Benchmarking Working Group Online Minutes #18

Thursday 12th December 1pm GMT (5am Pacific Standard time (sorry Lucie), 8am Eastern US, 2pm Europe, 12am Australian Eastern time (sorry Lisa!)

#---
Attending: Kate Willett (KW), Rachel Warren (RW), Victor Venema (VV), Claude Williams, Lucie Vincent, Robert Dunn Ian Jolliffe, LIsa Alexander

Apologies: Peter Thorne (PT), Matt Menne (MM) (at least your apologists are not latent with their actions!)

Actions from last meeting:
ACTION KW: Get the pad open earlier.
DONE

ACTION KW: Utilise the pad for discussion opportunities outside of calls.
DONE - well, actually lets just use this one - edit and comment away please!

ACTION KW; Revamp Terms of Reference and circulate - discuss at next call. Finalise by Steering Committee call in January.
DONE

ACTION: Assimilate concepts paper comments and circulate next week
DONE

ACTION Matt to provide adjustments file from PHA to Peter. Peter to assess structure of changes found.
Matt DONE, Peter - Done :-)
https://drive.google.com/folderview?id=0B2kUTTI1RC0XRkVxaFhCQ2dDV2s&usp=sharing
PT: All you really need to read and take in is:
https://docs.google.com/document/d/1s1LpLjVBXNbWBAI2cuA3uhEilQTvbCETHYSJozRyZFc/edit?usp=sharing
But the directory contains 'pretty' pictures
Suggest discuss further on next call? Although you are more than welcome to discuss how plain dumb I am in my absence, obviously ...
KW: Summary is that our present estimate of ~4 permanent and ~6 temporary changepoints per century is about right - more temporary than permanent.
PT: Conditional upon the assumptions, yes, arguably evidence for an even larger ratio of 'temporary' to ‘permanent’ than 60:40. Looking at the '100 series' plots it does look like in the real world there are a number of cases where there is a change, then someone goes 'oh my giddy aunt' and fairly quickly reverses the change. Which I guess makes sense but raises the Q as to whether because the US network is relatively actively managed and analyzed the set may be atypical of the rest of the world.

KW: This sounds like the platform breaks Peter Domonkos talks about - so relatively common at least in Europe. We haven't got any deliberate occurrences of these there-and-back-again Types. We also haven't specifically added in short changepoints but by increasing the frequency we can hope to do this by accident.
PT: An analysis of return periods in theory would be possible using the breakpoint files Matt provided but I think would need to limit to say the 30 year time chunk 1975-2005 otherwise would convolve network station drop in / out unacceptably. Might be worth looking at a histogram of break gaps in the US breaks file Matt provided recognizing that this will be an underestimate of the real world break frequency because we know that the operational algorithm (pairwise homogenisation algorithm) is conservative particularly with respect to the soft centre of small breaks.
PT: I would not say that the evidence is that they are 'precisely' there and back. They are close to there and back but to model them as returning precisely to where they are would be a mistake I think. Instead they are up then down or down then up with the order of magnitudes of each step being somewhat similar.
ACTION KW: Move docs to permanent storage and link from the website:

ACTION KW: Talk to Robert about ensuring cross-correlations are not too poor initially because they are built on unclean data and that they are not eroded unrealistically when errors are added.
PENDING
VV: Station Cross-Correlations of synthetic clean world should match homogenised real world (Where the station density is sufficient for good homogenization).
Station Cross-Correlations of synthetic error world should match inhomogeneous real world
IJ: But we're producing several worlds some of which will match real world better than others.
KW true

ACTION LV to talk with CW? About retrieving seasonal cycle information from the data?
CW: Work suggests seasonal cycle is not actually a sine curve - no idea what shape it is yet. Evidence of a cycle but not sine in shape.
RL: paper suggests need at least 4 harmonics - for full cycle (not just inhomogeneities)
VV: Just looking at seasonal cycle in the inhomogeneity
RL: Probably an increase in the number of harmonics would improve the fit to the seasonal cycle in the inhomogeneities
LV: We need a percentage of inhomogeneities that are equally affected over the year and then another percentage have more change in winter vs summer (and vice versa). Keep it simple.
RL: Given variability in winter is greater would expect variance of inhomogeneity to be larger - whether size of inhomogeneity is larger or not.
LV: Need variation depending on closeness to equator and perhaps other things (elevation/proximity to water bodies etc.)

ACTION LV to send parallel data to Victor.-
https://ourproject.org/moin/projects/parallel
PENDING

ACTION: Kate to work with IJ/VV? CG on reworking the flow chart.
This is me starting to do that - please look through these discussions below Colin/Victor/Ian

ACTION: ALL TO EMAIL KATE WORDS THAT SHOULD BE IN THE ISTI GLOSSARY – OR PUT THEM IN YOURSELF:
https://docs.google.com/document/d/1xItD6yeQTxqwVnbfx-ZwUsh7hKJu1HqJVEf-OKstS4Y/edit
ONGOING - MORE PLEASE
PUT YOUR WORDS HERE
Some words that appear in the paper like e.g. GCM and downscaling
discontinuity
See RWs comments in the paper.
RW: Here they are for the record (apologies if these are really obvious to everyone else): SI, micro-climate
ACTION: KW to put the word list and words from RWs comments in Concepts paper into ISTI glossary + GCM, downscaling, discontinuity, SI, micro-climate

Actions from this meeting:
ACTION: Post meeting minutes early
ACTION: KW to put the word list and words from RWs comments in Concepts paper into ISTI glossary + GCM, downscaling, discontinuity, SI, micro-climate
ACTION: KW clarify position of maybe (but not definitely) submitting a paper on the regional summaries at some point within the Terms of Reference
ACTION KW: tidy up Terms of Reference, post on website and submit to Steering Committee for sign off.
ACTION KW: shorten Concepts paper if possible.
ACTION KW: add definitions for uncertainty, inhomogeneity, changepoint and bias into Concepts paper - make sure we're consistent in talking about it
ACTION LV to second review paper.
ACTION KW: make the equation bit clearer - explanation, XOB, D is also variance, extra equation in GCM discussion?
ACTION KW: tidy up figures 6 and 7 and improve explanation in figure caption.
ACTION KW: reference ISTI glossary in paper
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.
ACTION KW: Move docs to permanent storage and link from the website:

#--
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/probability of occurrence for applying changepoints

1) Terms of Reference

Anything to discuss on circulated draft?
Length of benchmarking cycle

- aim for release of first Benchmarks Summer 2014 (in time for SAMSI workshop)
- aim for wrapping up Summer 2016?

Change length of benchmarking cycle from 3 to 2 years?
Too short - not enough people involved
Too long - benchmarks become old and redundant

VV: These are mainly arguments for the duration during which people can homogenize the data, which should not be too long. We also need time to analyse and to build a new benchmark. If we do the same benchmark, I would argue that not much happens in homogenization in 2 years. If we extend our goals, for example go to Tmin and Tmax or to daily, or do QC and gridding, we may need more time for the development. This would require additional research to be able to make a realistic benchmark. Much of the information we would need is not available in the literature. At least as long as it is a volunteer organisation. Given people feedback on the benchmark metrics we may not need that much time, but the analysis of all the other levels including the realism of the old benchmark would take more time and is important information for the next cycle.

PT: Main issue is that as unfunded effort you have to ensure that you can turn the crank on your bits in the timescale posited. Two years feels like a lot of gratis FTE effort on the part of WG members?

KW: Happy to go with three years - this was in response to discussion of speeding up the cycle at the NCDC workshop.

PT: Beware the workshop enthusiasm syndrome where the gap to actual resources is maximal and the resulting realism of target timescales hits a distinct minima :-(

KW: Guilty!

Vote to accept Terms of Reference: KW,LVA,RW,cw,IJ, VV,rl,RD,LV, (ALL PRESENT ON CALL)

ACTION: KW clarify position of maybe (but not definitely) submitting a paper on the regional summaries at some point within the Terms of Reference

All are agreed ToR is ok

ACTION KW: tidy up Terms of Reference, post on website and submit to Steering Committee for sign off.

2) Concepts Paper sign off

Peter is happy apparently :)

PT: It was sunny and snowy at the weekend so definitely happy :-). Now grey and about 8C so hurry up before I change my mind o:-)

LV: 3. In sections 2 and 3, two equations are presented and it is important that they are well defined: Equation 1: X = C + V + M, Equation 2: XOB = C + V + M + D
VV: It looks like these equations still come from a time when some where thinking to generate the clean data stochastically.
KW: The equation is only to represent a concept, not a formula in the strict mathematical sense of the word.
VV: I wonder whether it would make the paper clearer if the equations relate to what we actually do: X = GCM + downscaling, XOB = GCM + downscaling + inhomogeneities.
KW (in response to comments from LV): I had hoped that the paper explained the equation but it obviously needs a bit more work, V is to capture all of the low frequency variation - so any long-term trend (which isn't necessarily linear) and also the interannual and interdecadal variability. This low frequency stuff can be taken from a GCM which generally does ok at representing realistic natural variability and long-term trends. M is the micro climate which will have some autocorrelation and some cross-correlation. It is the residual once V and C have been removed. With V included there is stronger autocorrelation and cross-correlation but this will be provided by the smoothed GCM time series for each gridbox.

KW: Does that help at all or just make things worse?

KW: I used XOB because I wanted it to be clear that XOB, or the inhomogeneous time series is very similar to X (the clean time series). It is the 'OBserved' time series as opposed to the true time series. Y would be fine to use but I thought XOB might be more intuitive to non-statistical types who switch off when random letters are used. I can make this justification clearer in the text, or just use Y.
RW: XOB makes more sense to me - we've got quite a few letters so it's probably best to keep what they mean as intuitive as possible

KW: D can also just be a change in variance. I should also make that clear in the paper.
KW: I would rather keep it as it is – leave the equation in, in its current form. We are explaining what we need to replicate and then one possible method of doing so using a GCM and some VAR magic. I could add an extra equation transforming Eq. 1 into our methodology.

LV: 4. In section 5, four assessment levels are presented but very little is said about levels 3 and 4. Since we don't really know now about the assessment process, and since the paper is already long, should levels 3 and 4 be removed from the paper?
VV: I thought we wanted to have a limited number of measures from level 1 as our benchmarking measures. They are the ones with which we will determine how good the homogenization algorithms are for our stakeholders. These should be mentioned in some detail. I do think it is important to also mention that there are many more reasons to analyse the homogenized data and that we are interested in many more measures and indicate in which direction we are thinking there.
KW: I would quite like to keep the description of levels 3 and 4 in just because they are quite important in terms of getting the most value out of the benchmarks. As we

don't have the time/resources to focus on them it is nice to try and encourage others to do so as a research project.
ACTION KW: shorten Concepts paper if possible.

KW: Can we show figures 2 and 8? I think we can as long as they are referenced correctly but we could just put a link to the webpage they appear on and the reference/figure number?
VV: Probably have to ask for permission, otherwise would not expect a problem.
PT: Both are covered under governement copyright and you can use as you wish surely?

KW: Suggestion from LV to state our specific definition of uncertainty, inhomogeneity, changepoint and bias up front in the paper. I really think this is a very good idea - a little unconventional perhaps but it is so important that we communicate as clearly as possible. I will add these definitions into the text of the paper.
ACTION KW: add definitions for uncertainty, inhomogeneity, changepoint and bias into Concepts paper - make sure we're consistent in talking about it
changepoint is a date, inhomogeneity is a size/shape of change, bias - do we really mean bias?
IJ: 2 instances of uncertainty does need clarity checking.
ACTION LV to have a look over Kate's 'final' edits of the paper prior to submission to internal review
ACTION KW: make the equation bit clearer - explanation, XOB, D is also variance, extra equation in GCM discussion?
ACTION KW: tidy up figures 6 and 7 and improve explanation in figure caption.
RW: Could we also reference our glossary somewhere in the paper (once it's finalised)?
KW good plan
ACTION KW: reference ISTI glossary in paper

LVA query over journal - who are our readers? Is this journal likely to reach those people?
http://www.geoscientific-instrumentation-methods-and-data-systems.net/
KW: Perhaps not ideal but it is open source and searchable from Web of Science/Google Scholar - onus on all of us to tell the world about the paper: Twitter?Blog? Presentations at international conferences
Internal review necessary for KW, LV, CW - try to get it out around Christmas then submit mid-Jan

3) How many changes to apply and where to apply them:
NB. (see also Notes 1. below)
PERMANENT changepoint = a change applied to the entire record preceding the changepoint
TEMPORARY changepoint = a change applied only to the homogeneous sub-period (HSP) in question

This is needed for permanent abrupt changepoints (layer 2a) and temporary abrupt changepoints (layer 2b)
PT: We really need to see what the US PHA results show about relative propensity before we can winnow this down I think. We know hypothetically both types are possible the question is to what extent are the two types prevalent in real-world observations (at least in the US).
KW: Ok but we can still sort out how these numbers/locations will be allocated
PT: See above links from prior actions. Looks like assuming somewhere 50-80% temporary is 'reasonable' if US stations are a viable proxy for the rest of the world.

Option 1: assign a number of change points and then assign their location - exponential or binomial for number, geometric for location e.g.

- Use a binomial distribution with a mean of 4 PERMANENT changepoints per century
    - Use a geometric distribution to allocate these over time
- Use a binomial distribution with a mean of 6 TEMPORARY changepoints per century
    - Use a geometric distribution to allocate these over time
- For 30% of stations in a region assign 1+ (mostly 1 but can be more) GRADUAL changepoints
    - Use a geometric distribution to allocate these over time

PT: You need to get some good stations and some awful stations so not just mean return but distribution from good to rubbish between stations ideally. You want to pre-condition some types of changes. We know automated (semi-automated) measure transition occurred largely since 1980s, we know early period standardization and we know propensity to move to airports etc. Can you model this in option 2 as a combination of the flat prior with a second set of probabilities that have sharp maxima?
KW: Good to rubbish stations would be governed by the quality level assigned to that country, for that world. e.g,

QL 1 = excellent quality = 0-20% of stations contain changepoints (could also reduce frequency of changepoints per century)
QL 2 = good quality = 20-40% of stations contain changepoints
QL 3 = moderate quality = 40-60% of stations contain changepoints
QL 4 = poor quality = 60-80% of stations contain changepoints
QL 5 = terrible quality = 80-100% of stations contain changepoints (could also increase frequency of changepoints per century)
I would probably give most countries QL3.

KW: Pre-conditioned types of changes come under layer 1 - the clustered/known changes for each country. Within this there will have to be some allocation of changes across a small spread of years. So the discussion here is still relevant - e.g, the case of automation in the USA over the 1990s. For the isolated changepoints, using the Type Shelter1 (described below) would add in occurrences of Stevenson Screen to AWS in a more adhoc manner but only from the 1970s onwards (or whatever date we choose the probability of occurrence to be greater than 0).

VV: Maybe just personal preference, but I would find continuous changes in station quality more elegant than a large number of classes. How about setting the global frequency of breaks to (e.g) 6 per century. Then draw a random number from N(0,1) to perturb this frequency per country and again draw a random number from N(0,1) to perturb this frequency per station?
KW: Ok - sounds doable - so most countries would have zero perturbation and a few would have a very large reduction/increase in frequency.
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.
KW: Is Option 1 what was played with at the NCDC workshop? We looked at allocating a location for several gradual changepoints over time using some r code for geometric?

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 0.005 of having a TEMPORARY changepoint if all months are identical

KW: Or is Option 2 what was played with at the NCDC workshop? We looked at allocating a location for several gradual changepoints over time using some r code for geometric?
KW: Colin/Robert/Ian/Thordis and Victor (and anyone else too for that matter - sorry) - please help. You had some great ideas about coin flipping and super simple ways of doing this.
IJ: There's an awful lot of stuff here and on the flow chart, spreadsheet etc. and I'm feeling pretty swamped by it all, especially as new stuff keeps appearing.
This is my attempt to isolate the main points. It seems there are three aspects to corruption: (1) identifying where changepoints will occur; (2) deciding which type of changepoint has occurred; (3) determining the size and (where relevant) duration of a change. I'll confine my comments to Stages (1) and (2) which could be done separately or together. For either I think the way to do things is the simple procedure based on Kate's Option 2, which is closely related to choosing a geometric distribution for gaps between changepoints, but doesn't need to invoke that or any other distribution. 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.

PT: This is pretty much exactly what I did in creating the USHCN breaks that are described in Williams et al., 2012. Except I had several loops to account for different break types - e.g. I used a restricted time horizon window of 1982-2000 (ish) to define the location of an 'MMTS' break and applied it to 70% of stations (not allowing a station to be selected twice). As you say this is easiest done just by creating effectively random number vectors and using occurrences over / under some threshold that gives the desired exceedance from these to sample locations. For some breaks a flat prior is fine, for other types we will need to create more restricted window priors either by multiplying by a second probability (square wave, peaked, gaussian) or by creating several prior vectors to cope with this. But its really just a perturbation.

IJ: 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.

IJ: 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

VV: This would be my favourite option. Simple and we have no empirical evidence that breaks are more complex as such a Poisson process.

4) I'm still advocating changepoint types - here is my effort to convince you.

VV: To convince me, you would need to 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?

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 (with subtypes) 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 global 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/small change?, strong seasonal cycle (not biased to be warmer?) 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 through 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.
KW: Ah ok - have crossed out the 'no'

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).

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 of occurrence of each Type and subtype:

PT: I am a bit lost here as there are lots of %ages and I'm not sure what is a percentage derivative of another %age. Any chance of clarifying for those of us with little brain like myself?

KW: Sorry - I mean % chance of occurrence - so for any assigned permanent changepoint there is a 50% chance that it will be Type Shelter, 20% change it will be Type Time and 30% chance will be Type Move. Should I have done this in probability - so 0.5, 0.2, 0.3?

PT: Do you mean that these are compound probabilities? e.g is there a 100% chance of an instrument change post-1970, a 50% chance or some time integrated probability function. e.g. what proportion of the network are you intending to apply Type shelter 1 to? I could presently infer 50%, 100% or some (unknown) number <50% and that ambiguity makes it a little hard to provide salient input.

KW: For any assigned temporary changepoint there is a 40 % chance it will be given a Type Instrument and a 60% chance it will be given a Type Move (of which Type Move4 is the only temporary Type Move so it is given 100% of the 40% - sorry, confusing I know.

Type Shelter = 50% of permanent changepoints

Type Shelter1 = 100% for post 1970 of Type Shelter changepoints

Type Shelter2 = 100% for pre-1970 of Type Shelter changepoints

PT: 100% seems unduly large. We know only 70% of US network moved to MMTS for example.

KW: 100 % of all asigned Type Shelter changepoints which make up 50% of permanent changepoints

PT: Okay, as a conditional probability it makes more sense :-)

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

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
Type Gradual2 = 30% of Type Gradual changepoints
Type Gradual3 = 30% of Type Gradual changepoints
NB. Higher probability for abrupt change at end of Type Gradual1?

KW: I think 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?

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, its 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.
2. 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.
3. 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.
4. 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.
5. 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 magnitude 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!
6. 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.
7. Why change bias for Equator in world O3?
VV: Do not know anymore, maybe to be different from blind worlds?
8. Why have two distributions for the global random size/shape for world B3? (cell E8 on spreadsheet)
9. Why have a biased part of the random changpoints 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 – here are a number of processing steps:

1) Set up Type settings (size/shape) for each world/countrye
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 1(excellent - 0% of stations have changepoints) to 5 (terrible - 100% of stations have changes) – or scaled as suggested by VV
5) Layer 2a permanent (non-zero mean (Types Shelter, Time and Move) with some zero-mean (Type Move4))
  - how many changes from 0+, centred on ? (binomial/exponential?/option2)
  - locate these changes? (geometric/option 2)
  - for each change assign a Type based on probability
  - pluck a size/seasonal cycle from the Type distribution and apply to all pre-changepoint data
6) Layer 2b temporary zero-mean (Types Move4 and Instrument)
  - how many changes from 0+, centred on ? (binomial/exponential/option2)
  - locate these changes? (geometric/option2)
  - for each change assign a Type based on probability
  - pluck a size/seasonal cycle from the Type distribution and apply to HSP
7) 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 or 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.

5) AOB
More ISTI glossary terms?

6) Next Meeting: 18th of December

#--
Notes:

1) 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
```

2) 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.

3) 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.

4) 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

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.