Benchmarking and Assessment Working Group
2015 Progress Report
October 2015

Membership

Current Members:

Kate Willett (Chair)	Met Office Hadley Centre, UK
Claude Williams	National Centers for Environmental Information, USA
Ian Jolliffe	Exeter Climate Systems, University of Exeter, UK
Robert Lund	Department of Mathematical Sciences, Clemson University, USA
Lisa Alexander	Climate Change Research Centre, University of New South Wales, Australia
Stefan Brönnimann	University of Bern, Switzerland
Lucie A. Vincent	Climate Research Division, Environment Canada, Canada
Steve Easterbrook	Department of Computer Science, University of Toronto, Canada
Victor Venema	Meteorologisches Institut, University of Bonn, Germany
David Berry	National Oceanography Centre, Southampton, UK
Rachel Warren	Met Office Hadley Centre, UK
Giuseppina Lopardo,	Istituto Nazionale di Ricerca Metrologica (INRiM), Italy
Renate Auchmann	Oeschger Center for Climate Change Research & Institute of Geography, University of Bern, Switzerland
Enric Aguilar	Centre for Climate Change, Universitat Rovira i Virgili, Spain
Matt Menne	National Centers for Environmental Information, USA
Colin Gallagher	Department of Mathematical Sciences, Clemson University, USA
Zeke Hausfather	Berkeley Earth, USA
Thordis Thorarinsdottir	Statistical Analysis, Pattern Recognition, and Image Analysis (SAMBA), Norwegian Computing Centre, Norway

Peter Thorne (ex-officio) – Maynooth University, Ireland

New Members:

Robert Dunn	Met Office Hadley Centre, UK

Ex-Members:
NA

October 2014 to October 2015 Objectives:

1) Advocacy of ISTI and the benchmarks and support for users
2) Up to date reference list of work on homogenisation/benchmarking: https://sites.google.com/a/surfacetemperatures.org/home/benchmarking-and-assessment-working-group?pli=1#Reference%20Literature
3) Create software to produce analog-clean-worlds on the global scale
4) Produce enough clean worlds to make the open and blind error worlds

5) Submit clean-world methods paper
6) Finalise distribution and probability framework for analog-error-worlds
7) Create software to produce analog-error-worlds on the global scale
8) Submit error-worlds methods paper
9) Produce the official first version of analog-error-worlds
10) Finalise validation concepts
11) Create software and score system/intercomparison tables to run the validation on a proof of concepts scale
12) Produce a validation methods paper
13) Produce first version validation software ready to perform on real returned benchmarks
14) Release first version benchmarks (blind and open worlds)
15) Create a benchmarking support webpage to host the benchmarks, validation results, feedback

Objectives (Partially) Fullfilled:

1) Advocacy of ISTI and the benchmarks and support for users
July 2015 – Kate Willett presented a poster at the Copernicus Climate Change Service workshop on Climate Data, ECMWF, Reading, UK:
The International Surface Temperature Initiative.

April 2015 – Victor Venema presented a poster at EGU, Vienna, Austria:
Benchmarking and Assessment of Homogenisation Algorithms for the International Surface Temperature Initiative (ISTI).

February 2015 - Kate Willett presented at the WCRP Grand Challenge on Data for Extremes workshop, University of New South Wales, Sydney, Australia (and also CSIRO, BOM):
The ISTI: Land surface air temperature datasets for the 21st Century.

December 2014 - Kate Willett presentation at the University of Bern, Switzerland:
The International Surface Temperature Initiative and Benchmarking for Homogenisation Algorithms.

2) Up to date reference list of work on homogenisation/benchmarking:
https://sites.google.com/a/surfacetemperatures.org/home/benchmarking-and-assessment-working-group?pli=1#Reference%20Literature

This has been updated with all articles published this year.

3) Create software to produce analog-clean-worlds on the global scale
Software has been written in Python and R to run the analog-clean-worlds from ISTIv1.0.1 (July download) and should now be able to be run on any future versions in the same format. All code is stored in the GitHub SurfaceTemp organisation in the ISTI_Clean_Worlds repository: https://github.com/SurfaceTemp/ISTI_Clean_Worlds. A wiki has been created (https://github.com/SurfaceTemp/ISTI_Clean_Worlds/wiki) and a clear README documentation of how to run step by step has been written (Figure 1). Validation has been tested using similarity of climatology, standard

deviation, autocorrelation, cross-correlations at lag 0 and lag 1 and difference series autocorrelation and standard deviation.

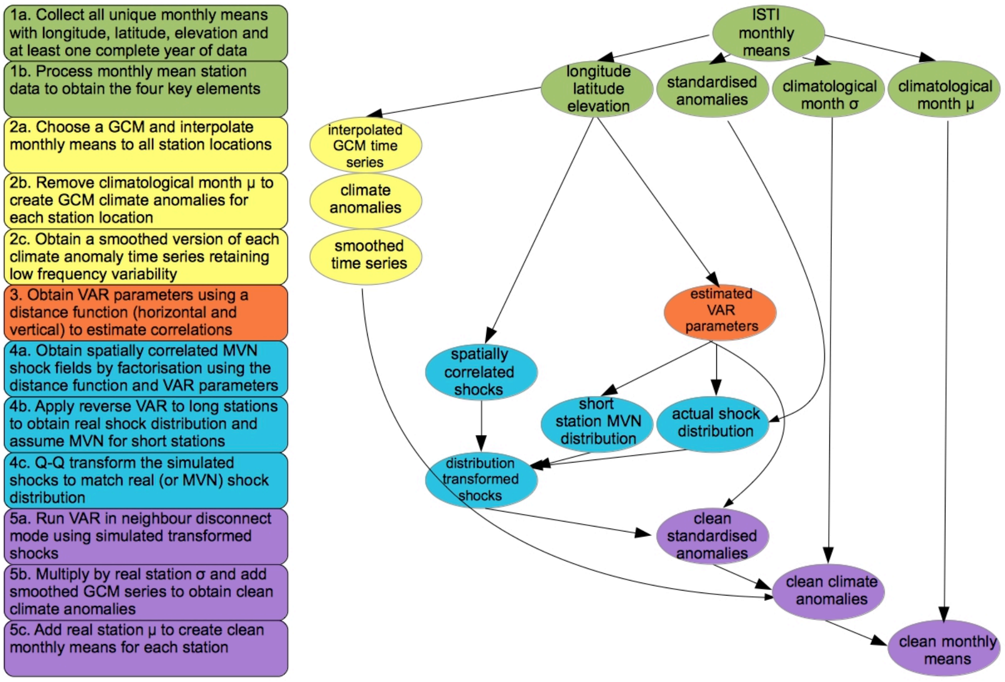
1a. Collect all unique monthly means with longitude, latitude, elevation and at least one complete year of data
1b. Process monthly mean station data to obtain the four key elements
2a. Choose a GCM and interpolate monthly means to all station locations
2b. Remove climatological month μ to create GCM climate anomalies for each station location
2c. Obtain a smoothed version of each climate anomaly time series retaining low frequency variability
3. Obtain VAR parameters using a distance function (horizontal and vertical) to estimate correlations
4a. Obtain spatially correlated MVN shock fields by factorisation using the distance function and VAR parameters
4b. Apply reverse VAR to long stations to obtain real shock distribution and assume MVN for short stations
4c. Q-Q transform the simulated shocks to match real (or MVN) shock distribution
5a. Run VAR in neighbour disconnect mode using simulated transformed shocks
5b. Multiply by real station σ and add smoothed GCM series to obtain clean climate anomalies
5c. Add real station μ to create clean monthly means for each station
ISTI monthly means
longitude latitude elevation
standardised anomalies
climatological month σ
climatological month μ
interpolated GCM time series
climate anomalies
smoothed time series
estimated VAR parameters
spatially correlated shocks
short station MVN distribution
actual shock distribution
distribution transformed shocks
clean standardised anomalies
clean climate anomalies
clean monthly means

Figure 1. Flow chart of analog-clean-world creation.

5) Submit clean-world methods paper
PARTIALLY MET
A first draft of this paper has been written and circulated around co-authors. Response from co-authors is still awaited and revisions (possibly methodological) are likely prior to submission.

6) Finalise distribution and probability framework for analog-error-worlds
The selection of open and blind worlds, which ones will be mandatory (versus optional), and the distribution and probability framework for the analog-error-worlds have now been finalised. One point of note is the agreement on each inhomogeneity having a random and biased component. This is described in Figure 2.

7) Create software to produce analog-error-worlds on the global scale
PARTIALLY MET
All code is stored in the GitHub SurfaceTemp organisation in the ISTI_Error_Worlds repository: https://github.com/SurfaceTemp/ISTI_Error_Worlds. A wiki has been created to describe the work and code/code structure is well under way.

8) Submit error-worlds methods paper
PARTIALLY MET
An outline for the paper has been drafted and we have written a working report on the error worlds, but completion will take place after completion of

the error world code. This will likely include an assessment of inhomogeneities found in real world data such as GHCNv4.

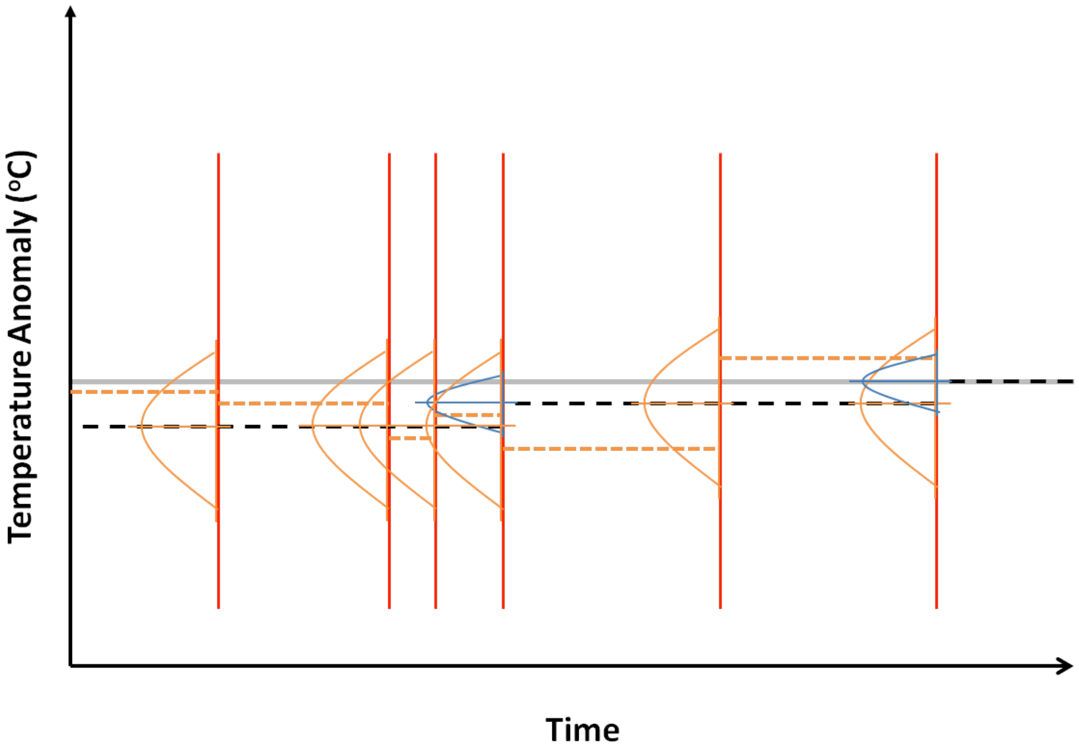

Figure 2 Diagrammatical example of how inhomogeneities will be implemented in a station time series. The vertical red lines show the changepoint locations. The horizontal grey line shows the mean of the reference period – the most recent 'homogeneous' period. The dashed black lines indicate the bias (permanent) component of each inhomogeneity and the dashed orange lines indicate the random (temporary) component of each homogeneity. The bias component sizes are selected from the distributions shown in blue and are always centred around the mean of the previous bias component (working backwards from present day). The random component sizes are added to the current bias component and selected from the distributions shown in orange which are always centred around the mean of the current bias component.

10) Finalise validation concepts
PARTIALLY MET
There has been quite a bit of discussion on this in teleconferences but the validation measures have not yet been finalised. Rachel Warren's PhD on daily benchmarks is nearing completion and includes a well developed validation framework. We envisage building on this to a large extent. One key issue for the global benchmarks is allowing for regional submissions and refusal to homogenise some stations that are too short, contain too much missing data or are of too poor quality initially.

Objectives Not Fullfilled:

4) Produce enough clean worlds to make the open and blind error worlds

The clean world code has yet to pass through methodological sign off. The official open and blind worlds cannot be created until the methods have been finalised.

5) Submit clean-world methods paper
Partially met – see above

7) Create software to produce analog-error-worlds on the global scale
Partially met – see above

8) Submit error-worlds methods paper
Partially met – see above

9) Produce the official first version of analog-error-worlds
This requires both finalisation of the clean world methodology following review from three statisticians and completion of the error world code.

10) Finalise validation concepts
Partially met – see above

11) Create software and score system/intercomparison tables to run the validation on a proof of concepts scale
There has been no progress but Rachel Warren's PhD on daily benchmarks is near completion with a well developed validation framework and code which we can use as a basis for our own.

12) Produce a validation methods paper
No progress has been made.

13) Produce first version validation software ready to perform on real returned benchmarks
No progress has been made.

14) Release first version benchmarks (blind and open worlds)
This has not been done. This can be done after sign off of the clean world methodology and completion of the error world code.

15) Create a benchmarking support webpage to host the benchmarks, validation results, feedback
No progress has been made.

Other Efforts and Achievements:

- Rachel Warren's PhD to develop a benchmarking process for daily data and work in collaboration with ISTI is nearing completion. Daily benchmarks for four regions of the USA and 3 worlds have now been released as blind worlds. Eight methods have been run on the benchmarks and homogenised data returned. A validation framework has been designed and the results analysed. Rachel is in the process of writing up results.

- Victor Venema visited the Met Office which was very helpful to progress work on the error worlds.

2015 Annual Overview:

Progress during 2015, as in previous years, has been slower than hoped and the benchmarks are still not completed. This is mostly due to the complexity of getting the clean world methods up and running on the global scale and balancing other commitments. We have held fewer teleconferences than would be ideal, but this is largely due to lack of progress between calls. The completion of working code for the clean worlds and a first draft of the paper have been major successes but some revisions are envisaged following review from three ISTI statisticians. Progress on the error world code has been good but substantial effort is still required to complete this work. Although progress is slow there is no reason to think that the benchmark working group will not succeed in creating complete benchmarks eventually. The latest deadline for the benchmarks was summer 2015. We are now aiming for summer 2016 and plan to use the latest version of the ISTI databank during January 2016.

Objectives for October 2015 and Beyond:

Table 1. Suggested timeline and plan for achieving objectives.

Objective	Description	Responsible Members	Deadline
Advocacy of ISTI and the benchmarks and support for users	Distribute our ISTI flyers and presentation of concepts and progress at relevant conferences and events	All	Ongoing
Up to date reference list of papers on homogenisation/ benchmarking: https://sites.google.com/a/surfacetemperatures.org/home/benchmarking-and-assessment-working-group?pli=1#Reference%20Literature	Keep the list of references up to date with new papers as they are published	All	Ongoing
Analog-clean-worlds global scale production	Produce analog-clean-worlds for all blind error worlds and submit methods paper 2	Team Creation – code run and data hosted by Kate Willett	January 2016
Analog-error-worlds open worlds	Create software to produce analog-error-worlds for at least the open worlds and submit methods paper (if desired?)	Team Corruption – lead by Victor Venema & Claude Williams	March 2016

Analog-error-worlds blind worlds (official benchmarks)	Produce analog-error-worlds from the analog-clean-worlds ready for distribution as official benchmark data	Team Corruption – lead by Claude Williams &Victor Venema. (parametric choices made by Kate Willett and kept blind from all potential users)	June 2016
Validation concepts finalised (including regional and incomplete submissions)	Decide upon number and type of tests with which to perform validation	Team Validation – lead by Ian Jolliffe	December 2015
Validation proof-of-concept	Create software and score system/intercomparison tables to run the validation on a proof-of-concept scale and submit methods paper (if desired?)	Team Validation – lead by Ian Jolliffe	July 2016
Validation global scale production	Produce software and framework ready for running on the global scale – automated or manual	Team Validation – lead by Ian Jolliffe	December 2016
Benchmark Cycle Release of analog-error-worlds	Release first official benchmarks – publicise widely	All – lead by Kate Willett	July 2016
Benchmarking Platform Design	Update the benchmarking webpage to include step-by-step ‘How to benchmark’ with appropriate links to data, validation and intercomparison tables, feedback to be invited through the benchmarking blog, a private list of participants contact details will be maintained offline.	All – lead by Kate Willett	July 2016

Deadline for submission of benchmark results	Homogenisers to submit their homogenised benchmark data and a set of specified statistics	Dataset creators	July 2018
Benchmark Cycle – release of the ‘answers’	Release the ‘answers’ (analog-clean-worlds) for the blind worlds	All – lead by Kate Willett	July 2018
Return of assessment of benchmark homogenisation	Supply all appropriate statistics to the dataset creators	Team Validation led by Ian Jolliffe and working group	January 2019
Organise benchmark cycle 1 wrap-up workshop	Plan and run a workshop, perhaps in conjunction with full ISTI meeting or other conference? Resource dependent.	All – lead by Kate Willett	Summer 2019
Publication of benchmark results and assessment of the cycle		Benchmarking working group	December 2019
Release of second benchmark cycle	Some improvements made from previous cycle and different issues explored with the error worlds. Potential extensions to other variables, daily data, and other dataset creation issues.	Benchmarking working group	To be decided