Benchmarking Working Group Online Minutes #19

Tuesday 28th January
3pm Greenwich Mean Time
10am Eastern Standard Time
4pm Central European Time
2am Australian Eastern Daylight Time - sorry Lisa!

Aims:
- Implementing inhomogeneities by Type or by Character?
- Team Creation - where are we now

#---
Attending: Kate Willett (KW), Rachel Warren (RW), Victor Venema (VV), Enric Aguilar (EA), Ian Jolliffe (IJ), Robert Dunn (RD), Claude Williams (CW), Matt Menne (MM) , Zeke Hausfather (ZH)

Apologies: Peter Thorne (sorry, something came up), Lisa Alexander (sorry will be in bed :-)

Actions from last meeting:
ACTION: Post meeting minutes early
DONE (hopefully)
ACTION: KW to put the word list and words from RWs comments in Concepts paper into ISTI glossary + GCM, downscaling, discontinuity, SI, micro-climate
DONE
ACTION: KW clarify position of maybe (but not definitely) submitting a paper on the regional summaries at some point within the Terms of Reference
DONE
ACTION KW: tidy up Terms of Reference, post on website and submit to Steering Committee for sign off.
DONE
ACTION KW: shorten Concepts paper if possible.
DONE
ACTION KW: add definitions for uncertainty, inhomogeneity, changepoint and bias into Concepts paper - make sure we're consistent in talking about it
DONE
ACTION LV to second review paper.
DONE
ACTION KW: make the equation bit clearer - explanation, XOB, D is also variance, extra equation in GCM discussion?
DONE - no extra equation
ACTION KW: tidy up figures 6 and 7 and improve explanation in figure caption.
DONE
ACTION KW: reference ISTI glossary in paper
NOT DONE - not entirely sure this is necessary.

ACTION KW: amend table to include perturbation of changepoint frequency (Option 2) instead of using Quality Levels to infer potential changepoint frequency in a percentage of stations.
DONE
ACTION KW: Move docs to permanent storage and link from the website:
PENDING

Actions from this meeting:
ACTION VV: Prepare slides/presentation on slope shape for gradual inhomogeneities for next call.
ACTION KW: Kate redo the spreadsheet with character. - share
ACTION KW: Doodle poll and minutes (past and present)
ACTION: KW contact enric/richard about gamma filtering for Team Creation
ACTION IJ: to send some emails and think about key topic for next call.

Summary from this meeting:
SUMMARY 1a locating changepoints/missing data periods containing changepoints:
- allow allocation of changepoints within missing data periods
- push all missing data period changepoints to end of period (first data point present)
- we will not increase likelihood of changepoint occurrence at the end on a missing data period because there is a higher likelihood of one occurring there because of the chance of allocating in the previous missing data. This may allow multiple changepoints, especially for longer periods and some abrupts/graduals will have simultaneous abrupts allocated.
- may add Wars to regional summaries

SUMMARY 1b use random number generator to allocate changepoint type/character:
- All happy with this approach and runif

SUMMARY 1c world, country and station quality allocations:
- Agree to have a global frequency and perturb the country frequency. A random draw is taken for each country. e.g., 4 changepoints global frequency results in 2% of regions with no changepoints and 2% with 9 changepoints.
- Allow the station changepoint frequency to be governed by how the threshold/random number generator works out - which is a poisson process where mean=country frequency.

SUMMARY 1d gradual changepoints:
- gradual changepoint locations should be done by allocating the mid-point of the period of gradual change to prevent edge effects and allow changes to have started before the series record or continue afterwards (essentially begin/end though)

- period before and after can be assigned by randomly generating a number for each month progressively away from the mid-point. When the number first crosses a threshold then the end point is reached.

- for example, for a 20 year average gradual change period (10 years before and after) within a 100 year time series the threshold should be1/120 (10 years * 12 months)
- once an endpoint is found add a probability that an abrupt changepoint occurs
- midpoints could occur within another gradual change period in some cases
- optimise towards fewer longer gradual changes rather than frequent very short ones.
- discuss the shape of gradual inhomogeneities next call - linear? seasonal cycle? some other curve?

SUMMARY 2 character verses type method for applying inhomogeneities:

- Go with the Character Method
- add semipermanent changepoints that persist for 2+ HSPs (homogeneous subperiods)

#---

AGENDA

https://docs.google.com/drawings/d/1Hn7lDQJivDcKfZpkTyuIBzvYFE9MSyc8IPBNMwdPJCA/edit - an example of a station data object and its attributes

https://docs.google.com/drawings/d/1fb-kjL2J1oG1KmR2c_4hWhlSvV3Ats7xDGXP45aAU7w/edit - a flow chart showing how inhomogeneities might be added - not yet updated

https://docs.google.com/spreadsheet/ccc?key=0Al6ocsUAaINSdHVKbVBoLWxjQU1Ib2c5bXV0RXZITHc#gid=0 - spreadsheet with example size/shape/seasonal cycle/probabilty of occurrence for applying changepoints

1) Some reminders and remaining Q's about locating changepoints:

a) We have decided to locate changepoints based on a probability for each month that takes into account the desired frequency (Option 2 - Notes 1.). This will be done for clustered/knowns (layer 1), abrupt permanents (layer 2a - which may then have a temporary on top), abrupt temporaries (layer 2b). For example: to assign ~4 permanent changepoints per century each month would have 0.003 probability of a changepoint occurring. Using a random number generator, a changepoint will be assigned if the random number is less than or equal to 0.003.

PT: Yes, in a temporally complete station but for a temporally incomplete / shorter segment sample how are you going to populate the breaks sensibly? If I have a 30 year station then do you want to force 1 break into the series (on the average)? In which case presumably you may want to actually look in that station for cases LE (what does LE stand for?) something more like 0.01 to get the right chances (or actually maybe I'm being thick - entirely possible, nay, depressingly likely as I have yet to have a second coffee ...). If you seed with 0.003 into the whole temporal series and then apply the obs mask all your seeded breaks for the station may be when no records exist. Also, how then do you propagate cleanly the break locations to team validation? If I have station 1890-1930 and 1950-2012 and we seeded a break in

1940 we are never going to get the right breakpoint location. The best any algorithm can / will do is assign a break across the data junction. BTW I utterly mucked this part up in doing the USHCN benchmarks - its the hardest part of the problem to work out how to implement properly. Personally, in hindsight, I would have seeded based upon the data mask (so I can only ever place a break where there is data in the station record itself) even though this is computationally harder and more expensive and forced a break at the resumption any time there is a break in series >2 years under the assumption that this pretty much always constituted station shut down and start up and that something would have changed as a result.

IJ: Is it realistic to assume that the frequency of changepoints increases in the non-missing part of a series with missing chunks, compared the frequency in complete series? This is the implication of this suggestion. Clearly missing data are a problem? How big a problem? Is it essential to include them in the benchmarking?

*IJ: Heresy perhaps, but let's repeat the last question. With a May deadline looming, perhaps the time has come to simplify things in the first cycle (such as not having missing data if it makes things more complicated) so that we have *something* useful in May.*

PT: The issue is that if a station really stopped and went quiet for two years there is a very high probability that the new station is one or more of: in a new location, renovated, run by a different entity, observed by a different observer etc. etc. - all of which, a priori, would radically increase P(break) so its probably best to just model in a break whenever there is a 2-year+ hiatus in the record. KW: Very predictable though. PT: Yes, but then e.g. PHA I believe does this. From a measurement standpoint it makes sense to if not put a break always to put a break very often at the resumption of the series (say p=0.8-0.9).

KW: I like the idea of initially treating the series as complete. If a changepoint is allocated in a missing period then push it to the end of that missing period with the caveat that we don't really want several breaks occurring within a few months.

KW: I think at this point it may be best not to worry about whether stations with missing data are more or less likely to have changepoints. We can just focus this time around on how much missing data it takes to break the algorithms? There will be a wide range of station qualities in the mix anyway so it likely be possible to compare a station with the same number of changepoints but different amounts of missing data.

VV: Good point that short time series should have a lower break frequency because the beginning of the time series is already a break inhomogeneity. That should be fairly easy to implement.

VV: Missing data: I would personally also include breaks in the missing data periods, just because this part of the data is not digitised is no reason to assume that there were no breaks in this period. On the contrary, as Peter indicates above, a missing data period is an indication that there was a change at the station, especially longer missing data periods are likely not there for nothing. Thus if we do not increase the break frequency in missing data periods, we should at least include them in these periods.

VV: That the date of the break is not defined in case of a missing data period, is something which team validation should take into account. They could move all *breaks to the end of the missing data period or do something more sophisticated in the computation of the contingency scores.*

KW: We agreed previously to move any changepoints within missing data periods to the end. Is there any difference to doing that verses just doing that for validation purposes? I don't think there is.

VV: At least it makes it easier for the people providing data, we would not need to specify to them to put such breaks at the end. They could also put them at the beginning or where ever they like. I have no idea if it could help IJ i.e. mathematicians. It would also allow for multiple breakpoints in one missing data period, which could have a combined larger effect.

KW: If we decide to put all changepoints at the end of a missing data period then I suppose that is predictable, and someone allocating it at the beginning would not really be wrong. It makes validation quite complicated with needing loops to say 'if data missing then treat like this' but its not impossible. I had envisaged being able to validate from some simple tables of adjustment locations and amounts rather than having to interrogate the data again.

VV: You could have a pre-processing step in the validation that put the breaks at the end of missing data periods. I expect that we have to interrogate the data anyway, you can ask for a table with break dates and times where a gradual inhomogeneity begins and ends, but the size of the break is hard to specify in a table, especially if people contribute methods that make the corrections weather dependent (such as Auchmann and Brönnimann). Thus we probably have to interrogate the data to compute the break sizes ourselves (the contingency scores should be a function of the magnitude of the breaks).

KW: I doubt that we would have any global scale weather dependent adjustments although they could scale with temperature only.

KW: The tables would contain the mean shift either for the whole HSP - that's what we've suggested in the concepts paper.

VV: Do not forget the gradual inhomogeneities, such tables could become horribly complicated even if we are lucky on no one makes weather dependent corrections in this cycle. (And at some cycle they will certainly start doing so, why not be prepared?)

KW; The tables are reduced to very broad brush info - so not ideal for in depth assessment but I think we do need something like this. There is no way Team Validation can analyse 10 times 30000 stations for each team that submits. Even my 5-10% of committed time wouldn't really cover that.

VV: In HOME we had tables with: data break, begin gradual inhomogeneity, end gradual inhomogeneity, and outliers.

KW: That sounds like a good start.

KW: I still don't like the idea of allocating changepoints to missing data periods though. It doesn't really make sense as the change is when the station began reporting again. We've also already agreed on this so its a step back. I agree that we should think about validation allowing the changepoint to occur at the beginning of the missing data period although you could argue that that is wrong in reality.

VV. We had also already agreed on having zero-mean and non-zero mean breaks and had even fixed the sizes for them in the different worlds and then our chair started a discussion about 5 different types of break inhomogeneities! :-) KW: Yes agreed but that was to try and decide how we were going to apply those zero-mean and non-

zero mean breaks that had different season cycles/likelihoods of other changepoints co-occurring etc. All of that is trying to stick to the original amounts of bias/random but within a framework that has the appropriate variety. So, we're absolutely sticking with your method, I'm just trying to work out how we apply it. Lets get to that later as we really need to sort that out. Happy to go with your method but I need more explanation to see how we can code it. VV: I added a smiley, I have no problems with opening the discussion again and coming up with better solutions. I just think that "having agreed on something before" is not such a strong argument. KW: Fair point. I'm panicking because we/I have very little time to work on this before May which is when we're meant to be pretty much there so trying to move things along. I also think in this case our initial decision was valid. Let's discuss in the call anyway. (Thanks for the smiley!)

IJ: Probably a silly question, but does most of this matter. If a series has a big chunk of missing data and there is a break within it, no-one can tell where that break was so all that can be assessed with respect to location is that it occurred *somewhere in the interval. Certainly true for abrupt breaks, but perhaps exact location does matter for gradual breaks - where did they start?

CW: For offset type breaks, I agree. The following summary by Enric and Peter would sum up the problem nicely - if the validation folks agree..... For gradual inhomogeneities, they are the hardest (for us) BECAUSE of the variation in our Pairwise solutions for the begin/end of the graduation(?). Have not been able to solve that one.... yet.

EA:

1) Missing data increases the chance of a break, as we go from states A) to status B). Easy example: Spanish Civil War ... lots of missing for 1936-1939, after 1940, new regime, new system, etc, lost of BPs ...

2) I think that allowing break points in missing data is not a good idea, even Victor's reasoning about future digitization is true, there is no way to detect such values

3) Could we move any BP occurring in a missing value to the first (to the past) available value, this is a BP in 1953, with 1952-1955 missing would go to 1951, which is what a good algorithm should detect

KW: I agree but should it be the beginning or the end of the missing data? I would have thought the end of the missing data period as that is when the 'new' station/instrument etc starts reporting. In reality it probably doesn't matter as long as we don't penalise for it.

PT: For validation you presumably simply give a +ve score for at either end of such a break. What matters is that you found a break between the two segments and made an adjustment. In theory validation could be on a compressed series with no mdi then you get a 1-timestep location error if you put the break at the wrong end of a long data_mdi gap in the series.

ZH: Are any changepoints regionally correlated?

KW: Yes - we have regionwide changepoints applied at times of known region-wide changes and we're making some up for regions where we don't have info.

VV: If I remember correctly, 2 changepoints per century regionwide, 5-6 in total, including random/biased/permanent/temporary.

KW: PHA found more changepoints for humidity - 8-10 per century.

CW: Should we have a Type War? We know the major wars that have occurred and that they have resulted in missing data (SB: and non-digitised data) and widespread changes.
KW: Not sure about a type (complex) but could probably add this to the regional inhomogeneities.
MM: Within a missing data period many changepoints may occur and so should be collated.
KW: OK because the layers account for having multiple changepoints at any one period they are then combined during a missing data period to the end or beginning.
IJ: Do we need missing data at all?
VV: We do need missing data because it is a very real and large challenge.
VV/MM preference for keeping missing data same as in real world.
SB: Will be higher chance of allocating a changepoint because of the time that elapsed.
VV: Could increase the threshold within the missing data period.
MM: As we don't really know lets keep it all the same throughout - just leave.

SUMMARY:

- allow allocation of changepoints within missing data periods
- push all missing data period changepoints to end of period (first data point present)
- we will not increase likelihood of changepoint occurrence at the end on a missing data period because there is a higher likelihood of one occurring there because of the chance of allocating in the previous missing data. This may allow multiple changepoints, especially for longer periods and some abrupts/graduals will have simultaneous abrupts allocated.
- may add Wars to regional summaries

b) We can then use the value of the random number (that is below 0.003 - so some value between 0.0 and 0.003) to assign a type/character - is 0.003 too small to do this depending on the precision of the random number generator? *[PT: The issue is going to be less precision and more getting a real random number generator - lots of things that claim to be aren't - I'd look to our stats experts for getting a good generator here that is as close to truly random as possible*] If it is then we can just pick another random number and use its value to assign a type or character.

IJ: Although there's no problem in choosing break presence and break type using one random number, it is perhaps easier to explain and doesn't need to use so many decimal places, if it's done in two stages.
RW: In R the command runif(n) will give n random numbers from a uniform 0,1 distribution to around 8 decimal places which should be sufficient I think. runif also allows you to specify the range of the uniform distribution - so you could also look at a uniform distribution from 0 to 0.0003, though I'm not sure if this would then do quite what we want it to
IJ: I'm sure it will be OK without changing the range of the uniform distribution

KW: Hopefully the runif(n) is random enough for our purposes. I did have to use a random number generator for Team Creation. I looked into a few and did initially write something based on Robert's code but went with runif() in the end.
VV. There is a nice and readable discussion about random number generators in the book "numerical recipes ". I would expect that for our application any number generator would be fine, we are not doing cryptography or spanning high dimensional spaces.

SUMMARY

- All happy with this approach and runif

c) The frequency of changepoints will depend on the world, country and station. At the country level quality can be assigned by picking a number from a Gaussian distribution with a zero-mean and standard deviation of 1 N(0,1) which is added to the World average changepoint frequency (e.g. global average of 4 permanent changepoints per century).

*PT: In reality there is probably something akin to a y=x*2 country distribution (y=breakpoint frequency and x is some nominal range). At -ve values of x you have the countries where expertise and geopolitics have made long-term measures all but impossible. At +ve values of x you probably have the tinkerers in chief (US springs to mind) and in the middle x~0 you may have many of the rest where the propensity to tinker may have been lower and expertise and station continuity reasonable.*
At the station level quality can be perturbed further by picking a number from a Gaussian distribution with a zero-mean and standard deviation of 1 N(0,1) which is added to the country average changepoint frequency (assigned earlier).
*IJ: This is not what I expected. I expected the levels to have been adjusted up/down based on knowledge that particular countries/stations have more/fewer changepoints than the global average. This idea of *randomly* changing things with a mean of zero seems unnecessary - there should be enough randomness in the number of changepoints at different stations without it. Given a mean of 4 changepoints per century, and assuming a Poisson process (as we are pretty much doing in our proposed method of generating changepoints) about 2% of stations will have no changepoints at all in a given century, and a similar proportion will have 9 or more changepoints. Do we really want more random variability between stations than this?*
KW: I agree at the station level this may be unnecessary but it has its uses as the country level to make all stations in a country more likely to have more frequent changepoints or more likely to have less frequent changepoints.
IJ: I think it’s better to deal with between-country variation by changing the mean for different countries rather than adding an extra variable.
PT: Presumably the random numberness in the random generator buys you this station to station propensity anyway so agree this additional nesting is not required. Some random draws with the same underlying distribution will give 5 or 6 breaks per station some one or none with an average of 3 (assuming complete stations - but see above).

VV: I try to avoid comments with "in HOME we", but in HOME we changed the frequency of breaks at the station level. The average number of breaks was 5 per century, but it varied between 3 and 8 breaks per century. This does make a difference, it makes the distribution broader as for Poisson. Some variability of the break frequency may make sense physically, stations near cities may have more breaks, or volunteer stations may move more often.
VV. However, I also see the variability at the network/country level as the more important one. That is more likely the one that influences the power of the homogenization algorithms, because the networks with many breaks in both candidate and reference will be significantly more difficult as the ones with an average number of breaks.
VV: My personal solution would be to have an average break frequency for the entire world, say 5 or 6 per century, perturb this per network by N(0,1) and perturb this again per station by N(0,1).
IJ: So perturb the mean - OK by me.
PT: Surely you get a spread just because a random number draw for <threshold may on average have n cases but the whole thing is distributed around that in some manner. Say I look for occurrences <0.003 in random draws of 1,000 length on average I will get 3 cases but I'll have a very few with none, a very few with even say 10 and then the bulk having 2-5 cases found.
VV. Yes, in that case you would get spread and it would follow a Poisson distribution. The question is whether we need a broader distribution. I guess there is almost no empirical evidence. My recommendation to perturb the mean is based on physical expectations. If we want to simplify, we could also keep the rate constant.
IJ: If different countries (or other groupings) of stations have different means, then the 'good' ones will have rather more cases with none and fewer cases with 10 or more than the 'average' stations - conversely for 'bad' stations. If you believe there are good and bad groups, then a single Poisson won't be dispersed enough to model it.
PT: Yes, you just perturb the threshold by 'country / group'. If you put it closer to zero you'll get more with very few breaks but still a handful with many, you increase it to say 0.06 you'll get far fewer with none and far more with lots.
IJ: Yes changing the threshold is the same as changing the mean for that group
KW: So to have a country that has a good proportion of stations with zero changepoints say we would adjusted the global changepoint frequency to 1, using the random number approach we would then have to have a number <0.0012 or there abouts? Do we feel that this gives us the right spread of frequencies across the country?
IJ Poisson distribution with small mean means no negatives and small chance of going very high.

SUMMARY:

- on second reading I'm actually a little more confused by this
- Agree to have a global frequency and perturb the country frequency either by a known amount or by using the global frequency as the mean of a Poisson distribution. A random draw is taken for each country. e.g., 4 changepoints global

frequency results in 2% of regions with no changepoints and 2% with 9 changepoints.

- Allow the station frequency to be governed by how the threshold/random number generator works out - which is a poisson process where mean=country frequency.

d) Can gradual changepoints be assigned in the same way as abrupts? Here we need a start point and a length whereas abrupt changepoints continue until the next abrupt changepoint or end of series. So we could have the same probability method but then when a changepoint is assigned it also has to have a length assigned which could be some proportion of the remaining series.

- For an average of 1+ GRADUAL changepoints per century:
- each month has a probability of ? of having a start point for a gradual
- a random number between 0,1 decides the proportion of the remaining time series for which the gradual change persists?
- during the gradual change, probability of another gradual change drops to 0?

RW: Causes of gradual changes: urbanisation? Instrument drift? How likely are these to co-occur in the real world - if they're likely to not occur at the same time then not allowing a gradual change to start within another gradual change is probably reasonable - we don't have to get everything perfect first time - this could be something we could consider adding in the second cycle - on the other hand if it wouldn't be too complicated to implement in the coding then there's no harm in adding some changes in like this (maybe just in one region) and seeing if the algorithms can cope.
EA: I think we can think of as many combinations as our imagination allows. One example: urban effects (+ trend) + trees planted at the same time around the garden (- effect) : so at the end, homogenizations would most likely adjust a trend of the resulting magnitude ... Do we want to specifically code this? Probably not, but if the probability is low for this combination to happen, I wouldn't worry about it.

- we want to have some stations with multiple gradual changes (at separate times as opposed to simultaneously) and some with 1 and 70% with none.

IJ: I think you can use exactly the same method to decide the start point of a gradual change as for abrupt changes. You could also use the same basic idea to determine when it ends: for each month after it starts draw a random number between 0 and 1 and end the gradual change if that value is below some threshold (the probability it ends in any given month). The probability of ending would be higher than the probability of starting if length of gradual changes was thought to be less than the gap between starts of changes, on average. The idea of using a random number to decide for what proportion of the remaining time the gradual change persists implies that the average length of gradual changes decreases as time passes (and does so rapidly towards the end of the series). I assume this is unlikely to be true
PT: Agree with Ian. The wrinkle really with the gradual changepoints is inevitably going to be ensuring that they are consistent with the data mask. This will be harder for gradual than abrupt. Abrupt is a simple logic check for not missing value - gradual seems like it will be a little harder to code to ensure we don't project out over huge

swathes of missing data. I'm sure its possible but my brain hurts just thinking about it.
KW: I didn't think it was that complicated. Assuming we model the graduals as a linear change, possibly with some seasonal cycle then that change just goes straight through the missing data period - because the city keeps growing. The assessment would be on characterising the slope correctly although for level 1 our main focus would be are the large scale features (climatology, trends etc) ok so its not a problem as long as they've done something to remove the slope.
PT: But I think you need to at least start and end the gradual change during periods with data for this to prove tractable. Now, the real world may not be tractable and here I simply mean tractable to validate in a meaningful manner. Take a station 'A'. Station 'A' exists from 1900 but with gaps in 1910-1930 and 1970-1980. The surrounds of station A started feeling a UHI encroachment in 1915 (when station A was not recording) and the UHI encroachment finished in 1972 (also when station A was not recording). The best any algorithm would do is assign a step function in 1930 relative to 1910, then a slope like adjustment 1930-1970 then another step like in 1980. i.e. no algorithm could actually replicate the applied slope segment break exactly. If you don't need data at start and data at end then I would fear that you will utterly screwball team validation's efforts to make a reasonable validation of such a case and we penalize good algorithms for doing the best that is actually possible?
*VV. To avoid edge effects, you could use the same method for breaks for the *middle* of gradual inhomogeneities and then in addition draw a number for the length of the period.*
VV: Agree with EA that modelling all gradual inhomogeneities is simplest solution and likely realistic.
VV: Agree with KW that we can just ignore the missing data periods in implementing the gradual inhomogenities. In the validation, we will have to compute the trend based on the data where is data. Like in the discussion above, I do not think we can validate a table with the adjustments people claimed to have made, that table would be much too complex and also be more likely wrong as the data itself (people are used to produce homogenized data, they will have to implement such a table in haste before our deadline).
KW: use thresholds for allocating the middle of the changepoint.
IJ: Then just use method above - a threshold based random draw again forwards and backwards - geometric process
MM: Better to skew on the longer period rather than lots of very short.
VV: ACTION next call shape of gradual changes.

SUMMARY

- gradual changepoint locations should be done by allocating the mid-point of the period of gradual change to prevent edge effects and allow changes to have started before the series record or continue afterwards (essentially begin/end though)
- period before and after can be assigned by randomly generating a number for each month progressively away from the mid-point. When the number first crosses a threshold then the end point is reached.

- for example, for a 20 year average gradual change period (10 years before and after) within a 100 year time series the threshold should be 0.1 (120 months/1200 months?) HELP - IS THAT RIGHT?
- once an endpoint is found add a probability that an abrupt changepoint occurs
- midpoints could occur within another gradual change period in some cases
- optimise towards fewer longer gradual changes rather than frequent very short ones.
- discuss the shape of gradual inhomogeneities next call - linear? seasonal cycle? some other curve?

2) Assigning a Type or Character to apply at each changepoint.

We can either assign the change based purely on stacking up probabilities of certain features or we can specifically define types from which to sample and give them nice names.

Victor, please can you explain your method here - I've called it Character method because it directly assigns the character of an inhomogeneity (positive/negative/mean-zero, seasonal cycle) all by probabilities I think - please correct:

IJ: Type vs. character. This is really a decision for the climate people. Statistically, at the end of the day the way the series are generated is pretty much the same whichever approach you take. But a random confusing thought: I guess you want to have your corrupted series mimicking reality as closely as possible. It might be easier to do this with the 'character' approach because there may be observed changepoints that can't be replicated by your list of 'types'. But hang on - how do you know that an observed apparent changepoint is real unless you can explain why/how it occurred i.e. you need to assign it a type?

PT: I think Ian's point is key - that from a construction viewpoint this is semantics because at the end of the day the code is only going to care about the logic and not the names. So, I would pick some way that makes sense to you to describe these and run with it at this stage. What the code needs is (Prob(break)) -> (break_locs) + (break_mags) -> (corruptions to add to series). As its all numbers it cares not a hoot what terminology you use. Terminology only matters to the humans so you should choose whatever will be most obvious way of describing what you are trying to mimic to your target audience(s). Key Q then is who is your key target audience here? And what will they most cleanly understand?

VV: The main problem I see with having multiple characters (relocation, change of observer, etc.) is that we do not know the break frequencies and the distributions of the break magnitude for all these different types. We would thus have to guess enormously, which makes it easy to attack our work. We do have relatively solid numbers for the frequency and magnitude averaged over all breaks. That is why I would suggest to keep it simple. Furthermore, if you make it complex, you need to have a reason for it, I fail to see how a large number of inhomogeneity types would make the benchmark more or less difficult to homogenize.

Kate's Type method (with previous comments in italics):

VV: To convince me, you would need proof that we can estimate the frequencies of these classes and for every class the break frequency, its bias component and its *random component. Furthermore, such a complication only makes sense if it would change the results of the homogenized datasets.*

KW: An additional temporary (can we use that word instead of random?) changepoint can be added quite easily onto some Types - e.g., a shelter change (permanent for the sake of argument) is more likely to have some simultaneous change that we may consider to be temporary such as an instrument change. Time of observation is arguably more likely to occur by itself, although not always.

KW: Well, I'll keep trying :) I disagree that we need to understand each type of inhomogeneity perfectly before we can implement it. I think we are ok just to make best guesses and approximations to the best of our ability. So, the Types I have described are not all-encompassing and I don't think they need to be. I have called them Shelter/Move/Time etc but could just as easily call them Dave/Barry/Susan etc. Although Type Shelter1 is designed to try and replicate a move from Stevenson Screen to AWS it actually encompasses a number of other inhomogeneity types because it is a distribution to be sampled from. In a way, I'm just using the names to help get my little brain around this big topic. I can see that it may be simpler in some ways to just toss a dice (weighted perhaps) to decide whether this permanent changepoint is warm-bias/cool-bias/non-bias, and then toss it again to decide what type of seasonal component (perhaps weighted based on the previous toss and the date?), and then flip a coin to see if a temporary component will also be added at the same time, there will then be more dice tossing to decide characteristics of that changepoint. Happy to go down that route - but I see it as almost identical to the Types route anyway. This is because we want to have weightings based on things we know a little about in the real world - e.g., there are likely to be more station moves (with additional instrument changes) than time of observation changes (which are most likely not to incur a simultaneous instrument change or other Type).The Types route just has the advantage of laying out all of our choices about dice weighting in a clearer manner I think.

KW: Any success Victor or is it time for me to give in?

VV: Maybe the dice are one reason for the disagreement. I think more in terms of random distributions as in yes/no or a limited number of categories.

VV: That being said, if we call the inhomogeneity types Dave/Barry/Susan, or more scientifically 1, 2, 3, 4, and 5, and give them all the same break frequency, magnitude and seasonal cycle, we might be getting close to an agreement, then we would not have to guess new statistical parameters for Dave/Barry and Susan. Having a few different types may make sense. For example the effect of a relocation would be semi-permanent, that is until the next relocation. Thus implementing (part of the inhomogeneities) as having a type, may make sense to be able to implement such semi-permanent inhomogeneities, next to the non-zero permanent inhomogeneities (random walk) and the temporary (noise) inhomogeneities.

VV: Question to the experts: does this make sense for more categories as just the relocations?

KW: Some types may be more likely to have a strong seasonal cycle than others but could be ok just to model this by coincidence of how the probabilities work out. An instrument change or calibration correction could have no seasonal cycle. A relocation/shelter change is likely to have a large and not necessarily sine curved seasonal cycle.

KW: Some types would be more likely to have multiple simultaneous changepoints occur - but again this could just be left to chance.
KW: I'm happy to go either way on this - I would just like some consensus from the group.
CW: What the homogenization algorithms will give you are what they know - segment length and location, change in mean, slope, seasonal cycle (in order of complexity). How we have defined them in the modelled world will only be conjecture to the programs because there may be more than one change type involved within close (i.e. not detectably separate) proximity. Unknowns and confusion are the heart of Station Histories. Still, we should impose on the modelled worlds the best understanding we have of the changepoint types, but I don't think we can ask the algorithms to know anything more than their solutions to the problem.

Once a changepoint location has been allocated, it can then be assigned a Type based on probability of that Type occurring at that time.
Clustered and Isolated changepoints can have the same Types of causes
Changes can be divided into 4 Abrupt Types for simplicity (there are more but these are the most common causes of inhomogeneity and their distributions will most likely cover the behaviour of other types that are not explicitly included) and 1 Gradual Type (with subtypes).
Each type will have a different probability of occurrence (perhaps depending on date/region? or underlying world criteria such as bias) and each subtype (1,2,etc) will have a probability of occurrence.
The length/frequency, size, shape and seasonal cycle for each type can be altered depending on the underlying world criteria - although this is tricky to ensure global/regional averages hold true.

1. Shelter change - permanent - non-zero mean/zero mean - seasonal cycle:

Type Shelter1 (zero mean/slightly cooler?, strong seasonal cycle, e.g., Stevenson Screen to AWS),
VV: I am not sure whether this is universally true, especially for mechanically ventilated AWS.
PT: Ventilation efficiency tends to have cancelling effects on Tx and Tn AIUI. Poor ventilation tends to yield enhanced maxima and minima and vice versa. So, DTR increases with poor ventilation and vice versa. What the sum impact is on the Tm series as Victor says is not a priori at least obvious.
KW: Sorry - you did point this out in the last call and I noted it but then got it wrong again here - how about the above amendment?
KW: Does this mean that we think a Type Shelter1 would be a very small change to the mean but a large change to the seasonal cycle?
PT: Possibly. It largely depends upon how well naturally ventilated the instruments are or rather how f(delta(ventilation efficiency)) is changed trough changing from screen to automated measurements. If you change the ventilation efficiency through contact heating / cooling the max/min are artificially enhanced when there is stilling around the instrument. In the US MMTS transition there is substantial evidence that the effect is slightly asymmetric such that the change in Tx is greater than Tn so Tm slightly decreases but in (ABS) Tm<Tn<Tx

KW: Ok - I don't think we need to worry about getting this perfectly right as it will differ station to station/country to country anyway. So if we can be just about satisfied with a small zero mean distribution and a large-ish seasonal cycle, and a high chance of having a simultaneous temporary change then BINGO.
EA: Regarding the size of the AWS-CON (Stevenson Screen/Conventional) inhomogeneity, the mean differences really vary as many factors come to play (sensor,screen in which the sensor is placed, particularities of the station, etc.), but more than a half of the cases (for tx in tn) result in a negative value for AWS-CON. Please, notice this is done with daily values

Type Shelter2 (cooler, especially in summer (skewed seasonal cycle) e.g., wild screen/north wall to Stevenson Screen)
PT: Effects I believe maximized in summer season.

2. Time of Observation change - permanent - non-zero mean - seasonal cycle:
Type Time1 (warmer e.g., early to later),
Type Time2 (cooler e.g., later to earlier)
VV: TOB has a strong seasonal cycle, related to the size of the diurnal cycle relative to the daily variability and the time of minimum temperature.

3. Station move - permanent or temporary - zero mean or non-zero mean - seasonal cycle:
Type Move1 (permanent/non-zero mean warmer e.g., rural to city), VV: probably a rare type of move, mountain to valley could be a better example of a warming move.
Type Move2 (permanent/non-zero mean cooler e.g., city to airport),
Type Move 3 (permanent/zero-mean e.g., random move),
Type Move 4 (temporary/zero-mean e.g., random move)
VV: Maybe moves should be semi-permanent, keep the bias fixed until the next move (or end of the time series).
KW: Move 4 is temporary so this is accounted for already.

4. Instrument change - temporary - zero mean - no seasonal cycle:
Type Instrument (e.g. random instrument change)

5. Gradual - temporary, non-zero mean or zero-mean, seasonal cycle
Type Gradual1 (temporary, non-zero mean (warmer e.g., urbanisation), small seasonal cycle)
PT: Type 1 gradual may be correlated with Type move 2 e.g. Reno, NV. The airport starts out way out of town and then the urban environment encroaches over time. Perhaps these two types could therefore be conditionally modelled somehow?
KW: This is partly done in the spreadsheet with % chance of an abrupt changepoint occurring at the beginning and % chance of an abrupt changepoint occurring at the end. There could be a much higher chance of a Type Move2 changepoint occurring at the beginning.

Type Gradual2 (temporary, non-zero mean (cooler e.g., increased vegetation/irrigation), large seasonal cycle (some interannual variability too but probably too complex))

Type Gradual3 (temporary, zero-mean (any, random incremental change), small seasonal cycle)

NB. Higher probability for abrupt change at end of Type Gradual1?

PT: Effectively are you saying that the site may be relocated out of town, and if so are you in effect arguing to model an increased prior for Type move 2 at end of Type Gradual 1?
KW: Yup - just need to extend the spreadsheet to make this more explicit by Type as at the moment it just has % chance of a changepoint occurring at the start and %chance of a changepoint occurring at the end - with no preference for which Type that abrupt changepoint would be.

Probabilities (% chance) of occurrence of each Type and conditional probabilities (% chance) of occurrence for each subtype:
e.g., for any assigned permanent changepoint there is a 50 % chance it will be a Type Shelter. Of those Type Shelters there is a 100% it will be Type Shelter1 if the date is after 1970 and 100% chance it will be Type Shelter2 if the date is before 1970.
EA: thanks for this explanation. It was not clear to me the meaning of the %. Now it is.

Type Shelter = 50% of permanent changepoints (TOO LARGE??)
Type Shelter1 = 100% for post 1970 of Type Shelter changepoints
Type Shelter2 = 100% for pre-1970 of Type Shelter changepoints
Type Time = 20% of permanent changepoints
Type Time1 = 20% of Type Time changepoints
Type Time2 = 80% of Type Time changepoints
Type Move = 30% of permanent changepoints, 60% of temporary changepoints (TOO SMALL??)
Type Move1 = 20% (40% pre 1950) of permanent Type Move changepoints
Type Move2 = 50% (0% pre 1950) of permanent Type Move changepoints
Type Move3 = 30% (60% pre 1950) of permanent Type Move changepoints
Type Move4 = 100% of temporary Type Move changepoints
Type Instrument = 40% of temporary changepoints
Type Gradual = 30% of stations contain 1+
Type Gradual1 = 40% of Type Gradual changepoints (TOO LARGE??)
Type Gradual2 = 30% of Type Gradual changepoints (TOO LARGE??)
Type Gradual3 = 30% of Type Gradual changepoints
NB. Higher probability for abrupt change at end of Type Gradual1?

We can add probability of a temporary changepoint occurring simultaneously with an assigned permanent changepoint. This probability will be different for different types. For example, there is a high chance that any Type Shelter would have a simultaneous Type Instrument or Type Move3 changepoint.
Look at the spreadsheet of example probabilities/size/shape/seasonal cycle for each type using the UK as an example country.
This contains the global statistics as specified by Victor and then the statistics for each type for the UK

We may need to create a similar set of statistics for each country - many countries would be identical e.g. within Europe.
https://docs.google.com/spreadsheet/ccc?key=0Al6ocsUAaINSdHVKbVBoLWxjQU1Ib2c5bXV0RXZITHc#gid=0
Qs:
1. A lot of the size/shapes are very small - what is the smallest size of inhomogeneity we believe we can reliably detect?
EA: which is the precision of the measurement? If we're measuring in 1/10th of degree, I think that very small sizes such as 0.05 might be detected, but can/should they be corrected?
*PT: People don't tend to go around deliberately adding breaks - so we have to assume there may be many breaks which are small - small relative to precision and small relative to statistical power for any algorithm to explore. This is why I have been saying that validation needs to in some sense reward and penalize based upon the size of the breaks. A missed break of 2K, a found break of 2K but adjusted to 1.75K, or a false positive introducing a break of 2K are far far far worse than missing a break of 0.05K. So, I believe when we come to validation we need to have a penalty function that penalizes almost based upon the (break_mag(error))*2 if that makes sense?*
EA: It makes sense. I had complaints from people using HOMER of breaks detected with size 0.02 ºC (with ACMANT). I wonder how realistic can this breaks be and how much effort should method developers in finding/correcting such problems. Knowing the uncertainty of corrections in the future, will help.
IJ: With respect to our Levels of Validation, at Level 2 where we assess whether or not changepoints are detected, we can tabulate hit rate etc separately for different sizes of break, and would be increasingly worried by poor performance as size of break increases. Level 1, where we compare 'true' and 'corrected' series will to some extent reflect performance in correctly estimating the size of the biggest breaks. Directly evaluating this latter performance is, I guess, part of Level 3?
VV: The distribution of detected breaks suggests that breaks smaller than 0.2 are quite often not detected. With ACMANT you could detect a clear break in the seasonal cycle and the also correct a small break in the annual means. Does not have to be a problem, but maybe you should also compute how accurate the correction constant is (without multiple testing, just t-test).
VV: The sizes reported are probably network means or parallel measurements. That is they are the biases, which are small, we would have an additional random component, which are much stronger.
PT: We should put realistic breaks in and not worry about making the problem a priori tractable. Realism is more important than tractability of problem. We make it unrealistic (in either direction) and it becomes of limited value. So worry only about whether they are plausibly realistic and not whether plausibly solvable by current marketplace algorithms.
VV: Agree
KW; We can call the Types Bananas, Apples and Elephants for all I care - its just a way of choosing a warm/permanent/seasonal cycle vs a cool/permanent/seasonal cycle vs zero-mean/permanent/no seasonal cycle vs any other Type we decide to include. We could just have a decision tree which chooses warmer/cooler/zero-mean

then seasonal cycle/no seasonal cycle. However, it's difficult then to try and replicate things that we know explicitly. We know that station moves are likely to incur a strong seasonal cycle change but instrument changes may not. We know that station moves are much more common than time of observation changes. In the grand scheme of things, it probably doesn't really matter. We just need to get some vaguely realistic errors in there and be able to summarise what exactly we have put into each error world.

PT: Certainly value in ensuring requisite types and phasings are in there. I think we probably all agree on this and are to some extent lost in the semantics of the precisely how to achieve it?

2. Make the sizes too big and we'll have massive biases - make them too small and none of them will be detectable - very difficult to ensure we're getting the global or regional average bias per century correct.

3. Do you agree with the probabilities of occurrence?

PT: Not sure I understand the compound probabilities sufficiently.

KW: Percent chance of any changepoint occurring?

PT: Yes, clearer now with your edits above that this was compound probabilities and not straight probabilities.

4. How big should a normal seasonal cycle be - half the size of the break on average? This may vary depending on latitude.

VV: Half the size fits well in Europe. No idea about elsewhere, it is not just a function of temperature or insolation, but can also depend on DTR.

5. Not really sure about B8 anymore - what would we do here? Could swap for fewer/larger biased changes?

VV: Learn what is difficult and should thus be studied in more detail and maybe included in the next cycle.

PT: I would like to see one world that is pushing the limits of current algorithms.

6. Are zero mean distributions really Gaussian? This would mean the most likely value is 0. We might wish the most likely value to be jointly -0.2 and 0.2 - so a symmetrical bimodal distribution?

VV: Would fit well to the distribution of detected inhomogeneities, but that is because the small ones are not detectable. These small ones are very important for the detection of the detectable ones. Empirical evidence based on metadata shows normal distribution. A bimodal distribution could be an option for the open worlds, to study the importance of the undetectable inhomogeneities.

PT: People aren't trying deliberately to introduce inhomogeneities so zero magnitude breaks would show well managed (but undocumented) change and as such I think we should allow a preponderance of c.0 breaks that reflect real world efforts not to introduce breaks. When you then assess skill you need to weight skill by size of breaks and penalize much more missed large breaks than missed small ones (break mag squared or similar).

VV: Just an idea: From a difficulty of homogenization perspective the zero-mean breaks are there to make the non-zero mean breaks harder to find and correct. That would suggest weighing the non-zero breaks stronger.

KW: Ok sounds like its best to keep a Gaussian distribution - good, simple!

IJ: I agree that it's not a good idea to go to a bimodal distribution. The discussion is straying into validation territory. You may want to tabulate success in detecting a *break in terms of (a) whether the mean is zero or not; (b) the size of the variance of the breaks (if this is varied); (c) the actual sizes of the breaks, as opposed to properties of their distribution, such as mean and variance.*

7. Do we need a best guess open world or would that be giving away too much?

VV: I also worry about that. Maybe the open ones should be more idealised.

8. Why change bias for Equator in world O3?

VV: Do not know anymore, maybe to be different from blind worlds?

9. Why have two distributions for the global random size/shape for world B3? (cell E8 on spreadsheet)

10. Why have a biased part of the random changepoints for world B7? (cell E13 on the spreadsheet)

VV: In the definition of this speadsheet, bias and random are opposites. Does not make sense.

This will still work in our three layers:

1) Set up Type settings (size/shape) for each world

2) Choose Country (which will also have its own Type settings for each world - see spreadsheet for UK example)

3) Layer 1 clustered (known and estimated for the unknown countries)

- can utilise all Types as appropriate but probabilities of occurrence will be specific to each region
- mostly PERMANENT and non-zero mean Types

4) Choose Quality level of Country (perturb global average by a random number picked from N(0,1))

5) Choose Quality Level of Station (perturb country average by a random number picked from N(0,1))

6) Layer 2a permanent (non-zero mean (Types Shelter, Time and Move) with some zero-mean (Type Move4))

- assign probability (P) of any month containing a changepoint based on station frequency of changepoints
- locate these changes by picking a random number (N) between 0 and 1 for every month - apply changepoint if N <= P
- for each changepoint assign a Type using either the random number or a second random number
- pluck a size/seasonal cycle from the Type distribution and apply to all pre-changepoint data

7) Layer 2b TEMPORARY zero-mean (Types Move4 and Instrument)

- assign probability (P) of any month containing a changepoint based on station frequency of changepoints
- locate these changes by picking a random number (N) between 0 and 1 for every month - apply changepoint if N <= P
- for each changepoint assign a Type using either the random number or a second random number
- pluck a size/seasonal cycle from the Type distribution and apply to HSP(homogeneous sub-period) only

8) Layer 3 gradual changes

- probability of applying to a station is 0.3 (30% of stations contain a gradual changepoint)

- how many changes from 1+? binomial or exponential - decreasing likelihood of more than 1?
- assign length/locations of changes - geometric
- pluck size/seasonal cycle from the Type distribution and apply to HSP
- probability of applying a temporary changepoint at the beginning? - if yes begin layer 2b again.
- probability of applying a temporary changepoint at the end - if yes begin layer 2b again.

Character method: (I'm just guessing here so feel free to correct/edit)
Global Statistics:
Total size/frequency/seasonal cycle (likelihood/size) of permanent biased changepoints
St dev/frequency/seasonal cycle (likelihood/size) of permanent non-biased changepoints
St dev/frequency/seasonal cycle (likelihood/size) of temporary non-biased changepoints.
Likelihood of simultaneous changepoint
Frequency/length/size/seasonal cycle (likelihood/size) for gradual changepoints with likelihood of an abrupt changepoint occurring at the start/end
Do any of these have dependency probabilities? e.g., if permanent biased changepoint is it more or less likely to have a seasonal cycle than a permanent non-biased changepoint? If it has a seasonal cycle is it more or less likely to have a simultaneous abrupt changepoint?
ACTION: Kate redo the spreadsheet with character., share
VV: What about semi-permanent? Carries on for 2+ changepoints and the returns to baseline. Shelter change, station moves, all I think.
KW: Is semi-permanent more realistic than permanent in all cases - in some cases it would sustain until the end of record anyway?
VV: Risk of going to more of a random walk than noise if all are semi-permanent.
CW: Region by region question?
VV: This is where Type would come in useful - some types would be semipermanent until the same type came up again.
KW: Can we just choose a number of changepoints for this inhomogeneity to carry on for? 2+

SUMMARY:

- Go with the Character Method
- add semipermanent changepoints that persist for 2+ HSPs (homogeneous subperiods)

3) Team Creation:
I will try and get the VAR code working on multiple gridboxes at once.

Any update from Robert on station/GCM comparison work?
Any comments from Enric having looked at the VAR code?
EA: I am looking at the code by looking at each function and trying to replicate it with my own coding to make sure
that it does what it should do. I exchanged a few comments with Kate on potential problems (small) and still looking at it. I will continue interacting with Kate on this.
EA, update: a couple weeks ago I finished reviewing Kate's code. I found no major problems and replicated the whole code achieving similar results. Recently, I have found a trick with R to speed up large codes. It is very simple: removing as many files, objects as possible and call the the function gc(), which immediately does the "Garbage Collection" and frees memory.
KW: Great stuff Enric. I now have the code running on clusters of gridboxes at a time and it still seems ok. I need to expand this to bigger regions and so the gc will come in useful.
KW: I'm still struggling to see how we will smooth over the regions though to make neighbour difference series work over 100s of km. Richard Chandler mention something like Gamma smoothing or filtering windows that could be implemented in R. Does anyone have any idea about this?
ACTION: KW contact enric/richard about gamma filtering for Team Creation

4) AOB
None

5) Next Meeting: Mid-February
ACTION KW: Send around doodle poll
- think about shape of changes for graduals
- seasonal cycle of changes for abrupts
- validation

IJ: What stats can we get from homogenisation algorithms?
ACTION IJ: to send some emails and think about key topic for next call.
What was done in HOME?
Mandatory results: start/end dates of abrupt and gradual changepoints, mean shift?

#--
Notes:
1) Ways of allocating changepoint frequency and location
Option 2: each month has some probability of having a changepoint applied (could be the same or seasonally/time varying) e.g.,

- For an average of 4 PERMANENT changepoints per century:

- each month has a probability of 0.003 of having a PERMANENT changepoint - if all months are identical

- For an average of 6 TEMPORARY changepoints per century:

- each month has a probability of 0.005 of having a TEMPORARY changepoint - if all months are identical

- For an average of 1+ GRADUAL changepoints per century:

- each month has a probability of ? of having a start point for a gradual

- a random number between 0,1 decides the proportion of the remaining time series for which the gradual change persists

- during the gradual change, probability of another gradual change drops to 0?

- we want to have some stations with multiple gradual changes and some with 1 and 70% with none.

IJ: For each time point independently you have a (multi-faced) coin tossing procedure. If Stages (1) and (2) are separated you first have the overall rate of changepoint occurrences, say 0.05 (any figures I use are not meant to be taken as realistic). For each time point, generate a random number between 0 and 1. If it is <0.05, a changepoint occurs, otherwise no changepoint. To decide which type of changepoint suppose for simplicity there are only 4 types, which occur with percentages 10%, 20% 30%, 40%. Generate another random number between 0 and 1. If it is <0.1 changepoint is of the first type, if between 0.1 and 0.3, then type 2, if between 0.3 and 0.6 type 3, otherwise type 4. You can combine the two steps (simpler?) and generate a single random number – if it is less than 0.005, you have a type 1 changepoint, between 0.005 and 0.015 type 2, between 0.015 and 0.030 type 3, between 0.030 and 0.050 type 4, and greater than 0.05 no changepoint.
I can think of a couple of variants on this basic procedure. First, as it stands you can only have one type of changepoint at a particular time point. You could allow the possibility of more than one type simultaneously by generating 4 random numbers rather than one in my simplified example. In that example, if the first random number is less than 0.005 type 1 occurs, if the second is less than 0.01 type 2 occurs, if the third is less than 0.015 type 3 occurs and if the fourth is less than 0.02 the fourth occurs.
The second variant would allow the probabilities of changepoints to vary with time, for example seasonally. The probabilities would presumably also vary with country/region

2) All changes can either be permanent (apply to all data prior to changepoint) or temporary (apply only to HSP):
STATION X

____PRESENT

STATION X WITH ONE PERMANENT BIASED BREAK IN THE POSITIVE DIRECTION (makes earlier period more negative relative to present) e.g. stevenson screen to AWS

```
                                                  ________PRESENT
1____________________________________1
```

STATION X WITH TWO PERMANENT BIASED BREAKS IN THE POSITIVE DIRECTION e.g. Time of observation bias

```
                                                  ________PRESENT
                1_______________________1
2__________2
```

STATION X WITH TWO PERMANENT BIASED BREAKS IN THE POSITIVE DIRECTION AND A PERMANENT RANDOM BREAK IN THE NEGATIVE DIRECTION (non-platform - applies to whole period prior)

```
                                                  ________PRESENT
31              ----------------3_____________1
2-----------------2
```

STATION X WITH TWO PERMANENT BIASED BREAKS IN THE POSITIVE DIRECTION AND A TEMPORARY RANDOM BREAK IN THE NEGATIVE DIRECTION (platform with length until next changepoint)

```
                                            ________PRESENT
1              3----------3______________1
2__________2
```

3) Things we've agreed on:

- Changepoints occurring in periods of missing data forced to occur at beginning of missing data period for ease of assessment
- Changepoints allowed to stack on top of each other - multiple changes can happen at once - considered as one changepoint for assessment though

KW: This may not be possible depending on how numbers and locations of changepoints are assigned

- Changepoints allowed to occur close to each other and within first and last two years of data - realistic problem.
- No random degradation of the data to mimic poor quality data - assume everyone has (will in the real world) conduct a reasonable standard of quality control
- QUALITY RATING: Probability of being a terrible to excellent station/country grouped from 1 to 5 - 1= no breaks, 2=very few breaks, 3=moderate breaks, 4=quite a few breaks, 5=terrible

PT: Arguably some of the more actively managed networks will have more frequent breaks. Paradoxical?

KW: Good point, annoying.

VV: We could also implement the quality as a continuous random variable that determines the break frequency and magnitude.

- Apply changepoints in 'reverse' because the homogenization process (invariably?) homogenizes relative to most recent segment?
- Specify the size of a break (mean) using a Gaussian distribution with mean and st dev specified for that type

PT: Are these things truly normal in all cases or is this a necessary evil assumption to make the whole problem tractable?

KW: Normal is nice! Mean and standard deviations may be sufficient here.

VV: The NOAA study on breaks know in meta data showed that averaged over all break types, the normal distribution is quite good. For individual break types we have no information. The lack of information could be a reason to assume a normal distribution, for its simplicity.

- Specify seasonal cycle shape (sine curve) with a mean and st deviation based on type of inhomogeneity - could be of opposing directions or same direction across year or only applicable for part of the season.
- Probability of abrupt temporary break occurring simultaneously with a permanent/biased break - not specified but it is allowed to happen?
- Probability of metadata being present? This may link to 'quality rating' of station/country.

PT: In realistic worlds also not just availability at issue but whether in some machine readable format of course. Realistically until we get our house in order in the real world will be hard to make use of more than US metadata.

VV: Currently mainly linked to the use of English in country.

4) Blind World Reminders:

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.

#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.

5) overview_error_worlds Table:
See also:
https://docs.google.com/spreadsheet/ccc?key=0Al6ocsUAaINSdHVKbVBoLWxjQU1Ib2c5bXV0RXZITHc#gid=0

		Bias ****		Random Breaks			Local Trends		
		West	Equator	Size*	Length	Seasonal Cycle	Length	% stations	Warming rate***
		°C/100yr		°C (sigma)	years	°C (sigma)	years	%	°C/100yrs
Statistical Inhomogeneities									
B1	Best guess for the west everywhere	0.2	0.2	0.7	15	0.35	25	30	1(-2 to 4)
B2	B1+ no spatially correlated CPs	0.2	0.2	0.7	15	0.35	25	30	1(-2 to 4)

B3	B1+more/smaller random CPs	0.2	0.2	80%=0.5, 20%=0.7	10	0.25	25	50	0.5
B4	B1+fewer/larger random CPs	0.2	0.2	1	20	0.5	25	10	2.5
Physical Inhomogeneities									
B5	Best guess everywhere	0.2	1	0.7	15	0.35	25	30	2
B6	B5+more/smaller bias CPs	0.2#	1	0.7	15	0.35	25	30	2
B7	B5+different background climate	0.2	1	0.7(biased:0.35)##	15	0.35	25	30	2
B8	B5+ national/exotic CPs###	0.2	1	0.7	15	0.35	25	30	2
B9	B5+seasonalAuchmannBronnimann	**	**	**	**	**	25	30	2
B10	B5+complex seasonal	0.2	1	0.7	15	0.6	25	30	2
O1	No inhomogeneites	0	0	0	0	0	0	0	0
O2	Only large CPs	0	0	-1 to 1	15	0	0	0	0
O3	No seasonal cycle	0.2	0.4	0.7	15	0	25	30	2

VV/KW: O4: No missing data periods inside

Allow CPs in first and last two years

* The sigma of the unbiased breaks and the sigma of the random part of the biased breaks

** What equations produce

VV: Auchmannn and Brönnimann would only produce the biased breaks, we would still need aditional random ones, would suggest to do so like the default word.

*** Average warming rate in °C per century during affected period (average over benchmark will be smaller; every station will have its own value)

**** Bias in °C per century

# Average bias per century the same, but double number of biased breaks with half the size of bias.

## Size of random breaks 0.7, random part of biased breaks 0.35°C.

### Exotic world has different inhomogeneities in many countries, the other countries will have the default values, listed here.

5) IJ: Some notes on Poisson, geometric, exponential and other distributions

The confusion here is partly because I was thinking of time as continuous. Of course, given that data are monthly, there is only a finite number of months in a station's time series, so things are discrete, not continuous. However, if the number of months is large it may be convenient to consider time as continuous, and move any chosen change time to the nearest month, for example.

Continuous time – the Poisson process. If changepoints for a station are equally likely to occur anywhere in time and are independent of each other, then changepoints follow a Poisson process. The number of changepoints in any fixed interval of time has a Poisson distribution and the time between one changepoint and the next has an exponential distribution.

Discrete time – if each month in the series has the same probability of being a changepoint, then the number of changepoints in a fixed number of months has a binomial distribution and the number of months between one changepoint and the next has a geometric distribution.

Thus, if time is considered continuous we can generate positions of changepoints using successive exponential variables. For discrete time we can use successive geometric variables, though it’s equivalent and simpler just to independently choose each month to be a changepoint or not, with a given probability. The same distributions (exponential for continuous, geometric for discrete) could be used to generate the length of a gradual change, although other distributions with shorter tails (e.g. Gaussian) might be preferred for length of gradual change. I can’t see a direct use for Poisson variables.