ISTI Benchmark Working Group Day #3
3rd July 2013

Attending: Kate Willett (KW), Peter Thorne (PT), Claude Williams (CW), Matt Menne (MM), Robert Lund (RL), Colin Gallagher (CG), Zeke Hausfather (ZH), Enric Aguilar (EA), Jared Rennie (JR)

Online attendees: Ian Jolliffe (IJ), Victor Venema (VV), Renate Auchmann (RA), Thordis Thorarinsdottir (TT)

Big Qs to keep in mind:

- How do we assess different style products: grids vs stations?
- What if stations are missed out?

- maybe because they are not of interest
- maybe because they are 'too hard' to homogenise

- What do we want back from users?
- Coordinate with VALUE downscaling group in terms of assessment terminology, tests and software?
- What about very close changepoints - how are these assessed?
- Length of record choices for assessment?
- Size or nature of changepoint that is important?
- See NOTES at end for Ian's and Victor's thoughts on these and more and the notes sent round from our meeting in S. Korea (Notes_from_12IMSC_Jun2013.doc)

Agenda:
Start teleconference: 9am – 10am: catch up with benchmarking WG members
9:00 – 9:10 – Kate Willett summary of day 2
Blindness: Kate to be the handle turner, basics of what is in the worlds will be known but specifics of what and where will not be. Try to speed up the cycles if possible to prevent release of specifics after assessment.
VV: Deadline for contributions and then start assessment.
IJ: How long do we give the users to run their algorithms on the benchmarks?
KW: Long enough to work on completely new projects
KW: Ian's idea of having a population of 15 worlds, with replacement, and then randomly sampling only 10 is nice
ZH: What if the random sampling missed out a key world?
IJ: Can have less random sampling chosen by Kate
VV: If we do this, I would prefer to have 15 and to make sure that all 10 worlds are included as that explores everything explicitly. In the analysis we want to compare different worlds with each other. Especially if one of the default worlds would be missing, most of the questions could not be answered.
KW: Lets see how difficult it is to generate the worlds?
CONCLUDE TO STICK WITH 10 FOR NOW

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

How to build: Bottom up station decision tree with layering.
Layer one - apply any known/clustered changepoints based on region

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

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

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

ACTION: CW/LV/ZH to report findings and email homogenisation list.

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

ACTION: Email Blair Trewin, what about Hawaii and Florida?

Fixing parameters:

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

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

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

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

Apply location of changepoints using a geometric distribution - all points are equally likely to have a changepoint applied with a probability that optimises on 7 per century P(7/(12*100)).

MM: Assessment period - even if everyone homogenises the full period we should assess over different time periods to take into account the different densities across the period.
KW: We had the idea of asking for a specific subset to be homogenised as a minimum e.g 5 regions of 100 stations from 1970-2012.
ZH: Some algorithms do better with more stations
KW: What is the optimum number of stations above which there is no benefit of more stations?
EA: 100 may be too many for some smaller groups that are national in focus.
VV: Smaller regions better for semi-automatic methods - not generally fully robust algorithms - should only select long nearly complete stations.
KW: Then its not a thorough test.
ZH: Two separate tracks - one for regional and one for global?
EA: Good to encourage those working on smaller networks by providing the smaller networks.
EA: What about some automated software for validation, especially of the open worlds.
TT: This could be done with an R package downloadable from a website
KW: A very nice idea - lets consider this a non-essential but very desirable extra.
ACTION: Kate keep in touch with Douglas Maraun about how this is being done for VALUE as they are planning on an online automatic assessment.

9:10 – 9:30 - Intro to Team Validation lead by Ian Jolliffe
ISTI Glossary
ACTION KATE start a document in googledocs

9:30 - 10:00 – Team Validation discussion – what to focus on while here, make some decisions
(Both validation measurements and what do we want returned to us – stations/grids, dates/size of breaks found?)
VV: To ease the discussion, I have merged Ian's discussion points with mine. No white line between comments indicate similarity. IG#: Ian General, IS#: Ian Specific, V#: Victor

IG5. Vocabulary. I think we should be very careful in defining the terms we are using, to reduce ambiguity and possible confusion as much as possible. Perhaps an ISTI glossary might be useful.

IG4. Blindness – discussed yesterday. I think it is very important to make as much blind as possible and that those parts that are not blind are equally available to all developers. I have another suggestion for increasing blindness which I suspect will be shot down by Team Corruption. Developers can easily find out the general structure of our 10 error worlds. Why not have more than 10 (say 15) and choose 10 randomly from these (preferably with replacement) so that developers will not know which have been selected, or even whether there might be repeats among the 10.

IG1. Which Levels do we validate:
Level 1 – correspondence between truth and homogenised series – Yes;
Level 2 – whether or not inhomogeneities are found or falsely identified – probably, but possibly not as part of main reporting exercise;
Level 3 – detailed identification of nature of inhomogeneities (size, duration …) – probably not;
Level 4 – how realistic/unrealistic, easy/difficult are the various error worlds – probably not.
V1. I would propose to limit the benchmarking to the station data and the global mean signal. All other scales and Level 0, 2, and 3 measures could be simply part of a normal analysis of the results. Without fixing them in advance and without using them for ranking.
VV: Importance of level 4 for the next round of the benchmarks
MM: Level 4 would form part of the assessment of how well we have done for a paper write up at the end
KW: Focus on level 1 and 2 as critical right now. Level 3 will come later.

IG3. How are results from individual stations or gridpoints to be combined to give a regional or global measure of performance? Could be done differently for different validation measures.
VV: Aggregate results from the station scale. Some things are more sensible at the larger scale - trends, RMSE (suggested by Robert with formulae later on too I think)
IJ: Order of taking roots and means and squares may be important
EA: Some errors will cancel others when aggregating over regions - so we will lose some information.
PT: Depends on end user of the product - some will want global/regional, some will want very local. One product may be better at the station level but not so good over the large scale average.
KW: Quite feasible to do station differences (histograms of differences) and regional differences (numbers)
VV: Which regions?
Giorgi? Hemispheric? Continental? CORDEX downscaling regions? Köppen climate classification?
KW: Keep things simple and manageable here but hope that this wealth of data will be used for far more detailed analyses afterwards.
KW: Continental! Stations aggregates (histograms/RMSE) and regional averages

Station data
- Level 1, user oriented measures
IG2. What do we ask for from the developers? Stations/gridpoints, subsets spatially and temporally? Requirements too rigid = no-one comes to play; too flexible = difficult to make useful comparisons. As well as what we ask for, there is the question of its format to fit with our software. Is this straightforward or tricky? These matters depend to some extent on our choices of verification/validation measures below
V2. To solve the problem that some (or likely all) homogenized contributions will not homogenize part of the data (not only stations in data sparse regions, but likely also some data sparse periods), we could ask everyone to return all data homogenized as well as possible, but indicate which data one would have removed in the real dataset. We could then use all stations and periods in the station based validation for a fair comparison. In the computation of the regional, continental or global

means, we could leave the marked bad stations out.
V3. We have the problem that the absolute level is not defined in homogenization; homogenization only improves the temporal consistency of the data. Convention is to use the last homogeneous subperiod (HSP) as reference for the absolute level. If we would adopt such a convention as well, there would be the problem that there may be a break near the end. One solution would be not to insert breaks near the end in the benchmark (but they would be there in the real data). Another option would be not to analyse those stations (for station level validation measures) where we know that there is a break in the last few (2 to 5?) years.
KW: Decided to allow changepoints at the end and preferentially assess using anomalies where we can to get around this problem.

IS1. Level 1. Want to compare properties such as variances, autocorrelations and spatial correlations of the truth and homogenised series. Currently no plans for doing this quantitatively. To compare the time evolution of two time series we could use (R)MSE, MAE or correlation. Comparison of trends will depend on whether we want to simply look at linear trend or something non-linear.
VV: How do we compare nonlinear trends?
KW: Preference for linear trends as these are a key communication tool.
VV: Smooth time series to decadal scale and compare RMSE
KW: This might be a good way to deal with the seasonal cycle that we have or have not put in
TT: Difficult to rank variances from different products
VV: What scale are we comparing the variance?
KW: All months? All years? On the seasonal cycle?
PT: More important for homogenisation algorithms that make some time varying/seasonally varying adjustments. Did any HOME algorithms adjust the seasonal cycle?
VV: CLIMATOL (J. Jaguijarro) and QM (X. Wang) likely the only ones on the monthly scale
VV: Take away some smoothed dedadal filter and assess variance on remaining high frequency data.
KW: Just use standard deviation?
IJ: ANOVA?
ACTION: Ian to think about use of ANOVA?

V4. We could analyse the temporal behaviour of the seasonal cycle in the same way as the annual mean, by computing an annual time series with the size of the seasonal cycle. Additionally we could have a look at the shape of the seasonal cycle.
V5. In addition to linear trends, we could study the mean value over predefined time slices. Alternatively we could compute the RMSE for a smoothed (decadal) time series. I guess they would be proper scores that would be hard to hedge.
Level 2, detection of inhomogeneities
V6. Keeping in the line of keeping it simple, my personal favourite would be to study the hit rate and false alarm rate defined as a/n and b/n, which n simply the number of data points or the number of breaks we put in. (we may need to come up with a new name to avoid confusion, as the definition is a little different from the usual one).
I have the feeling that the homogenization community is mainly interested in how many breaks are found and how many false breaks were put in. Combined scores are probably seen as either too complicated or as adding up oranges and apples.
This would us allow us to study to the HR and FAR in more detail. For example make a histogram of these rates as a function of the real break size, where it would be possible to see whether good algorithms actually find more small breaks or whether they are better because they find the large and important breaks more consistently. And it would allow us to vary the definition of a hit or false alarm (the imprecision of the timing). I think such analysis would be very insightful.
Would it make sense to compute a ROC curve from this kind of HR and FAR values?

IS2. Level 2. I'm afraid I don't like Victor's b/n. I have a number of thoughts, both on which measures are candidates to use and also what to do about correctly identified breaks whose timing is not right. To discuss this would take too much time in a brief call. I'll circulate something in the next few days. Things get more complicated when examining gradual changes than for simple breaks. Also, as well as choosing some mathematical/statistical measures, we need to decide whether to compute any measures based on costs/losses associated with the two types of mistake (inhomogeneity present, not found; no inhomogeneity present, but 'found'), or whether this is too user-specific.
ACTION: Discuss in future conference call.

IS3. Level 3. There may be things here we would like to examine, such size of breaks. We may or may not want to assess how well the size is estimated, but more likely we will want to stratify Level 2 results according to size. Break size may also be important in determining costs/losses in 2.
V7. Good algorithms are often able to find smaller breaks. These breaks have a much higher uncertainty in the location of the break. In the validation of the detected breaks this should be taken into account explicitly, otherwise good algorithms may have a handicap.
VV: Argument for stratifying assessment by breaksize - how big is the window with which we accept a correctly located changepoint? Smaller window for big breaks and bigger window for small breaks.
ACTION: Victor to put some numbers on that. UPDATE: proposal Victor, lets discuss this first in a future conference call.

V8. Most homogenization algorithms (maybe even all at the global scale?), correct gradual inhomogeneities by inserting breaks. Thus maybe we should exclude periods where we know that we inserted gradual inhomogeneities from the computation of the contingency scores.
We could also exclude periods with missing data. Alternatively we would have to "move" the breaks to the beginning or end of the missing data period.
*Global mean signal**ACTION - DISCUSS TODAY!!!!*
V9. In the comparison of the global (or large-scale) means, the only objective signal would be the model average. For every other choice we would be validating our own choice of gridding and averaging. Also for a model-based average, we would still need to define what the global mean is. Which regions are used (everything, everything except Antarctica, or everything below 60°?).
KW; Compare both:
Model vs Homogenised - would give uncertainty against the 'true' global measures
Clean world stations vs homogenised - would isolate uncertainty in the homogenisation algorithms
MM: Difficult for Berkeley
PT: For Berkeley just assess their regional statistics against the GCM model
KW: Or compare their regional statistics with regional average statistics from the clean worlds - accepting that our two different gridding techniques will be a factor. Best probably to do both. Could subsample Berkeley grids to where we can grid the clean data for clean-homog comparison and leave spatially complete for model-homog comparison. GHCN can do both gridding and station which will help further analysis of this assessment.
We will ask for stations and then grids if you only have grids (or want to submit grids) and then an ascii list of all breaks found, their location and size.

V10. For those contributions that provide only station data, we could do the averaging ourselves to also be able to compute global means and biases in the trends for these algorithms. Maybe we could ask BERKELEY to do so, without use of their cutting algorithm. I expect that they are overconfident about their uncertainty, but their mean is likely very accurate, optimally computed.

End teleconference/webex

10:30 – 12:00 – various groups carry on with Team Validation concepts and also Team Creation and Team Corruption

RL: A method for comparing closeness of time series.
Simple correlation of 'clean' time series with homogenised series - r=1 is perfect, r<0 is terrible
*1/Numberofstations * SUM (correlation scores)*
or
*SQRT(1/Numberofstations * SUM (MSE scores))*
or 1 - SSE/SST (Sum of sq errors over sum of sq totals)
*SQRT(1/Numberofstations * SUM (no. CPs - found CPs)^2) add some location cost or flexibility, exponential decay from 1 to 0? May be worth asking Jaxk Reeves about this.*
*SQRT(1/Numberofstations * SUM (trend - homog trend)^2)*
*SQRT(1/Numberofstations * SUM (homog trend % recovery somehow)^2)*
SQRT makes it sigma (st dev) rather sigma squared (variance)
A nice way of summarising station statistics over a continent.
Seasonal cycle - use euclidian distance:
Svec is [Jan, Feb...Dec]
SvecH is the homogenised [Jan, Feb...Dec]
*SQRT(1/Numberofstations * SUM matrix(Svec-SvecH)'matrix(Svec-SvecH))*
ZH: Trend bias recovery percent. What percentage of the way did you get between the error and the clean world with your homogenisation algorithm?
PT: Think about using all of this info (e.g., 4 algorithms run on the 10 benchmarks gives us 40 pdfs of efficiency plus 4 real worlds) to then assess the uncertainty in the real world trends as part of our round up final analysis?
See:

Thorne, P. W., P. Brohan, et al. (2011) "A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes" Journal of Geophysical Research - Atmospheres, doi: 10.1029/2010JD015487

Basically assume blind world 1 is correct and algorithm 1 moves 40% of the way (in terms of long-term trends) from bad data to truth (% trend recovery) then conditional on world 1 and method 1 our best estimate is that we should move the results for method 1 applied to real world data 1.5 as far again as the method implies to get the true real-world trend. Give it a meaningful pdf. Then repeat for nworlds x nmethods and recombine the whole set. This gives a consensus estimate of the true global mean trend and its conditional uncertainty for example.*

Potential groups for global:
NCDC
Berkeley
NIST?
Hadley Centre/CRU?
GISS - unlikely as they use GHCNM pre-homogenised?
JMA?
Finn Lindgren? - not stations
Potential groups for local:
Environment Canada
Missing stations - penalised for station analysis but not for regional.
UPDATE: David Hendry from Nuffield/Oxford Uni has a novel method and may be interested in coming to play

CG: Got some code running to implement the gradual changepoints.
Need to know distribution for duration (mean=25years, sd = ?, cannot be less than 1 month) and

size (mean=1,spread= -2 to 4? degrees C per century)
CW: Are gradual changepoints really linear? More likely a lot of very small steps.
CG: Could use a pareto (heavier tails than a normal), make it symmetric about the mean and sample from that for the size.
VV: Colin, what is your email? I could send you some Matlab code for a jumpy trend using a fat tailed distribution for the increments. We are just working on this for the validation study with NCDC and Zeke.
ACTION CW: Going to look at PHA to try and characterise some of these slope things - slopes and lengths.
RL: Duration to be modelled with a Poisson which is discrete (integer).
So this now becomes its own layer and we just flip a coin to decide whether it begins/ends with an abrupt or not.
VV: A gradual inhomogeneity ending with a break seems natural to me. A gradual inhomogeneity starting with a break is probably rare.
Layer 1 = gradual (may add an extra abrupt outside of the 7 per century) - not worried about that
Layer 2 = abrupt
Layer 3 = known/clustered

Team Creation!
PT: Create the station difference series (station - reference series) - where the reference series is some kind regional average but not just the average of all of the stations as otherwise there will implicitly be anticorrelation.
RL: Just do station-GCM – no problem
KW/PT: Not clear how that will work given that there is no reason why a GCM would be simultaneously similar to the station.
RL/CG: Statistical standardised anomalies have a time-varying mean and seasonal cycle removed and are then divided by their standard deviation.
KW/PT/MM: Climate anomalies can be many different things – just a seasonal cycle removed, sometimes divided by the standard deviation. It depends what purpose they are intended for.
ACTION: RL to try station-GCM method
ACTION: KW to work on the 50/50 method to see what it does across the globe – improve by removing loess before subtracting climatology and dividing by the standard deviation other wise we're introducing greater variance at the ends of records and smaller variance in the middle? Also roll out VAR(2+) method so that more persistence in the stations is captured. This may not overlay with low-frequency variability from the GCM directly if the smoothed curve removed is too sensitive. Still needs some thinking about really. Regional averages is not obvious – you are modelling the difference series which have very low cross-correlations and can be anti-correlated.

SUMMARY:

Regional aggregates of station stats (histograms/RMSE) and regional averages at continental and global scale.

Assessment to take two tracks – regional and global/continental – to include national (non-global) homogenisers.

Assess different time periods to account for the changing station density over time.

A webpage hosted or downloadable R package to perform automated assessment would be very useful to provide quick feedback, especially for the open worlds where users may want multiple attempts at the problem.

Do include changepoints within the last two years – can compare stations with and without – can assess in anomaly space.

Ideas formulating for assessing climate measures (level 1) but not how to compare variance – IJ to think about ANOVA

Compare climate measures-trends from homog-model and homog-clean (subsampled to match spatial coverage of both). Trends can be measured using %trend recovery measure between homogenised and clean world then combined across all worlds and methods to give a measure of uncertainty on our current global trend estimates.

Decision model to implement changepoints in error worlds based on Pareto or Poisson for gradual trend duration.

Standardised anomalies involve removal of some time-varying mean or linear trend to make them truly stationary – this is not what climate scientists consider to be anomalies – climate anomalies often only have a time-stationary seasonal mean removed and are sometimes divided by a time-stationary seasonal standard deviation.

Team Creation methods still need proving – 9010 method reduces to non-vector AR, Robert to try direct station-GCM methods.