Benchmarking Working Group Call #9
Friday 7th June 4pm BST (GMT+1), 3pm GMT,

Attending: Kate Willett (KW), Victor Venema (VV), Ian Jolliffe (IJ), Peter Thorne (PT), Rachel Warren (RW), Robert Lund (RL), Matt Menne (MM), Lucie Vincent (LV), Claude Williams (CW), Stefan Bronnimann (SB), Enric Aguilar (EA).

Not attending: Renate Auchmann

ACTIONS FROM PREVIOUS MEETING:

ACTION SB to check availability for attending NCDC workshop in July. - NOT ATTENDING
ACTION: ALL to pass around blogpost to anyone who may be able to help and submit any info they are aware of. - DONE
ACTION: KW list everyone who has responded and email round list and blogpost link. - DONE
ACTION: RW and KW to do the summarising and post to benchmark/ISTI blogs. Should be forwarded to homogenisation email list. - BEGUN
ACTION: VV to tidy document and circulate round particularly to Claude/Peter and others not on call - two weeks time. - PENDING
ACTION: KW to add RA to the email list. - DONE
ACTION: KW email Enric Aguilar and double check with silent members whether they still wish to be bothered by emails. - DONE
ACTION: IJ to recirculate Team Validation document in prep for next call (Fri 7th June) - DONE

ACTIONS FROM THIS MEETING:
KW: to invite potential new members
IJ: to continue trying to get new validation experts on board.
KW: add link on our website to VV's post on benchmarking
KW: Circulate minutes
KW: Circulate call details for next week
KW: to see about guest speakers at next call - Lucie (Canada), Manola (Spain?), Andrea (Meteomet project update)?
IJ: tidy up discussion on Team Validation
MM: Circulate meeting agenda when its ready
KW: sort out the next Team Validation call

AGENDA:

1) Welcome to new members:
Enric Aguilar
Renate Auchmann
Matt Menne

2) Other potential new members:

Ingeborg Auer (Head of EUMETNET?)?
Gregor Vertacnik?
Dan Hollis?
Extra validation person to be chased by Ian:
KW: ACTION to invite potential new members
IJ: ACTION to continue trying to get new validation experts on board.

3) Team Validation - Ian Jolliffe
Document circulated for discussion
Clarify what Team Validation should be doing in the coming weeks/months
From the proposal by IJ: There are three things that we might want to validate:
Validatation metric 1. Correspondence between 'truth' and homogenised series.
PT: Presumably we want to know this:
1. At various scales - station, local/gridded/gridpont?, regional, global
2. For various diagnostics - linear trend, ARMA (or similar) behaviour, mean, variance (possibly higher order as well?)
KW: careful with ARMA behaviour as we may be getting too close to how the data are simulated whereby weakness in the simulation would lead to artefacts in results testing ARMA differences?
3. Consistency across error worlds - we want to know if hypothetically method A does well in cases X, Y, Z but falls down against test cases P, Q, R whereas method B does okay across all tells us presumably something useful?
VV: Will we define the absolute level by stating that the last homogeneous subperiod (HSP) is the right level? If yes, what do we do in case of an inhomogeneity in the last 5 years? In case of a break in the last 5 years, I would suggest using the second last HSP.
KW: Given the need for intercomparability and the problems with HOME where not everyone used the same reference period it is probably worth specifying this explicitly - that the most recent period is the reference period?
KW: second last HSP will differ depending on the algorithm so this may not be ideal for intercomparison
VV: I would take the real second last HSP which we know from the truth.
KW: Why second last? Seems a little arbitrary.
VV: It is close to the last HSP. :-)
PT: If the last break is too close to end-of-series there is no real hope of anyone finding it. Say the last break was in May 2103 we wouldn't expect to find in June 2013 the break. Then it makes sense to look at the last reasonably long HSP.
VV: That is a problem. Another option would be not to evaluate the series where there is a break in the last 5 years.
KW: But then we're not doing our job very completely.
VV: Yes, that would bias the results and make the job more easy.
PT: Some folks may only put out anomaly products and not explicitly calculate climatologies / absolute values. Most will provide sufficient metadata to do so, but do we risk penalizing them by doing so ourselves when they did not design the algorithm to be able to look outside anomaly space?
KW: Are we confusing reference period meaning? I mean the period with which all other inhomogeneous chunks end up being adjusted to. Not a climatology period - that is another good point though, different people will use different climatology periods but we can renormalise anomalies I suppose.

VV: The reference period here would be last HSP (or second last), not related in any way to climatological 30a periods.
PT: The fundamental issue is whether we evaluate efficacy of the algorithms in absolute space or anomaly space or both. If someone produces solely anomalies what do you do if some portion of the assessment is considering absolute values?
VV: Homogenization algorithm aim to improve the temporal consistency of the climate data. The absolute level is not defined.
PT: Exactly, so we should assess in anomaly space primarily?
VV: I would say so and then we should thus define how the anomaly will be computed in advance. In HOME we subtracted the mean of all series prior to comparison, but that leads to artificially small "errors" in the middle to the time series. It is also interesting to compute the temporal behaviour of the errors and simply subtracting the means does not do that.
MM: Some data-creators may only provide a gridded product?
KW: Who are our likely users:
Berkeley - only likely to output a gridded product
GHCNM - station and gridded
NIST?
CRUTEM?
GISS?
Possibly: MASH and ACMANT (although these are algorithms rather than groups - would be really good to encourage a submission of global scale using these methods though)
PT: Post-processing choices will also effect comparisons
VV: Should specify how we will deal with issues of not homogenising all stations or all parts of stations or providing grids rather than stations
KW: Can we park this problem until later in the day - easier to understand once we have established who is coming to play? We can then figure out how much post-processing needs to be done. Ideally, as little as possible. For now, lets concentrate on which 3-5 metrics we wish to compare between the 'true' and homogenised data.
VV: Secular/Linear trends? Easier to deal with single values but these do not describe the data very well.
KW: propose keeping things simple for round 1 - linear trends are simple. Can make more complex for round 2?
MM: Could have different start dates - that might help.
KW: this may also be important as users will choose to use different time points.

SUMMARY OF DISCUSSION ON VALIDATION METRIC 1:
All agreed VM1 needs to be a part of validation. Concern over how to conduct fair comparisons given likely different submissions (station, sub-sets, gridded (with/without interpolation), different temporal periods etc.). Agree that a range of scales should be compared from station to global average where possible. Agree that trends should be compared - linear ok for round 1 but not perfect. No consensus on other metrics yet. Advise 3-5 things to validate.

Validation metric 2. Detection of whether an inhomogeneity has occurred at a particular time.
PT: This is really 2 issues - how many we detect and the number of false positive detections. The cost of these may differ. If false detections yield a normal distribution with zero mean and small sigma then we can 'discount' these and go more

aggressively after small breaks. But if false break assignments yield either non-zero mean and / or large sigma adjustments then we get into problems.
PT: Presumably we might want to stratify this in various ways:
1. Breakpoint size (how well does it get the big fish, the medium fish and the minnows)
2. Breakpoint proximity (how well we do when breaks are adjacent vs. distant)
3. Network density (does it do better where we have more stations)
4. Break type (slope vs. step vs. seasonal etc. if applied in that error world)
5. Proximity to series end / series break (latter for discontinuous series only obviously ...)

SUMMARY OF DISCUSSION FOR VALIDATION METRIC 2:
All agreed that VM2 needs to be part of validation - likely through some kind of contingency table (hit, false alarm, no event, miss) but no decision on whether successful locating of changepoints should be stratified by changepoint characterisitcs (size, frequency, proximity to endpoints, type, network density). These could be used to apply weightings to the contingency table.

Validation Metric 3. How well has the correct nature of an inhomogeneity (e.g. size, duration) been identified?
PT: Fundamentally most algorithms are a two step process. 1. Identify breaks, 2. apply adjustments. We need to be sure that we capture which of these aspects a given algorithm does well / badly at.
EA: Important to seperate size and time/position
IJ: part three is quite complex, would there be the information needed to assess this? Would data-creators provide this?
VV: which worlds are more realistic? Should this be taken into account during assessment?
KW: Can we weight skill scores by how difficult we view the error world in the first place?
IJ: how easy is it to be objectively sure which error-world is easier/harder or more/less realistic?
MM: May be in a better place to rank 'reality' in error worlds after the event
KW: Do we want to combine results from 1, 2, and 3 into a single skill score? Probably not - different users will be interested in different aspects seperately. Those interested in long-term trends/robustness of data more interested in no. 1. Algorithm creators more interested in 2 and 3.
PT: 1 does need to cover a range of metrics though - mean and variance
IJ: What level to publish? Basics, everything? In between?

SUMMARY OF DISCUSSION FOR VALIDATION METRIC 3:
Not fully discussed - important to separate time and size of changepoint but concern over how this information is obtained. Not all players will provide homogenisation diagnostics beyond the homogenised data. We can request specific diagnostics but having too many requests will put some off. We will ultimately have to do the best with what we have, recognising that not all players will give us everything we would like. Recognition that the 4 VMs will be of interest to different parties. VM1 more of interest to data-product creators and users in terms of uncertainty remaining in climate data. VMs2 and 3 of more interest to homogenisation algorithm developers in terms of algorithm development.

Validation Metric 4. Realism of the benchmark, of the inserted inhomogeneities - a ranking of complexity?

SUMMARY OF DISCUSSION FOR VALIDATION METRIC 4:
See comments for Metric 3 - VM4 should inform all users by scaling performance against difficulty of the task in hand - not entirely sure this is a validation metric though - rather something with which to scale validation metrics by to make on overall comparison?

VV: 3 fundamental validation problems
A. Some people may not homogenize all stations. If a station is too inhomogeneous, it is best to remove it from the dataset. As the ISTI not only has a benchmark, but also a real dataset, where such a removal is important, we should also count on people not homogenizing some stations and make our validation fair to such differences in the number of stations homogenized.
PT: Also some people will only analyze sub-regions of the globe and we want to be able to assess them and compare favorably at least to the global algorithms and maybe even (somehow!) to algorithms that are run on completely different subsets of the databank holdings.

B. The variability of the time series is different for the various climate regions. Do we want to normalise for that?
KW: I think this will inform how uncertainties due to inhomogeneity are different in different regions as a result of both station density/quality and natural variability.
VV: I expect that we want to do both. No normalization for absolute values that a climatologist can use and normalized values, which are more informative for how much of the errors could be removed by homogenization. Same for fundamental problem C.

C. The size of the bias and the random component of the inhomogeneities will different as well. Do we want to normalise for that?
VV: Practical problem: How to quantify the reproduction of a secular nonlinear trend?
PT: Another practical probem: If someone comes along having produced an ensemble product solution (something e.g. that NCDC is likely to do) what are we going to assess, the individual runs, the ensemble mean, their best guess product or something else?
KW: Run away? Ideally we'd assess the median/best guess and the spread - that's going to be computationally fun.
PT: Yes, you could run percentiles assessed in some way. Ask the producer to rank by say GMST trend estimates arising and solely send a set of runs corresponding to say 10 percentiles.
VV: Assessing the entire ensemble including its spread is optimal. However, if there is only one contribution doing so, we still need a method to compare this method with the others.
VV: Yet another practical problem: Will we also accept contributions, which only make an estimate of the mean global temperature signal, but do not produce values for every station?

4) Workshop update

Attendees:
Zeke Hausfather
Robert Lund
Colin Gallagher
Kate Willett
Enric Aguilar
Peter Thorne
Claude Williams
Matt Menne
Jared Rennie

Monday morning: various talks on ISTI, databank, benchmarking activities that have happened / are in action.

Monday afternoon: Team Creation

Tuesday early morning: Webex telecon with working group members who can join to update and discuss - start Team Corruption (Renate, Stefan and Victor can join)

Tuesday: Team Corruption

Wednesday early morning: Webex telecon to update and discuss - start Team Validation

Wednesday: Team Corruption/Team Validation and loose ends

5) AOB
KW: I updated the website - look - http://www.surfacetemperatures.org/benchmarking-and-assessment-working-group We now have a 'what is benchmarking' section and I have set up a 'working' page with headers for each working group. Let me know if you would like to add anything to this at any time.
VV: Does that new section link to my post on benchmarking? ;-) http://variable-variability.blogspot.com/2012/01/what-distinguishes-benchmark.html
ACTION: KW to add link to VV's post on benchmarking

6) Next call
Friday 14th 4pm (BST) 11am (EST) - Team Corruption
ACTIONS:
KW: Circulate minutes
KW: Circulate call details for next week
KW: to see about guest speakers at next call - Lucie (Canada), Manola (Spain?), Andrea (Meteomet project update)?
IJ: tidy up discussion on Team Validation
MM: Circulate meeting agenda when its ready
VV: We should probably fix the date for the second next call soon. At least I would only be available Monday to Wednesday on that week, so it would be very shortly after the next call.
KW: Next call (during week of 17th-21st June) will be about Team Validation.
ACTION:

KW to sort out the next Team Validation call

NOTES: