

15 February 2012 Databank WG Conference Call Summary Notes

1300-1500 UTC

Members In Attendance

Jay Lawrimore (NOAA NCDC, USA, Chair)
John Christy (University of Alabama, Huntsville, USA)
Waldenio Gambi de Almeida (CPTEC/INPE, Brazil)
Byron Gleason (NOAA NCDC)
Albert Klein-Tank (KNMI, Netherlands)
David Lister (Climatic Research Unit, East Anglia, UK)
Matt Menne (NOAA/NCDC, USA)
Colin Morice (UK Met Office)
Madeleine Renom (IFFC, University of the Republic, Montevideo, Uruguay)
Jared Rennie (NOAA NCDC, USA)
Matilde Rusticucci (Univ of Buenos Aires, Argentina)
Steve Worley (National Center for Atmospheric Research, USA)
Peter Thorne (CICS-NC, USA, Chair, Surface Temperatures Initiative Steering Committee, ex officio)

Regrets

Meaghan Flannery (Australia Bureau of Meteorology)
Albert Mhanda (ACMAD, Niger)
Barbara Tencer (Univ of Buenos Aires, Argentina)
Koji Ishihara (Japan Meteorological Agency)
Vyacheslav Razuvaev (Russian Research Institute of Hydrometeorological Information)
Jeremy Tandy (UK Met Office, Exeter, UK)

Agenda

- I. Introduction of new member to Databank Working Group (Colin Morice, UKMO)
- II. Review action items from 20 Oct 2011 conference call (Lawrimore).
- III. Update on activities of the Int'l Surface Temp Initiative Steering Committee and ITS9 Conference (Thorne) (**attached** is final version of Databank conference paper.)
- IV. Update on data collection activities at the Regional level (All)
- V. Overview of algorithm development for Stage 2 to Stage 3 Merge (Rennie, Gleason) Refer to **attached** document (Databank-MergingMethodology-15Feb12.pdf)
Jared and Byron placed additional information and graphics on the wiki under the Merge Program heading at http://editthis.info/intl_surface_temp_initiative/Main_Page .
- VI. Determining appropriate thresholds for Data Overlap Tests (Morice)
- VII. Crowd sourcing task team update (Thorne)
- VIII. Summarize Activities and Next Steps for coming months
- IX. Other issues

I. The WG welcomed new member Colin Morice from the UK Met Office!

II. Action Items From 20 Oct 2011 conference call.

SW to share ICOADS white papers on merging process with NCDC.

Provided and made available on the Databank wiki -

http://editthis.info/intl_surface_temp_initiative/ICOADS_Merge_and_Duplicate_Elimination_Processing

JR to investigate possibilities for automatic notification of wiki changes to working group members who are interested.

Automatic notification is not possible but Jared will notify WG members when new information is added.

JR and others to continue working on merging process for Stage 3 Monthly data and to report back to the WG on strawman.

To be discussed during Agenda item V.

MF to continue working with Pacific Island partners and to report back on progress next call.

MF provided JL with an update and will be included under Agenda item IV.

MF to follow up and determine what data sources provided by NIWA can be included in the Databank and made freely available.

MF provided JL with an update and will be included under Agenda item IV.

JL to work with Graphics team on developing a certificate for country recognition. Completed and now available.

PT to coordinate at WCRP OCP to determine if co-signatories can be made by CCI, GCOS, and WCRP on the certificates.

The certificate was signed by Pres of CCI, Chair of GCOS Steering Committee, and Director of WCRP.

JL to prepare letter for high level request for support from Brazil INMET. WGA to follow up with JL on the issue.

Letter completed and JL to continue working with WGA.

DL to follow up with IEDRO to determine if they would be able to assist with imaging of the forms in Phil Jones' office.

DL to provide update under Agenda item IV.

DL to follow up with JL and JR on other sources from the UK.

DL to provide update under Agenda item IV.

JC to send Uganda digitized data as it becomes available.

JC has sent data for 3 stations and will send all stations by end of Feb.

SW to continue to pursue data associated with NCAR library digitization efforts.

SW to provide update under Agenda item IV.

AKT to discuss Surface Temperature Initiative with participants at the ETCCDI workshop in December and investigate potential new participation and possible data sources.

AKT to provide update under Agenda item IV.

III. Update on Steering Committee activities.

PT provided an update on steering committee activities.

<http://www.surface temperatures.org/steering-committee> . Noted general updates including ITS9 presence and the progress report.

JL noted the Databank ITS9 conference paper was completed and thanks to everyone for their contribution to the paper.

IV. Update on data collection activities at the Regional level.

DL made contact with IEDRO about the data held at CRU. Hard copy material. Contact with BADC. Hopeful of additional data series for the UK. **DL to continue to work with BADC to identify new source data.**

JC: UK itself or UK former territories?

DL: Perhaps some outside the UK.

JC: Close to completing station data for Uganda. **JC will send to Jared Rennie and Jay Lawrimore (32 stations with at least 4 years of data).**

JC: Included GHCN what to do there?

JL: GHCN is a source.

JL: How does the merging activity JC did interact with the rest? Should it be that we trust these as higher sources? Put it as a higher priority source.

JL: Stage 3 is a consolidated version of stage 2 sources. Stage 2 that is merged already may be given a higher priority.

PT: trust that source as higher.

Additional discussion needs to be made. **JL to follow up with WG members and report back to group.**

AKT: Monthly updates for Europe and SE Asia. New data in SE Asia in last month can be added. Raised at workshop in Gambia. Approx. 80 unique stations that could be submitted. Work in progress. Reluctance from data holders, but somewhat hopeful. Would be submitted directly to the databank? Possible. Workshop paper in progress keeps this going.

JR will double check to ensure new stations within the ECA source are being pulled through when they are added – not just pulling in existing stations.

WA: Have converted more data to databank format. Lost some people in the last month or two which has complicated things. Stalled presently but looking for additional resources. Brazilian NMHS request letter will be sent shortly. Some weeks time we should have an answer.

JL to send completed letter.

WA will send a second letter from CPTEC to INMET, reinforcing the importance of the participation in the project - to obtain position on data release from INMET.

MRu: Argentinian data discussions with Agricultural national research institute, not NMS. They can provide series, 20-30 years and some experimental stations. We had discussions in Dec. Will restart now that it is past the holidays so will restart contact to try to put this on. Daily data in general. Not yet touched on monthly.

MRu will continue discussions with Agricultural national research institute to acquire daily data.

MRe: Has been talking with NMS about the data. Letter to send to NMS would help. **Will request meeting and reflect importance of the initiative.**

JL to complete a letter of request in support of MRe efforts.

SW: Looked through archives and identified a number of sources. Greenland 1987-> sent, fairly large world monthly which is a composite of many sources. 1738->. Captures quite a lot of sources (DS570). Other assets being prepared and sent in. Antarctic stations may be an opportunity. Some location problems exist in some of the datasets. Working on these.

SW to continue efforts to obtain other assets including Antarctic stations.

JR: Update on Databank Stage 1 and Stage 2 sources. We have the two sources SW referred to. Seven new sources since last call. Four more about to go in. UW-Madison source is being chased presently as a very promising lead.

JR to continue incorporating new stations into Databank as they are acquired.

MF provided an update via e-mail:

- Discussions with NIWA are continuing. Jochen Schmidt (chief NIWA climate scientist) is considering the request to supply historical NZ data to ISTI global databank, however he has stated that NIWA's IP Policy and EULA does not allow the re-hosting of their data. The BoM climate program is hosting Errol Lewthwaite (NIWA) for 2 weeks so I intend to get him inside re: ISTI goals and objectives. I am hoping that we don't have to wait until 2015 for congress to ratify the amendment to resolution 40 demanding that nations supply their historical data. In the intervening period I intend on forwarding to Jochen the draft amendment of resolution 40 (when it becomes available) if negotiations continue to stagnate.
- Work is progressing on an agreement with Vanuatu to host their historical data. I am chasing the agreement protocol now through the BoM's Executive and International Affairs EIAB branch. This is necessary as EIAB must approve all agreements that come under the WMO umbrella. Janita Palahad (manager of COSSPac) and Amanda Amjadali (new manager of the PCCSP) are travelling to Fiji this April and intend on presenting/discussing this agreement with four Meteorological Bureau Directors from Vanuatu, Fiji, Tonga and Samoa, with the focus on Jotham Mapat (Director of Vanuatu Met. Service)

MF to continue progress toward agreement with Vanuatu.

MF to meet with Errol Lewthwaite (NIWA) to get him inside the ISTI goals and objectives

MF to forward to chief NIWA climate scientist the draft amendment of resolution 40 if negotiations continue to stagnate.

V. Overview of algorithm development for Stage 2 to Stage 3 Merge

First thing is to prioritize the sources then undertake tests with geolocation metadata, data overlap, and data non-overlap to decide whether unique or not and whether to merge. There is a withhold option.

DL: Is this just monthly? Do you plan to look at daily?

JL: Daily is a longer term aim, we'll learn a lot from monthly.

The prioritization of the 36 sources we currently have in Stage 2 is not fixed. WG input on order of prioritization is encouraged.

PT provided an overview of the prioritization strawman as available on the wiki at http://editthis.info/intl_surface_temp_initiative/Source_Hierarchy .

SW: Question regarding why “Data sources which contain max / min data, have had no QC / homogenization applied with poorly known provenance” is a higher priority than T Avg data with Known provenance.

JL: Agree this is one that should be moved higher in precedence since Known provenance is a high priority of the Databank.

PT: First iteration should be max/min, second should be average. That way we only have a max/min priority.

AKT: How are you going to do max and min?

PT: Suggest mirror merging decisions made for maximum temperature to minimum temperature and then look at T Avg. Avoids an issue with potentially assigning to different stations max and min. Also saves compute time.

We aren't looking to inter-mingle which could create issues.

JC: Is there an issue with different data resolution (tenths vs. hundredths, e.g.?)

JL: Need to consider this in prioritization.

SW: What you really want to know is the accuracy of the reading. What was the original accuracy? A tenth means can you infer hundredths? When in F originally tenths.

SW: How accurate was the original number is the key question?

CM: Averaging to monthly may be able to be more accurate.

JR summarizes results of initial merging effort. Metadata comparisons lead to decision to test. Overlap and non-overlap tests. Overlap = 5 years+ overlap. 3 tests were considered. Normalized RMSD, paired t-test, index of agreement – somewhat similar to r^2 .

Paired t-test was rejected.

Two versions of the merge being worked on. Eventually will settle on either RMSD or Index of Agreement.

Data thresholds are being worked on in collaboration with CM.

When you change calculations you change decisions regarding the station mix. This is not a certain science. Decisions will impact the outcome.

Withholding is an important aspect. These consist of stations for which the automated process cannot determine if a candidate station should be merged or is unique. This is a necessary part of the merge process because it is not possible to confidently determine every possible case using only automated methods. And it is not possible to look at every potential pair manually.

Will be working with Non-overlapping data merge process following completion of the overlap tests. May use tests such as difference in means and variance.

DL: Need to be concerned with series that have been homogeneity adjusted. Many stations will have some parts of series adjusted and others not adjusted. This will affect the ability of the test to identify stations for merging.

JC: Might there need to be country by country choices?

JR: Certainly.

AKT: Entirely different answers in merge decisions.

JR: May be down to arbitrary thresholds. CM has been working on thresholds and this will modify the results.

VI. Determining appropriate thresholds for Data Overlap Tests

CM provided an overview of the tests he has developed for identifying appropriate thresholds.

Two tests. Unique, same, two nulls are required to test each with the grey area in the middle.

Below Threshold H1 the master and candidate stations will be identified as definitely different. Above Threshold H2 the master and candidate stations will be identified as definitely the same.

Trained using difference series. Set thresholds accordingly. Bootstrap using difference series with null criteria, different overlap periods -> inference of distribution to set thresholds. Mean and variance assigned to two nulls.

Will enable us to more directly compare the NRMSD and Index of Agreement methods soon and then we'll have a better sense of the test approaches. Thresholds can still be tuned.

JC: Will geographical issues come into play – one set of thresholds work in one region and not another?

PT: Possibly for h1, but h2 should be geographically invariant.

Group generally comfortable with the merge prioritization

Opened up discussion for comments from every workgroup member on merging process – particularly the issue of whether the merge process should err on the side of merging too many stations that should be uniquely identified or erring on the side of making too many stations unique that should have been merged.

SW: set up pretty well and happy with progress. Err on side of keeping things unique.

JC: Want to see some case studies from E Africa to make a decision.

JL: Encourage everyone to help us to validate what we do in regards to their regions of expertise. Will be working toward continual improvement of the merge process.

AKT: Simpler is better, robustness of thresholds. Will look more closely at Netherlands stations. Too many unique is preference. Gut feeling.

MR: Err towards unique

Madeline: Err towards unique

JC: What if we have two stations that should have been merged end up in there unique?

PT/JL: They will be separate ids, user would need to decide.

Waldenio: May be country by country. Sometimes data is wrongly identified. Need to think more.

DL: Err on side of unique assignment. Withheld bin shouldn't be trashed. Comparison of overlap using arithmetic differences by eye is sufficient.

PT: Note that we will release the code. Others will be able to assess sensitivity.

Discussion of need to publish article on the Databank. General agreement that we need two articles. One as a general overview of the Databank and philosophy and a second more technical paper specifically focused on Stage 3 merge methodology. BAMS + JAOT strategy.

VII. Task team updates.

Crowd sourcing. <http://www.surface temperatures.org/databank/data-rescue-task-team>.

PT provided an overview of the crowdsourcing task team recent activities. Continues to run into problems associated with the need for funding for imaging and digitization of the many sources available.

Outlined that we could do this, just need to crack the funding nut.

VIII. Summarize Activities and Action Items

JR and others will continue to work on merge process for overlap and non-overlap conditions.

CM to work with JR and BG on determining appropriate thresholds for merge and unique stations.

DL to continue to work with BADC to identify new source data.

JC will send to Jared Rennie and Jay Lawrimore (32 stations with at least 4 years of data).

JL to follow up with WG members and report back to group on how to handle situations where a data provider submits raw data and data that was merged or processed using expert judgement.

JR will double check to ensure new stations within the ECA source are being pulled through when they are added – not just pulling in existing stations.

JL to send completed letter to Brazil INMET.

WA will send a second letter from CPTEC to INMET, reinforcing the importance of the participation in the project - to obtain position on data release from INMET.

MRu will continue discussions with Agricultural national research institute to acquire daily data.

MRe: Will request meeting with NMS and emphasize importance of the initiative.

JL to complete a letter of request in support of MRe efforts.

SW to continue efforts to obtain other data including Antarctic stations.

JR to continue incorporating new stations into Databank as they are acquired.

IX. Other issues.

None.