ISTI Benchmark Working Group Day #2
2nd 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), Robert Dunn (RD), Renate Auchmann (RA), Stefan Brönnimann (SB), Renate Auchmann (RA)

Agenda:
Start teleconference/webex:
9:00 – 9:10 – Kate Willett summary of day 1,
Big Qs for Team Corruption:
1. How blind is blind enough?
VV: From our discussions in Korea: Blindness – should let everyone know everything the team knows i.e. shape of distributions applied, type of breaks, that background worlds will be changed. The members of the BAWG should not have an advantage over other homogenizers of the benchmark and feedback from the homogenization community on our dataset could be useful. But choice of GCM, the clean world specifics, generation of exact size/number/type of breaks for each station will be done randomly within the code and so only person turning the handle (Kate) will likely have any prior knowledge of this. Not discussed, but there is another aspect of blindness. If possible, the person doing any non-automated part of the process, such as writing up comparisons between results for different algorithms should ideally be blind to which results are from which algorithm.
PT: How timely is assessment going to be? Too quick and its difficult to make the assessment blind, too slow and there isn't so much reward for their work.
VV: 3 years seems to be a long time. Can give immediate feedback for open worlds.
KW: Conclude that ideally assessment would be blind that depends on how quickly we can recycle.

2. How many different background 'clean worlds'?
VV: A problem with having multiple worlds with the same inhomogeneities or similar statistical properties of the inhomogeneities is that this could theoretically be exploited to get better results. Thus maybe we should study the influence of the background GCM in the open datasets. Such worlds could be similar to the default world with respect to their error models, but should use slightly different settings.
KW: Users wouldn't know which world is which so I'm not sure that this is a problem.
KW: Formal assessment from blind worlds only so if this is removed from blind worlds it won't form part of the assessment.
VV: For the worlds we would like to compare with each other, it may be good to have the same background GCM. #B1 to #B4 could have one, and #B5 to #B10 could have one background GCM. #B7 will have a third background GCM.
PT: Important that even if we use the same GCM the overlying station pink noise structure is different.
ALL AGREED
VV: Can play with more GCMs in the open worlds

3. How on earth do we build these things?
VV: For the statistical worlds: draw from normal distribution of the mean, draw from the distribution of the size of the inhomogeneity, generate a fitting shape with these random numbers (something sinusiodal for the defaults worlds, something more limited to one or a few months for

the difficult seasonal cycle world).
VV: For physical inhomogeneities the inhomogeneity follows from the covariates. We should check, however, whether the results look realistic.
VV: Maybe it would be a good idea to ask the homogenization list for help on the seasonal cycle of inhomogeneities. Especially for low and high latitude regions.
PT: Very important to avoid locating a changepoint in the middle of a missing data block - start or end is ok.
RL: Can you apply the breaks to temporally complete data and then put in the missing data?
CW: Causes problems for metadata
PT: Better to apply missing data first and fix code to ensure no changepoint where data are missing.
RL: How are we distributing changepoints through time?
MM: Did we agree to apply seasonal cycle to a percentage?
VV: Yes - something like only 50% of changepoints to have a seasonal cycle
CW: Not convinced its so prolific.
ACTION: CW/LV conclude findings on seasonal cycle in changepoints - we can then put a figure on this.
VV: Homogenised results probably underestimate the size of the seasonal cycle - too much noise and generally in the HOME benchmarks the actual seasonal cycle applied was mostly underestimated. 50% value came from a study in Romania.
CW: Looking at each of the raw segments for an offset in the seasonal cycle compared with its neighbours.
MM: Don't apply a perfect sinusoidal seasonal cycle to the changepoint - will depend on background climate.
PT: Even for 'flat' changepoints we should add some white? noise because this will also depend a little on physical climate - rarely a pure step.
EA: What drives these? Difference in the stations?
PT: Change in the variance as well as the mean
CG: Add constant shift plus change in the variance.
KW: This change in the variance is likely to be seasonal - warm/cool, cloudy/sunny, wet/dry months will drive it.
VV: And then its likely to be cross-correlated
ALL: Sigh!
PT: Some metrological (not typo - honest! measurement science) (instrumental) component in this too.
EA: Some types of changes will have more affect on the variance than others.
KW: Can we use a simulated station nearby for a station move? Might be tricky as we don't want to duplicate stations so perhaps we do need to be able to simulate stations that have never existed.
ZH: Can we study this from documented changes?
EA: Station moves can be anything - so don't need to characterise explicitly.
VV: Too much work to really do this perfectly - for volunteers.
KW: Are their changes to a station that would not change the variance?
- calibration
CW: Difference between changes in variance that are noticeable and those that are not detectable - within the noise.
PT: Suspect there are bigger changes in very poor and very rich (tinkering) countries.
ACTION VV to ask Homogenisation list for experience on seasonal cycle.
EA: Tropics - are the changes bigger? Seasonal cycle likely to be lower because annual seasonal cycle is smaller.
KW: Should then relate to the climatology/variance of the actual station.
VV: Something due to wetting/wind?
MM: Vegetation cover and soil moisture?

KW/EA: Can we ask Blair Trewin what he sees in Northern Australia? Southern Florida but not looked at this.
ACTION: Kate talk to Blair
KW: Can we apply changepoints with a specific character: shelter change, station move (to airports, to rural), time of observation, urbanisation). Can we have percentage of each type?
SB: This is related to network wide changes - how are these going to fit in the blind worlds? Are all worlds using the real regional statistics on this?
VV: Can link this to biased (non-zero distribution of changepoint size) and non-biased (distribution with mean zero) changepoints as some causes of changepoints we know to be in a certain direction always.
SB: Temporal clustering of network wide changepoints - do we want to reflect this?
VV: Yes the collection of regional inhomogeneities will be applied such that the blind worlds are 'best guess' for the world to some extent.
PT: Very likely more changepoints than we think - certainly more than detected.
EA: Many changes were global and some are regional - should apply these as we have found them.
VV: 2 per century a reasonable estimate and one world will have 4 per century.

4. How many different things to assess in one world?

5. What do we ask people to homogenise as a minimum?
VV: Return all data (fully automatic methods), return global mean signal (if you can) or return small networks of selected long stations in Giorgi regions (partially automatic methods). In addition return a file specifying the data of breaks, outliers, begin and end of gradual inhomogeneities. In addition return a file specifying which periods and stations you would not have homogenized in case of a real dataset that would go to users.

Technical questions for Team corruption
We currently have plans for 9 worlds. Do we have ideas for a 10th, or shall we limit ourselves to 9 blind worlds.
What should be do/discuss today?
A - Fix parameters (table of previous telecon on corruption) for default world
B - Discuss parameterization (distributions to use)
C - What kind of inhomogeneities do we want to put in the experimental world (#B8)
D - How to implement the seasonal cycle?
E - We may still have one undefined world (#B7 - currently B5 with different GCM). We could also limit ourselves to 9 blind worlds.
F - what should people be asked to homogenise.
G - How to apply gradual changes?
I - How to package up the error world data?
J - Instructions on how to come and play (will need input from Team Validation too)
End teleconference/webex

Talk from yesterday: A proposal for error models that challenge homogenization methods to their limits
Victor Venema, Matt menne, Zeke Hausfather and Claude Williams
USHCN benchmarks don't have trend/gradual inhomogeneities - so here are some new benchmarks that explore that
When are things so inhomogeneous that you can not homogenise them
'When good homogenisation algorithms go bad'
Station sparsity
How messy are the data - changepoint frequency and with/without trend biases
10 steps of increasing difficulty but also increase the number of worlds with increasing density - up

to 600 worlds.
Nice figures of stations (y axis) vs time with location and size of trends shown by colour
Nice figures of homogenised perturbation (y axis) vs time (x axis) - gradual trends are not purely flat - a little jumpy.
Not yet built so potential to use the ISTI benchmark clean data.
PT: So are we decided on only focussing on Tmean rather than Tmax, Tmin and Tmean?
KW: Not sure we can cope with more than Tmean right now.
RL: Should be easy to do
EA: Apply same changepoints to all?
EA: Important to have Tmax and Tmin for the daily.
ZH: Are there changepoints that affect Tmin and Tmax oppositely such that there is no effect on Tmean?
PT: Yes.
ZH: If an algorithm is good at Tmean then can we expect it to perform as well for Tmax and Tmin anyway?
MM: Doing Tmax and Tmin would be a much bigger problem for Team Corruption.
RL: Just a dimensionality problem so don't see a problem with it.
PT: PHA work show that more changepoints are found in DTR over Tmean despite DTR having greater standard deviation - so not a signal to noise ration thing but the fact that Tmax and Tmin are compensating and so fewer shifts are detected in Tmean.
VV: I would suggest to extent the benchmarking in the second cycle 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)). He shows that average sizes of breaks in Tx and Tn is likely larger. Their sizes are anti-correlated, which is the reason why the break in the Tm is typically smaller. How these relations are for other types of breaks, I do not know.

Team Corruption talks and discussion (including Victor Venema and other interested parties on telecom) - finalize error model proposals (e.g., seasonal cycle) for base worlds and scenarios for special runs
What should be do/discuss today?
A - Fix parameters (table of previous telecon on corruption) for default world
I think all worlds will use the regionally known dates/types of inhomogeneity as far as possible - intended to be 'realistic' and 'best guess'
EA: The poorer quality of data in the Tropics may be confounding detection of inhomogeneities in the tropics - suspect they are not necessarily larger than elsewhere.
MM: We're really just guessing
KW: Why do we think changes in the Tropics should be larger?
VV: Stronger insolation which influences Tx and less problems with Tn at night due to higher humidity. As in the mid-latitudes the breaks in Tx and Tn are anti-correlated, the break in Tm is not that large. In the tropics Tn may not compensate for Tx. Unfortunately this is just a hypothesis. I have not seen any data on this.
PT: Larger humidity and lower winds could drive this.
RL: What about the poles?
PT: Suspect these would be larger too.
VV: Could go both ways. Less radiation problems, but maybe ventilation could be a problem with ice and snow on the instruments.
KW: Do we need to add issues of quality - random errors/outliers
PT: Should be adding some 5 sigma events
CW: But then we're testing the QC algorithm at the same time as the homogenisation which complicates the assessment.
MM: This was done in HOME and wasn't very successful.

VV: It did not influence the results significantly, at the same time it was not elegant as many algorithms assume that the data is QC-ed in advance.
KW: AGREED that clean worlds are created assuming that a good QC process has been conducted on the data.
BE EXPLICIT THAT THIS IS THE CASE

B - Discuss parameterization (distributions to use)
Assume values in the table are the mean of the distribution rather than 1 st dev of adjustments will be within this range.
CG: Peter's graphic of PHA on the ISTI databank doesn't imply a bias of 0.2 degrees
PT: That's because various biased types are aveaged out over the globe.
CG: So really need a percentage and distribution for each changepoint type.
CG: Would apply a Laplace (back-to-back exponentials?)
PT: Conflate a fat and low normal with a tall and skinny normal?
RL: That won't give you what you want
VV: Do we have any empirical evidence for Laplace? We do for Gaussian in one network.
KW: Which network?
VV. The USHCN has a Gaussian distribution if you look at all the breaks from known metadata. All other datasets with detected inhomogeneities, I have seen look Gaussian, except that the small breaks are missing as they are harder to detect. Thus I would expect that if you see a fatter tail, that is may be because you have mixed multiple normally distributions.
KW: Fits tend to not do well over the fatter tails. Ok so still working on a non-Gaussian overall though.
CG: Pareto would give fatter tails than Laplace if needed.
MM: Actually maybe we should go Gaussian - so missing the middle and everything other than the tails really.
All coming around to this idea now.
ACTION: need standard deviations on the table
CW: Can changepoints be added at any time over the period of record? - YES
PT: So no minimum homogeneous sub-period
KW: Are we allowing changepoints in the last two years?
PT: Changepoints that are drawn randomly could result in very close changepoints. How this is assessed comes down to Team Validation?
ACTION: move to Wednesday.
EA: HOME considered a HIT if changepoint found within 6 months.
KW: I think we should allow these close changepoints to occur - Team Validation cope with it with the two types of assessment - skill at preserving the 'nature' of the station and skill at detecting the location and character of the changepoint.

Types of changes:
Station moves to airports from cities
Station moves to airports from rural (e.g., Canada)
Station moves from airports to rural
Wild shelter to Stevenson screen
Cotton Region Shelter to Stevenson screen
North facing wall to stevenson screen
Stevenson screen to AWS small cyclindrical shelter
Observing time change from morning to evening
Observing time change from evening to morning
Daily Tmean algorithm change to increasing number of hours sampled during the day
Instrument
– Zero drift, shrinking glass initial years

– Calibration errors
– Response, integration time
– Temperature out of range
– Quicksilver thermometers: T < -39°C

Change surrounding
– Urbanization, growing vegetation, irrigation

Shelter type
– Radiation & wetting protection
– Natural or forced ventilation
– Snow cover
– Plastic screen: insolation on hot days (plastic may hang through)
Definitions
– Computation mean temperature
Maintenance procedures
– AWS: Icing, damage detection
– Painting & cleaning schedule
Digitisation & database
– Minus sign forgotten
– Station names mixed up
– Pre-homogenised data
KW: Some of the above are Quality issues which we would like to ignore for the purpose of these benchmarks - assume everyone has run a nice QC algorithm first to reduce assessment to power of homogenisation only.
VV: There is a grey region, but I would argue that it depends on whether they are detectable as QC problem, i.e. large enough. Especially if you do not have the daily data any more, many of the "QC" will be hard to see in monthly or annual data. And if they are persistent, I would call them inhomogeneity. For me QC is about single strongly wrong values.

Columns H and J could be simulated with an exponential
RL: What about geometric - assume IID (independent and identically distributed)
CG: Determines both length and number of changepoints through some probability
P(X1 = k) = p(1-p)^k k=1,2,3
*p= 1/(n*lamda) n=15 lambda=12 so 15years*12months*
*RL: Called the renewal process - rgeom - give it a number for average length/return period 1/ (n*lambda)*
Average time between points is 1/p
CG: More likely to get changepoints that are close than far apart
So - not saying that most (the mean) of breaks occur at 15 years apart - just saying that 15 years is the average of the geometric distribution and more breaks are more frequent.
KW: So we can use this for types of changes that can occur at any time throughout the record: instrument change, station move. Use other methods to apply specifically located changes e.g., automation occurs within a specific few year period for many regions.
PT: If we know a station is now an airport (US: AP = airport) then we know it likely moved there from somewhere else if that station is old. Similarly if a station existed before 1987 and is MMTS then it must have changed from manual.
PT: Can pull out metadata for all US COOP stations to find which are now MMTS. May be able to do something similar for Australia/Canada/Spain (AIRPORT/AERO/AEROPUERTO) and UK (Kate's excel file).
PT: Use info we have and guesstimate everything else.
KW: Balance between - may be easier to just use approximate percentage of stations effected over

what range of years. Will see how the coding goes - if possible use explicit knowledge about airport stations. If not, use approximate distributions.
PT: What about someone that uses a probability of change based on the fact that a station is MMTS?
KW: If we can do it we will.
Liquid in glass ~1500
MMTS = maxmin temperature sensor - manual ~3200 sites (changed from LiG - some now changed back)
AWS = automatic weather station - ~1200 US airport sites (changed from LiG)
Not the same thing -
Metadata could just have dates of changes 'known change here' and include null metadata where no actual change occurs.
Qs over gradual trend length
RL: Can overlay different types of moves with the rgeom method so can have additive probability of changepoint locations of different types but then wouldn't be able to back out which changepoint was which type.
CG: Poisson thinning? Generate all changepoints then partition those into a variety of types
EA: What about specifically located changepoints?
KW/CW: Apply those afterwards/separately?
EA: Apply in layers: airport moves, move to automated, randoms and then do some checking for sensible locations of the randoms.
VV: How do we apply a seasonal cycle to the gradual inhomogeneities? Use multiplicative model for that?
KW: What is a multiplicative model?
VV: 50 % of the mean inhomogeneity in that year, e.g.
EA: Could have a lower limit for length of gradual trends e.g. 10 years?
MM: PHA on USHCN does show very short periods of gradual inhomogeneities - actually due to wetting and drying from wet years.
KW: Instrument drift and then correction?
MM: Looks like gradual trends may be more frequent and shorter than estimated. 50% of changes found were slopes and many very short (often sawtooth with abrupt change in the middle). Have modified the table to increase percentage of stations and decrease average length.
CW: How much of this is natural variability in areas of complex topography? Changes in atmospheric circulation and hydrological cycle.
ACTION: Kate to add Colin G to Team Corruption - can help with R coding and distributions. Enric can also help with the coding.

Allow changes in last two years?
VV: Convention is to use the last homogeneous subperiod (HSP) as reference for the absolute level. If we would adopt such a convention as well, there would be the problem that there may be a break near the end. One solution would be not to insert breaks near the end in the benchmark (but they would be there in the real data). Another option would be not to analyse those stations (for station level validation measures) where we know that there is a break in the last few (2 or 5?) years.
KW/CW: Reluctant to not do something that is a real issue just because it's difficult to assess. We feel that we should allow these end of record changepoints to occur. Homogenisation of these stations is then unlikely ever to be perfect - but we know why.
EA: Problem of adjusting to last section as a reference. So adjustments will not be very good.
PT: Have a world where we preclude changepoints in the last 2 years.
EA: How about B4?

RL: Do we allow an abrupt change in the middle of a gradual change?
KW: Important point - we do need to add changepoint types in layers - at least abrupt and gradual

separately, otherwise we will never have sawtooth type features which definitely do exist.
Discussion of how to build an error model:
Bottom up - start with a station, what changepoints to apply?
Top down - start with a network, what distribution of changepoints to apply?
BOTTOM UP DECISION TREE:
Station X
A gradual station? (p=0.3)
NO:
Apply abrupt changepoints approximately 1 every 15 years (or 7 per decade)
*Use Geometric distribution with P(1/(15*12)) = approximately one every 15 years/7 per century*
*Gm(P[1/(15*12)]) => k changepoints located at T1, T2,...Tk*
For each changepoint T a size is chosen from a normal distribution with a mean=0.2/7 and a st dev 0.7.
YES:
Apply abrupt changepoints approximately 1 every 15 years (or 7 per decade)
*Use Geometric distribution with P(1/(15*12)) = approximately one every 15 years/7 per century*
*Gm(P[1/(15*12)]) => k changepoints located at T1, T2,...Tk*
For each changepoint T a size is chosen from a normal distribution with a mean=0.2/7 and a st dev 0.7.
For each changepoint T also decide whether there will be a slope change too. Probability of there being a slope is uniform across the changepoints (as likely in the later ones as in the early), probability of one occurring after one has already been found is much much lower. Tiny probability of there being no slope is possible.
If there is a slope pick the length from a ? distribution and the size from a normal distribution
Apply known times/character of changepoints as a separate layer -

Layer 1 – apply known/clustered changepoints
*Layer 2 – apply locations of abrupts geometrically with probability converging on 7 per century P(7/12*100)*

- *for each abrupt apply a type – probability of each type occurring*
- *apply appropriate size by normal distribution*
- *apply appropriate shape/character by distribution*

Layer 3 – apply locations (if any – 30% of stations yes) of gradual trends based on poisson (discrete)?

- *yes/no apply additional abrupt at beginning*
- *yes/no apply additional abrupt at end*
- *Apply a duration (1 to n years, mean 25 yrs – pareto=fat tailed?)*
- *apply a rate for the slope (-2 to 4, mean of 3 deg C per century)*

C - What kind of inhomogeneities do we want to put in the experimental world (#B8)
D - How to implement the seasonal cycle?
E - We may still have one undefined world (#B7 - currently B5 with different GCM). We could also limit ourselves to 9 blind worlds.
F - what should people be asked to homogenise.
G - How to apply gradual changes?
I - How to package up the error world data?
J - Instructions on how to come and play (will need input from Team Validation too)

How to apply seasonally varying changes?
How to apply gradual changes?
How to package up the error world data?
VV: What do you mean by that?

Instructions on how to come and play (will need input from Team Validation too)

SUMMARY:

Blindness: Kate to be the handle turner, basics of what is in the worlds will be known but specifics of what and where will not be. Try to speed up the cycles if possible to prevent release of specifics after assessment.

Number of clean worlds: 3 different GCMs for blind worlds (1:4, 5:10 (not 7), 7) but each world will have its own 'station pink noise'. Open worlds will use different GCMs and a unique station pink noise overlay for each world.

How to build: Bottom up station decision tree with layering.

Layer one - apply any known/clustered changepoints based on region

Layer two - non-clustered - if not a 'gradual' station apply abrupts - allocate type (by probability) location and size, if a 'gradual' station apply abrupts (location and size) and for each decide whether to apply a 'gradual' and for how long/size

Do not locate changepoints within missing data periods - force to be start or end of missing period

Seasonal cycle needs more thinking - 10-50% of stations, changes in variance likely joined to temperature, radiation, humidity, wind

Check changes in Tropics with northern Australia as a proxy - are they larger?

Fixing parameters:

Gradual trends likely to be in more stations (30%) and shorter/more frequent? (25 yrs), abrupt changes occur at beginning and during gradual trends

Not including random errors - assume a good QC process has been run on the data - isolate assessment of homogenisation algorithms

Assume Gaussian for size of changepoints - we're missing the middle and much of the sides in our current ability to detect

No minimum homogeneous sub-period and allow changepoints in the last two years

Apply location of changepoints using a geometric distribution - all points are equally likely to have a changepoint applied with a probability that optimises on 7 per century.

NOTES:

Studying the influence of biases

Statistical inhomogeneities

#B1. Best guess world for the West everywhere. A mix of random and biased abrupt breaks with some gradual inhomogeneities, some spatially correlated breaks, seasonally varying breaks, realistic missing data (see Section 2).

#B2. Best guess world (#B1), but no spatially correlated breaks.

#B3. Best guess world (#B1),but more and smaller unbiased breaks and gradual IH. The properties of the biased breaks stay the same.

#B4. Best guess world (#B1),but fewer and larger unbiased breaks and gradual IH. The properties of the biased breaks stay the same.

Physical inhomogeneities

#B5. Best guess world, with a bias of 0.2°C per century at high- and mid-latitudes and 1°C near equator. Implemented by making the bias a function of insolation and log(humidity) (or net IR surface flux at night), if they are capable of producing biases).

#B6. Best guess world (#B5),but instead of ~2 breaks per century with a bias, it has ~4 breaks per century with a bias on average. Total trend bias the same, thus the 4 biased breaks only have half the bias size.

Random and biased breaks.

#B7. Best guess world (#B5), but exploring different background climate?
#B8.Best guess world (#B5), with national more exotic inhomogeneities. Next to the typical exposure and relocation based inhomogeneities, there are many less frequent causes that have their own specific signature. They typically happen in just one network and by implementing them only in a small number of countries, we can try many different inhomogeneity problems.
Studying the influence of the seasonal cycle
These two blind worlds should be analysed together with #B5 and #O3 (a world without an annual cycle in the inhomogeneities).
#B9.Best guess world (#B5) where the biases are implemented by using the equations of Auchmann and Brönnimann (2012) taking insolation, humidity, wind and snow cover into account.
#B10. A more difficult seasonal cycle that only affects a small number of months, up to one season. As in all cases with a seasonal cycle, this would include occasions where breaks were in opposing directions for different parts of the seasonal cycle.