ISTI Benchmark Working Group Day #1
1st July 2013

Attending: Kate Willett (KW), Peter Thorne (PT), Claude Williams (CW), Matt Menne (MM), Robert Lund (RL), Colin Gallagher (CG), Zeke Hausfather (ZH), Enric Aguilar (EA), Jared Rennie (JR)

Online attendees: Victor Venema (VV), Renate Auchmann (RA), Ian Jolliffe (IJ), Robert Dunn (RD), Lisa Alexander (LA), David Parker (DP), Stefan Brönnimann

---

Agenda:

9.00 – 9.15 Welcome and opening remarks – NCDC Director, Tom Karl

Strong support for ISTI, hoping to move to daily

VV: I would also love to go to daily, but think that at least for the benchmarking of homogenization algorithms the next cycle would be much too early for global benchmarking. The current correction algorithms are deterministic, but should be stochastic. We have no idea about typical statistical properties of daily inhomogeneities, how the distribution is affected, how measurement errors cross-correlate between stations, how the inhomogeneities are related to the weather. Thus personally, I would suggest to extent the benchmarking from Tm to Tx and Tn. That may already be quite difficult as one would have to know how these breaks cross correlate, which is only known for some transitions from parallel measurements (for example in Parker (1994)).

Funding support - difficult to get Statistics funding in the USA for climate related things

Not so much in the UK

ACTION: Kate to push for research proposals on this?

VV: The advantage of daily may be that that would be a topic for which we may be able to get funding. And if we actually have people working dedicatedly on the project, it may be doable, but it would have to be quite a large project.

Comercial interest?

What about output? Need a concepts paper to lay out the plan and refer back to.

ACTION: KATE TO GET THIS OUT.

9.15 – 9.30 The International Surface Temperature Initiative basics – Peter Thorne

WMO, BIPM and ISI TIES endorsement

9.30 – 9:45 The ISTI monthly databank construction and characteristics – Jared Rennie

Databank first version open alongside code used to conduct merge to stage 3 product - includes GHCN-M and GHCN-D

Preference for stations with Tmax and Tmin over Tmean - Tmean then created from Tmax/Tmin rather than taking Tmean. Keeps things consistent where possible.

VV: Herman Mächel at the German Weather Service (DWD) is currently studying whether the Tm from Mannheimer Stunden is better or worse than the one from Tx and Tn.

Merge based on weighted metrics: distance in space, height difference, station name similarity using the Jaccard Index (cannot distinguish Kyoto and Tokyo though), start year of record if record <10 years. Create a final probability - >50% then move on to comparison of data within station records. Data metrics include overlap and index of agreement over that period with a look-up table for probability of station uniqueness/station match. Decide based on thresholds. Estimate 97% correct merge. Harder in areas of greater station density.

VV: A better merge algorithm could also be a topic for a research proposal. A more accurate method should probably compute the cross correlations between stations taking into account that data may be homogenized and thus homogenize the difference time series again before computing how well they fit.

KW: This may be taken into account in the 'index of agreement' - paper is in review. I will ask.
VV: I am on the paper: currently inhomogeneities in the difference time series are not taken into account for the index of agreement.
PT: All code etc. will be released (already is up on the ftp area with beta 4). So, anyone can come in and help produce a new release with an improved method. If we evaluate it and like it we'll increment the databank version to version 2. It's not an easy problem set though so you would need to be brave :-)
VV: Yes, it would be a lot of work, that is why I wrote that it is a research project. Not something to do on the side.
Now on Beta 4 - as of June 2013
Have a BLACKLIST of 'bad' stations - bad metadata?
Stage 3 data are in netCDF files now as well as ascii (ISTI and GHCN format).

9.45 – 10.00 Benchmarking overview status and plans – Kate Willett
Kate outlined how three teams will be key to the success of benchmarking. Creation, corruption and validation.
Creation: Suggested these be built from low-frequency / spatial detail of a climate model and then add realistic high frequency cross-correlated data structures. Can explore different scenarios by looking at different climate models / run types. Showed that it is possible to create plausible data structures. We need a single person create these so they are blind?
Corruption: Talked about the concept of open and closed worlds. Discussed how the breaks may be inferred and applied.
VV: Slide 2: The Youtube video of IEDRO on data rescue was seen 144 times in 2 years. We should do more PR and not leave that to the non-traditional participants of the climate debate.
VV: Slide 3: Assessment: To solve the problem that some (or likely all) homogenized contributions will not homogenize part of the data (not only stations in data sparse regions, but likely also some data sparse periods), we could ask everyone to return all data homogenized as well as possible, but indicate which data one would have removed in the real dataset. We could then use these stations in the station based validation for a fair comparison. In the computation of the regional, continental or global means, we could leave such bad stations out. KW: good point but leave this until tomorrow.
VV: Yes more for day 3, but was in your slides.

10.20 – 10.40 Benchmarking the USHCN – Matt Menne / Peter Thorne / Claude Williams
GCM plus noise to approximate station series - did consider cross-correlations, added a variety of error-models to that.
World 8 - control GCM (no climate trend), perfect data (clean)
World 5 - big breaks, perfect metadata - easy
World 7 - mixed break sizes, some clustering
World 1-4 - clustering and sign bias family
World 6 - very many small breaks - very hard
Ran the PHA 100 times exploring different tweaks of the 17 tunable parameters of the PHA
Big breaks, no sign bias is easy, small frequent breaks and sign bias is hard. PHA always moved data closer to 'truth'. In real data scenario trend for Tmax is increased (homogenised vs raw), for Tmin it is reduced.
Subsequent factor analysis has led to resolving the optimal settings for the PHA and a revised version of 100 restricted settings members has yielded a better estimator of 'truth.
Now have an R coded Bayes approach to PHA on USHCN - Zhang, Zheng and Menne, 2012, Journal of Climate. vol. 25
Berkeley approach also tested on the CONUS benchmarks (same as the Williams et al. 2012 paper)

Start teleconference/webex:

11:00 – 11:10 Morning round up – Kate Willett

Big Qs for Team Creation:

1) Is our method going to work over data sparse regions and data dense regions?

2) How good is good enough?

VV: Variance should be right within a few percent, cross-correlations with nearest (and 10 nearest) within one percent (on average). Reasonable (small) decadal variability in difference time series.
VV: Added later: in case absolute homogenization algorithms will also participate, we should also have realistic variability for the global (and large scale) mean signal.
PT: Being careful about level of GCM low frequency input/loess smoothing - this could add something unrealistic
KW: Agree - we should test statistics of the difference series between neighbours.
KW: worried about incorporating what look like inhomogeneities in the 'clean' data through our synthetic creation methods.
PT: Will always find breaks that are not real - false positive. But different errors have different costs to different end users. Problem specific
IJ: Is it the size rather than the nature of the errors that is most important?
KW: Good Q Ian - but I'll copy that one to tomorrow for Team Corruption/Validation problems as that is what we'll talk about tomorrow.
ACTION: shift to tomorrow
PT: discussed with experts which GCMs are better to use - better realisations of variability.
ZH: Problems with reanalyses as they have station data in them?
PT: Yes and also they are of limited period.
KW: And they are not homogeneous if they change their satellite inputs and other data streams.
DP: Could use 20CR as this is homogeneous(ish)
SB: 20CR does contain severe inhomogeneities in the Arctic in the early period.

3) Time period needed?

SB: Long time period preferable
VV: Make as long as possible - important to include the early period of record where homogenisation is a major problem due to data sparsity. Another good reason to use GCMs is their provision of simultaneous wind/solar/radiative information which may be a good way to add inhomogeneities.
EA: Long is better but station density is a problem for simulation? PT: Planning to mimic ISTI databank entirely - so back to 1850ish if we want. EA: Probably not usefully homogenising anything prior to 1880.
KW: As long as GCMs go back then fine - PT: 20th C runs start in 1850 so ok.
RL: More data = good
KW: Team Validation will need to cope with these different choices of length of record. Really want to compare apples with apples so a 1880-2010 comparison with 1970-2010 isn't going to work. Need to have a common period of record for the overlap but provide extra assessment for those -
ACTION: Shift to Wednesday session
CW: Further back the better - can we simulate more data to test the uncertainty further back in time?
KW: As long as station existed at some point we can simulate it back to whenever - and Open World with complete data back to 1880.
VV: Agree going before 1880 would be good. Enric is right that results may not be good with just statistical homogenization, but that is also something we would like to know quantitatively.

4) How do we ensure smooth cross-correlations globally across matrix boundaries?
Gibb sampling?

5) How do we compute this efficiently?
R package invoking fortran – speedy but complex and less transparent/tweakable
R code with parallelisation and bigmatrix – transparent/tweakable but complex, computer intensive
R code without parallelisation and bigmatrix – transparent/tweakable but very computer intensive

LA: consider using Python
R or Python
RD: I think that there are interfaces between R and Python, so you could use any existing R code from within Python - RPy - though Python might have most of the routines required by now

6) Who is going to set up the webpage framework/databank storage framework?
JR: very doable
PT: you need to provide some blurb
VV: Also need a framework for results upload with format checking.

7) Who is going to be the handle-turner?
Probably Kate?

11:10 – 11:30 Team Creation progress – Robert Lund
See four files sent around
140 MCDW stations for the USA
DATA.eps
Infilled missing data with climatology
Proof of concept with 4 but have run ~50 successfully
Anomaly.eps - standardised anomlies - Guassian distribution describes the data well.
DP: What about the Pinatubo effect? Are similar features in the synthetic data?
CW: Seems like the variance is larger in the synthetic data
SB/DP: May need longer serial correlation as some of the persistence isn't replicated.
RESULAR1.TXT
This was done with Vector autoregression for lag 1 (VAR(1))
Station cross-correlations are very similar between real and synthetic
Station autocorrelations at lag 1 - very similar for each station, cross-autocorrelations a little out - not so much to be a shocker though.
But we're missing the persistence
RESULAR2.TXT - same thing but for order 2 VAR(2)
Numbers suggest that there is persistence at least out to order 2 (if not further)
DP: Spotted an error in RESULAR2.TXT - actually printed out real data instead of simulated.
Oops.
Still demonstrates that we probably need more than order 1 if we're going to model full standardised anomalies.
PT: Can we get some of that for free by getting some of that from a GCM? So an VAR(1) might be ok for method using a GCM?
RL: found that VAR(8) is optimal
CG: Are you sure the standardised anomalies are stationary as some of these orders will be trying to account for that?
MM: What if your real data have breaks?
RL: Will cause problems and mess with the VAR a little.
VV: Probably inhomogeneities more of a problem out to higher orders, probably not for order 1
IJ: Can you test how much better is the second model doing than the first with a likelihood ratio test. So a good method for testing different methods anyway?
PT: Let GCM do most of the work and VAR(1 to 2) to deal with the high frequency.

End teleconference/webex

13.30 – 13:40 Call progress summary – Kate Willett
Kate showed some slides from her IMSC talk. Take a gridbox, fit a smoothed function with loess filter.
PT: Why smooth? Surely you want to retain all the physics? Then add white noise with some spatio-temporal autocorrelation function?
KW: So use the model gridbox as is and simulate the station deviations from that areal-average? then simulate the pink noise of the stations on top of that. Still an issue of how wide is this correlation decay distance in the pink noise. 100s km for anomalies, hopefully a lot less for a difference series?
KW: NEED TO KNOW:
How realistic are GCM spatial correlations? Compare GCM gridbox with CRUTEM4 or GHCNM grids?
What is the correlation decay distance of the 'pink' noise/microclimate that we wish to put on top of a gridbox mean?

Our options so far:
1) 100% statistics using VAR(?) - reasonable job at station characteristics but questionable how well this will do globally over 30000 stations.
VV. Covariates such as insolation, wind, snow cover would be missing. That is information we only have in the model.

2) 50/50 GCM/statistics using some low frequency filter from the GCM and simulating up to VAR(2) for the stations - still questionable how globally realistic this will do.
VV: idem.

3) 90/10 GCM/statistics using GCM gridbox as is plus VAR(1) simulation of pink (with some spatio-temporal correlation structure) noise station characteristics on top. Would involve removing a regional average from the stations and simulating what is left. Does this not need to be done at the same scale as the GCM gridbox that we will add in? Big problem for sparse grids as subtracting gridbox average will end in zero difference.
PT: Essentially focussing on recreating the characteristics of station-reference difference series because this needs to be realistic to make it a fair test on the homogenisation algorithms.
So the idea is that you can take a regional average - what size and how is that affected by adding back many different gridbox regional averages? Are you introducing more differences then? Need to look at this with weighted regional averages and on gridbox by gridbox for comparison. At this high frequency noise correlation decay distances will be far smaller than for the anomalies. May only be robust on a small gridbox scale. Another option is to use the 1by1 degree gridboxes from something like the Berkeley estimates? May be a problem choosing one existing data-product in this way? Worth considering this approach though.
Assuming these 'difference'/pink noise series are stationary.

Station length - should we have at least one world back to 1750? May as well for an open world? Blind worlds 1880 onwards. There are 7 GCMs that start in 1750.

Missing data - should there be worlds with spatially complete data? May as well for an open world?

Are we going to allow changes where there are missing data? Not really fair - should be at beginning or end of missing data period - not in the middle.

ISTI databank - all stage 3 data are at least 2 years. Can we simulate stations that are this short

inferred from statistics from other long nearby stations? How much missing data can we cope with? There may well be stations that we cannot homogenise?
To cope with the fact that we may not be able to simulate all stations we could have an open world where some simulated stations are artificially shortened/missing data added to test how well these are dealt with.

Coping with coding: hoping to pull on advice and help from Stephen Moscher (contact from ZH), Finn Lindgren, Doug McNeall, Richard Chandler...)

SUMMARY:
High level support for this work but funding is difficult - not seen as innovative enough for statistical funding bodies
The properties of the station-neighbour/reference is key. If station and difference series serial correlation and station cross-correlations are similar then we're ok. Smoothing of the model means that we're unlikely to get realistic global spatial patterns of modes of variability though.
We should produce stations for as long a time period as we can - most GCMs start in 1850 and some go back to 1750. Provides more options for Team Corruption.
R or Python are still software of choice - Kate will be handle turner so no-one else knows the 'truth' of the clean worlds.
Three solutions posed for Team Creation - converging on 90/10% GCM/statistics where GCM provides most of the climate signal and persistance and a VAR(?) models the topography/instrument driven microclimate of the stations on top. Actual station climatologies and variance will be use
Test different models using likelihood ratio test to decide which fits the data best.