ISTI Databank Working Group call

Discussion notes collected before and during the conference call via etherpad. A summary of the actions resulting from this call are at the end of the document.

30 Sep 2015 @12Z (14 Europe, 13 UK, 8 EDT)

Present: Jay Lawrimore (JL), Peter Thorne (PT), Jared Rennie (JR), Steve Worley (SW), Victor Venema (VV), John Christy (JC), Albert Klein-Tank (AKT), Colin Morice (CM), Kate Willett (KW)

Apologies in advance: Matt Menne (MM)

Agenda

I. Welcome and review action items from the previous Databank WG conference call (Lawrimore)

Action: JL and others to make decision regarding whether to increment as v1.1.0 or v2.0.0 with an accompanying journal article.

Decision is to release as version 1.1.0.

Action: JL to initiate a letter of support for PT's Irish grant proposal.

Letter of support completed.

PT: Decision from grant body still pending at this time

Action: JL and JR to follow up with Kate regarding the version of Databank for benchmarks.

Some discussion regarding Kate's use of databank and feedback provided by Kate.

PT: Note that Kate is close to getting clean worlds out. Need to decide whether to spin these off of v1.0.0 or v1.1.0

VV: The clean worlds will not be published, so it should be no problem to use v1.1.0 for this. When the error worlds are published would be the important deadline.

MM: There is also the issue of potentially using the stage 4 data to gen the benchmarks since some spurious bad values may be impacting the station variances in some cases

Action: JL to contact Marc Prohom regarding climate series for the Pyrenees.

E-mail sent. Data along with other new sources to be incorporated in version of databank subsequent to v1.1.0.

From Marc - The database was built jointly with other partners from Spain, Andorra and France (a EU POCTEFA project). I have to confirm if the data is freely available, but I think that there is no problem, especially for research activities. The data we developed is on monthly resolution, both for Tmax and Tmin series.

Apart from this punctual collaboration, my institution (Meteorological Service of Catalonia) would be interested in taking part of the International Surface Temperature Initiative. In fact we are already involved in the Parallel Measurements project, and we think that we could collaborate in other issues. For example, we are in charge of more than 170 AWS (with more than 25 years of data) and we are

involved in data rescue and homogenization of historical series (back to 1780) that could improve the temperature databank for this area of the Mediterranean (you can have a look on our website, although is only available in Catalan by now: http://www.meteo.cat/wpweb/climatologia).
JL: If no exceptions will ask if he would like to join the WG.

Action: JR and others to determine if we lost coverage in the early years (pre-1950) because of our removal of the 4 homog. sources.
MM: Just a note, the data sources were ommitted from consideration in latest (v1.1.0) merge not because they were homogenized sources per se, but because they had threaded station records (a join of two or more stations) and higher ranking sources had these separate components already. The merge algorithm cannot be expected to resolve such issues (and the best outcome is to withhold the data, which doesn't always occur). This issue had come up repeatedly with other sources in previous merges, and it means that there is a limit to thoughtlessly throwing everything into the merge pot and hoping for a good outcome. It has become clear that intellectual curiosity and an informed background on what is being included and why is necessary to building a credible database.

Action: JR to post summary document to Databank website on surfacetemperatures.org.
JR: Will do when v1.1.0 released

Action: JR to review potential data from Chile as recommended by AKT.
Still pending
See http://lacad.ciifen.org/

Action: JR to share analysis with the group before making a final decision.
Covered in current call

Action: PT to add Victor to mailing list and to change worley.steven@gmail.com to worley@ucar.edu.
Done

Action: VV to make POST list of parallel datasets available to the databank WG members.
Done

Action: MM and AKT to determine if any potential sources for addition to databank merge are already in ECA&D.
Add as an action for subsequent to this call.

II. Status of Monthly Databank and review of updates for v1.1.0 (Rennie)

CM: CRUTEM4 is now updating from some ISTI sources, which adds support to removal from the recommended stage 3 merge. See
http://www.metoffice.gov.uk/hadobs/crutem4/data/CRUTEM.4.4.0.0_release_notes.html

Will sources ommitted from the recommended merge be maintained in stages 1/2 of the databank?
JR: Yes, they are still on the Databank FTP

PT: Encouraged that new version has better coverage throughout record and does not suffer a drop-out in recent years.

PT: Assuming WG decides to okay release we should give some thought to a comms strategy. At a minimum we should write a post on the blog and update the databank webpage. Is there other stuff we should do?
VV: Is there something new that would make it newsworthy? More than "just" more data.
VV: If not, GHCNv4 might be the better moment to put our energy in.
PT: Happy to limit to a blogpost and website update if that is general feeling. We should also tweet to raise awareness (those of us with twitter accounts).
MM: Agree that not much publicity is required on this one.
VV: I am not the "general feel", was just a thought. Is there a scientifically interesting change?

PT: What are next steps to release? Can we release now(ish)?
JL: Will be released by October 15th (between 8th and 15th).

PT: Can the tech note be made public? Would be great if it could.
JR: It will be provided with release, as part of versioning control

PT: When will the variants be available?
Will take some time to run the variants given server limitations at NCEI. Difficult to give a specific date but not too distance future.

PT: Some sources easy to add in - particularly those sitting around for 1 to 2 years. Can this be considered?
JL: Yes will determine if possible.

SW: would be good to add so that when create the new merge will be easier to sort out the differences.
SW: when looking at changing the threshold: one area of globe influenced more than another?
JR: had looked on the global scale. nothing coming to mind regarding one particular region standing out.

III. Status of Daily Databank (Menne)
MM: As Jared's tech document notes, GHCN-Daily v3.21 was released last June that includes the element (TAVG) - average daily temperature - as well as 1400 stations new to GHCN-Daily from the RBCN. The TAVG daily data are generally not computed as (TMAX+TMIN)/2 but rather are calculated using fixed hours of the day according to national tradition (and supplied by the data providers) or synoptic hours. TAVG was added since a daily average of traditional fixed hours pre-dates the use of maximum/minimum thermometers in many countries. TAVG from synoptic hours for RBCN sites

provides some protection for issues associated with late or erroneous CLIMAT reports in monthly updates. In addition, the monthly merge algorithm was unable to cleanly merge the CLIMAT sources into the stage 3 data, so the solution to this problem was to ensure that all ~3000 RBCN stations are represented in GHCN-Daily. This meant that a simple id match could be used to match the monthly CLIMAT data sources to the stage 3 data via the GHCN-Daily WMO ID cross reference.

GHCN-Daily v3.22 as also released ealier this month. Version 3.22 used a different (more authorative) source of daily SNOTEL data (a network of monitoring stations in the mountainous western U.S.), which also includes the addition of 191 new SNOTEL stations, including 60 from Alaska. Unfortunately, these new values will not be represented in ISTI v1.1.0.
PT: would be in 1.2 if you go that route.

CM: Is there documentation available for the v3.21 update (changes in global coverage, coverage in earlier record etc.)?
MM to follow up with CM

IV. Update on activities of the Int'l Surface Temp Initiative Steering Committee and Benchmarking progress (Thorne)

Relatively little progress of the initiative as a whole since the last call. The next milestone will be the release of the benchmarks at which point there may well be a further flurry of activity. We still need to start working on a products portal and instigate and underlying group to work on it.
PT: No current portal lead or group instigated. SC has task to instigate the group. Really need funding to support it. If Irish grant comes through there are funds to support it.
VV: The BAWG has started coding the error worlds. You can see the progress here:
https://github.com/SurfaceTemp/ISTI_Error_Worlds
VV: Everyone is welcome to join, that is why we are on GitHub.

KW: The code is now up and running for simulating the clean worlds. I have written a paper presenting the methodology using v1.0.1 (downloaded July 2015). We plan to rerun for the real benchmarks anyway so upgrading the ISTI versioning is actually a good thing. Hopefully I'll get the paper submitted by Christmas. The focus is now error-worlds and validation.
I have had issues with:
- cannot simulate stations shorter than 3 years anyway ~2000
- 30 stations which appear to be ships - I've not included these
- a few stations which look identical (but could be rounding of r to 2 significant figures only?) - not actually a problem for me as long as their locations differ - they will be simulated differently.
- 510 stations which have identical locaitons - real or due to lon/lat precision - I think I can deal with these
- a number of stations which don't have elevations - I have obtained a proxy using a digital elevation model

- poor quality station data (e.g. USS0004B02S) which makes the variability very large in the simulated world - 0s as missing data, December goes from -11.65 to +35.73

- - Matt has suggested I use stage 4 data instead which I could do but found it hard to find access to.
- - JR: USS0004B02S is not in the QCU data, did not pass the 10 year threshold

PT: In going from stage 3 to stage 4 (qcu in NCEI speak) do we lose stations compared to stage 3? If so would we need to reinject stage 3 not in qcu?
JL: qcu is certainly better.
JR: We lose 10,000 stations which are short. qcu is alpha.
PT: If someone wants to use every single station they would not have that ability if the v4alpha dataset was used (~26000).
AKT: Thinks it's a good thing to use the 26K stations - can't simulate the very short stations.
PT: the qcu is already quality controlled -- if running benchmarks off qc then no way of injecting bad data??
KW: Our benchmarks are purely focussed on testing ability to deal with systematic error. We though that random error should be dealt with seperately.
PT: will carry this over to the benchmarking WG. decisions will need to be made soon.
KW: when will data be available?
JL: qcu data will be available on the 15th or 16th of Oct on the ncei ftp site in beta form.
KW: Tempted to publish, methods based on stage 3 but run real benchmarks on qcu/stage 4.

I am still waiting to hear back on my Irish grant application which if successful would be a substantial force multiplier.

I still have on my radar the desirabitlity of having an ISTI workshop to bring folks back together and re-energise. If anyone has any funding suggestions I am all ears.
VV. Good idea. For the BAWG the most important moments would be before we publish the error worlds or after people return the homogenized data.
VV: With funding is better, more people would show up, but could we also do it without funding?

Progress report from Databank WG is due this coming month please.

V. Update on activities of the Parallel Observations ScienceTeam (POST) and kick-off conference call (Venema)
ISTI-POST is now on GitHub. We keep our code here to make collaborative coding easier.

BTW, also ISTI-BAWG clean worlds and error worlds projects and the daily benchmark are there as subprojects of the ISTI
https://github.com/SurfaceTemp
May also be an option for the databank group, if NOAA code review rules permit.

To share the data *before* publication, we use a cloud server or FigShare, at the discretion of the project leader.
http://figshare.com/
After publication the full dataset will be published at NOAA datacentre, connected to the ISTI and at ECA&D. Maybe FigShar would be an option for the data for a specific paper. They also give doi's.

We have set up projects by now.
POST-temp: The transition from conventional observations to AWS for temperature (lead by Enric Aguilar; members: Renate Auchmann, Petr Stephanek, José Guijarro, Alba Gilabert, Theo Brandsma, Victor Venema)
POST-precip: The transition from conventional observations to AWS for precipition (lead by Petr Stephanek; members: Renate Auchmann, Victor Venema)
POST-early: The transition to Stevenson screens for temperature (lead by Theo Brandsma; members: Peter Domonkos, Renate Auchmann, Victor Venema)
POST-move: The influence of relocations (lead by Alba Gilabert; members: Blair Trewin, Jenny Linden, Manuel Dienst, Bert Heusinkveld, Jared Rennie, Renate Auchmann, Victor Venema)
POST-humid: Changes in humidity measurements (lead by Kate Willett)

We had a meeting for POST-temp and POST-precip in Brno and also fleshed out more general details of the data processing there. There was a talk on both projects at EMS2015, At the Data Management Workshop (St. Gallen) Enric will give a talk on POST-temp and Petr on POST-precip. We have about a dozen datasets now, most in Europe and South America. Africa and Asia are hard to find data; suggestions very welcome.
Theo has started gathering data for POST-early and for POST-move we will soon have our first telecon.

POST-humid still has to start, which the chair does not mind, we are sufficiently busy. There was a suggestion for a project on wind, but we do not have enough people and especially person power for that. Other suggestions are naturally welcome.

Our software converts all data in a standard format and computes averages at daily, monthly and annual scales. It now includes QC and manual detection of inhomogeneities (series are split, not corrected). Enric has been working very hard to make this happen.

We want to implement the indices of the ETCCDI and the WMO ET-SCI (sector specific indices). The ET-SCI software is on GitHub as well:
https://github.com/ARCCSS-extremes/climpact2/blob/master/ClimPACTv2_manual

Time plan. We are behind in not having coded the indices yet, but have made good progress in building up groups to work on specific topics and have worked more than planned on finding data.

To get more data maybe also some publicity may be good. A blog post is planned after the Data Management Workshop, to make more concrete what we are working on. That could be adverticed on

the homogenization list and CLIMLIST. Do we have something like a ISTI newsletter for (bi-)annual updates? Maybe also a short article in EOS, EGU News or even BAMS would be possible. Any other ideas?

Related news. Jared has written a blog post on his new study comparing parellel measuremens of COOP and USCRN.
http://variable-variability.blogspot.com/2015/06/COOP-United-States-Climate-Reference-Network-USCRN-stations.html

Getting permissions is less hard than many had expected. Parellel data is seen as experimatal data, not operational. Up to now we can publish all the daily data, except for two cases where we can still publish the indices. In one case we could "only" get 5 years of data for a very long dataset. North America sits on their data; first want to publish themselves, but then we could have it.
ATDD, Atmospheric Turbulence Data Division.
NWS testbed Sterling facility
May both have parallel measurements
Jay to follow up with NOAA colleagues at these facilities.

VI. Other Business (All)

VV: State multi-element database

MM: This may be a topic for the next meeting. We are slowly laying the ground work to vertically integrated the hourly/daily/monthly data. In doing so, there will be implications for how we use or don't use the currently monthly merge technology. In any case, it is proving difficult to manage the ISTI monthly temperature databank because the merge algorithm leads to many results that are difficult and time consuming to investigate (withheld stations or sources that should be added, duplicate stations that shouldn't have been added, etc.). Ironically, the merge decisions, though ultimately tractable, are not all that transparent! These issues that we have spent many months dealing with are much more easily resolved and managed using the GHCN-Daily merging and reprocessing paradigm, which I am recommending at this point that we use for future merges. Making this change will also allow us to manage station metadata, which we will have to do as part of the vertical integration. More on this later.
VV: I could imagine that the merge algorithm could be made much more specifc if it used the variance of "homogenized" difference times series and cut away 1% of the largest differences. That would detect pairs that belong together, but where one has been (partially) homognized or QC-ed.

JL: Daily CLIMAT data in BUFR format: Creation of a daily CLIMAT message for the dissemination of summary of the day climate observations. The need was highlighted at the first meeting in Exeter. Have been working with WMO and developed a BUFR template for a once-a-month transmission of the summary of the day observations for the past month (temperature, precip, snowfall, snow depth). Producing a test each month and sending to Colin for review and validation.

CM: Have other things out of the way and will be able to begin evaluating in the coming month.

VII. Summarize Activities and Next Steps for coming months(All)

PT: How many new data sources are now queued, how many stations in them, do any potentially fill data voids, and what is the plan for v1.2 / v2? (MM: see above)
JR: Close to 15 at this point.
PT: Any in areas where there is poor coverage?
JR: Need to look at it more -- and will focus on ID'ing sources that would help fill in.
PT: certainly prirority would be those easy to add and those maybe not easy to add but would help fill in data sparse areas.
PT: could we have maps of the sources and PORs so WG could get a sense.
JR: yes. can do.

PT: Given the potential timescale of a complete redo of the merge algorithm should we plan to release a version using current algorithm and as many new sources as we can reliably add in a few months? Risk of holding everything back is that it puts off further new submissions if previous submissions get sat on and not incorpated?
JL: yes will give this more consideration.

ACTIONS resulting from this call:

- JL, JR, and MM to consider whether it's possible to add some existing new sources using the current merge methodology while recognizing need for a revamped methodology along the lines of the methodology used for GHCN-M.

- JR to provide maps and PORs for each of the sources that could be added to the databank.

- KW to follow up with Benchmarking WG on whether to use GHCN-M v4 qcu beta or whether a qc'd version of the full 36,000+ databank station set is needed.

- JR to make v1.1.0 Tech report available along with data and create a post regarding the release.

- JL to follow up with ATDD to ask whether they have an archive of historical parallel measurements that could be used by the POST.

- MM to respond to CM's question about the new additions to GHCN-Daily.

- CM to begin testing and evaluation of the Daily CLIMAT data.