Benchmarking Working Group Call #8
Wednesday 29th May 4pm BST (GMT+1), 3pm GMT,

Attending: Kate Willett (KW), Victor Venema (VV), Ian Jolliffe (IJ), Peter Thorne (PT), Rachel Warren (RW), Renate Auchmann (RA)

Not attending: Robert Lund

ACTIONS FROM PREVIOUS MEETING:

ACTION: RL Needs 100 stations - KW to send MCDW data to Robert. DONE
ACTION: KW to set up googledocs/okfn? wiki? blogpost? DONE
PT: Should we provide an update that summarizes the email traffic that did not pass through the blog and again mirror across benchmarking and main blogs?

ACTIONS FROM THIS MEETING:

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

NEXT MEETING:
Friday June 7th 4pm BST/GMT+1, 11am Eastern USA, 5pm Euro time.

AGENDA:

1) NCDC Workshop: Matt Menne
PT: Basically we have some limited funding to facilitate a workshop to advance the benchmarks. That money is not yet all spent so some additional participants could request support. It'll be held over July 1st to 3rd in Asheville. At the present time the agenda and aims have yet to be fully fleshed out and we should aim over the coming week or two to clarify aims, objectives and outcomes. Working with NCDC staff.
Attending: PT, CW, KW, RL, SB is a possibility

ACTION SB to check availability for attending.
IJ/VV unable to attend.

2) Team Creation update: KW
KW: Robert has been working with the 100+ MCDW data stations. Progress ongoing.

3) Team Corruption update: VV and others

3i) Response to call for regional inhomogeneity info: KW/VV/PT
KW: Excellent response mostly via email to Kate/Victor.
ACTION: ALL to pass around blogpost to anyone who may be able to help.
ACTION: KW list everyone who has responded and email round list and blogpost link.
ACTION: RW and KW to do the summarising and post to benchmark/ISTI blogs. Should be forwarded to homogenisation email list.

3ii) Victors Worlds by Question
Main Q - issue of biases
Q - seasonal cycle due to insolation and process errors
KW seasonal cycle could be positive/negative in different seasons
SB Processing errors can cause opposing errors in different seasons for example negative numbers incorrectly recorded. These should be included in our error models.
Q - platform inhomogeneity
Q - length of data
Q - data sparsity
VV: Default to mirror characteristics of real ISTI data including missing data, station length, density, correlation structure.
PT: benefit of using identical underlying error structure on top of different background worlds - i.e. background climate trend, different methods in modelling errors (physics based vs pure statistics based)

Question: Allows us to study the influence of station density without being convoluted with changes in the properties of the inhomogeneities. What is uncertainty after homogenization?

Question: Study whether we could detect changepoints if the bias in the low-latitudes were large. What is uncertainty after homogenization?

Question: Study how important the presence of spatially correlated breaks is. What is uncertainty after homogenization?

Question: The seasonal cycle could vary more than in reality. Tests how well homogenization can handle variability in the homogeneity from year to year, and improve the seasonal cycle.
RA: Diurnal cycle not so large, depends on resolution. Subdaily data averaged to daily - smaller in the daily means, depended on the region - some cancelling out in the daily. Need for more research to gain understanding in this. Variability year to year? Appears to be quite stable.

[PT: In the US at least the NCDC algorithm says there tends to be one break /decade on average and even that may be an under-estimate for break propensity - maybe the US network is anomalous in this regard though?]. VV: This is just about the biased breaks, the will be additional unbiased breaks and gradual inhomogeneities. [PT: NCDC only find biased breaks]
VV: The ensemble mean of inserted unbiased breaks is zero, they do not affect the global temperature trend. The ensembl mean of inserted biased break is non-zero, these breaks have the ability to change the global trend. Detected breaks will typically be non-zero, althoug the annual mean break can be zero and the break was detected due to its annual cycle.
Question: Is homogenization more difficult if the bias is spread more evenly?

Question: Can homogenization even correct biases if the biased breaks are minute? This world is probably not realistic any more, but not in a way that can be exploited. For the uncertainty estimate it is better if the world is blind.

Question: What is uncertainty after homogenization? Is removing the trend bias more difficult if there are more an larger inhomogeneities?
[PT: The USHCN benchmarking in Williams et al. suggests otherwise - that breaks that are small SNR are the ones that are problematic. Large SNR breaks are easy to detect and deal with - where easy is relativistically defined!]
VV: Yes, that was a mistake, should be "more and smaller" in #B7 and "less and larger" in #B8.

Question: What is uncertainty after homogenization? Is removing the trend bias more difficult if there are less and smaller inhomogeneities?
KW: Could this be 'more' instead of fewer as we have established that we probably underestimate the amount of small breaks that are present in the data.

Question: Does the shape of the seasonal cycle matter for the remaining error after homogenization?
VV: may not be a sine curve - may just be a few months that are affected.
KW: #B9 is kind of addressed nicely in #B4 I think – useful to cluster worlds that explore related questions.
VV: Could be.
Related to #B4, #O3

~~#B10. Best guess world, but with more short platform type inhomogeneity pairs.~~
~~Generate the dataset by inserting 10 additional platform pair break inhomogeneities (20 breaks) per century. They are perfect platforms, i.e. go up and down the same amount. The length is determined by a uniforms distribution to the power 3 and multiplied to get values between 0 and 5 years. The size (standard deviation of a Gaussian distribution) depends linearly on the length of the platform. A platform with 5 year length has the same size as the unbiased breaks; a platform with 0 year length is twice as large. Not sure~~

~~this is as useful as say missing data vs non-missing, nature of seasonal cycle, biased+random, gradual, with/without background trend issues.~~
~~Question: Some people think that real datasets have more short inhomogeneity pairs that look like a platform. This world would allow us to investigate whether this would affect the error after homogenization and thus whether their existence should be studied in more detail. What is uncertainty after homogenization?~~
KW: We have decided not to include this as a blind world – this issue can be effectively explored using smaller networks, possibly with code made available by us alongside the open worlds

Question: Lindau and Venema (2013) showed that some multiple breakpoint methods work well for time series of about 100 years, but may be over or under predicting for longer and shorter datasets. Such a short dataset would investigate this for the participating homogenization algorithms.
[PT: Yes, endpoint effects almost certainly matter. But why use the last 40 years when network density dominates? Its the early period when the density is poor that this is more likely to be problematic? I'm assuming most methods use some kind of neighbour based expectation framework?]
VV: Good idea, that would also be interesting and a more challenging period.
KW: #B11 could be combined with a non-missing vs missing which is also a very important factor in terms of algorithm skill. Non-missing could be uniform lengths of 100 years. Missing should reflect real network so will contain many shorter stations. May want shorter non-missing and shorter missing.

Question: Investigate importance of data rescue. Add one or two additional stations (if possible (not on islands)) to stations that have less than 2 neighbours within a certain distance (500 km?).
KW: I think #B12 is already addressed with the Best guess for the West everywhere in #B1.
KW: #B12 difficult to do for Team Creation in the first place as all stations will be modelled on actual stations in ISTI databank so creating 'new' stations that do not exist in reality is a little more tricky.
VV: Good point, let'sskip this one, we can also probably estimate the effect in the returned data by analysing the results as a function of station density.

OPEN error worlds for homogenization
#O1. No inhomogeneities.
Question: How much does homogenization make the data worse, if there are no inhomogeneities?

#O2. Best guess world for the West, but the random component of the inhomogeneities is not given by a normal distribution, but either -1 or +1 degree. No gradual inhomogeneities.
Question: Sanity test, if you cannot solve this one, you are in serious trouble.

#O3. Best guess world for the West, but no seasonal cycle.
Question: Interesting in comparison with standard seasonal cycle and difficult seasonal cycle (#B9) to study importance of seasonal cycle. (It is impossible

to do this one blind, at least someone analysing the results would notice any way and could change settings accordingly.)

#O4. Shall we also have one realization of the best guess world in the open for testing the algorithms?
Question: Is there a difference between blind and open contribution. It would provide a realistic world for playing and testing the algorithms before they are applied to the blind section. On the other hand, a disadvantage would be that people could tune their algorithms to any less realistic aspect of the best guess world and thus perform better on the benchmark as on real data.

KW: #B3 easiest world? Good to have some idea of hierachy.
VV: Probably, could also be #B4 (Auchmann) or #B8 (fewer breaks)

KW: Not so sure about #B5, #B6 or #B10
VV: #B5 is quite important, we need to know whether it makes a difference whether the inserted bias is caused by 2 breaks with a large bias or 4 with a smaller one. If this makes a difference, and I think that may well be, then we would need to study the biases in more detail.
I also doubt about #B6 as it is rather extreme and unrealistic. On the other hand, if the results are good for such a world, that would increase our confidence a lot.
World #B10 is included to be inclusive. And we said we wanted to be sure that no one would find later that real inhomogeneities are more difficult as a world in the benchmark. Detection platform inhomogeneity pairs is difficult and whether they are detected also influences the detection of other IH with are more important for the homogeneity of the data.
KW: #B10 could be addressed with a smaller network - if we release error model code then others can reproduce error models of their choice.

KW: Open worlds look good.

KW/IJ - testing different background trends, underlying natural variability could be done with open worlds. Could also combine with physical. Perhaps use #O2? Then if we don't finds its a problem is this because its too easy in the first place?

THESE WORLD DESCRIPTIONS SHOULD BE CIRCULATED AROUND homogenisation email list once finalised.

PT: important to think about analysis, validation and write up - clear storylines will make this easier.

ACTION: VV to tidy document and circulate round particularly to Claude/Peter and others not on call - two weeks time.

3iii) Example Blind error worlds by cascading complexity:
Example set of 10 error worlds and what they will test.

TEST: false alarm rate, does homogenisation do damage?

TEST: Can homogenisation algorithms get the basics right? Are we detecting/adjusting for the largest inhomogeneities?
TEST: Can homogenisation algorithms get a moderately 'bad' world right? Are we detecting/adjusting the small/gradual inhomogeneities?
TEST: Can homogenisation algorithms cope in the 'real' world but with/without the complexity of a non-stationary climate?
TEST: The unknowns - if the world is worse than we thought do the algorithms cope?
Something like above starts to let us say something about the nature of different confounders of homogeneity algorithms and would allow us to at least nominally place where candidate algorithms 'fail'. Algorithms that rate 2 will be given less weight than those 7 where the rating is the world at which they begin to show seriously deleterious performance characteristics. If no algorithm is >3 then we at least know that and interpret the ensemble of algorithms applied to the real-world data accordingly?

4) Team Validation discussion: IJ
PT: Implications from global to local level. Quantify effect of missed and false alarms.
IJ: How close the homogenised series is to the 'truth'? Are the locations inhomogeneities detected? Is the character of the inhomogeneities correctly diagnosed?
PT: Undertake these measures on all scales.
KW: This should be the focus for the next meeting.

ACTION: IJ to recirculate Team Validation document in prep for next call (Fri 7th June)

5) Last Meeting Minutes sign off: OK

6) AOB:
PT: Generic ISTI call on Tuesday 4th June, 12Z - all welcome.

PT/VV: Potential new members: *Zeke Hausfather and Enric Aguilar?*

ACTION: KW to add RA to the email list.
ACTION: KW invite new members and double check with silent members whether they still wish to be bothered by emails.

Notes