Robust Benchmarking of Homogenisation Algorithms for the Surface Temperature Initiative

Kate Willett, Robin Chadwick, Lisa Alexander and Peter Thorne and members of the Benchmarking and Assessment Working Group

1. Why Benchmark?

21st Century requirements on Climate Data Products:
- **long-term**
- **spatially widespread**
- **high resolution (in space and time)**
- **traceable to instrument type / original record**
- **robust to varying non-climatic influences**

HOMOGENISATION IS ESSENTIAL

- **Multiple data products can increase confidence in a result e.g. long-term surface warming**
- **BUT Methodological choices can be diverse, with different strengths and weaknesses (Fig. 1)**

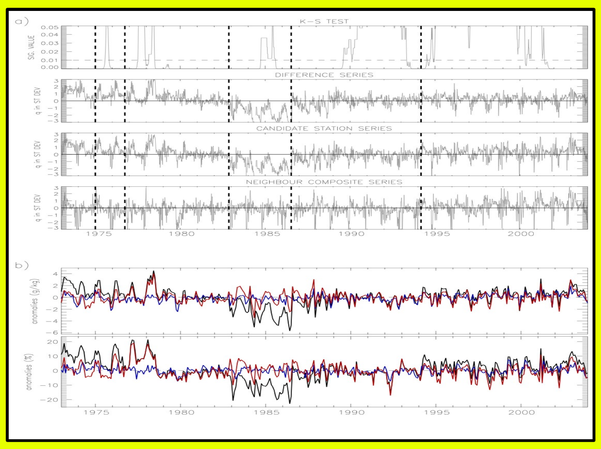

Fig. 1 Station data (black line) can have multiple discontinuities of different sign and magnitude from non-climatic influences. Homogenisation algorithm skill(red line) may be high for some discontinuity types but low for others.

- **We cannot be absolutely certain that even homogenised data are free from all errors**
- **BUT we can better understand the strengths and weaknesses of methodological choices**
- **Benchmarking of homogenisation algorithms will provide a quantifiable measure of uncertainty and facilitate algorithm development**

2. The Benchmarking and Assessment Working Group

Purpose: To facilitate use of a robust, independent and useful common benchmarking and assessment system for temperature data-product creation methodologies to aid product intercomparison and uncertainty quantification.

- review current understanding on non-climatic discontinuities affecting the surface temperature record – journal paper

- define the error models to be included in the benchmark datasets with which to test homogenisation algorithms (capturing all known real-world discontinuities)

- create these benchmark datasets as analogs to the consolidated master database (see section 5) that reflect real-world characteristics/noise

- coordinate 3 year working cycles: create new benchmarks (beginning of first year), algorithm testing (year 1-3), release of 'world-truth' and wrap up workshop (end of third year)

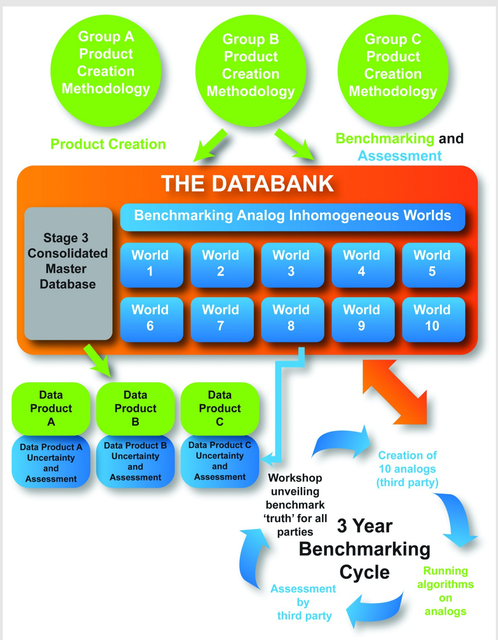

Fig. 2 Schematic of how the benchmarks and benchmarking cycle will work. Benchmarks will be available as part of the Surface Temperature Databank for data-product creators to test their algorithms on.

Members: Kate Willett (UKMO Hadley Centre) (Chair), Claude Williams (NCDC), Ian Jolliffe (Exeter Climate Systems, Uni. of Exeter), Robert Lund (Dep. Mathematical Sciences, Clemson Uni.), Lisa Alexander (Climate Change Research Centre, UNSW), Olivier Mestre (Meteo France), Stefan Brönniman (University of Bern), Lucie A. Vincent (Climate Research Division, Environment Canada), Aiguo Dai (Climate and Global Dynamics Division, NCAR), Steve Easterbrook (Dep. Computer Science, University of Toronto), Chris Wikle (Dep. Statistics, University of Missouri), Victor Venema (Meteorologisches Institut, University of Bonn)

3. Benchmarking Concepts

Create ~10 pseudo-worlds (error models) - apply breaks to the analog 'truth' reflecting true physics of real changes (e.g., Instrument changes, Location changes, Systematic instrument degradation, Urbanisation, Land use change, Local environment change, etc.). Break characteristics are dependent on radiation (time of day, time of year, cloudiness) and wind speed.

i. Create a globe of analog stations (identical to structure of consolidated master database) that reflect real-world characteristics for that station location (mean, variance, autocorrelation, missing data and co-variance with neighbours) without systematic bias – **the known 'TRUTH'**.

$$X_{TRUTH(t,l,h)} = S_{t,l,h} + T_{t,l,h} + \varepsilon_{t,l,h}$$

X = benchmark analog station at time *t*, location *l* and height *h*
S = seasonal cycles
T = trends (long-term signal, local effects, ENSO, NAO, Volcanoes, Solar Cycles etc.)
ε = random error at time/place/height (recording error, instrument error etc)
With a realistic temporal autocorrelation and spatial covariance structure

ii. These can be purely synthetic based on statistical models (above) or part synthetic based on downscaled physical models (GCMs) (Fig. 3). Either way, real-world characteristics are essential – algorithms must be able to cope with real-world noise. For each analog station, these characteristics can be drawn from the consolidated master database.

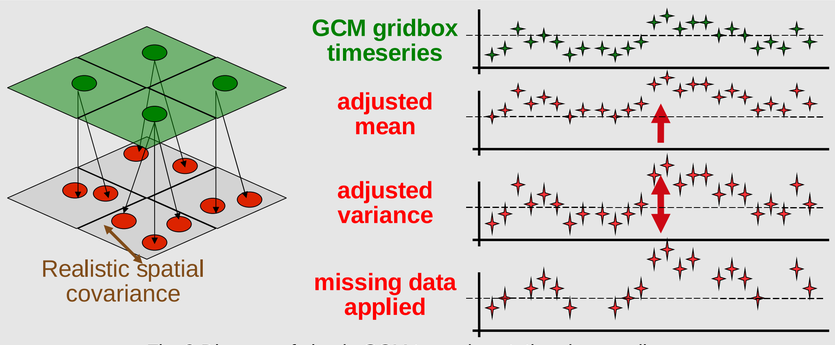

Fig. 3 Diagram of simple GCM to analog station downscaling.

$$X_{ERROR\_WORLD(t,l,h)} = S_{t,l,h} + T_{t,l,h} + B_{ERROR\_WORLD(t,l,h)} + \varepsilon_{t,l,h}$$

B = break at time/place/height (abrupt, gradual, seasonal, clustered, variance changes etc)

iii. A suite of error models (Fig. 4) can then be applied to create a number of analog 'worlds' that encompass all known possible breaks on a sliding scale from overly optimistic (e.g., few large breaks) to overly pessimistic (e.g., many breaks of differing magnitudes in addition to gradual changes, changes in the mean and the variance, seasonally varying changes) – **the known 'ERRORS'**.

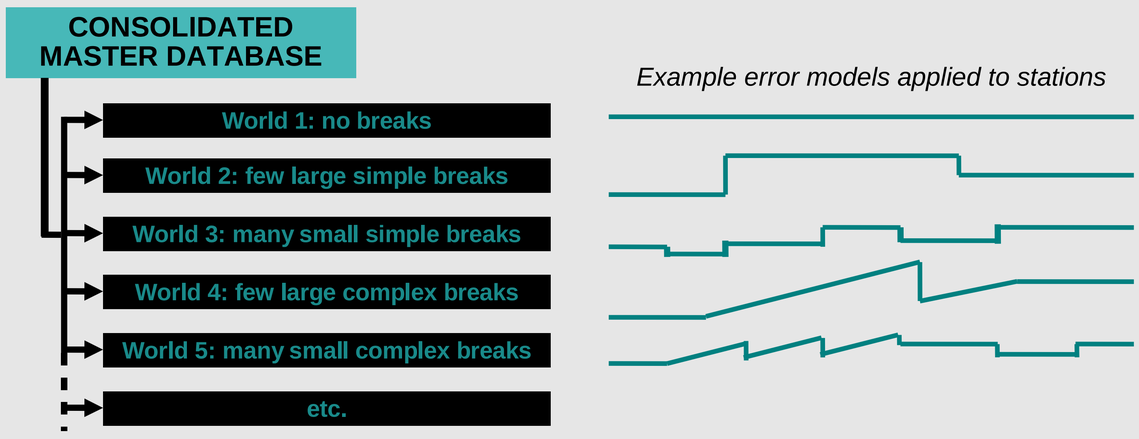

Fig. 4 Simplistic diagram of error model structure.

4. Assessment of the Benchmarks

The two components of benchmarking assessment:
1. **hit rate verses false alarm rate** - taking into account correct sign, location and magnitude within an acceptable range of error
2. **proximity of homogenised world meanstate to 'truth' meanstate** – how similar are region climatologies, variance, background trends, station autocorrelation, neighbour covariance?

i. Component 1. could be assessed using contingency tables (Fig. 5a) and ROC scores (Fig. 5b) although these would have to be weighted to take into account closeness of detected break location/magnitude to actual break location/magnitude and that a large break false alarm is worse than a small break miss.

a)

	Changepoint	No Changepoint
Detected (within +/- 3 months)	5	3
Not Detected (within +/- 3 months)	2	42 (potential detections given period of data)

Percent Correct Hit Rate: 90%
Heidke Skill Score = 61%
Probability of Detection hit rate = 71%
False Alarm Rate = 37%

b)

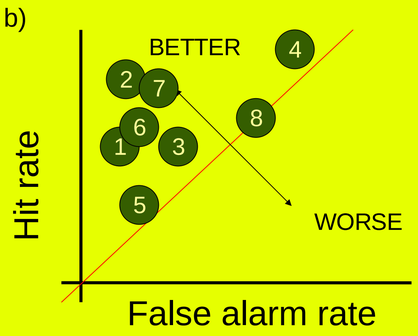

Fig. 6 Simplistic example of scoring hit rates verses false alarm rates for assessing the benchmarking.

ii. Assessing component 2. is more complex in terms of obtaining a quantifiable and comparable measure.

5. The Surface Temperature Initiative

To facilitate development of robust, high quality and traceable (to known reference standards and origin) monitoring products from hourly to century timescales and from location specific to the global mean.

- Endorsed by the 15th WMO Commission for Climatology Symposium
- First workshop – September 2010, UK Met Office, attended by climatologists, metrologists, statisticians, economists and IT experts.
- Data rescue: digitisation and open transfer of near-real time data.
- Comprehensive databank (Fig. 7): free, traceable, version control
- Benchmarking and Assessment: to assess product fitness for purpose, to enable cross-comparison and aid methodological advancements.
- Data-product portal hosting all databank related products with a suite of visualisation and cross-comparison tools.

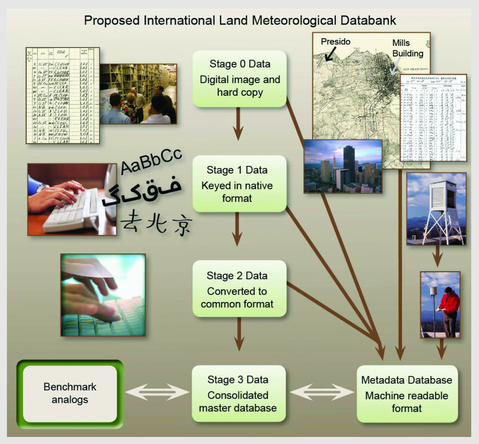

Fig. 6 Planned structure of the Surface Temperature Initiative Comprehensive Databank.

6. Useful Links

Website of the Benchmarking and Assessment Working Group:
Related documents, progress updates and membership.
http://www.surfacetemperatures.org/benchmarking-and-assessment-working-group

Blogsite for the Benchmarking and Assessment Working Group:
A place for open discussion – only members can post threads but anyone can comment.
http://surftempbenchmarking.blogspot.com

Website for the Surface Temperature Initiative:
http://www.surfacetemperatures.org

Met Office Hadley Centre, FitzRoy Road, Exeter, Devon, EX1 3PB United Kingdom
Tel: 01392 885680 Fax: 01392 885681
Email: kate.willett@metoffice.gov.uk