Geosci. Instrum. Method. Data Syst. Discuss., 4, 235–270, 2014
www.geosci-instrum-method-data-syst-discuss.net/4/235/2014/
doi:10.5194/gid-4-235-2014

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access

Concepts for benchmarking of homogenisation algorithm performance on the global scale

K. Willett[1], C. Williams[2], I. Jolliffe[3], R. Lund[4], L. Alexander[5], S. Brönniman[6], L. A. Vincent[7], S. Easterbrook[8], V. Venema[9], D. Berry[10], R. Warren[11], G. Lopardo[12], R. Auchmann[6], E. Aguilar[13], M. Menne[2], C. Gallagher[4], Z. Hausfather[14], T. Thorarinsdottir[15], and P. W. Thorne[16]

[1]Met Office Hadley Centre, FitzRoy Road, Exeter, UK
[2]National Climatic Data Center, Ashville, NC, USA
[3]Exeter Climate Systems, University of Exeter, Exeter, UK
[4]Department of Mathematical Sciences, Clemson University, Clemson, SC, USA
[5]ARC Centre of Excellence for Climate System Science and Climate Change Research Centre, University of New South Wales, Sydney, Australia
[6]Oeschger Center for Climate Change Research & Institute of Geography, University of Bern, Bern, Switzerland
[7]Climate Research Division, Science and Technology Branch, Environment Canada, Toronto, Canada
[8]Department of Computer Science, University of Toronto, Toronto, Canada
[9]Meteorologisches Institut, University of Bonn, Bonn, Germany

[10]National Oceanography Centre, Southampton, UK
[11]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK
[12]Istituto Nazionale di Ricerca Metrologica (INRiM), Torino, Italy
[13]Centre for Climate Change, Universitat Rovira i Virgili, Tarragona, Spain
[14]Berkeley Earth, Berkeley, CA, USA
[15]Norwegian Computing Center, Oslo, Norway
[16]Nansen Environmental and Remote Sensing Center, Bergen, Norway

Received: 27 February 2014 – Accepted: 21 May 2014 – Published: 4 June 2014

Correspondence to: K. Willett (kate.willett@metoffice.gov.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

GID
4, 235–270, 2014

Abstract

The International Surface Temperature Initiative (ISTI) is striving towards substantively improving our ability to robustly understand historical land surface air temperature change at all scales. A key recently completed first step has been collating all available records into a comprehensive open access, traceable and version-controlled databank. The crucial next step is to maximise the value of the collated data through a robust international framework of benchmarking and assessment for product intercomparison and uncertainty estimation. We focus on uncertainties arising from the presence of inhomogeneities in monthly surface temperature data and the varied methodological choices made by various groups in building homogeneous temperature products. The central facet of the benchmarking process is the creation of global scale synthetic analogs to the real-world database where both the "true" series and inhomogeneities are known (a luxury the real world data do not afford us). Hence algorithmic strengths and weaknesses can be meaningfully quantified and conditional inferences made about the real-world climate system. Here we discuss the necessary framework for developing an international homogenisation benchmarking system on the global scale for monthly mean temperatures. The value of this framework is critically dependent upon the number of groups taking part and so we strongly advocate involvement in the benchmarking exercise from as many data analyst groups as possible to make the best use of this substantial effort.

1 Introduction

Monitoring and understanding our changing climate requires freely available data with good spatial and temporal coverage that is of high quality, with remaining uncertainties well quantified. The work described herein forms part of the wider efforts of the International Surface Temperature Initiative to enable robust assessment of means, trends and variability of the historical climate.

GID
4, 235–270, 2014

The ISTI (www.surfacetemperatures.org; Thorne et al., 2011) is striving towards substantively improving our ability to robustly understand historical land surface air temperature change at all scales. A key recently completed first step has been collating all known freely available land surface meteorological records into an open access, traceable to known origin where possible, and version controlled databank (Rennie et al., 2014). To date the focus has been on monthly temperature time series, so far achieving a database of 31 999 unique records in the first release version as it stood on 14 November 2013 (Fig. 1).

There are multiple additional steps that must be performed subsequently to transform these fundamental data holdings into high quality data products that are suitable for robust climate research, henceforth referred to as climate data records (CDRs). At present a number of independent climate data groups maintain CDRs of land surface air temperature. Each uses its own choice of methods for a range of necessary processes (e.g. quality control, homogenisation, averaging, and in some cases interpolation). ISTI's second programmatic focus is to set up a framework to evaluate these methodological choices that ultimately lead to structural uncertainties in the trends and variability from CDRs. This paper focuses on evaluation of homogenisation methods, termed benchmarking and assessment, to reduce the uncertainty in trends and variability caused by inhomogeneity in the data and methods used to account for it.

The objective of this paper is to lay out the basic concepts for developing a comprehensive global benchmarking system for homogenisation of monthly land surface air temperature records. Section 2 discusses creation of spatio-temporally realistic analog station data. Section 3 discusses realistic but optimally assessable error models. Section 4 explores an assessment system that meets both the needs of algorithm developers and data-product users. Section 5 lays out a proposed benchmarking cycle to serve the needs of science and policy. Section 6 concludes.

CDRs should represent points in space, and be free from any non-climatic influences thereby providing a clean, homogeneous record. The unknown degree to which they do not represent true climatic changes hampers robust understanding. This has

GID
4, 235–270, 2014

consequences for informed decision making since observational records underpin all aspects of our understanding of climate change. With a handful of exceptions historical measurements have not been made in an SI (International System of units) traceable manner. Even the present day standard of a screened thermometer may still contain biases compared to the "true" WMO recommended standard of shaded free air temperature (WMO, 1992, 1998; Harrison, 2010, 2011). However, for analysis of changes in climate, achieving this WMO standard is less important than the long-term continuity of a given station and its practices. Unfortunately, change has been ubiquitous for the majority of station records (e.g. Lawrimore et al., 2011; Rohde et al., 2013). The dates of these changes (known as changepoints) are in many (very likely most) cases unknown and their impacts (known as inhomogeneity) either poorly quantified or more often than not entirely unquantified.

Climate observations made at individual stations exhibit multi-timescale variability made up of annual to decadal variations, seasonality and weather, all modulated by the station's micro-climate. Inhomogeneities can arise for a number of reasons such as station moves, instrument changes and changes in their exposure (shelter change), changes to the surrounding environment and changes to observing/reporting practices. While in the simplest cases a station may have one abrupt inhomogeneity in the middle of its series, which is relatively easy to detect, the situation can be far more complex with multiple changepoints leading to diverse inhomogeneities. For example, inhomogeneities may be:

- geographically or temporally clustered due to events which affect entire networks or regions;
- close to end points of time series;
- gradual or sudden;
- variance-altering;
- combined with the presence of a long-term background trend;

GID
4, 235–270, 2014

– small;

– frequent;

– seasonally or diurnally varying;

and often a combination of the above. A good overview of inhomogeneities in temperature and their causes can be found in Trewin (2010). Identifying the correct date (changepoint) and magnitude for any inhomogeneity against background noise is difficult, especially if it varies seasonally. Even after detection a series of decisions are required as to whether and how to adjust the data. While decisions are as evidence-based as possible, some are unavoidably ambiguous and can have a further non-negligible impact upon the resulting data. This is especially problematic for large datasets where the whole process by necessity is automated.

In this context attaining station homogeneity is very difficult; many algorithms exist with varying strengths, weaknesses and levels of skill (detailed reviews are presented in Venema et al., 2012; Aguilar et al., 2003; Peterson et al., 1998). Many are already employed to build global and regional temperature products used in climate research (e.g. Xu et al., 2013; Trewin, 2013; Vincent et al., 2012; Menne et al., 2009). While these algorithms can improve the homogeneity of the data, some degree of uncertainty is extremely likely to remain (Venema et al., 2012) depending on methodological choices. Narrowing these bands of uncertainty is highly unlikely to change the story of increasing global average temperature since the late 19th century. However, large scale biases could be reduced (Williams et al., 2012) and estimates of temperature trends at regional and local scales could be greatly affected.

The only way to categorically measure the skill of a homogenisation algorithm is to test it against a benchmark. In our context, a benchmark is a set of station data where the "truth" is known, as are the changepoints and inhomogeneity characteristics. The ability of the algorithm to locate the changepoints and adjust for the inhomogeneity, ideally returning the "truth", can then be measured.

GID
4, 235–270, 2014

Previous homogenisation efforts have used as homogeneous as possible real data or synthetic data with added inhomogeneities, or real data with known inhomogeneities to test homogenisation algorithms. Although valuable, station test cases are often relatively few in number (e.g. Easterling and Peterson, 1995) or lacking real-world complexity of both climate variability and inhomogeneity characteristics (e.g. Vincent, 1998; Ducre-Robitaille et al., 2003; Reeves et al., 2007; Wang et al., 2007, 2008a, b). Relatively comprehensive but regionally limited studies include Begert et al. (2008) who used the manually homogenised Swiss network as a test case.

The European homogenisation community (the HOME project; www.homogenisation.org; Venema et al., 2012) is the most comprehensive benchmarking exercise to date. HOME used stochastic simulation to generate realistic networks of ~ 100 European temperature and precipitation records. Their probability distribution, cross- and autocorrelations were reproduced using the so-called surrogate data approach (Venema et al., 2006). Inhomogeneities were added such that all stations contained multiple changepoints and the magnitudes of the inhomogeneities were drawn from a normal distribution. Thus, small undetectable inhomogeneities were also present, which influenced the detection and adjustment of larger inhomogeneities. Methods that addressed the presence of multiple changepoints within a series (e.g. Caussinus and Lyazrhi, 1997; Lu et al., 2010; Hannart and Naveau, 2012; Lindau and Venema, 2013) and the presence of changepoints within the reference series used in relative homogenisation (e.g. Caussinus and Mestre, 2004; Menne and Williams, 2005, 2009; Domonkos et al., 2011) clearly performed best in the HOME benchmark.

Recent studies have generated synthetic data test cases with varying degrees of real world characteristics (e.g. variance, station autocorrelation, multiple changepoints within a station record and a variety of inhomogeneity types) on larger scales (e.g. Menne and Williams, 2005; DeGaetano, 2006; Titchner et al., 2009; Williams et al., 2012). However, none offer sufficient complexity of test data with sufficient comprehensiveness of inhomogeneities. Furthermore, none are part of an internationally recognised system that could provide universally useful results.

GID
4, 235–270, 2014

For robust climate analysis and comparison of independent climate data products, it is necessary to agree on a global benchmarking and assessment system. The issue is becoming increasingly important, because policy decisions of enormous societal and economic importance are now being based on conclusions drawn from observational data. In addition to underpinning our level of confidence in the observations, developing and engendering a comprehensive and internationally recognised benchmark system would provide three key scientific benefits:

1. objective intercomparison of data-products,
2. quantification of the potential structural uncertainty of any one product,
3. a valuable tool for advancing algorithm development.

The Benchmarking and Assessment Working Group was set up during the Exeter, UK 2010 workshop for the ISTI. Its purpose is to develop and oversee the benchmarking process for homogenisation of temperature products as described here. Further details can be found at www.surfacetemperatures.org/benchmarking-and-assessment-working-group and blog discussions can be found at http://surftempbenchmarking.blogspot.com. The Benchmarking and Assessment Working Group reports to the Steering Committee and is guided by the Benchmarking and Assessment Terms of Reference hosted at www.surfacetemperatures.org/benchmarking-and-assessment-working-group.

2 Reproducing "real-world" data – the analog-clean-worlds

Simple synthetic analog-station data with simple inhomogeneities applied may artificially award high performance to algorithms that cannot cope with real world data. A true test of algorithm skill requires global reconstruction of real world characteristics including space and time sampling of the observational network. Hence, the ISTI benchmarks will replicate the spatio-temporal structure of the ~ 32 000 stations

GID
4, 235–270, 2014

in the ISTI databank stage 3 as far as possible (http://www.surfacetemperatures.org/databank; Fig. 2; Rennie et al., 2014) available from ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/monthly/stage3/.

The benchmark data must have realistic trends, variability, station autocorrelation and spatial cross-correlation. Conceptually, we consider individual station temporal variability of ambient temperature x at site s and time t as being able to be decomposed as:

$$x_{t,s} = c_{t,s} + l_{t,s} + v_{t,s} + m_{t,s,} \quad (1)$$

where:

- c represents the unique station climatology (the deterministic seasonal cycle). This will vary even locally due to the effects of topography, land surface type and any seasonal cycle of vegetation.
- l represents any long-term trend (not necessarily linear) that is experienced by the site due to climatic fluctuations such as in response to external forcings of the global climate system.
- v represents region-wide climate variability. That is to say interannual and interdecadal variability due to El Niño and La Niña events, annular modes (AO and AAO), or multidecadal variations such as the Pacific Decadal Oscillation or Atlantic Multidecadal Oscillation. Such modes have regionally distinct patterns of surface temperature response e.g. a positive AO yields warm winters over Northern Europe.
- m represents the station micro-climate (local variability). Such station-specific deviations are oftentimes weakly autocorrelated and cross-correlated with nearby stations, but tend to be more distinct on a station-by-station basis than the remaining terms in Eq. (1).

GID
4, 235–270, 2014

These terms are strictly additive in this conceptual framework. Such a decomposition is necessary to be able to subsequently build realistic series of $x_{t,s}$ on a network wide basis that retain plausible station series, neighbour series and regional series characteristics including mean, variability and cross-correlations. Below, a discursive description of the necessary steps and building blocks envisaged is given. A full description of the methods will be forthcoming when the benchmarks are released.

Most algorithms analyse the difference between a candidate station and a reference station (or composite). Crucially, temperature anomalies (where the seasonal cycle, c, has been removed) are used to create the difference series. The large-scale trend, l and variability, v, are highly correlated between candidate and reference series and so mostly removed by the differencing process. It is thus critical that the variability, autocorrelation and spatial cross-correlations in m are realistic.

For the benchmarking process, Global Climate Models (GCMs) can provide gridded values of l (and possibly v) for monthly mean temperature. GCMs simulate the global climate using mathematical equations representing the basic laws of physics. GCMs can therefore represent the short and longer-term behaviour of the climate system resulting from solar variations, volcanic eruptions and anthropogenic changes (external forcings). They can also represent natural large-scale climate modes (e.g. El Niño–Southern Oscillation – ENSO) and associated teleconnections (internal variability). However, the gridded nature of GCM output means that GCMs cannot give a sufficiently realistic representation of fine-scale meteorological data at point (station) scale. Hence, they cannot be used directly to provide the m and c components at the point (station) level. However, the l and v components are expected to vary very little between stations that are close (e.g. within a gridbox). There are two advantages of using GCMs to provide l and v. Firstly, they provide globally consistent variability that can be associated with ENSO-type events or other real modes of variability with large spatial influence along with at least broad-scale topography and its influence. Secondly, there are different forcing scenarios available (e.g. no anthropogenic emissions, very

GID
4, 235–270, 2014

high anthropogenic emissions) providing opportunities to explore how different levels of background climate change effect homogenisation algorithm skill.

The annually constant c component in Eq. (1) is straightforward to calculate for each station and then apply to the synthetic stations. The m and v (if not obtained from a GCM) component can be modelled statistically from the behaviour of the real station data. Statistical methods such as vector auto-regressive (VAR) type models (e.g. Brockwell and Davis, 2006) must be invoked to reproduce the spatio-temporal correlations but limitations exist where stations are insufficiently long or stable enough to be modelled. Balancing sophistication of methods with automation and capacity to run on $\sim 32\,000$ stations is key. Ensuring spatial consistency across large distances (100s of km) necessitates high-dimensional matrix computations or robust overlapping window techniques.

Ultimately, while analog-clean-world month-to-month station temperatures need not be identical to real station temperatures, real station climatology, variability, trends, autocorrelation and cross-correlation with neighbours should be maintained. Analog-clean-world station temporal sampling can be degraded to varying levels of missing data as necessary.

3 Devising realistic but optimally assessable error models – the analog-error-worlds

The analog-error-worlds will be based on a series of analog-clean-worlds and will be created by adding inhomogeneities from predefined error-models. These error-models should be designed with the three aims of the ISTI in mind i.e. to aid product intercomparison; to help quantify structural uncertainty; and to aid methodological advancement. There will be both *blind benchmarks*, where the answers/analog-clean-worlds underlying the released analog-error-worlds will not be made public for a time; and *open benchmarks*, where the answers/analog-clean-worlds will be publicly available immediately.

GID
4, 235–270, 2014

Blind benchmarks will be used for formal assessment of algorithms and data products. By being blind they prevent optimisation to specific features. While certain features will be widely known, it should not be known which world explores which type of features or the exact changepoint/inhomogeneity magnitude. For the most part these blind worlds should be physically plausible scenarios based on our understanding of real world issues. Their inclusion of the control case of a homogeneous world will enable the assessment of the effect of false detections and the potential for algorithms to do more harm than good. Ultimately, they should be designed to lead to clear and useful results, distinguishing strengths and weaknesses of algorithms against specific inhomogeneity and climate data characteristics. They need to achieve this without completely overloading algorithm creators from the outset either with a multitude of complexities in all cases or with too many analog-error-worlds to contend with.

The *open benchmarks* will enable algorithm developers to conduct their own immediate tests comparing their homogenised efforts from the analog-error-worlds with the corresponding analog-clean worlds. These open worlds will also be useful for exploring some of the more exotic problems or alternatively those straightforward issues that do not require a full global station database to explore.

To ensure focus on homogenisation methods, benchmarks will not include random error due to isolated instrument faults or observer/reporting mistakes. For monthly averages, random errors at observation times will often average out. Given a reasonable level of quality control, an essential step in any CDR processing, these errors are not thought to impact long-term trend assessment although for individual stations this may not be the case. Regardless, the assumption here is that all data will have been quality controlled to some extent prior to homogenising. Hence, users will not be required to quality control the analog-error-worlds although they are strongly recommended to quality control the real ISTI databank. In future versions of the benchmarks, specific error worlds could include known types of random error to test how this affects the homogenisation algorithm skill.

GID
4, 235–270, 2014

Systematic errors are the key problem for station homogeneity and the prime focus for these benchmarks. These are persistent offsets or long-term trends away from the true ambient temperature (metrologically speaking an artefact which causes the measurement to differ in a sustained manner from the true value of the measurand). Sets of different systematic inhomogeneities can be added to the analog-clean-world stations to create inhomogeneous analog-error-worlds. Conceptually, for any analog-station x as denoted by Eq. (1) a d term can be added to represent an inhomogeneity at time t and location l to give an observed value x' which differs from the true value (x):

$$x'_{t,s} = c_{t,s} + l_{t,s} + v_{t,s} + m_{t,s} + d_{t,s}. \quad (2)$$

At any point in time, d may be zero, a constant (possibly with some seasonal or climate related variation e.g. an instrument change may yield a warm bias in winter and a cool bias in summer if not well ventilated) or a value that grows/declines over time as e.g. a tree grows or urban areas encroach. Experience with current benchmarks over restricted regions (Williams et al., 2012; Venema et al., 2012) suggest that several artefacts exist in most stations such that the $d_{t,s}$ term may change several times during the period of record of a station (roughly every 10 to 30 years or more often).

By necessity, homogenisation algorithms have to make an assumption that a given station is at least locally representative at some point in its record. For convenience, and because the major interest is change in temperature rather than actual temperature, the most recently observed period is treated as the reference period by the majority of algorithms. Any adjustments are made relative to this period. Hence, our assessment will assume all stations are representative in the most recent part of their record such that d is zero at present day and develops backwards. In a perfect case, a homogenisation algorithm would detect d in the analog-error-world correctly, remove it, and return x' to its true ambient temperature x from the original analog-clean-world (Eq. 1).

These d elements should be physically plausible representations of known causes inhomogeneity (e.g. station moves, instrument malfunctions or changes, screen/shield

GID
4, 235–270, 2014

changes, changes to observing practice over time) as summarised in Table 1. A range of frequencies and magnitudes should be explored. Ideally, they should take into account the effect on temperature from the change in climate covariates (e.g. rainfall, humidity, radiation, windspeed and direction) as accurately as possible at present, accepting that in the current state of knowledge this will in many respects be an assumption based on expert judgement. Many complicated examples of covariate impacts on d exist. For example, in soil moisture limited regions changing vegetation between wet years and dry years increases variability compared to a more constant soil moisture environment (B. Trewin, personal communication, 2013; Seneviratne et al., 2012).

Inhomogeneities added should be both abrupt and gradual, including the effects of land use change, such as rural-to-urban developments, which are important for some applications. They should explore changes that vary with season which can result in changes in variance as well as the mean. Some should be geographically common, reflecting both region-wide changes, and others isolated. Isolated changes may arise due to the need to replace broken equipment or when stations are maintained by individual volunteers or groups. Region-wide changes tend to occur in networks that are centrally managed or owned.

Some inhomogeneities are reasonably well understood and apply to a given period and region e.g.:

- north wall measurements to Stevenson screens in the 19th century (Böhm et al., 2001),
- French screens to Stevenson screens around 1900 (Brunet et al., 2006),
- wild screens to Stevenson screens in the mid 20th century (Auchmann and Brönnimann, 2012),
- Stevenson screen with liquid in glass thermometers to electronic thermistors (Maximum/Minimum Temperature System) in the USA in the mid-1980s (Quayle et al., 1991; Menne et al., 2009),

GID
4, 235–270, 2014

– tropical sheds to Stevenson screens in the Tropics during the early 20th century (Parker, 1994).

These (or similar) could be included in one or more of the analog-error-worlds. However, the inhomogeneities are commonly undocumented and unknown and could be of any magnitude, frequency, clustering or sign and are likely a combination of all these. Current efforts are ongoing to collect together times and types of changes known to have occurred for each country (http://www.surfacetemperatures.org/benchmarking-and-assessment-working-group#Working Group Documents). It is envisaged to replicate what we believe to be realistic regional distributions of inhomogeneities within at least some subset of the analog-error-worlds.

Metadata have been used by homogenisers as a useful tool for improving the detection of changepoints. Substantive metadata are digitally available for the US Cooperative Observer Network which comprises the bulk of US station data. Elsewhere, digital holdings are rare but will likely be made available digitally in the future. Therefore, alongside the analog-error-worlds some changepoints should be documented, some should not be and some should have documented changes where no actual temperature change is effected. The latter could relate either to an inconsequential change in instrumentation/procedure or a false metadata event in the record.

A selection of error-models should be chosen to explore different features of both the type of inhomogeneity (e.g. size, frequency, seasonality, geographic pervasiveness etc.) and characteristics of the real-world observing systems (e.g. variability, trends, missing data etc.). Worlds should incorporate a mix of inhomogeneity types discussed above and the set of worlds should be broad, covering a realistic range of possibilities so as not to unduly penalise or support any one type of algorithm or too narrowly confine us to one a priori hypothesis as to real-world error structures. They should methodically address key questions by testing skill under these situations (e.g. changepoint clustering vs. sparsity; proximity of changepoints to the end vs. the middle of station records; large vs. small inhomogeneities; a combination of both; and the presence of strong vs. no background trend) as shown in Fig. 3.

GID
4, 235–270, 2014

4 Developing an assessment system that meets all needs

Any data-product creators utilising the ISTI databank and undertaking homogenisation will be encouraged to take part in the benchmarking as a means of improving the uncertainty estimation (specifically homogenisation uncertainty) of their product. This will involve running their homogenisation algorithms on the blind analog-error-worlds to create adjusted analog-error-worlds, just as they have done for the real ISTI databank stations. To take part they must submit homogenised benchmark data and results to the Benchmarking and Assessment Working Group for assessment. In time this process could be automated through a webpage which would also assist users of the open benchmarks.

There are two components of assessment: how well are individual changepoints located and their inhomogeneity characterised and how similar is the adjusted analog-error-world to its corresponding analog-clean-world? An algorithm may do very well at retrieving the climatology or trend behaviour without necessarily performing well in detecting individual changepoints/inhomogeneities, or vice versa. Algorithms may perform well at characterising long-term regional trends but have markedly different performance characteristics at sub-regional and shorter timescales.

The assessment can be split into four different levels:

Level 1 – difference between analog-clean-world and homogenised series analog-error-world climatology, variance and trends.

Level 2 – measures such as hit and false alarm rates for correct detection of changepoints and inhomogeneity character.

Level 3 – detailed assessment of strengths and weaknesses against specific types of inhomogeneity and observing system issues.

Level 4 – reality of the various analog-error-worlds assessed by comparing characteristics of inhomogeneities found in real data to that found in the analog-error-worlds. This will help improve future benchmarks.

GID
4, 235–270, 2014

For Level 1 assessment of large scale features (e.g. c, l and v in Eq. 1), a perfect algorithm would return the analog-clean-world features across a range of space and time scales. Algorithms should, ideally, at least make the analog-error-worlds more similar to their analog-clean-worlds. This information can be calculated for stations, regional means or global averages from each adjusted analog-error-world. Similarity can be measured in terms of proximity in °C for mean and linear trend approximations and standard deviation for variance. This can be presented as percentage recovery (after Williams et al., 2012). An example is shown in Fig. 4 for linear trend approximations with further explanation. Although linear trends do not describe the data perfectly, they provide a simple measure of long-term tendency that can be compared. This method does not indicate algorithms that result in a linear trend of the wrong sign (positive or negative). This may be seen as a more serious problem than a linear trend being over or under estimated and so would need to be treated separately. Other scores, such as the squared error or the absolute error, could also be used to measure differences between adjusted analog-error-worlds and analog-clean-worlds.

Level 2 and 3 measures, such as the hit/false alarm rates, could also be split into accuracy of changepoint location and the accuracy of adjustments applied. Furthermore, a sliding scale may be used to penalize close but not exact hits rather than assigning them as misses. Care should be taken though considering that some algorithms may adjust the inhomogeneous data well, performing highly in the level 1 assessment, while not locating changepoints accurately or vice versa. For example, many small inhomogeneities may be homogenised by locating a single large amplitude inhomogeneity. Similarly, a large inhomogeneity may be homogenised by applying many small adjustments. Large inhomogeneities are easier to detect than small ones so assessment could be split into inhomogeneity size categories (e.g. Zhang et al., 2013). This information is of importance to algorithm developers. Arguably, adjusting for detected inhomogeneities that are not actual inhomogeneities (false detection) adds error to the data and so could be scored more negatively than missing a real inhomogeneity. However, this critically depends upon the size of the adjustments applied. If adjustments for

GID
4, 235–270, 2014

false detections are small there will be little change in climatology and trend statistics, hence the cost of false detection diminishes.

Such assessments of detection and adjustment skill could be done through contingency tables (Table 2) where numbers of hits, misses, false alarms and "correct misses" are counted and used to construct various skill scores (Menne and Williams, 2005). Defining the number of "correct misses" is not straightforward, especially where a sliding scale is used to define a "hit", and needs to be investigated. Alternatively, measures that consider only hits, misses and false alarms may be used. The ideas used to assess detection skill can be adapted to investigate size-of-adjustment skill, as shown in red in Table 2. This could be visualised for each data-product using an adapted form of an ROC (Receiver Operating Curve) plot (Fig. 5) where each analog-error-world result is positioned according to its hit rate and false alarm rate. Users can quickly see on which worlds that particular data-product/algorithm scores highly, and which worlds are problematic. This can be used to infer applicability of data-products for a specific use or intercomparison with data-products created from different algorithms.

Levels 1 and 2 are of primary focus for assessing uncertainty and comparing data-products. Level 3 is of more importance to algorithm developers than data-product users, informing where best to focus future algorithm improvements. Level 4 is mainly aimed at the working group. For the first benchmark cycle, assessment will focus only on levels 1 and 2 to provide a quick response to the benchmark users. Ultimately, all worlds and results from the assessment will be made publicly available, ideally alongside any associated data-products. This will allow further bespoke assessment as required by interested analysts.

It is important that this process is made easy to encourage participation. In a perfect world all participants would submit a homogenised version of all stations in each analog-error-world and a list of detected changepoints and applied adjustments (e.g. Fig. 6). This would enable assessment of all levels. However, it is more likely that different groups will select different stations based on their desired end-product. These may be limited to long stations only or limited to specific regions. Some groups may

GID
4, 235–270, 2014

wrap their homogenisation into a fully gridded product such that they are unable to provide individual homogenised stations or a list of adjustments. This would prevent any level 2 and 3 assessment.

Given the above, while it is important to specify an ideal set of return items (e.g. homogenised stations, changepoint locations and inhomogeneity adjustments) to submit as part of the benchmarking assessment it is also important to have the capacity to accept a wide variety of submissions. This may be done for level 1 by performing assessment both at the station scale and also at the regional average scale, accepting that some component of differences found will be due to station selection and gridding methods. For stations and regional averages participants could be asked to submit their best estimates of specified statistics such as the climatology, variance and linear trend (e.g. Fig. 7). A set of regions could be specified such as the Giorgi regions commonly used within many aspects of climate science (Giorgi and Francisco, 2000) in addition to hemispheric and global averages. It would also be possible to specify a minimum subset of stations to be homogenised to allow fair comparison across the regionally focussed products, some of which may use manual homogenisation methods and so be unable to tackle global scale homogenisation (cf. Venema et al., 2012). An important distinction could be made between best estimates of clean-world regional statistics and statistics calculated on all stations within that region. In some cases it would be a wise decision to remove a station that has too many missing data or that is too poor quality, but comparisons using all stations in the analog-clean-world compared to a participant's best estimate may penalise such approaches.

5 Providing a working cycle of benchmarking to serve the needs of science and policy

A repeatable cycle of blind benchmark release (analog-error-worlds), homogenisation period, assessment period, release of the underlying analog-clean-worlds/answers (changepoint locations, size and shape of inhomogeneities added) and a wrap-up

GID
4, 235–270, 2014

workshop would encourage people to use the benchmarks and allow for sequential improvement of the benchmarks and investigation of different homogenisation issues. Not all issues will be able to be covered in the first cycle. This could be a 3 year cycle, overseen by the Benchmarking and Assessment Working Group (Fig. 8).

If the cycle is too short then there are risks that not enough people will get involved, reducing the usefulness of product intercomparison. If the cycle is too long then the benchmarks become out of date. Additionally, much may be learned about each analog-error-world from homogenising without release of the underlying analog-clean-world. This runs the risk of algorithms becoming over-tuned to these specific worlds.

The wrap-up would bring together users and creators of the benchmarks to assess how they were useful and how they can be improved for the next cycle. This will likely be in the form of a workshop and overview analysis paper. The databank will develop over time as will algorithms and the benchmarks will need to be updated both in terms of station coverage and methodologically.

The focus here is limited to monthly mean temperature data but it is envisaged that maximum and minimum temperatures and, subsequently, daily temperature records will be included in the future. Also, the current framework is only set up to assess the homogenisation algorithm skill. There are many different aspects of data-product creation including quality control processes, station selection and interpolation and gridding methods. The benchmarks created here could also be used to assess some of these but at this time it was thought advantageous to focus only on the homogenisation element in order to make relatively rapid progress. We hope that the provision of this benchmarking framework will broaden in the future to include these other important aspects of data-product creation.

GID
4, 235–270, 2014

6 Concluding remarks

An international and comprehensive benchmarking system for homogenisation of global surface temperature data is essential for constraining the uncertainty in climate data arising from changes made to our observing system. The International Surface Temperature Initiative is in a unique position to undertake this work and provide testing alongside the provision of the raw climate data. A repeating cycle of benchmarking assessment has been proposed including concepts for creation of benchmark data and assessment. The task is large and will take time to accomplish. However, this will for the first time enable global scale quantification of uncertainty in station inhomogeneity, which is one of the least understood areas of uncertainty associated with the land surface air temperature record.

Assessment of skill against the benchmarks will enable meaningful inter-comparison of surface temperature products and assessment of fitness for purpose for a broad range of end-users from large scale climate monitoring to local-scale societal impacts analysis. Such a detailed and global testing of homogenisation algorithms will also be a significant aid to algorithm developers, hopefully resulting in vastly improved algorithms for the future. These benchmarks can also be used to test other aspects of climate data record production such as station selection and interpolation. If successful, this work should significantly improve the robustness of monthly surface temperature climate data records on a range of spatial scales. This will improve the accuracy of assessment of recent changes in surface temperature and associated uncertainties to end-users.

Ultimately, the value of these benchmarks will only be as great as the number of groups participating in the exercise. The authors therefore strongly advocate development of new approaches and climate data records by new groups. The value of the new records will be greatly enhanced by undertaking benchmark testing as well as by using ISTI databank data.

GID
4, 235–270, 2014

Acknowledgements. The work of Kate Willett was supported by the Joint UK DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). Renate Auchmann was funded by the Swiss National Science Foundation (Project "TWIST").

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T. C., and Wieringa, J.: Guidelines on Climate Metadata and Homogenization, WCDMP 53, WMO-TD 1186, World Meteorol. Organ., Geneva, Switzerland, 55 pp., 2003.

Auchmann, R. and Brönnimann, S.: A physics-based correction model for homogenizing sub-daily temperature series, J. Geophys. Res., 117, D17119, doi:10.1029/2012JD018067, 2012.

Begert, M., Zenklusen, E., Haberli, C., Appenzeller, C., and Klok, L.: An automated procedure to detect discontinuities; performance assessment and application to a large European climate data set, Meteorol. Z., 17, 663–672, 2008.

Böhm, R., Auer, I., Brunetti, M., Maugeri, M., Nanni, T., and Schöner, W.: Regional temperature variability in the European Alps 1760–1998 from homogenized instrumental time series, Int. J. Climatol., 21, 1779–1801, 2001.

Brockwell, P. J. and Davis, R. A.: Time Series: Theory and Methods, 2nd Edn., Springer, New York, NY, 2006.

Brunet, M., Saladié, O., Jones, P. D., Sigró, J., Aguilar, E., Moberg, A., Lister, D., Walther, A., Lopez, D., and Almarza, C.: The development of a new dataset of Spanish daily adjusted temperature series (1850–2003), Int. J. Climatol., 26, 1777–1802, doi:10.1002/joc.1338, 2006.

Caussinus, H. and Lyazrhi, F.: Choosing a linear model with a random number of change-points and outliers, Ann. I. Stat. Math., 49, 761–775, 1997.

Caussinus, H. and Mestre, O.: Detection and correction of artificial shifts in climate series, J. Roy. Stat. Soc. C-App., 53, 405–425, 2004.

DeGaetano, A. T.: Attributes of several methods for detecting changepoints in mean temperature series, J. Climate, 19, 838–853, 2006.

Domonkos, P., Poza, R., and Efthymiadis, D.: Newest developments of ACMANT, Adv. Sci. Res., 6, 7–11, doi:10.5194/asr-6-7-2011, 2011.

GID

4, 235–270, 2014

Ducré-Robitaille, J.-F., Vincent, L. A., and Boulet, G.: Comparison of techniques for detection of changepoints in temperature series, Int. J. Climatol., 23, 1087–1101, 2003.
Easterling, D. R. and Peterson, T. C.: The effect of artificial changepoints on recent trends in minimum and maximum temperatures, International Minimax Workshop on Asymmetric Change of Daily Temperature Range, College Park, MD, 27–30 September 1993, Atmos. Res., 37, 19–26, 1995.
Giorgi, F. and Francisco, R.: Evaluating uncertainties in the prediction of regional climate change, Geophys. Res. Lett., 27, 1295–1298, 2000.
Hannart, A. and Naveau, P.: An improved Bayes information criterion for multiple change-point models, Technometrics, 54, 256–268, 2012.
Harrison R. G.: Natural ventilation effects on temperatures within Stevenson screens, Q. J. Roy. Meteorol. Soc., 136, 253–259, doi:10.1002/qj.537, 2010.
Harrison R. G.: Lag-time effects on a naturally ventilated large thermometer screen, Q. J. Roy. Meteorol. Soc., 137, 402–408, doi:10.1002/qj.745, 2011.
Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S., and Rennie, J.: An overview of the Global Historical Climatology Network monthly mean temperature dataset, Version 3, J. Geophys. Res.-Atmos., 116, D19121, doi:10.1029/2011JD016187, 2011.
Lindau, R. and Venema, V. K. C.: On the multiple breakpoint problem and the number of significant breaks in homogenisation of climate records, Idojaras, 117, 1–34, 2013.
Lu, Q., Lund, R. B., and Lee, T. C. M.: An MDL approach to the climate segmentation problem, Ann. Appl. Stat., 4, 299–319, doi:10.1214/09-AOAS289, 2010.
Menne, M. J. and Williams, C. N.: Detection of undocumented changepoints using multiple test statistics and composite reference series, J. Climate, 18, 4271–4286, 2005.
Menne, M. J. and Williams Jr., C. N.: Homogenization of temperature series via pairwise comparisons, J. Climate, 22, 1700–1717, 2009.
Menne, M. J., Williams Jr., C. N., and Vose, R. S.: The United States Historical Climatology Network monthly temperature data – version 2, B. Am. Meteorol. Soc., 90, 993–1007, 2009.
Parker, D. E.: Effects of changing exposure of thermometers at land stations, Int. J. Climatol., 14, 1–31, 1994.
Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E. J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., and

GID
4, 235–270, 2014

Parker, D.: Homogeneity adjustments of in situ atmospheric climate data: a review, Int. J. Climatol., 18, 1493–1517, 1998.

Quayle, R. G., Easterling, D. R., Karl, T. R., and Hughes, P. Y.: Effects of recent thermometer changes in the Cooperative Station Network, B. Am. Meteorol. Soc., 72, 1718–1723, 1991.

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q.: A review and comparison of changepoint detection techniques for climate data, J. Appl. Meteorol. Clim., 46, 900–915, doi:10.1175/JAM2493.1, 2007.

Rennie, J. J., Lawrimore, J. H., Gleason, B. E., Thorne, P. W., Morice, C. P., Menne, M. J., Williams, C. N., Gambi de Almeida, W., Christy, J. R., Flannery, M., Ishihara, M., Kamiguchi, K., Klein-Tank, A. M. G., Mhanda, A., Lister, D. H., Razuvaev, V., Renom, M., Rusticucci, M., Tandy, J., Worley, S. J., Venema, V., Angel, W., Brunet, M., Dattore, B., Diamond, H., Lazzara, M. A., Le Blancq, F., Luterbacher, J., Mächel, H., Revadekar, J., Vose, R. S., and Yin, X.: The International Surface Temperature Initiative: global land surface databank: monthly temperature data, version 1, release description and methods, Geosci. Data J., in press, 2013.

Rohde, R., Muller, R. A., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., Wurtele, J., Groom, D., and Wickham, C.: A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011, Geoinfor. Geostat., 1, 1, doi:10.4172/2327-4581.1000101, 2013.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X.: Changes in climate extremes and their impacts on the natural physical environment, in: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, edited by: Field, C. B., Barros, V., Stocker, T. F., Qin, D., Dokken, D. J., Ebi, K. L., Mastrandrea, M. D., Mach, K. J., Plattner, G.-K., Allen, S. K., Tignor, M., and Midgley, P. M., A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), Cambridge University Press, Cambridge, UK, and New York, NY, USA, 109–230, 2012.

Thorne, P. W., Willett, K. M., Allan, R. J., Bojinski, S., Christy, J. R., Fox, N., Gilbert, S., Jolliffe, I., Kennedy, J. J., Kent, E., Klein Tank, A., Lawrimore, J., Parker, D. E., Rayner, N., Simmons, A., Song, L., Stott, P. A., and Trewi, B.: Guiding the creation of a comprehensive surface temperature resource for 21st century climate science, B. Am. Meteorol. Soc., 92, ES40–ES47, doi:10.1175/2011BAMS3124.1, 2011.

GID
4, 235–270, 2014

Titchner, H. A., Thorne, P. W., McCarthy, M. P., Tett, S. F. B., Haimberger, L., and Parker, D. E.: Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments, J. Climate, 22, 465–485, 2009.

Trewin, B.: Exposure, instrumentation, and observing practice effects on land temperature measurements, WIREs Climate Change, 1, 490–505, 2010.

Trewin, B.: A daily homogenized temperature data set for Australia, Int. J. Climatol., 33, 1510–1529, doi:10.1002/joc.3530, 2013.

Venema, V., Bachner, S., Rust, H. W., and Simmer, C.: Statistical characteristics of surrogate data based on geophysical measurements, Nonlin. Processes Geophys., 13, 449–466, doi:10.5194/npg-13-449-2006, 2006.

Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P., and Brandsma, T.: Benchmarking homogenization algorithms for monthly data, Clim. Past, 8, 89–115, doi:10.5194/cp-8-89-2012, 2012.

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

Vincent, L. A., Wang, X. L., Milewska, E. J., Wan, H., Yang, F., and Swail, V.: A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis, J. Geophys. Res., 117, D18110, doi:10.1029/2012JD017859, 2012.

Wang, X. L.: Accounting for autocorrelation in detecting mean-shifts in climate data series using the penalized maximal *t* or *F* test, J. Appl. Meteorol. Clim., 47, 2423–2444, doi:10.1175/2008JAMC1741.1, 2008a.

Wang, X. L.: Penalized maximal *F* test for detecting undocumented mean-shift without trend change, J. Atmos. Ocean. Tech., 25, 368–384, doi:10.1175/2007JTECHA982.1, 2008b.

Wang, X. L., Wen, Q. H., and Wu, Y.: Penalized maximal *t* test for detecting undocumented mean change in climate data series, J. Appl. Meteorol. Clim., 46, 916–931, doi:10.1175/JAM2504.1, 2007.

Williams Jr., C. N., Menne, M. J., and Thorne, P.: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States, J. Geophys. Res., 117, D05116, doi:10.1029/2011JD016761, 2012.

GID
4, 235–270, 2014

WMO: WMO No. 182, International Meteorological Vocabulary, Geneva, Switzerland, 1992.
WMO: Final Report, Commission for Basic Systems, Working Group on Data Processing, Task Group on WMO/CTBTO Matters, 15–17 July 1998, available at: www.wmo.int/pages/prog/www/reports/wmo-ctbto.html, Geneva, Switzerland, 1998.
Xu, W., Li, Q., Wang, X. L., Yang, S., Cao, L., and Feng, Y.: Homogenization of Chinese daily surface air temperatures and analysis of trends in the extreme temperature indices, J. Geophys. Res.-Atmos., 118, 9708–9720, doi:10.1002/jgrd.50791, 2013.
Zhang, J., Zheng, W., and Menne, M. J.: A Bayes factor model for detecting artificial discontinuities via pairwise comparisons, J. Climate, 25, 8462–8474, doi:10.1175/JCLI-D-12-00052.1, 2012.

GID
4, 235–270, 2014

Table 1. Known inhomogeneities between observed air temperature and the ambient air temperature representative of a given location in terms of: problems; possible causes and effects; physical solutions; and possible implementations in modelling a benchmark.

Problem	Possible Cause	Possible Effect	Physical solution	Benchmark modelling
Reported air temperature is not measured air temperature	Errors in reporting, units, data transmission etc.	Abrupt change that is either constant over time or a function of temperature	Identify error and correct (difficult to adjust using an automated process because errors may be unique)	Draw from past experience. Apply blanket changes using a constant or simple formula as a function of temperature alone
Measured air temperature is not true air temperature	Instrument error (malfunction or change in type), calibration error	Abrupt (or gradual for some instrument malfunctions) change that is either constant or a function of temperature (or drifting for some instrument malfunctions) *(random errors should be removed by quality control process)*	Identify error and correct, using metadata where available	Statistically model distributions of typical size and frequency. Apply blanket changes using a constant or simple formula as a function of temperature alone
True air temperature is not representative ambient air temperature	Change in instrument shield, practice or microclimate (due to move of instrument)	Abrupt change that is likely to vary as a function of variables such as radiation, windspeed and soil moisture	Identify error and correct. Modelling energy balance of shield and microclimatic conditions	Statistically model distributions of typical size and frequency. Semi-empirical modelling of errors based on assumed changes in radiation, windspeed and soil moisture
Representative ambient air temperature is affected by local influences	Changes in station surroundings, urbanization	Gradual change that is likely to vary as a function of variables such as radiation, windspeed and soil moisture	Correction not desirable from a physical or monitoring perspective, but from a detection and attribution perspective. Modelling energy balance of shield and microclimatic conditions	Statistically model distributions of typical size and frequency. Semi-empirical and possibly numerical modelling of resulting trend and its high frequency characteristics due to changes in radiation, windspeed and soil moisture
Different ambient air temperatures are merged	Change in station location	Abrupt change that is likely to vary as a function of variables such as radiation, windspeed and soil moisture	Unmerge (correction not desirable from a physical perspective, especially for high frequency data, but from a low frequency large scale monitoring and detection and attribution perspective)	Change in spatial sampling from the analog-known-world to merge series
Changes in diurnal sampling affect statistics	Change in observation time	Abrupt change that is likely to vary as a function of variables such as radiation	Split (correction not desirable from a physical perspective) or correct (low frequency large scale monitoring and detection and attribution perspective)	Statistically model distributions of typical size and frequency. Change in temporal sampling from synthetic source data or in the case of low frequency GCM output use semi-empirical modelling of errors based on assumed changes in radiation

GID
4, 235–270, 2014

Table 2. Example contingency table for assessing changepoint location detection and inhomogeneity adjustment (option shown in bold) skill of homogenisation algorithms. Potential detections are the number of potential changepoints within the time period minus the total number of detections and misses. These are used to quantify those occasions where no changepoint is found and none is present. One way to do this is to assume that there is potentially a maximum of 1 changepoint every 6 months (some algorithms can only search for changepoints with 6 months of data either side) such that a 26 year period will have 52 potential changepoints.

	Changepoint	No changepoint present	TOTALS
Changepoint detected within ±3 months **(Inhomogeneity adjustment must be correct sign (±) and within ±1 °C)**	HITS: 5 **(4)**	FALSE ALARMS: 3 **(3)**	8 **(7)**
Changepoint not detected within ±3 months **(Inhomogeneity adjustment value incorrect sign or not within ±1 °C)**	MISSES: 2 **(3)**	CORRECT MISSES: 42 **(42)** (potential detections)	44 **(45)**
TOTALS	7 **(7)**	45 **(45)**	52 **(52)**
Heidke Skill Score		61 %	
Probability of Detection Hit Rate		71 %	
False Alarm Rate		7 %	

GID
4, 235–270, 2014

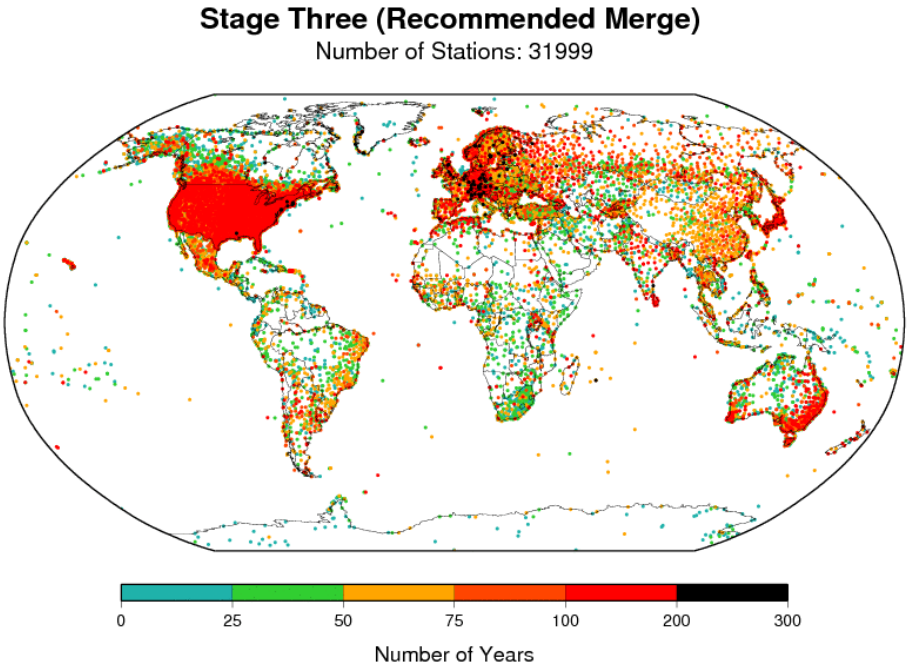


Figure 1. Station locations coloured by length of record for version 1 of the ISTI Land Meteorological Databank Stage 3.

GID
4, 235–270, 2014

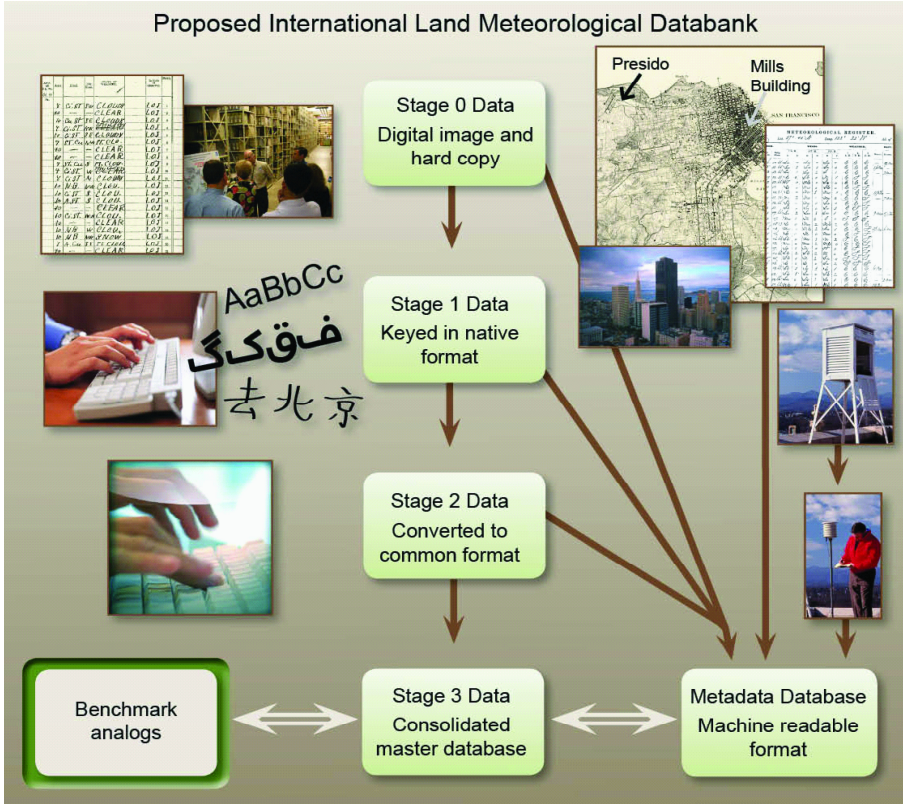


Figure 2. Structure of the Surface Temperature Initiative Databank.

GID
4, 235–270, 2014

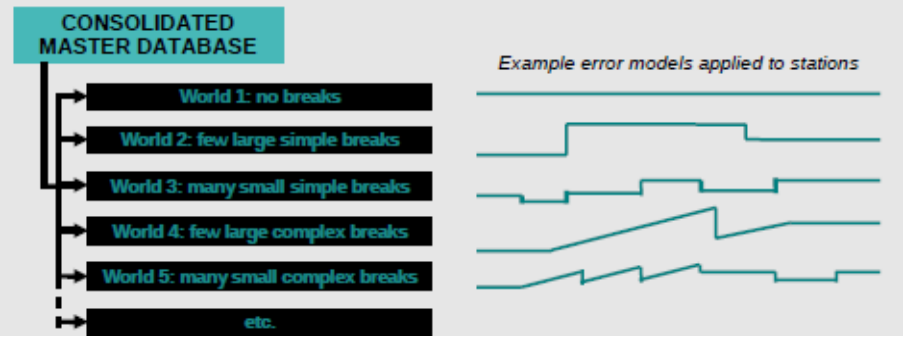


Figure 3. Example of a set of analog-error-world models.

GID
4, 235–270, 2014

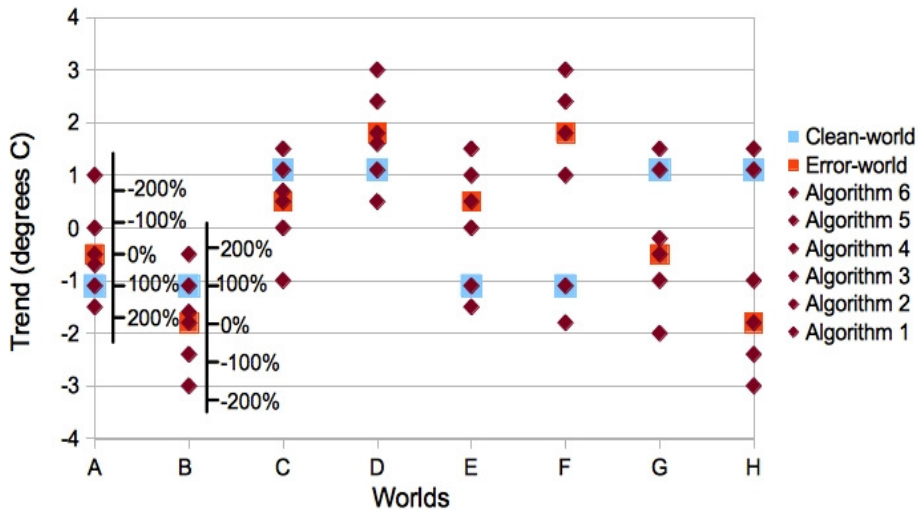


Figure 4. Example summary graph of algorithm skill for 6 hypothetical algorithms measured as trend percentage recovery. This uses the trends calculated from an adjusted analog-error-world scaled against the difference between the analog-error-world and its corresponding analog-clean-world. 100 % trend recovery would indicate a perfect algorithm. Greater than 100 % would be moving the trend too far in the right direction. Less than 100 % would be an algorithm that does not move the trend far enough towards the analog-clean-world. A negative percentage would indicate an algorithm that moves the trend in the wrong direction. This method does not indicate algorithms that result in a trend of the wrong sign (positive or negative). This may be seen as a more serious problem than a trend being over or under estimated and so would need to be identified separately.

GID
4, 235–270, 2014

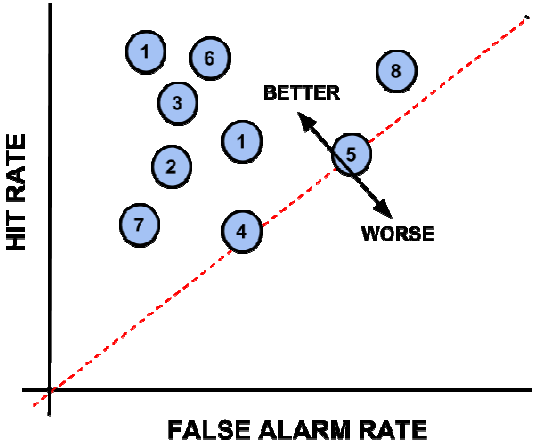


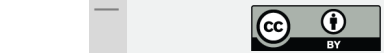
Figure 5. An adapted ROC plot summarising skill of a homogenisation algorithm across a comprehensive range of error-worlds. Note, in order to retain the value of the analog-error-worlds being unknown to users, these should be labelled as summaries of error-world concepts but not be traceable to the source error-world. So for example, 8 may refer to the "few large abrupt changepoints" world but it would not be analog-error-world 8.

GID
4, 235–270, 2014

STATION_ID	STN_Longitude	STN_Latitude	START_DATE	END_DATE	START_SIZE	END_SIZE	CHARACTERISTICS
037720-99999	-000.450	51.483	01051980-12-00	01121992-12-00	0.50	0.50	abrupt mean shift
037720-99999	-000.450	51.483	01121992-12-00	01031996-12-00	0.62	0.62	abrupt mean shift seasonal variance
037720-99999	-000.450	51.483	01071976-12-00	01031996-12-00	-0.21	0.10	gradual increase in mean
037840-99999	000.783	51.200	01051980-12-00	01121992-12-00	0.50	0.50	abrupt mean shift
037840-99999	000.783	51.200	01121999-12-00	01032003-12-00	-0.10	-0.10	abrupt mean shift

Figure 6. Example ASCII file containing detected changepoint locations and inhomogeneity adjustments for each station to be uploaded to the data-product portal for assessment by the Benchmarking and Assessment Working Group against the analog-clean-world(s).

GID
4, 235–270, 2014



STATION_ID	STN_Longitude	STN_Latitude	Err_World	STATISTIC	START_DATE	END_DATE	Jan	FEB	MAR	APR	MAY	JUN
037720-99999	-000.450	51.483	1	MEAN	01011960-12-00	12122010-12-00	2.1	4.5	6.0	8.2	10.4	14.2
037720-99999	-000.450	51.483	2	MEAN	01011960-12-00	12122010-12-00	2.2	4.6	6.1	8.3	10.5	14.3
037720-99999	-000.450	51.483	3	MEAN	01011960-12-00	12122010-12-00	2.1	4.5	6.0	8.2	10.4	14.2
037720-99999	-000.450	51.483	4	MEAN	01011960-12-00	12122010-12-00	1.9	4.3	5.8	8.0	10.2	14.0
037720-99999	-000.450	51.483	5	MEAN	01011960-12-00	12122010-12-00	2.5	4.7	6.2	8.4	10.6	14.4
037720-99999	-000.450	51.483	6	MEAN	01011960-12-00	12122010-12-00	2.5	4.7	6.2	8.4	10.6	14.4
037720-99999	-000.450	51.483	7	MEAN	01011960-12-00	12122010-12-00	2.0	4.4	5.9	8.1	10.3	14.1
037720-99999	-000.450	51.483	8	MEAN	01011960-12-00	12122010-12-00	1.8	4.2	5.7	7.9	10.1	13.9
037720-99999	-000.450	51.483	1	VARIANCE	01011960-12-00	12122010-12-00	2.5	2.5	2.7	2.4	3.2	3.5
037720-99999	-000.450	51.483	2	VARIANCE	01011960-12-00	12122010-12-00	2.4	2.4	2.5	2.3	3.0	3.2
037720-99999	-000.450	51.483	3	VARIANCE	01011960-12-00	12122010-12-00	2.5	2.5	2.7	2.4	3.2	3.5
037720-99999	-000.450	51.483	4	VARIANCE	01011960-12-00	12122010-12-00	2.7	2.9	3.1	2.8	3.4	3.9
037720-99999	-000.450	51.483	5	VARIANCE	01011960-12-00	12122010-12-00	1.8	1.8	2.1	2.1	2.2	2.5
037720-99999	-000.450	51.483	6	VARIANCE	01011960-12-00	12122010-12-00	2.5	2.5	2.7	2.4	3.2	3.5
037720-99999	-000.450	51.483	7	VARIANCE	01011960-12-00	12122010-12-00	2.4	2.4	2.5	2.3	3.0	3.2
037720-99999	-000.450	51.483	8	VARIANCE	01011960-12-00	12122010-12-00	2.5	2.5	2.7	2.4	3.2	3.5
037720-99999	-000.450	51.483	1	TREND	01011960-12-00	12122010-12-00	0.1	0.1	0.1	0.1	0.1	0.2
037720-99999	-000.450	51.483	2	TREND	01011960-12-00	12122010-12-00	0.0	0.1	0.1	0.0	0.0	0.1
037720-99999	-000.450	51.483	3	TREND	01011960-12-00	12122010-12-00	-0.2	-0.1	-0.1	0.0	-0.1	-0.2
037720-99999	-000.450	51.483	4	TREND	01011960-12-00	12122010-12-00	0.1	0.1	0.1	0.1	0.1	0.2
037720-99999	-000.450	51.483	5	TREND	01011960-12-00	12122010-12-00	0.0	0.1	0.1	0.0	0.0	0.1
037720-99999	-000.450	51.483	6	TREND	01011960-12-00	12122010-12-00	-0.2	-0.1	-0.1	0.0	-0.1	-0.2
037720-99999	-000.450	51.483	7	TREND	01011960-12-00	12122010-12-00	0.3	0.3	0.2	0.3	0.2	0.2
037720-99999	-000.450	51.483	8	TREND	01011960-12-00	12122010-12-00	0.1	0.1	0.1	0.1	0.1	0.2
037840-99999	000.783	51.200	1	MEAN	01011962-12-00	12122010-12-00	2.2	4.6	6.1	8.3	10.5	14.3
...												

Figure 7. Example ASCII file containing climatological monthly means, variance and trends for each station as calculated from each adjusted analog-error-world to be uploaded to the data-product portal for assessment by the Benchmarking and Assessment Working Group against stations/regions from the analog-clean-world(s).

GID
4, 235–270, 2014

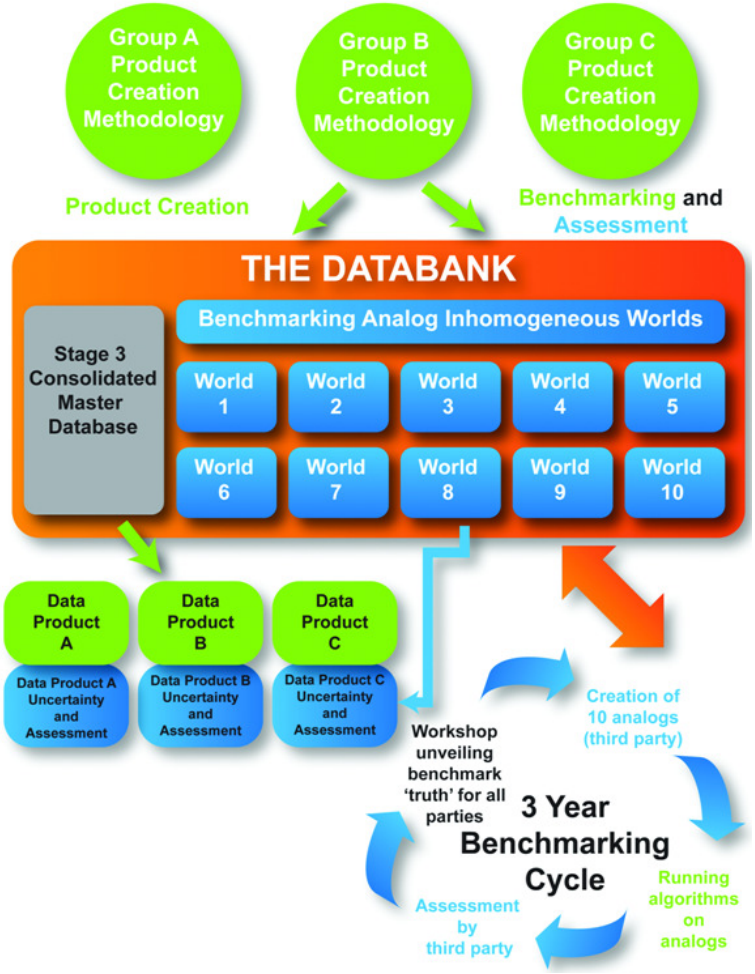


Figure 8. Schematic of the benchmarking assessment and benchmark cycle.

GID
4, 235–270, 2014