Benchmarking and Assessment Working Group
2013 Progress Report
November 2013

Current Members:

Kate Willett (Chair)   - UKMO Hadley Centre, UK
Claude Williams   - NCDC, 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   - College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK
Giuseppina Lopardo,   - Istituto Nazionale di Ricerca Metrologica (INRiM), Italy

Peter Thorne (ex-officio) – NERSC, Norway

New Members:

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   - NCDC, 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), Nowegian Computing Centre, Norway

Ex-Members:

Olivier Mestre   - Meteo France, France
Aiguo Dai   - Climate and Global Dynamics Division, NCAR, USA
Mike Finney   - Department of Mathematical Sciences, Clemson University, USA

October 2012 to October 2013 Objectives:
1) Benchmark Cycle concepts and plan formalised and submitted to JAOT or similar.
2) Create fully functioning open source software for producing the analog-clean-worlds and submit a methods paper.

3) Create a first suite of analog-clean-worlds.
4) Design methods and create software for producing the analog-error-worlds.
5) Design methods and create software for assessing the results of tests on the benchmarks ready for the Benchmark cycle
6) Create a platform for guiding users of the benchmarks in how to use them and how the assessment works
7) Publicise the aims and objectives of both the ISTI and the work of the Benchmarking and Assessment working group widely and engage with as many similar efforts as possible

Objectives Met:
1) Benchmark Cycle concepts and plan formalised and submitted to JAOT or similar.
This has been partially met. A second draft of the concepts paper has been circulated and the third draft will be circulated shortly . This paper is intended to be submitted to GIMDS (Geoscientific Instrumentation, Methods and Data Systems; http://www.geoscientific-instrumentation-methods-and-data-systems.net/home.html*) by Christmas 2013.*
2) Create fully functioning open source software for producing the analog-clean-worlds and submit a methods paper.
This has been partially met. R code has been written using the VAR method to create synthetic stations within a gridbox. Not all stations can be replicated due to poor quality and missing data. Expansion to large regions will be a computational challenge and achieving global scale spatial consistency is also not yet solved. Work on alternative methods is ongoing.
4) Design methods and create software for producing the analog-error-worlds.
A set of blind and open error worlds has now been defined. Progress has been made on defining distribution shapes and probability frameworks for assigning location and size of the errors. This needs further work to finalise the statistics and software is yet to be created. This also depends on there being a set of analog-clean-worlds to work with.
5) Design methods and create software for assessing the results of tests on the benchmarks ready for the Benchmark cycle
The assessment has been divided into four levels and progress has been made on the types of assessment to be conducted at those levels, the requirements of returned data/statistics and the statistical tools to be used. This needs further work to finalise and software needs to be developed.
7) Publicise the aims and objectives of both the ISTI and the work of the Benchmarking and Assessment working group widely and engage with as many similar efforts as possible
June 2013
The ISTI: Land surface air temperature datasets for the 21st Century
An overview of benchmarking homogenisation procedures for the ISTI
Two invited talks at the 12th International Meeting for Statistical Climatology, Jeju, South Korea
July 2013
Benchmarking Working Group 3 day workshop, NCDC, NC, USA
November 2013

The International Surface Temperature Initiative (ISTI): Dragging land surface temperature data kicking and screaming into the 21st Century
Invited talk on ISTI/Benchmarking for Reading University Meteorology Department

Objectives Not Met:
1) Benchmark Cycle concepts and plan formalised and submitted to JAOT or similar.
This has been partially met. See above.
2) Create fully functioning open source software for producing the analog-clean-worlds and submit a methods paper.
This has been partially met. See above.
3) Create a first suite of analog-clean-worlds.
This is pending completion of 1).
4) Design methods and create software for producing the analog-error-worlds.
Methods partially complete (see above), they need to be finalised and software developed.
5) Design methods and create software for assessing the results of tests on the benchmarks ready for the Benchmark cycle
Methods partially complete (see above), they need to be finalised and software developed.
6) Create a platform for guiding users of the benchmarks in how to use them and how the assessment works
No progress to date.

Other Efforts and Achievements:
- Rachel Warren's PhD continues to develop a benchmarking process for daily data and work in collaboration with ISTI.
- The work of the Benchmarking working group was presented at the 12th International Meeting for Statistical Climatology by invitation. This allowed face-to-face meetings between some members and also led to Thordis Thorarinsdottir joining the group.
- Funding was provided by Matt Menne for a 3 day workshop in NCDC to work on the benchmark data. This enabled Enric and Kate to travel from Europe and USA members also to attend. Other members were able to join in by internet and daily teleconferences. This enabled much detailed discussion of methodological choices not previously possible over the telephone. Much progress was made in terms of concepts but little hard programming was achieved. Two new members (Colin and Zeke) have now joined us. A meeting report has been written and published on our website.
- An ISTI glossary has been set up to define some of the complex jargon that has developed throughout the project. This is now online as an open document for anyone to read/update as necessary.
  https://docs.google.com/document/d/1xItD6yeQTxqwVnbfx-ZwUsh7hKJu1HqJVEf-OKstS4Y/edit?usp=sharing
- A list of known dates and types of inhomogeneities has been compiled for as many regions as possible with help from the homogenisation mailing list. This is now online as an open document for anyone to read/update as necessary.

2013 Annual Overview:
Progress during 2013 has been much slower than hoped and most of the objectives were not met in full. The major success of this year has been the funding (thanks to Matt Menne) and execution of a benchmark development workshop at NCDC, NC, USA. This face-to-face opportunity was hugely beneficial for discussing some of the many very complex aspects of this work and for ensuring accurate communication of concepts across all disciplines represented. Prior to this, weekly teleconferences were held to define the error world types and validation aspects as far as possible. In many ways, the workshop provided a useful catalyst to making progress. While the workshop was invaluable and resulted in a plausible set of actions for completing the project there was a long lull over the summer holiday period due to holidays and sick leave for Kate. As with last year, development of concepts has taken far longer than expected due to the complexity of the problem and the sheer amount of decisions that will have to be largely expert judgement based. Again, the failure to make real progress with the analog-clean-worlds (partly a result of delays in the release of a version 1 databank holdings which arose for several reasons detailed in the databank working group report) has had a knock-on delay to the analog-error-worlds and assessment schedules. Ideally, the analog-clean-world method will be completed during Autumn 2013 and both the overall benchmarking concept paper and analog-clean-world methods paper submitted by the end of 2013. Realistically, a complete benchmark set of data holdings will not be ready until Spring/Summer 2014. There is now a hard deadline of Summer 2014 to at least have the open world benchmarks released to provide test data for a recently successful ISTI workshop bid to SAMSI.

Objectives for October 2013 to October 2014:
1) Benchmark Cycle concepts paper submitted to GIMDS.
2) Finalise the fully functioning open source software for producing the analog-clean-worlds and submit a methods paper.
3) Create a first suite of analog-clean-worlds (open and blind).
4) Finalise error-world distribution statistics and probability framework and create software for producing the analog-error-worlds.
5) Create a first suite of analog-error-worlds (open and blind).
6) Finalise assessment statistics and statistical tools and create software for assessing the results of tests on the benchmarks ready for the Benchmark cycle.
7) Have a prototype working assessment framework by Summer 2014.
8) Create a platform for guiding users of the benchmarks in how to use them and how the assessment works
9) Publicise the aims and objectives of both the ISTI and the work of the Benchmarking and Assessment working group widely and engage with as many similar efforts as possible

Suggested timeline and plan for achieving objectives:

Objective	Description	Responsible Members	Deadline

Advocacy of ISTI and the benchmarks and support for users	Presentation of concepts and progress at relevant conferences and events	All	Ongoing
Up to date reference list of work on homogenisation/ benchmarking: https://sites.google.com/a/ surfacetemperatures.org/h ome/benchmarking-and-assessment-working-group?pli=1#Reference%20 Literature	Add to this as widely as possible	All	Ongoing
Benchmark Concepts Paper	Overview outlining the entire Benchmark concept and importance for climate science	All, lead by Kate Willett	Submit by December 2013
Analog-clean-worlds open worlds	Create software to produce analog-clean-worlds on a global scale, produce enough to create the open error worlds and submit methods paper	Team Creation – lead by Robert Lund and Kate Willett	December 2013
Analog-clean-worlds global scale production	Produce analog-clean-worlds for all blind error worlds	Team Creation – code run and data hosted by Kate Willett	Spring 2014
Analog-error-worlds concepts finalised	Using the defined set of blind and open worlds define the distribution and statistical probability framework	Team Corruption – lead by Claude Williams/Vict or Venema	December 2013
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 Claude Williams/Vict or Venema and coding by Kate Willett	Spring 2014
Analog-error-worlds blind worlds (official benchmarks)	Produce analog-error-worlds from the analog-clean-worlds ready for	Team Corruption – lead by	May 2014 (open worlds at

	distribution as official benchmark data	Claude Williams/Victor Venema	least), Summer 2014 (Blind)
Validation concepts finalised	Decide upon number and type of tests with which to perform validation	Team Validation – lead by Ian Jolliffe	December 2013
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	May 2014
Validation global scale production	Produce software and framework ready for running on the global scale – automated or manual	Team Validation – lead by Ian Jolliffe	Have ability to accept returned benchmark data and statistics by Summer 2014 but full scale assessment could be delayed until Summer 2015 if necessary
Benchmark Cycle Release of analog-error-worlds	Release first official benchmarks – publicise widely	All – lead by Kate Willett	Summer 2014
Benchmarking Platform Design	Create a webpage showing step-by-step 'How to benchmark' with appropriate links to data, validation and intercomparison tables with registration so that feedback can be provided and contact maintained	All – lead by Kate Willett	December 2014 (ideally earlier but more important to get benchmarks created first)
Deadline for submission of benchmark results	Homogenisers to submit their homogenised benchmark data and a	Dataset creators	September 2015

set of specified statistics			
Benchmark Cycle – release of the 'answers'	Release the 'answers' (analog-clean-worlds)	All – lead by Kate Willett	September 2015
Return of assessment of benchmark homogenisation	Supply all appropriate statistics to the dataset creators	Team Validation led by Ian Jolliffe and working group	March 2016
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 2016
Publication of benchmark results and assessment of the cycle		Benchmarking working group	December 2016
Release of second benchmark cycle	Some improvements made from previous cycle and different issues explored with the error worlds	Benchmarking working group	Summer 2017