a number of timescales, the most commonly-used being annual, monthly, daily and sub-daily. This raises issues such as:

- At which timescale can inhomogeneities most effectively be detected, and adjusted for? (this will not necessarily be the same for detection and adjustment).

- Should homogenisation be carried out on a 'base' set of data from which further variables are derived (e.g., a daily maximum/minimum temperature dataset from which daily, monthly and annual means can be derived), or should different variables be homogenised separately?

Most published work has involved detection at an annual or monthly timescale. Initial results from the HOME project suggest that this is likely to be a more effective approach than detection on shorter timescales, with signal-to-noise being an issue. Note that detection power is inversely related to sample size (series length); and larger uncertainty is usually associated with statistical estimates from smaller samples. Annual data series could be too short to detect changepoints with acceptable uncertainty.

While a number of existing datasets involve different variables being homogenised separately, a major issue with such an approach is that they may no longer be internally consistent (e.g. a monthly mean may not be the mean of the daily values, or a daily mean may no longer be consistent with the maximum and minimum).

The homogenisation of fixed-hour sub-daily observations is a largely unexplored scientific question, although some attention has been given to homogenising daily means which are based on fixed-hour observations, in those countries where daily means are calculated using that method, and there has been some limited work done on assessing sub-daily inhomogeneities using weather types.

<u>Scale of homogenisation and quality control efforts</u>

It is likely that homogenisation work will be most effectively carried out at the national/regional level, for a number of reasons:

- Researchers working within their own country or region are likely to have access to a greater range of data (e.g. additional stations, or more daily/sub-daily observations) than are available internationally. They are also likely to have access to a wider range of metadata (including pictures, results of calibration checks etc.), and language difficulties will be less of a factor in interpreting that metadata.
- National-level researchers are more likely to be familiar with the local geography and climate around their observing locations. This is particularly important in quality control in assessing what level of spatial variation between sites is reasonable.
- Homogenisation and quality control can be a very labour-intensive process, and carrying out the work on a national/regional basis limits the resources that need to be committed by any one institution.
- Carrying out work at a national/regional level will give national institutions greater ownership of the project, and may help smooth the path in resolving data policy issues.

The development of a consistent framework between nations/regions will be an important part of this project. The use of inconsistent methods between regions raises the possibility of inconsistencies between data sets once national/regional data sets are consolidated into a global set.

A significant question which will need to be addressed is that of the relative merits of fully automated homogenisation and quality control methods, and those with some level of manual intervention. Automated methods have the advantages of, for example, being much less resource-intensive than manual methods, being fully reproducible, and of being much more amenable to regular updating. It remains to be determined whether such methods are capable of matching, or at least approaching to within an acceptable level, the accuracy of more manually-intensive methods.

The potential for carrying out the work on a more distributed basis still (e.g. through volunteer individuals) is an interesting one which is worth exploring, but would require the development of suitable tools and training. Such approaches may be better deployed to areas such as digitisation in the first instance.

Uncertainty assessment

Very limited attention has been given to uncertainty assessment in existing homogenised data sets, although some, mostly theoretical, attention has been given to the question of what the minimum detectable inhomogeneity is in any given time series. This is unfortunate because no measurement is properly complete without a rigorous assessment of its associated uncertainty.

A proper assessment of uncertainty, including uncertainties arising from the data itself and those uncertainties associated with any homogenisation procedure, will be an important aspect of the development of homogenised datasets in this project. The uncertainty, properly quantified, will give significantly enhanced confidence in the data to researchers who use the data products and, probably more importantly, to those outside of the community. Potential approaches include:

- The use of multiple reference series (e.g. through using different combinations of neighbouring stations).
- The use of different detection and/or adjustment methods on the same dataset, or alternatively the comparison of two different methods used in neighbouring regions (e.g. cross-border comparisons of two different national data sets in adjoining countries).

Possible metrics include the proportion of techniques which detect a known (from metadata) inhomogeneity, or the spread of magnitude of inhomogeneities detected using different methods or reference series.

To retrofit a fully assessed uncertainty analysis to past data is very difficult, if not impossible, however a rigorous approach to quantifying the uncertainty will go a long way to improving confidence in the data. To do this the following will need to be performed:

- Understanding quantitatively the influence factors and their relative importance on the measured quantities
- Establishment of a uniform approach to quantifying uncertainty in data analysis following internationally accepted guidelines (e.g. the ISO Guide to Uncertainty in Measurement) (for example developing a transparent way of

identifying and addressing discontinuities in (non-contiguous) data sets and outliers)

Recommendations

- To use daily maximum/minimum temperature as the 'base' data set to which adjustments are made, with data at monthly and longer timescales derived from the daily data (adjusted where appropriate) rather than adjusted separately.
- To ensure that all detection and adjustment of inhomogeneities is fully documented, allowing reassessments to be made in the future (e.g. if new techniques are developed or previously unknown data or metadata become available).
- To carry out an objective evaluation of known methods for homogenisation/adjustment, in collaboration with the COST action;
- To establish a testbed of data for this purpose (see white paper 9);
- To seek to ensure that all sources of uncertainty are well quantified and defined.

References

Böhm, R., Jones, P.D., Hiebl, J., Frank, D., Brunetti, M. and Maugeri, M. 2010. The early instrumental warm bias: a solution for long central European temperature series 1760-2007. Climatic Change, 101, 41-67.

Brunet M, Saladié O, Jones P, Sigró J, Aguilar E, Moberg A, Lister D, Walther A, Almarza C, 2007: A case-study/guidance on the development of long-term daily adjusted temperature datasets. WMO/TD No. 1425, World Meteorological Organization, Geneva. Available online at http://www.wmo.int/pages/prog/wcp/wcdmp/wcdmp_series/documents/WCDMP_Spain_case_study-cor_ver6March.pdf.

Caussinus, H., and O. Mestre, 2004: Detection and correction of artificial shifts in climate. *Journal of the Royal Statistical Society*, Series C, 53, 405-425.

Lund, RB, and Reeves J, 2002: Detection of undocumented changepoints --- a revision of the two-phase regression model. *Journal of Climate*, 17, 2547-2554.

Peterson, T. C. and co-authors, 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. Int. J. Climatol. 18: 1493-1517.

Reeves J, Chen J, Wang XL, Lund R, Lu Q, 2007: A review and comparison of changepoint detection techniques for climate data. J. Appl. Met. Climatol., 46: 900-915, doi: 10.1175/JAM2493.1.

Trewin, B.C. 2010: Exposure, instrumentation and observing practice effects on land temperature measurements. *Wiley Interdisciplinary Reviews: Climate Change*, published online at http://www3.interscience.wiley.com/journal/123356866/abstract.

Wang, XL, Y. Feng, 2010: RHtestsV3 User Manual. Climate Research Division, Environment Canada. 28pp. Available at http://cccma.seos.uvic.ca/ETCCDMI/software.shtml

Wang, XL, 2008: Accounting for autocorrelation in detecting mean-shifts in climate data series using the penalized maximal *t* or *F* test. *J. App. Meteor. Climatol*, 47, 2423–2444. DOI: 10.1175/2008JAMC1741.1.