**ISTI steering committee / databank working group / benchmarking working group joint call**

11th December 2012 8am EST, 13Z, 14 CET

Dial in details circulated in advance by email

Present on call: Peter Thorne, Jay Lawrimore, Jared Rennie, Victor Venema, Steven Worley, John Christy, Kate Willett, Blair Trewin, Xiaolan Wang, Albert Mhanda, Ian Jolliffe, Dave Lister, Albert Klein Tank, Robert Lund, Lucie Vincent, Colin Morice, Madeline Renom, Matt Menne

Apologies in advance: Richard Chandler, Tom Peterson, Chris Merchant, Meaghan Flannery, Andrea Merlone, Rob Allan, Matilde Rusticucci (technical issues with joining)

In advance of the call were distributed:
* pdf file containing summary of merged beta release v.2
* databank WG report (to steering committee only)
* draft of the databank release paper
* word document of potential people to approach about developing products from the databank and using the benchmarks

Note: reviews of actions from previous calls were stayed until the next Steering Committee and Working Group calls respectively. This call was to approve moving forwards to first databank release and to discuss next steps more generally - a strategic call rather than a Business As Usual call. It was an opportunity to draw breath, assess where we are and next steps and engage across groups.

**Collated actions arising**

**ACTION**: Jared Rennie to host the merge process metadata on ftp and make its availability known through the initiative blog and databank ftp pages.

**ACTION**: Kate to put Jared in contact with appropriate ECMWF people to scope potential to interact the databank with reanalysis feedback files.

**ACTION**: Feedback on paper to Jared jared.rennie@noaa.gov by 12/17 @23Z. Jared to circulate revised draft to co-authors by 21st.

**ACTION**: Jared Rennie and Xiaolan Wang to follow up on getting the 400 Chinese daily resolution stations' data into the databank

**ACTION**: Jay Lawrimore and Kate Willett to work with their groups to facilitate the populating of a daily resolution parallel measurements database for understanding daily inhomogeneity properties

**ACTION**: Kate Willett to engage with benchmarking working group and provide update to steering committee on what is planned and a new timeline to release of the first version of the benchmarks. By no later than end of Jan 2013.

**ACTION**: Peter Thorne to convert the document of potential data product creators provided by SC members into a spreadsheet on googledocs which everyone can access and members of SC / WG to volunteer to approach the relevant individuals to encourage their participation in product generation from the databank and in the benchmarking exercise.

**ACTION**: Peter Thorne will liaise with GCOS secretariat and AOPC to ascertain how most appropriately to raise a request for regional national focal points to review the databank release and help us to start to improve the provenance of what we have.

**ACTION**: Peter Thorne and Andrea Merlone to discuss Tempmeko presence for ISTI.

**Full call notes**

**1. Databank beta release v2 summary - Jared Rennie**

See associated pdf file summarizing the status. Jared talked us through this presentation and there then followed discussion.

JRC: How will the images be archived? I have many for East Africa from various sources. Has this part of the Databank been set up and cross-referenced to the raw machine-readable datafiles?
JR: Sources we have stage 0 for we are hosting. We are happy to host any data, but it is on request / demand not an expectation. It adds to the management burden and costs $s.
JL: We won't house all the stage 0 (unaffordable), but the provenance flags can provide the provenance tracking.

BT: I think a useful functionality to add will be an accessible list somewhere of merging decisions (both those which have been done and those which haven't), and a mechanism for people to provide feedback on these - noting that locals will probably know the finer details of their network, metadata precision etc. better than we can hope to as an international operation. Relatively few countries will probably take this up but would be good to have the facility there. Definitely not something that has to be in place pre-launch, though.
PT: We do that as part of the metadata that is spat out by the merge process and it can be hosted if desired.
JR: May need to be reformatted for usability.
**ACTION**: Jared Rennie to host the merge process metadata on ftp and make its availability known through the initiative blog and databank ftp pages.

KW: Query HadISD inclusion over ISD - will take this up offline but noted here so I don't forget. Generally very impressed with the work though. Main issues: some stations already merged, significant QC already applied
PT: HadISD given low priority because of this but general feeling has been that it has substantive value over ISD due to extensive QC on HadISD and work on merging for things like WMO country ID first two number changes - ISD contains too many shoddy / dubious records to be included. This is also why we backed off from GSOD inclusion (which is derived largely from ISD).

KW: Reanalyses background fields - hosting of these as part of the databank has been mentioned - is this on the cards?
PT: feasible in theory would be a nice Value Added Product but not an immediate priority.
**ACTION**: Kate to put Jared in contact with appropriate ECMWF people to scope potential to interact the databank with reanalysis feedback files.

KW: Potential for hosting non-standard met data? i.e. ad-hoc city wide temperature studies?

PT: In the future this will be possible but right now its best to focus on monthly and then daily long-term records rather than campaign / mobile data. Undoubtedly these data may be valuable down the line and we shouldn't forget them.

LV: Regarding use of homogenized and raw data. What are you using? Have you used homogenized / raw data for EC. Can you use only raw data?
JR: We try to save everything available in stage 2, but for Canadian data we have used only the raw in the Stage 3 merging program. We have 53 stage 2 data decks but only 47 are in the merge – the remaining sources are withheld because they are either gross duplication (Canadian homogenized is a subset of the larger raw for example) or there are perceived quality issues either over the metadata or the data (e.g. GSOD), which make us uncomfortable to include at this time.

JRC: What is the key issue that determines the higher trend result viz-a-viz GHCNv3?
JR: At end the higher trends may in part be coverage as found in CRUTEM4.
PT: See recent blogpost. Most of the difference appears to be in areas (5 degree grids) that were already sampled.

VV: What is the provenance of GHCND? Why top?
JR: Its top because its the daily stage 3 prototype. 2 variants use GHCND as a lower priority.
PT: GHCND as top is forward looking towards a vision of a vertically integreated set of holdings where users can data mine monthly to daily (potentially to sub-daily) resolution for better dataset understanding.
MM: A lot of the GHCND dataset does have good provenance. A majority. This comes from large well documented NMS sources and data rescue efforts.

CM: HadISD although it may be lowest priority - it takes higher priority than average temperature sources and those with some average component
JR: Yes, its the last max / min source that goes through. CM is right that it still takes priority over average only sources.

VV would like to see a variant with only data with good provenance.
PT: This could be done either as one of the official variants or by downloading the code and data and running locally. We will discuss.

**2. Immediate steps to databank full v1 release - Jay Lawrimore**

The databank will remain in beta release through December. This is the last chance for both us and the external user community to comment before proposed lock down. We plan to release version 1 and submit the methods paper in Jan 2013.

---

**Formal approval process**

The following members on the call or in abstentia hereby formally approve release of the databank v1:
Peter Thorne
Steven Worley
Kate Willett (although would like an extra day to properly read the paper)
Blair Trewin
Michael de Podesta
Robert Lund
David Lister

All others implied after 5 days notice to have approved

The following members on the call or in abstentia approve pending the clarification / changes detailed
None

The following members on the call or in abstentia do not approve at this time
None

---

Paper co-authors should send any comments to Jared Rennie (jared.rennie@noaa.gov) by Dec 17th 23Z. Those who are not co-authors can also comment and their input and feedback will be appropriately acknowledged. Jared will circulate a revised version based upon feedback by Dec 21st.

**ACTION**: Feedback on paper to Jared jared.rennie@noaa.gov by 12/17 @23Z. Jared to circulate revised draft to co-authors by 21st.

Can paper co-authors please advise if they have internal review procedures that are required to be fulfilled.

BT (on behalf of MF) - need to complete forms but can use lead author's institution's internal review reports for this purpose, no separate review required.
Steven Worley - no internal review procedures at NCAR
KW/CPM (on behalf of MetO) - 2 week internal review period usually needed
JRC - No issues from UAHuntsville

**3. Future databank development in the short to medium term - Jay Lawrimore**

The plan is to get this version 1 release archived and a doi from NCDC.

The monthly databank will be updated on a monthly basis (at minimum) utilizing the version controls documented in the paper draft

- Will the doi remain constant through the update procedure? (Steven Worley)

JL: Think it would. Need to consider further and discuss w/doi folks at NCDC.
SW: I agree

We are still interested in finding more new data sources. The plan is to periodically release substantive updates after queueing for a while and taking the time to understand the data and documenting their impact through a technical note and increment the v1.n by +1 accordingly.

In the medium term we need to start turning our attention to daily data and its merging. The next 'cycle' (three year benchmarking - see the ISTI paper in BAMS) would ideally have an ISTI databank offering at daily and monthly resolution. This means a release in calendar year 2016/2017 of monthly v2 and daily v1. Over the coming 12 months we will begin brainstorming on what we have, the current GHCN-D method and other aspects but we do not envisage any concrete progress on development. At this stage the priority is on understanding what data sources we have or could possibly get hold of. So the priority is to focus on finding and gaining access to daily resolution data sources and discussion of approaches. Both WG and SC members are strongly encouraged to help us in this regard. We will have several Databank WG calls next year both to review the v1 product and to discuss next steps.

XW: China Meteorological Administration (CMA) has agreed to provide me daily data for 400 stations in China. They said that I can contribute the data to ISTI. I would like to know the procedure for submitting daily data to the databank.
**ACTION**: JR and XW to follow up on getting the 400 Chinese daily station data into the databank
VV: The current benchmarking concept for the homogenization of monthly data is well suited for the next cycle with daily data. For daily data, the downscaling may need some more attention. The main new problem for daily data is that we do not know the properties of inhomogeneities in daily data. How do they change the distribution, which co-variates should be taken into account, etc.
I am trying to set up an open repository with (sub-)daily parallel measurements. This resource could be used to study the statistical properties of daily inhomogeneities. Help in making an inventory of available parallel measurements is appreciated, https://docs.google.com/spreadsheet/ccc?key=0ArsJmg1UW9uGdC1WMmRvclBHZTlwTHYxRXV4a1JnUkE, as well as data contributions.
**ACTION**: Jay Lawrimore and Kate Willett to work with their groups to facilitate the populating of a daily resolution parallel measurements database for understanding daily inhomogeneity properties

PT: We will also attempt to re-engage the data rescue issue both through the data rescue task team and the crowdsourcing avenues as funding opportunities arise
KW: 5th ACRE meeting showcased new data rescue efforts but lots more to be done - possible focus on Middle East
BT: Need to fully understand the rich mosaic of data rescue efforts. Augment ACRE with non-pressure data.

AKT: WMO expert team on data rescue is being stood up. We should try to understand what its remit is.

**4. Plans for creation and promotion of benchmark analogs to the databank - Kate Willett**

Please note that an annual progress report is now overdue.

KW: Monthly benchmarks still work in progress. Slow progress by Kate coding up Robert's method to work on larger station networks. Still not 100% satisfied my code is doing a good enough job but it may have to do for version 1.

Clean benchmark data (Analog known world) IDL code now running on MCDW data for the USA. This now needs to be recoded in open source software that also improves the statistical components (R/Python) which is also slow going.

Concept paper drafted but not submitted - needs polishing and circulating.

Methods paper for analog known worlds half written - awaiting final proof of concept using MCDW data for the USA.

Work has not really begun on the addition of error worlds and validation work is awaiting progress on creation of the benchmarks.

- KW to complete progress report by January 2013, circulate to WG members and sign off by end of January 2013

- KW to arrange WG call ASAP - it will be easier to plan a timeline when I have got the analog-known-worlds up and running as this is taking far longer than I had hoped. I've put off the call until I had something concrete to work with.

Daily benchmark PhD focussing on GHCN-Daily.

RL: If having troubles with matrix algebra we can perhaps simplify.
KW: Can we have a follow on call?

PT: Need benchmarks by end of next calendar year

**ACTION**: Kate Willett to engage with benchmarking working group and provide update to steering committee on what is planned and a new timeline to release of the first version of the benchmarks. By no later than end of Jan 2013.

**5. Engaging analysts to create data products from the databank holdings - Peter Thorne**

A subset of the steering committee have started trying to work out how to engage potential groups with an interest in analyzing the databank holdings, create products and partake in the benchmarking exercise. If we don't get a critical mass interested and taking part then this will fail.

Firstly, NCDC will be creating its GHCNMv4 product from the databank and running it on the benchmarks (once produced). The v4 is likely to be released second half of 2013 and is already under development and testing.

What strategies should we follow for engaging groups who might be interested in creating products from v1 and also engaging the broader community? Suggestions and action owners are most welcome. Obvious avenues:
1. Personal contacts
VV: I think the document attached to the mail lists the most likely groups to be interested in homogenization of the ISTI dataset.

2. piece in EOS or similar/flyers or adverts that can be dropped at conferences or submitted to journals/magazines - something fairly snappy with a weblink rather than a ream of text.
3. CLIMLIST
4. Potential meeting (funds?)
5. Presentation(s) at pre-existing meetings and fora?
- Kate Willett and potentially Matt Menne are likely to present at the next IMSC meeting.
- A poster at the homogenisation session or even a splinter meeting at EGU, http://www.egu2013.eu/townhall_and_splinter/splinter_meeting_request.html ?
- Also possible at session "Taking the temperature of the Earth: Temperature Variability and Change across all Domains of Earth's Surface", further details can be found on the meeting website at: http://meetingorganizer.copernicus.org/EGU2013/session/12115
- EarthTemp Network (June, Copenhagen) -- interest in exploiting with satellite data as well as standalone

**ACTION**: Peter Thorne to convert the document provided by SC members into a spreadsheet on googledocs which everyone can access and members of SC / WG to volunteer to approach the relevant individuals to encourage their participation in product generation from the databank and in the benchmarking exercise.

KW - anyway of feeding into forthcoming funding calls i.e. NERC - my understanding is that there is sometimes a request for what recent burning issues require funding and then some funding calls are tailored to meet those needs - e.g. NERC water cycle funding?

RL: Bigdata funding in the US - ISTI fits. NSF. Submit a proposal.
VV: I am not sure big data people find our dataset to be big.
SAMSI theme on uncertainty quantification. PT to follow up on.

JRC-idea: I think to succeed there needs to be a responsible person per country or group of countries who devotes time to be the data guru for that country, to develop relationships needed to acquire the data we desire and guide the provenance descriptions for the Databank. Don't know how to do this apart from volunteers at this point.
BT: Need more local.
PT: WMO expert team / ISTI? Who leads?
AKT: Yes we should be trying to fill missing metadata. We have a role to play.
PT: Use GCOS focal points? Request to AOPC.
AKT: Also WMO
Work to create a stage 0 linkage for every data point.
MM: GCOS AOPC would be appropriate.
**ACTION**: Peter Thorne will liaise with GCOS secretariat and AOPC to ascertain how most appropriately to raise a request for regional national focal points to review the databank release and help us to start to improve the provenance of what we have.
SW: Can we spit out on a country by country basis?

JRC-Questions (don't know where to put these): Are global averages determined from 5x5 grids? Are smaller grid spacings possible? Are interpolation schemes available to infill all land grids?
PWT: Lets cover under AOB. But basically this is part of teh structural uncertainty we wish to capture. These things should not be prescribed per se as how we average / interpolate likely matters in the same way that choice of databank merging or homogenization approach does.
CPM: This would require stage 5 homogenized station data. Are there other efforts than at NCDC to provide these? Once available existing algorithms could be tested.
PWT: Basic idea is to try to get multiple stage 5 products and certainly having some standard processing options to isolate station qc/adjustment effects from averaging and interpolation differences would make a degree of sense.

**6. AOB**

Suggest next steering committee call at 8am EST, 13Z on Jan 3rd. We need to put together an annual report by end of Jan.

AM. I Suggest to give a general presentation of ISTI and progress at Tempmeko 2013 - Madeira - 14 Oct 2013. Large Symposium on temperature. A session on Environment will be included. 28 feb 2013 deadline for presenting abstracts. We can talk abaout at next call in January.
http://www.tempmeko2013.pt/dates.php
**ACTION**: Peter Thorne and Andrea Merlone to discuss Tempmeko presence for ISTI.