Benchmarking Working Group Online Minutes #16
Friday 15th November 2pm GMT

Attending: Kate Willett (KW), Peter Thorne (PT, cough willing - will mute), Victor Venema (VV), Thordis Thorarinsdottir (TT), Claude Williams(CW), Ian Jolliffe (IJ), Matt Menne (MM)

Apologies: Rachel Warren (RW), Enric Aguilar (EA), last minute issue)

Actions from last meeting:
KW: Draft Progress Report and circulate
DONE

KW: GCM data to Robert and complete CleanWorld code
DONE

LV/CW - may be able to provide something on seasonal cycles in changepoints later - see timescales listed below.
CW: Needs an algorithm to tease out the seasonal cycle from paired stations. Tried a couple (Levenberg-Marquardt) but no success - need phase, amplitude and possible trend.
ACTION: CW: To talk to Robert Lund/Colin Gallagher about ways of analysing seasonal cycles between pairs of stations.

KW: Email regional summaries link (https://docs.google.com/spreadsheet/ccc?key=0Al6ocsUAaINSdHpTREJzVkRZUTdfVjNPRlh0Q1V3WUE&usp=sharing#gid=0) to group and ask VV to email around homogenisation list
DONE, additions from Greece and amendments from Poland.

KW to email R Clean World code to Enric ASAP as now is a good time
DONE

KW: Find latest version of table of shift types that was edited in the workshop. Call needed on this to finalise beginning with a summary of where we got to in the workshop.
DONE

KW: Continue to keep track of VALUE progress.
ONGOING

KW: to email ISTI glossary link (https://docs.google.com/document/d/1xItD6yeQTxqwVnbfx-ZwUsh7hKJu1HqJVEf-OKstS4Y/edit?usp=sharing) around list and possibly ISTI main list
DONE - query of meaning of 'changepoint'. Does it cover both abrupt and beginning/end of gradual changes or just abrupt?
RW: I would say it refers to both the beginning and end of gradual changes as these are still points of change.
KW: All agreed changepoint covers everything.

ACTION KW: edit Glossary document with changepoint meaning

KW: Lets have a focussed call on methods of comparing assessment statistics over regions where we can summarise where we got to in the workshop.
DONE

IJ: Contingency table document put together a while back. Discuss and circulate for a future call.

KW: Link discussion of acceptable window of detection in with focussed call on comparing variances of different climate features \u2013 Team Validation call.

KW/RL: to sort Clean Worlds out and report back.
ONGOING

KW: Team Corruption changepoint distributions topic for the next call, summarise where we are with this.
DONE

KW: Recirculate concepts paper by end of this week - aim for GIMDS
DONE

KW; Send around Doodle poll for that week.
DONE

KW: Plan for next meeting:
- summarise progress from workshop
- find and circulate latest version of distributions - discuss shape and size.
- Concepts paper - ready to submit?
- Metadata - how to generate and how good/bad?
DONE

Actions from this meeting:
ACTION: CW: To talk to Robert Lund/Colin Gallagher about ways of analysing seasonal cycles between pairs of stations.
ACTION KW: edit Glossary document with changepoint meaning
ACTION KW: Chat with Colin Gallagher to see how much he can help here.
ACTION: Kate to try and draw up a flow chart and iterate with Victor. Some open source drawing program.
ACTION: KW to doodle poll and set up

#--
AGENDA
1) Progress Summary from workshop on error world statistics and distribution probability framework

Big questions for Team Corruption
i) How blind is blind enough?

ii) How many different background 'clean worlds'?
iii) How specifically do we build these things?
iv) How many different things to assess in one world?
v) Specific decisions to be made:
A - Fix parameters (table from overview_error_worlds.xls) for default world
B - Discuss parameterization (distributions to use)
C - What kind of inhomogeneities do we want to put in the experimental world (#B8)
D - How to implement the seasonal cycle?
E - What should people be asked to homogenise as a minimum (station subset, period etc..)?
F - How to apply gradual changes?
G - How to package up the error world data?
H - Instructions on how to come and play (will need input from Team Validation too)

i) Blindness:
The basics of what is in the worlds (list and table in Notes: section below) will be made public. Which world is which, and the specific date/character of each changepoint, and choice of underlying GCM will be kept secret. To minimise discovery of benchmark specifics over time we will:
VV: The benchmark is so large that it will be easy to see which world in the benchmark belongs to which scenario in our error-worlds paper. You should only put in the effort. Security by obscurity should not be our main defence against cheating. Making the worlds realistic will be much better protection.
- have an optimal length period for each cycle (too long and too much will be discovered, too short and we risk too few entries to make useful intercomparisons)
- EITHER perform assessment in one go at the end (slow return can deter entries if cycle is too long) and allow multiple entries

RW: Could we perform assessment as entries come in, but only return the results when we also release the truth, this may create less of a backlog than just analysing everything at the end?
- OR perform assessment adhoc and only allow one entry (which could be an ensemble of methods)
- Release Mid-2014 for 18 months, 6 months for assessment, cycle length 2 years.

KW: Assessment in one go at the end or ad hoc throughout?
PT: Learning how to perform the assessment is likely going to be an iterative process i.e. you won't get it right straight off the bat - so it would make sense to at least be developing assessment techniques as things come in (i.e. in house assessment and sharing of results and discussion of how to tweak assessment) but then apply the final assessment tools at the end. The risk of saying you'll assess at the end is that you never apply in anger posited toolset until then and then find the assessment tools need tweaking and then its too late? Submitters should only see the final assessment ideally but 'trusted' submitters may be able to help inform the toolset and feedback development so this is very definitely a shades of grey decision in my view.
VV: I would expect that most submit their data in the last week, some one week later.
IJ: I agree that this is best done iteratively. One question that occurred to me is how much developers are told about how the products will be assessed. There will clearly be some strong clues in the concepts paper for example, but we don't want to give too much detail because we will be evolving our methods and also we don't want to tempt developers to steer their algorithms towards our methods of assessment.

IJ: Another thing occurs to me. Should we asking users (of which developers may well be a subset) what properties they think are most important for homogenised series to possess, in case we're missing something.
VV: In HOME we did not do so, but theoretically specifying the assessment in advance belongs to benchmarking. I would suggest limiting this benchmarking to a small number of values (RMSE of the data at a number of time scales and trend errors) and see the rest (level 2 to 4) not as benchmarking but as validation. In praxis the problem will be small because people have a pretty good idea what constitutes well homogenized data. I thought specifying the assessment in advance was also the reason to write a paper on the assessment in advance.
IJ: Good point. I'd forgotten about the potential paper.
TT: I would argue that developers should have some idea of what will be assess and how. There is, of course, some danger that they will steer their algorithms to optimize the assessment. However, I think that is better than keeping them in the dark on the assessment methods.
PT: Even if they do then the fact that the error worlds span a range of 'scenarios' make it hard to game?
CW: Will there still be open worlds? KW: Yes
TT: Good to tell what we will focus on so that they can use that to perform their own assessments.
KW: discuss this at a later call - like the idea of Benchmark being Layer 1, validation being level 2 and 3
IJ: Like the idea of the paper specifying what the assessment will be. Plan for this in next Team Validation call. Ideally by next May.

ii) Number of clean worlds:
Ideally there will be 3 different GCMs used to provide a background climate (long-term trend and natural low frequency variability) for the blind analog-error-worlds (1:4, 5:10 (not 7), 7).
Each world will have its own 'station pink noise'/high frequency component overlaying/merged - a product of the Z_t shock terms randomly created during VAR clean world creation where $A_t = \Phi A_{t-1} + Z_t$.
Open worlds will use different GCMs and a unique station pink noise overlay for each world.
KW: All agreed?
IJ: I guess partly because I wasn't at the workshop, I don't fully understand a number of the points, including this one. I'll confine comments to those I think I do understand.

iii) How do we build these things?
We have established a bottom up station by station probability decision tree with layering.
IJ: I guess partly because I wasn't at the workshop, I don't fully understand a number of the points, including this one. I'll confine comments to those I think I do understand.
PT: Will this fully enable sampling the rich ecosystem of station issues such that some stations contain no or very few significant breakpoints whereas others may be really really bad? Its probably important to verify that we are getting some reasonable distribution (spread) of station by station issue frequency as well as intra-station statistics?
o Layer one - known (regional summaries)/unknown regionally clustered changepoints (nationwide, airport moves)
Some distributional probability system required here to populate within known/allocated distribution estimates
KW: These distributions and probabilities need to be defined.
PT: Discussion below wrt layer 2 may be more applicable here. My bad :-(
VV: For countries we do not know anything about the regional changes we need to estimate probable distributions for these.
ACTION KW: Chat with Colin Gallagher to see how much he can help here.

o Layer two - non-clustered abrupt changepoints

Allocate location using a Geometric model. Changepoint type (station move, shelter change etc) could then be assigned using appropriate probability distribution. Size and shape (seasonal cycle) of each changepoint can then be allocated using an appropriate probability distribution. Should one of these changepoints occur simultaneously with a previously assigned changepoint then the latter changepoint will not be applied.

PT: Why? In the real world it may be prevalent to have two (or more) breakpoints simultaneously. Classic example is the US transition from CRS to MMTS. Here you changed the instrument type and more often than not the location (governed by need for electrical power connection). The instrument type change changed Tx and Tn in a quasi-consistent manner (and could be modelled as a tight distribution) but the location change would be a random or quasi-random larger sigma on a station by station basis. I suspect that quite often there are changepoints which arise from a combination of such multiple effects and could / should be modelled as such.

KW: My gut feeling is not to apply double changepoints. This is because our estimate of what a change looks like is still largely guess work and so adding two together just complicates things unnecessarily at this stage. We have no idea whether what we have implemented is realistic and it also makes assessment very complex. Perhaps something we could bring in for version 2 but I would prefer not to for version 1.

PT: But this is simply a case of saying that we model some layer 1 and layer 2 breaks (on the assumption some layer 2 breaks share characteristics of layer 1 but our knowledge is imperfect) as a combination of 2 marginal distributions as the breakpoint is the sum of TWO impacts which have distinct distributions. These distributions are one tight PDF that proxies for the instrument change component effect which is broadly invariant and one broader PDF that proxies for the location change effect in the example of building a population of pseudo-CRS-2-MMTS breakpoints. Other change types which are combinatorial (mix more than one change type concurrently which is probably reasonably prevalent e.g. a station relocation may be concurrent with an instrument change) may combine different marginal distributions. Otherwise although individual breaks look okay the distributional aspects over at least subsets of the network as a whole may not be sufficiently similar. We know from USHCN that the adjustments over the MMTS transition taken as a whole look, on the average, like the known instrument change effects, but individually they do not. So they really are a combination of a systematic distribution with what looks in the net like a random distribution.

KW: Its more a problem for how we assess this. Programmatically it is easier to just have a maximum of one changepoint per time step (although it could have abrupt and gradual characteristics). Do we penalise someone for not finding two? Do we just count them as one?

PT: They are clearly one break in time. The issue is that they represent the combination of more than one effect and those effects are random draws from distinct distributions. So, an algorithm will only find one break per station series but you need the population of station breaks distribution to in some sense 'mimic' the bidistributional nature of the causes and impacts across a network as a whole if you are to test performance 'plausibly'.

VV: I would argue that the breaks in layer one will have a bias and a random component. Avoiding additional layer two breaks near layers one breaks does not seem necessary, except maybe on exactly the same date.

VV: Do we have any reason to assume that a Poisson process is wrong for breaks? If not I would go for the simplest case, i.e. Poisson.

PT: Need to make sure we have some stations with very few and some with very many in vaguely realistic proportions.

VV: And the same for networks such that some are really bad and some are really good. IJ: A different distribution per network.

KW: How do we define networks?

MM: Use big countries and then other places can be divided into regions rather than countries.

CW: Do we want to always assign the known inhomogeneities to the actual countries? i.e., MMTS in America applied to Africa?

VV: Good to always apply MMTS to America though as this is known and realistic.

VV Easiest to just use countries.

KW: Agreeing that whenever a changpoint is assigned there is a decision tree to assign level of complexity (single change, multiple changes) which will allow a much greater number of co-occurring changes rather than leaving this to chance. For assessment this will be considered as a single changepoint.

o Layer three - gradual changepoints

Apply either 0 or 1+ gradual changepoints optimising on 30% of stations having at least 1 using a Geometric distribution. There is a small chance of having more than 1 (probability to be decided formally).

Decide whether to assign an abrupt changepoint in addition at the beginning, end, both or neither and assign its size based on a Gaussian distribution.

Allocate the size of the gradual using a Gaussian distribution and duration using a Poisson distribution.

Should one of the abrupt changepoints occur simultaneously with a previously assigned abrupt changepoint then the latter changpoint will not be applied.

IJ: Here I've pasted in something I said in response to the BenchmarkWorkshopReport. I think it relates to the reference to Geometric in Layer 2 above and to the Poisson in Layer 3. One additional comment, hinted at in the final sentence is that the exponential distribution would be used instead of the geometric for gaps between changepoints, or lengths of gradual changes, if time is treated as continuous rather than discrete.

"Geometric vs Poisson. I wasn't involved in this part of the discussion, so I may be misunderstanding things, but it seemed to me that geometric and Poisson distributions were the wrong way round on pp 7-8 of the workshop report. Poisson distributions are appropriate for number of occurrences of an event in a fixed time period, so could model the number of inhomogeneities in a time period - essentially you would be modelling the occurrences of inhomogeneities as a Poisson process. In this case the time between inhomogeneities would be modelled as an exponential distribution. Geometric distributions are discrete analogues of the continuous exponential distributions and can model the time until something happens, such as the end of a gradual change. It looks to me as if 'Poisson' and 'geometric' have somehow got switched on pp 7-8."

IJ: May just be a geometric type argument rather than distribution.

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

o Applying some seasonal cycle to a changepoint as opposed to just a mean shift needs more thinking. We estimate that 10-50% of stations have changes in variance likely joined to temperature, radiation, humidity, wind. CW/LV and the homogenisation list are to provide further advice on this. Gradual changepoints may also have some seasonality or be stepped in nature.

KW: Agree this framework in principle?

KW: Can we have an example basic concept flow chart of the probability/distribution descisions process of how this might work - using B5 as an example?

KW: Can we split the worlds between a few of us to draw up a concept flow chart of probability/distribution for each of the layers?

ACTION: Kate to try and draw up a flow chart and iterate with Victor. Some open source drawing program.

iv) How many things to assess in one world?

Blind and open world general characteristics have been defined - see list in Notes.

KW: Agreed?

PT: I doubt you will ever get full agreement. The question is are they agreed sufficiently to deem them 'fit-for-purpose'?

VV: One difficulty with World 9 – Auchmann & Brönnimann type errors using GCM fields of humidity, insolation, wind etc to inform inhomogeneity characteristics – is that it mimics the transition from a Wild screen to a Stevenson screen. This has happened only once if at all. However, in the benchmark we will insert 2 or 3 of such biased breaks per century. Thus we would have to see this perturbation as a general improvement in radiation protection and apply it multiple times. That would not be very elegant, but acceptable and we do not have such methods for other biased breaks. MM: Could be too complex to build at this time.
KW: Lets see how we go and swap later if necessary. This could be swapped for a world without gradual changes or using a different method to build clean and error worlds or a world with random errors/poor quality/no QC performed or vice versa if we decide to add random errors in all – see discussion below.

v) Specific decisions to be made:
Gradual trends are likely to be in more stations than we first estimated - now 30%.
Gradual trends can be shorter/more frequent as demonstrated by findings from the PHA run on the USHCN (25 yrs).
Abrupt changes can commonly occur at the beginning/during/end caused by some change in vegetation or siting characteristics.
KW: Agreed?
VV; Would imagine there is a higher chance of having an abrupt change at the end of a gradual trend than at the beginning.
PT: Could just leave all abrupts to layer two but have greater probability of a changepoint occurring at endpoints of gradual changepoint.
TT; Could model it as an inhomogeneous point process but it gets quite complicated - so some time periods are more likely to have a change point occur.
KW: An abrupt computing module may work – from any layer there is a probability of an abrupt changepoint occurring so the abrupt changepoint module is invoked. If the dice roll calls for an abrupt change to occur then one is picked randomly from an appropriate distribution.

We will not include random errors (i.e., outliers) in the analog-error-worlds as this may contaminate assessment of homogenisation algorithm skill with how skilful the users choice of quality control tests are. We assume that a good QC process has been run on the data and will make this clear when releasing the benchmarks.
KW: Agreed?
VV: Yes, but for our uncertainty estimates from the benchmark to be valid for the ISTI dataset itself, this would mean that we need someone to do QC on the ISTI dataset.
PT: Agreed but we could also make a pair of open worlds where the sole difference is whether these are added or not. Developers could then use this a. to define whether their QC is 'good enough' and b. whether inadequacies in their QC / not applying QC may hurt their algorithm?
MM: We have our own NCDC QC - stage 4 will be QC'd data - so essentially we're starting with stage 4.
KW: How about sticking to our assumed QC product and people can use stage 4 if they don't want to QC. We should strongly be encouraging everyone to QC prior to homogenisation. They can QC the benchmarks and we should hope that this removes nothing because we haven't added any of these isolated random (QC issue) type errors although I would be surprised if that were the case.
KW: Testing ability of QC algorithms in addition to homogenisation algorithms is beyond our scope here.

--
CALL ENDED HERE AT 90 MINUTES

The distributions of abrupt and gradual changepoint size are Gaussian. Although published findings on detected breaks report fatter tails and a narrow middle with substantial missing middles (Brohan et al. 2006; Menne et al. 2010; Willett et al. 2013), a study of known changepoints from USHCN metadata shows the distribution to be Gaussian (Menne, NOAA) - suggests detection of unknowns misses more than just the very small ones and that multiple breaks that are combined to one may be the reason for the fat tails in the detected breaks.
KW: Agreed?

The frequency of abrupt changepoints can be modelled using a GEOMETRIC distribution where each time point is equally likely to contain a changepoint but the number of locations allocated is optimised on a given frequency per century through a given probability parameter.
PT: We know that certain breaks are clustered (layer 1) so I assume here you are talking layer 2?

The presence of a/multiple gradual changepoints can also be modelled GEOMETRICALLY. Whether an abrupt changepoint co-occurs at the beginning or end can also be modelled given a probability of that occurring. This may result in an extra changepoint being allocated in addition to the average X Per century stipulated. This is not of great concern. If an abrupt changepoint is allocated within the gradual trend that is also ok.
We assume that the length of a gradual inhomogeneity can be modelled with a POISSON, a discrete distribution. The distribution cannot be 0 or negative but it is likely to be two tailed.
RW: Poisson distributions can give a value of zero, but truncated Poissons can't
IJ: I repeat that unless there is some reasoning I've missed, Poisson distributions are not relevant here.

No minimum homogeneous sub-period will be specified, allowing changepoints to fall where ever, including within the last few years of the record. While hard for algorithms to find, comparisons can be made in anomaly space and between stations with and without changepoints at the end of the record to assess the impact of this.
KW: Agreed?
PT: Yes. In fact there is a logical argument that real breaks may tend to be clustered. I change my instrument, then I monitor for a while and decide say after a year that I am running 1K colder than I was compared to my three neighbours so I then make another change. Where instruments are actively rather than passively managed there may be good reason to expect a propensity for breaks to actually cluster in the real world? So, maybe you want to think about this possibility in one or more of the analogs. Certainly, such a propensity would tend to confound clean algorithm results.
VV: Maybe we could play with temporally correlated breaks and platform break pairs in the open worlds? (We already planned to insert platform pairs in some countries in world #B8.) Then we could study whether that makes a difference. For the default worlds, I have not seen any empirical evidence that we need something that complicated.
KW: What are platform pairs?

2) Using latest version of changepoint size/frequency distributions
(overview_error_worlds.xls - circulated) - discuss to finalise if possible
See table in Notes below
Size of changepoints (abrupt and gradual) in the Tropics. Fewer gradual in the Tropics?

Kate summarise from chats with Blair
KW: Agreed numbers of table?
PT: Lets start developing on these numbers and then review some prototype series that result (you could develop the bias series w/o needing to implement onto the clean analogs) and reassess. I'd like to see some series of the posited break structures to be applied visually which would be more easily interpretable than a table of dry numbers.

3) Concepts paper - ready to submit?
Second draft now circulated.
KW: Any major comments?
PT: I circulated a marked up draft.
VV: I would still love to go through the entire text once, may need a bit more time.

4) Metadata - how to generate and how good/bad?
If time - any thoughts?
KW: Metadata available for US stations? Just replicate this? No worlds specifically state metadata inclusion. Could include vs not include for some countries in B8 that have the same 'exotic' changepoint structure applied.
PT: Metadata may want to be more complete for a world we wish to be 'tractable'. Thinking politically here in terms of what messages we may wish to put out AFTER the benchmarking, saying that x% (I'm guessing 90%) of algorithms performed better when metadata was quasi-complete would send a strong message to the community and funders about how its also important to concentrate on metadata rescue, preservation, digitization and standardization? Regardless, perhaps in the open worlds produce a world with perfect metadata. This would allow users to test the verity of their adjustment component to their algorithm in an isolated sense in that they could prescribe the breaks absolutely. I guess if the world is open this is in theory calculable anyway though?
VV: The default should be the generate artificial meta data for those networks where the database also has real metadata and the quality of the information should match the real dataset as well. To create some political pressure to generate better metadata, we could have one world where metadata is available for all networks/countries. I would not make such metadata perfect, but just as imperfect as real metadata.

5) AOB
KW: presented ISTI/Benchmarking stuff at Reading University Met Department. Heard about CHARMe which is a European (FP7) funded project to standardise data-product metadata (publication, user info, quality assessment etc.). Kate is now on their email list and in contact with them. May be a useful way of linking ISTI Benchmarking Assessment Info to products in the future.

6) Next Meeting: week of November 25th-29th
Thanksgiving so Tuesday-Wednesday better.
VV: Main topics: seasonal cycle of inhomogeneities and artificial metadata. Possible topic: ideas for additional open worlds.
ACTION: KW to doodle poll and set up

#--
Notes:
Blind World Reminders:
Studying the influence of biases
Group 1: 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.

Group 2: Physical inhomogeneities
#B5. Best guess world, with a bias of 0.2℃ per century at high- and mid-latitudes and 1℃ 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.

Group 3: 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.

overview_error_worlds Table:

Blind worlds		Bias ****		Random Breaks	Random Breaks	Random Breaks	local trends	local trends	local trends	
Biases		**West**	**Equator**	**Size***	**Lenght HSP**	**Seasonal cycle**	**Length**	**Percentage stations**	**Warming rate*** **	
		°C/100a	°C/100a	°C (sigma)	a	°C (sigma)	a	Percent	°C/100a	
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 breaks	0.2	0.2	0.7	15	0.35	25	30	1	-2 to 4
B3	B1 + Hard: More and smaller	0.2	0.2	0.8*0.5°C+ 0.2*0.7°C	10	0.25	25	50	0.5	
B4	B1 + Easy: Dewer and larger	0.2	0.2	1	20	0.5	25	10	2.5	(exclude breaks in last 2 years)
Physical inhomogeneities										
B5	Best guess everywhere	0.2	1	0.7	15	0.35	25	30	2	
B6	B5 + Biased breaks smaller + more	0.2#	1	0.7	15	0.35	25	30	2	
B7	B5 + different GCM/ background trend	0.2	1	0.7 (biased: 0.35)##	15	0.35	25	30	2	
B8	B5 + Exotic inhomogeneities###	0.2	1	0.7	15	0.35	25	30	2	
Seasonal cycle										
B9	B5 + Auchmann and Brönnimann	**	**	**	**	**	25	30	2	
B10	B5 + Hard annual cycle	0.2	1	0.7	15	0.6	25	30	2	
Open worlds										

O1	No Inhomogeneities	0	0	0		0	0	0	0
O2	Only large breaks	0	0	(-1,+1)	15	0	0	0	0
O3	No seasonal cycle	0.2	0.4	0.7	15	0	25	30	2

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

** What equations produce

*** 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.