Benchmarking Working Group Online Minutes #20

Monday 17th February
1pm Greenwich Mean Time
8am Eastern Standard Time
2pm Central European Time
12am Australian Eastern Daylight Time - sorry Lisa!

Aims:

Confirm shape (variety of shapes) for gradual inhomogeneities
Confirm shape (variety of shapes) for abrupt inhomogeneities
Confirm things that we will ask benchmark users to return to us for validation
List of potential reviewers for concepts paper.

#---

Attending: Victor Venema (VV), Rachel Warren (RW), Kate Willett (KW), Ian Jolliffe (IJ), Thordis Thorarinsdottir (TT), Lisa Alexander (LA)

Apologies: Peter Thorne (will try to make next one - sorry) Enric Aguilar (in and out with students and lectures. Will try to follow the etherpad)

Actions from last meeting:

ACTION VV: Prepare slides/presentation on slope shape for gradual inhomogeneities for next call.
DONE
ACTION KW: Kate redo the spreadsheet with character. - share
NOT DONE
ACTION KW: Doodle poll and minutes (past and present)
DONE
ACTION: KW contact enric/richard about gamma filtering for Team Creation
DONE
Thanks to Richard Chandler. Kate and Richard to meet up and try to code this some time soon.
VV: I would need some more introduction to understand the text below. What is the problem to be solved?
PT: I think its to do with how to get realistic large area congruence of statistics in the 'clean' worlds whereby the regional patterns are suitably correlated?
KW: Correct - I can't run a 30000by30000 matrix. When creating the Clean World stations using the VAR code we need to break the stations down into small groups to run but then there will be issues of cross-correlations being far too low at group boundaries. The Gibbs Sampler is a way of smoothing over space I think - hopefully without over-smoothing and removing the tails of the distribution.
Gibbs Sampler
1. Split the N locations within a neighbourhood into M groups, each of which contains n<<N locations ("<<" is "considerably fewer than")
2. Choose some initial values at each location, for example by generating random numbers independently from normal distributions with the correct mean and variance. (Our VAR code?)
RW: Sorry if this is a silly question, but initial values for what?

PT: Again, my assumption here, but I think its perturbations to the underlying model-based field to represent geographical departures and instrumental noise?
I think the initial values are our initial synthetic versions of the standardised anomalies created using the VAR code. PT: As I said ... in non-tech terms my little brain can cope with :-) KW: Yup

3. Loop over all of the M groups: for each group, calculate the joint distribution of the temperatures within this group *conditioned* on the current configuration of temperatures in the neighbourhood (see below) of the group; and replace the current temperature configuration there with a sample from this joint distribution.
4. Repeat step 3 until the joint distribution has converged (see below).

neighbourhoods:
The idea here is that if you're working with a particular group, you can draw a "boundary" around it such that, if you know the temperatures for all locations within that boundary, the temperatures *outside* the boundary become irrelevant to all intents and purposes. Again, you may need some experimentation to determine this; but hopefully the intuition is clear. The formula for the conditional distribution given the values in the neighbourhood is, as always, on Wikipedia: http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions. Although there are scholarly alternatives to that as well ...

how to determine convergence:
The usual way of doing this is to produce "trace plots" of appropriate summary measures i.e. to plot the value of a particular summary measure (of the joint distribution) against the iteration number. The obvious summary to compute is the (log)-likelihood, but this itself requires N^3 operations. My inclination would be to take a sample of locations around the globe and to compute the log-likelihood at these locations (considered simultaneously) as a summary measure. If the sample was taken in such a way that for each site, there is at least one other site in the sample within a distance D (to be determined), then this would be a good way to ensure that the pseudo-realities have a realistic spatial correlation structure.

KW: Thordis is familiar with this method!
VV: Query whether we need VAR at all in this case?
KW: I believe that we do as we need a 'near-perfect' initial data to start it and make sure it converges to something vaguely similar to reality. If we just initialise it with white noise then we are likely to end up with something that does not have the right autocorrelation or cross-correlation.
TT: Try this out on Australia - an isolated region.
KW: May need somewhere with a little more data but point taken.
ACTION KW: Work with Richard Chandler and Thordis to get this working over at least a single region.

SUMMARY:
We will attempt to apply a Gibbs Sampler to provide spatial consistency on the global scale using smaller clusters of VAR reproduction of standardised anomalies to prevent computer meltdown due to very large matrices.

ACTION IJ: to send some emails and think about key topic for next call.
DONE

Actions from this meeting:
ACTION KW: Kate redo the spreadsheet with character. - share
ACTION KW: Work with Richard Chandler and Thordis to get this working over at least a single region.
ACTION: CW, LV, MM, VV - ultimately to decide/discuss on what we want to do. Include jumpy graduals? Do not include jupmy graduals? Explore in World 8 (exotic IH) only?
ACTION: TT to send around an email to help to explain variance verification problems.

Summary from this meeting:
SUMMARY Team Creation:
We will attempt to apply a Gibbs Sampler to provide spatial consistency on the global scale using smaller clusters of VAR reproduction of standardised anomalies to prevent computer meltdown due to very large matrices.

SUMMARY 1) Shape of changes for gradual inhomogeneities:
We do not know anything about the shape of gradual inhomogeneities.
CW/LV/MM/VV to decide/discuss how to proceed
a) Reproduce gradual inhomogeneities as a linear trend (with and without seasonal cycles) - some jumpiness (up and down) will occur anyway due to overlaid abrupt changes
b) Reproduce at least some gradual inhomogeneities as jumpy slopes (with and without seasonal cycles)
c) Use World 8 (exotic IH) to explore jumpy gradual inhomogeneities.

SUMMARY Team Validation issues:
Assess best guess changepoint date only
Use a weighted window for classifying a 'hit' around the true changepoint location that peaks at the changepoint date, reaches 1sigma around 6-12 months to reward exact hits more than near-hits.
Large inhomogeneities are easier to find - should weight such that smaller ones are rewarded more or assess different sizes seperately (small/medium/large).
Reward provision of uncertainty ranges (for changepoint location and inhomogeneity adjustment size) in a 'Strengths of algorithm' section on the assessment report.
Provide for optional return of uncertainty ranges when designing format for returned statistics (dates etc).
Problem with gaps - are the inhomogeneities or are they not? Berkeley treats them all as IH so would get a high false alarm rate in our validation.
Problem with validating adjustments for gradual inhomogeneities when some/most/all apply abrupt shifts.
- validate the absolute change applied across each homogeneous sub period/gradual change period

- have one world with no gradual changes (e.g, World 9 - this as seasonal cycle driven by solar/humidity changes which is a little complex to achieve)
- have one world with identical abrupts to world 9 and additional gradual change (e.g, World 5 - best guess everywhere)
- compare results between the two worlds to see how well gradual changes are dealt with.

SUMMARY What should benchmark users provide?:
Ask people to return all stations
Validate dates (start and end for gradual) and size (flat/linear change calculated by us based on provided dates and homogenised series) on station level
Validate climate characteristics on regional level
We will calculate the regional statistics and size of adjustments

SUMMARY potential reviewers for Concepts paper:
Richard Cornes (CRU)
Blair Trewin
Ben Santer
Douglas Maraun
Josef Steinebach

#----------

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) Shape of changes for gradual inhomogeneities (VV)
Linear or jumpy trends for reproducing gradual inhomogeneities - see document Linear_or_jumpy_trend_inhomogeneities.pdf

Jumpy trends - just adding a noisy value using the power law as opposed to a constant value - always jumping in direction of trend.

KW: Jumpy verses smooth - jumpiness may occur in some cases because we allow abrupt changes to happen during the gradual changes. We don't really know how common/uncommon this will be in our error worlds though - just random. I suppose this will also result in a jump in the opposite direction.
KW: Constrained jumps? Constrained by what? How do we know?
IJ: Any difference between jumpy and random abrupt?

VV: Jumpy are more common
KW: Balance between reality and ease of validation
KW: Could just add jumpy graduals into World 8 as an exotic inhomogeneity.
ACTION: CW, LV, MM, VV - ultimately to decide/discuss on what we want to do. Include jumpy graduals? Do not include jupmy graduals? Explore in World 8 (exotic IH) only?
KW: How would jumpy graduals be validated? Still as absolute change over the period?

SUMMARY:
We do not know anything about the shape of gradual inhomogeneities.
CW/LV/MM/VV to decide/discuss how to proceed
a) Reproduce gradual inhomogeneities as a linear trend (with and without seasonal cycles) - some jumpiness (up and down) will occur anyway due to overlaid abrupt changes
b) Reproduce at least some gradual inhomogeneities as jumpy slopes (with and without seasonal cycles)
c) Use World 8 (exotic IH) to explore jumpy gradual inhomogeneities.

2) Shape (seasonal cycle) of changes for abrupt inhomogeneities (if its not already 30 minutes in)
Passed over for item 3)

3) Results to be used for Validation (IJ)
IJ (in mail): Apparently at least one algorithm does do this for break point position, but it seems unlikely that this is something we will want to validate. Rather, varying the width of window for which position is deemed to be correctly located will serve a similar purpose.
VV: I do not understand the second sentence.
IJ: Consider (a) an algorithm says there is a break at time t - we say the algorithm is correct if the true location is in the interval t +/- w; (b) an algorithm says it is confident that there is a break within an interval t +/-w - we say the algorithm is correct if the true location is indeed in that interval. What I meant is that (a) and (b) are pretty much the same. Choosing a small w in (a) implies that we think algorithms ought to be able to confidently give a narrow interval in (b). Choosing a larger w in (a) means that we think that the algorithms would need to give a wider interval to be confident of including the true location.
PT: Alternative here is where data providers explicitly give a window and pdf (b) to use that returned set of values. Otherwise we take the deterministic breakpoint location (a) they gave and treat it as probabilistic for the purposes of a skill assessment by overlaying a normally distributed PDF with sigma width 12 months (or whatever half width you wish) renormalised to be unity at peak and centred upon the deterministic break location. Implication is that we expect a 95% probability of catching a break within 24 months (2 sigma) of its true occurrence? This has the advantage of at least giving some greater value to algorithms that consistently get closer than some 1/0 square wave filter approach that rewards being 12 months out the same as getting it 'right'?

IJ: Interesting thought - but I need to give it more thought. A question: is +/- 24 *months the sort of accuracy we think is realistically achievable. As an outsider it seems pretty unambitious.*
PT: Yes, its not greatly ambitious, but then if we get it matched with 24 months some peaked skill distribution although recognizing the 'detection' is marking it down to virtually zero skill compared to something that is much nearer the right location. Could go anything between 6 and 12 months for the 1 sigma range if normally distributed and make a reasonable case I think - no right answer. Certainly not matching within 24 months is the limit of reasonably acceptable match skill!
IJ: OK, thanks for the clarification
VV: If the SNR is sufficient, the detection will have almost no error. As soon as it gets difficult, the uncertainty grows very rapidly. There is not much in between.
PT: Well if you wanted to actually assess by something like SNR you could renormalise the pdf so at peak its ABS|(break mag)| - that way you penalize getting large breaks (high SNR) much more if they are wrong.
VV: We should definitely take the size of the breaks into account. Otherwise we would penalise methods that try to detect smaller breaks as well and the scores would look good for simple methods that restrict themselves to the most obvious breaks.
VV: On the positive side, even with breaks at random positions, you can already make a quite good homogenization. Thus missing the true position by some years is likely not a big deal.
KW: Its nice if algorithms provide a range for their detected changepoints. However, from a validation point of view it may be best/simplest/fairest to just validate their best guess changepoint. This is because this will have been the point at which any adjustment has been made. We can reward the provision of date uncertainty and size uncertainty in other ways. In the overall assessment algorithm attributes that are particularly desirable could be listed e.g., uncertainty estimates (data, size). Tricky because something may seem desirable such as 'ability to detect and adjust for gradual inhomogeneities' but actually if this is not done very well then it could make the algorithm worse. For this reason, desirable attributes should be things that can only ever be good. I'm not sure what other things this would include though - lists of neighbours for each candidate station?, automation/reproducibility, ...
VV: Could encourage return of uncertainties by setting up a return file format with space for providing uncertainties in changepoint and inhomogeneity.
TT: We need their best guess that was actually used for adjustment.

ZH (in mail): : In the Berkeley approach a gap of more than a year (> 12 continuous months) is considered a breakpoint. I believe it was a somewhat arbitrary choice, and as far as I know we haven't done any explicit analysis of how long a gap is optimal, but I'll double check with Robert Rohde. Robert added that: "The RMS shift at gaps is 0.95 C, at documented moves 0.60 C, at documented TOB change 0.58 C, and at empirical cuts 0.76 C. I should note that is a break can be justified for multiple reasons (e.g. a gap and a move at the same place)
VV: If the scalpel is already used to split a time series because of a gap, is there still a statistical test for a break point? In other words are multiple causes possible for a splitting? Otherwise we may have a problem, if we treat every split due to missing

data as a break point (like Berkeley does), they would likely have a quite high false alarm rate (all the gaps that are not a break). If we only validate the breaks that Berkeley sees as breaks (in the continuous parts of the series), they would have lots of misses (the breaks that were in the gaps). We probably should talk to Berkeley if they want to produce special output for the ISTI. Otherwise the former is most consistent, as Berkeley treats such gaps as breaks.
PT: Most people treat gaps of >12-24 months as breaks and that is probably a reasonable thing to do in reality as was discussed on the last call etherpad.
KW: Do they? We don't. I don't think PHA does.
VV: At least it is new to me. It does sound like a good idea for the future.
VV: Given that the RMS shift at gaps is even larger than the RMS shift for the other categories, it seems as if a gap is really a strong indicator of a higher probability of a (strong) breaks. Could that be a reason to reconsider our decision at the last call to keep the break probability fixed for breaks during a missing data period?
EA: How much? If we put the breaks over the complete dataset, then mask out missing values and finally assign a break in the missing section to the first non-missing value, the probability is already increased. Also, some missing periods will be due to causes which are not related to breaks, but to databank management and digitization issues.
*VV: That is indeed a difficult question and the main reason we decided not to increase the break frequency. The number of breaks that happen "automatically" is quite small, which homogeneous subperiods of 15 years, the probability of a break in one year of missing data would be 1/15. The large breaks founds by Berkeley in such gaps suggest the real value is much higher. I do not know the average length of the gaps that are longer than 1 year, if I would guess 4 years, we could have a break probability per year of 1/4 or as implemented per month of 1/(4*12).*
VV: A difficult and messy point for the validation will be what to do with the inserted and also detected breaks in the periods where we inserted a gradual inhomogeneity. Because most (probably all) contributions will homogenize such gradual inhomogeneities with a number of breaks of increasing size. Thus validation of breaks in these periods is somewhat weird. On the other hand, that would penalise almost all contingency scores equally (except if there would be contributions that actually use gradual in homogeneities).
KW: If we have at least one Blind world with no gradual changes, with a sister world that has the identical abrupt changes but additional gradual changes then this could be assessed quite nicely. We could swap B8 (exotic inhomogeneities) or B9 (insolation and humidity driven seasonal cycles) as these are more complex to create than the others and run a B5 with no gradual changes instead.

SUMMARY:
Assess best guess changepoint date only
Use a weighted window for classifying a 'hit' around the true changepoint location that peaks at the changepoint date, reaches 1sigma around 6-12 months to reward exact hits more than near-hits.
Large inhomogeneities are easier to find - should weight such that smaller ones are rewarded more or assess different sizes separately (small/medium/large).

Reward provision of uncertainty ranges (for changepoint location and inhomogeneity adjustment size) in a 'Strengths of algorithm' section on the assessment report.
Provide for optional return of uncertainty ranges when designing format for returned statistics (dates etc).
Problem with gaps - are the inhomogeneities or are they not? Berkeley treats them all as IH so would get a high false alarm rate in our validation.
Problem with validating adjustments for gradual inhomogeneities when some/most/all apply abrupt shifts.
- validate the absolute change applied across each homogeneous sub period/gradual change period
- have one world with no gradual changes (e.g, World 9 - this as seasonal cycle driven by solar/humidity changes which is a little complex to achieve)
- have one world with identical abrupts to world 9 and additional gradual change (e.g, World 5 - best guess everywhere)
- compare results between the two worlds to see how well gradual changes are dealt with.

Thanks very to all those who replied to my email on 'What will developers provide?' Some of the replies surprised me and I learnt a lot. I give below my original email, in italics, with summaries of your responses interleaved. I have a version with all your complete comments included if anyone wants to see it.
To summarise in a couple of lines, I think we want to validate positions of breaks (abrupt and gradual) at station level, climatology (mean, variance, spatial covariance, autocorrelation, trend, seasonal cycle) at regional level, and the homogenised series at both levels. If other things are produced by developers or derived ourselves we might consider how to validate them, but only as a lower priority.
Information given by station:

1. A possible required list is:

- homogenised series

- position of abrupt changes (with 2sigma(?) uncertainty range if available)

- position of gradual changes (start and end) (with 2sigma(?) uncertainty range if available)

- size of abrupt changes (with 2sigma(?) uncertainty range if available)

- size/shape of gradual changes (with 2sigma(?) uncertainty range if available)

- size of change between start and end of any missing data period greater than a given length*.

*Following our discussion on the last call regarding missing data, presumably there is no way of being more specific about what has happened in such a period and, even if there was, there are no real data to compare with during the period

There seems to be agreement that we only want the homogenised series, the positions of abrupt breaks and the start and end points of gradual changes. There is a suggestion that if we want the sizes of breaks we can estimate these ourselves, given the error world series and the homogenised series.

2. The list implies that a single number will be given for position and size. It would be much better practice to give a range of values (confidence or prediction intervals), but this would probably make the validation less straightforward. Are any of the algorithms likely to do this?

Apparently at least one algorithm does do this for break point position, but it seems unlikely that this is something we will want to validate. Rather, varying the width of window for which position is deemed to be correctly located will serve a similar purpose.

3. Do we ask, in addition, for climatology (mean, variance, spatial covariance, autocorrelation, trend) for stations? Which of these do we want to validate for stations?

If we want to validate some, do we calculate them ourselves, using the same methodology as for the clean worlds? This is a general point, mentioned again in a specific context below - the bulleted list above has things that can only be given by the developers. But, in addition, there are also a number of 'derived' quantities of interest (climatology, regional series). Do we want to ask developers to produce these, or do we want to construct them ourselves, or both?
It looks as if the emphasis on validation for these 'derived' quantities (an extra one in addition to those listed above is the seasonal cycle) will be for regional aggregates.

4. Are we really going to try to validate for every individual station?

Most breaks are station specific so the position of breaks is inevitably a station level problem. For the climatology parameters, reservations have been expressed about doing it for stations, but it's not impossible.

Information given by region:

1. Repeating the point in 3 above, do the developers provide regional series, or do we calculate regional quantities from station information provided by the developers, or perhaps both? Different developers might use different ways of calculating regional series from station series. Do we need to decide in advance, and announce, how we will derive regional series from our clean world station series?

There seems to be agreement that we should produce our own regional series.

2. What do we want? Simply the regional series and their climatologies, or might there be breaks that occur across a whole region for which we would want information similar that in the bulleted list above?

No-one has suggested that we will be looking at regional breaks.
Mean? Nonessential
VV Most important things would be trend and decadal variability
IJ: Temporal behaviour of homogenised series should track behaviour of the clean world
KW: Auto-correlation and spatial cross-correlations may not actually be that important to validate
TT: We need to know how well we can validate certain things - how do we rank methods by their skill in getting the variance right?
TT: Problems with saying that any method is better than another based on the variance. Can look at the second moment.
IJ: Second moment about the origin.
TT: Its because you need a consistent baseline - the mean may not be consistent but the origin is. E.g. the mean may be wrong so how do you validate the variance?
ACTION: TT to send around an email to help to explain this.

3. The concepts paper says that developers may be asked to produce climatology estimates for their regional series, but again there is a danger that different developers will use different estimation methods. We should probably do this ourselves, even if some developers do it, for consistency.

I think there is agreement on this.

4. One problem identified in the paper is whether all stations (or a subset of relatively complete stations) are used in averaging.

No comments on this.
KW: Do we need to specify a specific subset of stations that have to be returned?
VV: If a station is not returned then it will penalise the returned regional quantities - keep the region station network as set and compare the regions stats on full clean verses homogenised subset.
VV: Really need everyone to homogenise all stations but then just identify the stations they would not choose to include in a series?
KW: Could have a chosen continent for validation?
VV: Could still have short/too horrible stations
LA: Better to encourage more rather than less.
KW: SAMSI workshop is a good test bed for what people can return.

SUMMARY:
Ask people to return all stations
Validate dates (start and end for gradual) and size (flat/linear change calculated by us based on provided dates and homogenised series) on station level
Validate climate characteristics on regional level
We will calculate the regional statistics and size of adjustments

VV: What about assessing gradual changes - when some adjust with abrupts?
KW: Could swap world B9(seasonal cycle with insolation/humidity) for a B5 with no graduals.

TT: three issues - time, size, abrupt/gradual - can validate all three things.
VV: In reality there are probably no groups that will adjust a gradual change with a slope.
VV: Just focus on dates and size.
KW: agree for now just to quantify absolute change over the period (abrupt or gradual). As and when homogenisers are able to apply slope changes we can modify assessment to cope with that - later cycles.

4) AOB
KW: Concepts paper: Who to suggest for reviewers - homogenisation experts, climate monitoring experts, statisticians,?

5) Next Meeting: Mid-February PT: Think you mean early March? :-) Early March! - thanks Peter.
Team Creation - are we there yet?
Team Corruption - who/how to take this forward
Team Validation - validation of climate characteristics - mean, variance?, trends etc?

#----------
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.