Benchmarking and Assessment Working Group
Conference Call 2

Wednesday 30th March 2pm GMT

Attended by:
Kate Willett (KW)
Ian Jolliffe (IJ)
Lucie Vincent (LV)
Peter Thorne (PT)
Robert Lund (RL)
Steve Easterbrook (SE)
Claude Williams (CW)
Victor Venema (VV)

Unable to attend:
Chris Wikle
Lisa Alexander
Olivier Mestre (OM)
Stefan Bronniman

Purpose: Creation of the benchmark 'truths'

Actions list:
KW -circulate Implementation Plan and Terms of Reference
ALL -comment as appropriate on circulated documents

KW -to draft a Benchmarking and Assessment Working Group Terms of Reference and circulate for comment

KW -to draft a white paper documenting the benchmarking concepts and design and circulate for comment -discussion on blog too.

ALL -to consider potential funding -for hosting workshops

KW -to make some GCM model runs available on googledocs

ALL -next call end of May

1. Progress Update by Kate and anyone else who has anything

- *Kate gave a presentation and lead a discussion on benchmarking at NCDC (Kate Willett, Claude Williams, Robert Lund, Peter Thorne, Matt Menne)*
- *Kate has prepared a Benchmarking Poster - to be presented at EGU 2011 next week (downloadable from http://www.surfacetemperatures.org/benchmarking-and-assessment-working-group)*
- *The Steering Committee is working on an 'Implementation plan' and 'Terms of Reference'. These will be circulated around the group – please make comments - there are parts relevant to the Benchmarking and Assessment Working Group:*
  - *informal reports to the Steering Committee quarterly*
  - *formal report to the Steering Committee annually - October*

- deadline for benchmark truth and error model creation - ~ 8 month lag from the Databank version 1 release = April 2012

Discussion of benchmarking structure and cycle

Benchmark creation concepts (structure, cycle, 'truth' and error models) need to be in place by April 2012 - actual benchmark production should lag databank by 8 months - let people play/explore the databank first –it will take a while to establish what is in the databank.

LV: What about time between 2012-2014? What should we be doing then? (benchmark release of 'truths' and reset of benchmarks)
KW/PT: During this period all members will need to field questions, advocate usage of the benchmarks - conference posters/talks?, plan for future benchmarks –including improvements as users explore the benchmarks and find inevitable bugs and desirable changes.

PT: Steering Committee have asked that working groups create their own Terms of Reference.

ACTION: KW: to circulate Steering Committee Terms of reference as a starting point. Benchmarking Working Group will shortly take on drafting their own set – probably led by a rough draft by KW to start and then comments from all members after circulation.

VV: how often will databank be released? PT: ideally every 3 years, aligned with Benchmark cycle

Issue of databank updates in time and space - becoming unaligned with static benchmarks - need to avoid confusion over multiple versions - scope for annual updates at some point down the line? The databank may receive regular updates from some sources making the benchmarks out of date. However, to keep this manageable the benchmarks will (for now) remain static throughout the 3 year cycle. After official new releases of the databank time will be needed to properly assess use of/improvements to make to the benchmarks.

SE: What is the main purpose of the benchmarks?
PT/KW: To assess fitness for purpose when comparing products, to feed into methodological advancement (scientific paper on different algorithms on a consistent set of benchmarks), a tool to feed into creation of uncertainty estimates?

PT: Ethos – benchmarks are a way of encouraging publication and discussion of strengths/weaknesses of methods

Workshop 3 years after benchmark release - early 2015. Answers would be released 4-6 months prior to the workshop, so people can assess the performance of their algorithms.

SE: also, hold a workshop much earlier in the cycle - to bring people together to do initial trials of alogoritthms - exploratory. Online possible - Steve to report on personal experience or align with existing efforts. Other funding? FCO? NSF? SAMSI?

ACTION: All to consider potential funding opportunities.

2. Creating the benchmark 'truths' - these are meant to be an analog of the 'Consolidated Master Database' of global station data from the Databank as far as possible - ~10000 stations?

PT actually closer to c. 70,000 monthly resolution likely by version point. Mainly more dense where already dense

Concept: the benchmarks must include real world artifacts
PT/VV/OM: need to use random (unexpected) background trends to avoid over tuning to 'knowns'.

What background data?
GCMs include realistic teleconnections - although station covariance/autocorrelation will need to be tweaked to move away from the gridbox (downscaling a gridbox to multiple stations) – autocorrelation and inter-station characteristics are likely too strong/over correlated in GCMs.
GCMs allow use of different trends/up/down/no trend with different forcing runs (Control (no trend), So2 (natural only -little wa rming, some cooling), C20C (anthropogenic and natural - warming), A1B (strong warming).
GCMs will have some spatial correlation and autocorrelation
Available from CMIP3 archive - pcmdi portal : https://esg.llnl.gov:8443/index.jsp
Some of the CMIP5 data is already available: http://pcmdi3.llnl.gov/esgcet/home.htm
ACTION KW: make some runs available on googledocs
GCM use agreed

What additional features need to be added to the background data?
- Seasonal cycles?
- Trends (climate change, solar cycles, volcanoes, modes of natural variabilty [ENSO, PDO, NAO, MJO etc.,], local effects etc.,)
- Random error?
Who will do this/how?
Monthly to start - daily eventually.

OM: careful using known regional trends because the benchmarks are then 'known' to those using them. Random regional trends are better.

RL - this needs to be based on a stated statistical model - like the benchmark poster – see below

Basis for the benchmark ("the truth") consists of:

$$\mathbf{X} = \mathbf{S} + \mathbf{T} + \mathbf{R} + \mathbf{N}$$

X = benchmark analog station at time /location/height
S = seasonal cycles
T = trends (long-term/ local effects / ENSO/ NAO/ MJO etc., Volcanoes/ Solar Cycles etc.)
R = random error - due to residual measurement errors in the record) not systematic - that will come later
N = Nudge to station specific mean/variance
[this is described in the poster (linked from website), blog post and to some extent forms the basis of the pseudo worlds (also documented on website]

Some statistical check on autocorrelation and spatial covariance - by adding the Random Error component? There may be better statistical methods of doing this?
Some nudge to station specific mean and variance

Later, once the 'truths' have been designed, we will need to add systematic errors due to station move/instrument change etc. - this will be the creation of the error models. These should be designed to address specific questions but in a realistic way e.g., A single large break oversimplifies the problem. Error worlds will likely span from an overly optimistic view of the world to an overly pessimistic view of the world.

PT: Need to decide on a set of questions that you want these benchmarks will answer? e.g., Does method success depend on presence or absence of trend? Can't answer them all so alight on a subset you wish to answer.

PT: Need to ensure that you have realistic inter-station series characteristics, particularly AR1 and sigma. Most algorithms depend upon this for breakpoint and adjustment steps. Need to get the cross correlation matrix between all stations right. Downscaling from models will likely leave stations being too closely related and so some 'nudging' will be necessary.

ACTION: Kate to draft this in a white paper -to be edited by all until we're happy - further discussion on blog agreed.

How do we assess? Station, region, global, one does not follow the other. An algorithm may get the global mean but at expense of station realism.

Metadata is available for some real stations and this should be replicated in these synthetic worlds as much as possible. The content of the metadata will differ depending on error model but the presence of some metadata is important. This may not always match a detectable breakpoint, as is the case in the real world.

3. AOB

Next call in two months – end of May

4. Minutes Agreed by:

KW, IJ, LV, PT, RL, SE, CW, VV