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.

Benchmarking homogenisation algorithm performance against test cases

[1]Kate Willett, [2]Matt Menne, [2]Peter Thorne, [3]Stefan Brönnimann, [4]Ian Jolliffe, [5]Lucie Vincent & [5]Xiaolan Wang

[1] Met Office Hadley Centre, FitzRoy Road, Exeter, UK
[2] National Climatic Data Center, Ashville, NC, USA
[3] University of Bern, Switzerland
[4] University of Exeter, Exeter, UK
[5] Climate Research Division, Science and Technology Branch, Environment Canada, Toronto, Canada

Discussion Areas:

i. The need for pseudo-data to have full space and time sampling of the observational network
ii. Source data for pseudo-data construction
iii. Design of orthogonal discontinuity test cases that can optimally answer multiple questions
iv. Should discontinuity test case construction be truly blind to the dataset creators to avoid over-tuning of algorithms to the discontinuity test cases?

Introduction:

For climate change research it is essential to detect, attribute (if possible) and, adjust for discontinuities (where desirable depending on end-user) originating from changes in station location, instrumentation, recording practice or surrounding environment. Reconciliation of discontinuities using station metadata is preferable but these data are mostly undigitised or in many cases undocumented even on paper. The ability of chosen algorithms to achieve station homogeneity in the absence of complete and verified metadata (methodological uncertainty) has rarely been assessed comprehensively so as to enable useful dataset intercomparison. Algorithms vary in their skill at identifying discontinuities that are: multiple and geographically or temporally clustered; close to end points; gradual; variance based; in the presence of a background trend; and seasonally or diurnally variant. In addition, quantifying discontinuity magnitude (and hence the presence of discontinuities and adjustments required) against background noise (natural variability and white noise of erroneous observations remaining even after quality control) is very difficult, especially if discontinuity magnitude varies diurnally and seasonally. Even after detection a series of decisions are required as to whether and how to adjust. These decisions can have a further non-negligible impact upon the resulting dataset estimates. While decisions are as evidence based as is physically possible, some are unavoidably arbitrary. This is especially problematic for large datasets where the whole process by necessity is automated.

Benchmarking (measuring performance against a standard) against homogenous real or synthetic data series with known discontinuity test cases applied is increasingly common practice (Easterling & Peterson 1995; Vincent 1998; Ducre-Robitaille et al. 2003; Wang et al 2007; Begert et al. 2008; Wang 2008a; Wang 2008b). However, these pseudo-datasets are usually very simplistic e.g. a single timeseries lacking real world spatio-temporal sampling, climatology, variability and noise. Furthermore,

reconstructing realistic discontinuity test cases that capture all eventualities to the best of our knowledge is a very large task indeed, especially for datasets covering multiple regions. Recent studies (e.g. Menne & Williams 2005; DeGaetano 2006; Wang et al. 2007 and Wang et al. 2008a; 2008b) have generated synthetic pseudo-data with varying degrees of real-world characteristics (e.g. variance, autocorrelation and randomly assigned sign or position of discontinuity). Pseudo-data have also been generated from homogeneous series of real data using a re-sampling technique (e.g. bootstrap; Wang et al. 2010). Recent European efforts from COST-ESO601[6] have produced a monthly mean benchmark dataset compiling synthetic and real data with a number of discontinuity test cases. Titchner et al. 2009 created a monthly pseudo-dataset by sub-sampling the atmosphere-only general circulation model (GCM) HadAM3 and adjusting the variance, mean and white noise to match real world characteristics. Various discontinuity test cases were explored. Recent work between the Met Office Hadley Centre, University of New South Wales and the National Climatic Data Center has generated a number of daily pseudo-datasets and discontinuity test cases from the GCM HadCM3 under control, natural and all forcings scenarios. Sampling density, variance, climatology and autocorrelation are adjusted to match the HadGHCND daily temperature record (maximum and minimum), and then various scenarios of discontinuities added.

At present, no agreed global standard exists against which to benchmark multiple datasets. Worse still, existing benchmarks are usually created by the dataset creators themselves leaving potential for unintentional tuning of algorithms or test cases. The issue is becoming increasingly critical. At present no homogenised sub-daily product exists, yet there is an ever growing need for such data. We therefore conclude the need for a global collaborative effort to: create globally applicable pseudo-data that sufficiently represent real world characteristics; comprehensively capture real world characteristics of inhomogeneity for creation of realistic discontinuity test cases; and objectively test all candidate homogenisation procedures produced against this pseudo-data with a realistic suite of discontinuity test cases applied. This effort should be independent from any single group of climate dataset creators (but may include some people with expertise from various institutions).

Discussion of the main issues:

Reconstructing full space and time sampling of the observational network characteristics

The majority of existing algorithms have already been skill tested using simple synthetic pseudo-data with simple discontinuity test cases applied which runs the risk of artificially awarding high performance to algorithms that cannot cope with real world natural variability and noise and the variation in sampling density and data completeness/continuity. Climatology (including annual and diurnal cycles), variance (white noise and natural variability), autocorrelation, sampling density (network coverage and missing data) and presence of a trend are real world characteristics of high importance that can affect algorithm skill and generally are not all included in such cases. Hence, to truly test algorithm skill on real data we need global reconstruction of real world characteristics including space and time sampling of the observational network as far as possible. Pseudo-data should be created at a range of resolutions (sub-daily to monthly) where the underlying characteristics and discontinuity test cases applied are identical. This will be particularly relevant to those

groups considering algorithm temporal transferability e.g. of a monthly algorithm to the daily timescale.

Source data for creation of homogeneous pseudo-data

As discussed, pseudo-data should be physically based on real world characteristics with spatio-temporal correlation structures (e.g. ENSO variability and teleconnections) that exist in the real world. There are three realistic options:

- **High quality homogenous observational datasets**
  These have the benefit of being real observations but may differ from the majority of stations which due to poorer quality will likely retain more white noise even after quality control. Furthermore, whether these constitute truly homogenous stations where instruments have been regularly calibrated to metrological standards is questionable. Critically, coverage of such stations is likely very small and so of limited use for benchmarking data from across the globe.

- **4th generation reanalyses products**
  These are globally complete, high resolution (gridded), and represent real world characteristics. The 20th Century Reanalysis[7] continues back to the early 20th century and as it only assimilates sea level pressure does not suffer from some of the major discontinuities apparent at major transitions in the global observing system in some other products although time-dependent biases are still possible. The most recent ECMWF reanalyses appear to adequately reflect surface temperature changes over land at least on the monthly mean timescale (Simmons et al., 2010). However, spurious discontinuities are present from major changes in the observing network (e.g. at the beginning of the satellite era) and post-1997/98 in ERA Interim likely due to a source change to the NCEP operational SST product.

- **CMIP5 GCM output[8]**
  These are globally complete, high spatial resolution (gridded) and 3 hourly for surface variables although verification compared to real-world data is essential first. They are homogeneous (insofar as they are at least consistent with the basic physics underlying each model and the applied forcings). However, care should be taken to choose a GCM, or preferably multiple GCMs, that represent low-resolution natural variability (e.g. ENSO, PDO etc.) and teleconnections satisfactorily as not all do. There are a range of forcing scenarios including control, high emissions (e.g. A1B) and natural only which can be used to assess some range of dataset algorithm performance with and without underlying trends.

Spatially and temporally complete gridded fields from reanalyses or GCMs can be sub-sampled and nudged with real world climatology (allowing for algorithms that can work in either absolute or anomaly space), variance, white noise and adjustment for autocorrelation from the databank (see White Papers 3 to 6). Where possible, efforts should collaborate with and build on existing work.

Exploring all eventualities with optimum discontinuity test case design

Test cases should be designed in parallel with homogenisation assessment requirements and with the broader assessment discussed in White Paper 10. They should encapsulate physically plausible effects of changes in: station location; instrument; recording practice; or surrounding environment and analyses should lead to clear and useful results. Some discontinuities are well documented and pinned to a period and region (i.e. mid-1990s automation in the USA) and these should be included. However, far more are undocumented and unknown and could be of any magnitude, frequency, clustering or sign bias and are likely a combination of all and a mix of abrupt and more graduated discontinuities. While we can characterise the main features of real world inhomogeneities relatively comprehensively (Elliott 1995, Peterson et al. 1998; Vincent et al. 2007) metadata is vastly incomplete and undigitised. A thorough review is necessary prior to creation of the discontinuity test cases for benchmarking and also to algorithms design (see White Paper 8). This could be facilitated via a relevant conference such as RMS, EGU or AGU. Digitising all available metadata and adding it to the databank is also strongly recommended (see white paper 3).

Approximately 10 global discontinuity test cases should ideally be constructed for benchmarking, full assessment (see White Paper 10), publication and collation of data products (see White Paper 13). These should be physically plausible scenarios based on our understanding of real world issues that likely pertain and include the control case of a homogeneous world to asses the effect of algorithms giving false positive detections and adjustments. These should incorporate a mix of abrupt, gradual and seasonally/diurnally varying discontinuities. They should methodically address key questions by testing skill under situations of: discontinuity clustering verses sparsity; proximity to endpoints verses midpoints; large verses small discontinuities; a combination of both; and the presence of strong verses no background trend (i.e. taken from control, A1B, c20c and natural climate runs).

Homogenisation algorithms will produce a new estimate of large scale features (climatology, variance and trends) that will fall somewhere on a benchmark spectrum populated by the inhomogenous world (pseudo-data with the suite of discontinuity test cases applied) and the target truth (the homogenous pseudo-data) for the region/regions in question. A perfect algorithm would recreate the target truth across a range of space and time scales. Performance in terms of percentage of successful, missed and incorrect adjustments should also be noted. This assessment will feed into the broader assessment described in White Paper 10. Although successful homogenisation of sub-daily or even daily resolution data has not yet been demonstrated the Surface Temperature initiative will likely spur attempts at this and so the benchmarking dataset needs to be available at high resolution.

Avoiding prior knowledge of test cases and resulting over-tuning of algorithms to discontinuity test cases

The benchmark dataset (pseudo-data with discontinuity test cases) will be freely available although it is advisable that assessment is done by an independent group (see White Paper 10). In the interests of transparency we argue that all pseudo-data creation, discontinuity test case creation and benchmarking is done independently from any single group of derived dataset creators. A multi-institution group or groups including expertise from non-traditional participants could be set up via the September workshop of those already involved in or with an interest in pursuing such

work. Applying to funding bodies will be essential. The methodology underlying this work should be fully documented and published via peer review including the pseudo-data with discontinuity test cases but withholding the 'solutions' (the original homogenous pseudo-data). Having a single suite of test cases does still pose the problem of potential to over-tune through multiple iterations of dataset creation in order to reduce uncertainty although it is unlikely that any knowing malpractice of tuning to the test data will occur, especially as full audit trails and code for dataset creation will also be published (White paper 13). Furthermore, there is low likelihood of over-tuning as long as we use a wide range of discontinuity types across the test-cases that are physically based and fully represent real world eventualities.

Recommendations:

- Global pseudo-data with real world characteristics
- GCM or Reanalyses data should be used as source base with real spatial, temporal and climatological characteristics applied to recreate to a reasonable approximation the observational record statistics
- Review of inhomogeneity across the globe finalised via a session at an international conference (link with White Paper 8) to ensure plausibility of discontinuity test cases
- Suite of ~10 discontinuity test cases that are physically based on real world inhomogeneities and orthogonally designed to maximise the number of objective science questions that can be answered
- Benchmarking to rank homogenisation algorithm skill in terms of performance using climatology, variance and trends calculated from homogenous pseudo-data and inhomogenous data (discontinuity test cases applied) (skill assessment to be synchronised with broader efforts discussed in White Paper 10)
- Independent (of any single group of dataset creators) pseudo-data creation, test case creation and benchmarking
- Peer-reviewed publication of benchmarking methodology and pseudo-data with discontinuity test cases but 'solutions' (original homogenous pseudo-data) to be withheld

References:

[6] http://www.homogenisation.org

[7] http://www.esrl.noaa.gov/psd/data/gridded/data.20thC_Rean.html

[8] http://cmip-pcmdi.llnl.gov/cmip5/data_description.html?submenuheader=1

Begert, M., Zenklusen, E., Haberli, C., et al., 2008: An automated procedure to detect discontinuities; performance assessment and application to a large European climate data set. *Meteorologische Zeitschrift*. **17**, (5), 663-672.

DeGaetano, A. T., 2006: Attributes of several methods for detecting discontinuities in mean temperature series. *Journal of Climate*. **19** (5), 838-853.

Ducré-Robitaille, J.-F., Vincent, L. A. & Boulet, G., , 2003: Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology*, **23**, 1087-1101.

Easterling, D. R. & Peterson, T. C., 1995: The effect of artificial discontinuities on recent trends in minimum and maximum temperatures. International Minimax Workshop on Asymmetric Change of Daily Temperature Range, SEP 27-30, 1993 COLLEGE PK, MD. *Atmospheric Research*. **37**, 19-26.

Elliott, W. P., 1995: On detecting long-term changes in atmospheric moisture. International Meeting of Experts on Long-Term Climate Monitoring by the Global Climate Observing System, JAN 09-11, 1995 ASHEVILLE, NC. *Climatic Change.* **31**, 349-367.

Menne, M. J. & Williams, C. N., 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal Of Climate*. **18,** 4271-4286.

Peterson, T. C., Easterling, D. R., Karl, T. R., et al., 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *International Journal Of Climatology.* **18**, 1493-1517.

Simmons, A. J., Willett, K. M., Jones, P. D., Thorne, P. W. & Dee, D., 2010: Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets. *Journal Of Geophysical Research-Atmospheres.* **115**, D01110, doi:10.1029/2009JD012442.

Titchner, H. A., Thorne, P. W., McCarthy, M. P. et al. 2009: Critically Reassessing Tropospheric Temperature Trends from Radiosondes Using Realistic Validation Experiments. *Journal Of Climate.* **22**, 465-485.

Vincent, L.A., 1998: A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate*, **11**, 1094-1104.

Vincent, L. A., van Wijngaarden, W. A. & Hopkinson, R., 2007: Surface temperature and humidity trends in Canada for 1953-2005. *Journal of Climate,* **20**, 5100-5113. DOI: 10.1175/JCLI4293.1.

Wang, X. L., Wen, Q. H., and Wu, Y., 2007: Penalized Maximal *t* Test for Detecting Undocumented Mean Change in Climate Data Series. *Journal of Applied Meteorology and Climatology*. **46**, 916-931. DOI:10.1175/JAM2504.

Wang, X. L., 2008a: 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**, 2423–2444. DOI: 10.1175/2008JAMC1741.1

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

Wang, X. L., Chen, H., Wu, Y. et al., 2010: New techniques for detection and adjustment of shifts in daily precipitation data series. *Journal of Applied Meteorology and Climatology*. (accepted)