Overview

This was a combined workshop on homogenisation algorithm development and the culmination of the COST funded HOME action on developing and assessing homogenisation algorithms. This is largely a European venture but with algorithm input from the USA (NCDC) and some involvement from outside the EU.

The COST HOME action (www.homogenisation.org) has spurred on the development of a number of sophisticated algorithms (HOMER, ACMANT, MASH, PRODIGE, AnClim, iCraddock, Climatol etc.) which use a variety of techniques to detect inhomogeneities (e.g., SNHT, penalised maximal test, versions of pairwise comparison) and to adjust (e.g., ANOVA – seasonally varying, flat adjustments). The homogenization task can be broken down in to two smaller tasks, detection (including building a reference) and correction, but the main improvements are somewhere in between. They are in the use of correction methods without assuming that the reference is homogeneous and the handling of missing data (segments). Some of these are automated but some require manual input. There is high value to looking at the data for each station but this is not possible in large networks and may lead to subjective adjustments which are non-reproducible and not easily traceable. Furthermore, automatic algorithms are easier to validate, which may speed up their improvement.

A key part of this effort is the benchmarking of the algorithms. Small networks (up to ~15 stations?) of monthly benchmarks were created. These were made up of real data with known inhomogeneities, synthetic data based on white noise and surrogate data which attempted to reproduce real world variability. Seasonally varying errors were added, with perhaps an overly strong seasonal variation. The benchmarks were first kept blind to the algorithm creators and the results assessed. The pairwise method of NCDC (USHCN) was also tested on these benchmarks. Some creators then further developed their algorithms based on things learned from the benchmarks. This tended to improve their scores but this is essentially overtuning to the benchmarks.

Most of the methods performed quite well. This depended on how the methods were assessed. Those methods using a seasonal cycle to detect and adjust for breakpoints did better when assessed on the monthly scale than the annual scale. The USHCN method was quite conservative but had a very low false alarm rate compared to other algorithms. All methods struggled when the networks became very sparse.

First results for a validation of daily corrections methods based on surrogate data with known breaks were also presented. We will need to understand the nature of inhomogeneities better to perform a true benchmarking of daily homogenisation algorithms. If we would like to progress to such a benchmark for the second ISTI cycle, it would be valuable to strengthen research on this topic outside of the working group also and there are interested parties.

A lot can be learned from the development of the benchmarks and the assessment using those benchmarks:

- Importance of clearly stating what is to be assessed prior to benchmarking i.e. reference period, annual or monthly assessment?
- Importance of being clear about the implications of assessment – an algorithm that has a perfect detection rate may not adjust very well and vice versa, an algorithm doing well in the later, well observed period, may not do as well in a more data sparse period, an algorithm detecting and adjusting seasonally may do well in strongly seasonal inhomogeneities and not so well in flatter inhomogeneities – opposite to an algorithm looking for flat shifts and making flat adjustments.
- Blind benchmarks are safest – preventing overtuning to specific benchmarks. However,

there is value in playing with realistic benchmarks for algorithm development as long as the limitations of the benchmarks is understood

- International expertise are essential – speakers from different countries brought local knowledge of their networks and their problems to the table which will be important for global benchmarks.
- Make using the benchmarks as easy as possible – how do we make the analog-error-worlds easily accessible to all and getting the results back in a uniform and easily assessable manner.
- Work with other smaller more focussed research groups – any funded actions on daily benchmarks.

Notes for Team Creation:

Future benchmarks:

- should be realistic
- realistic outliers/random errors - assume a good QC has been undertaken
- insert random missing data (which we will have masked from the real stations anyway)
- study frequency and size of local trends (which will come from the climate models)

Adding the noise term – some of this will be uncorrelated with other stations – simple random errors, some of this would be the weather term although how this would play out on monthly timescales is unclear – persistent cold or hot events – these would be correlated across networks. Some kind of simple weather generator? Could this sort of thing be modelled from the real stations? Study periodicities in common or something like that? Could use geospatial statistics to get at spatial covariance and add 'weather' based on these underlying relationships?

May be worth storing some other information from the models to be used by team corruption – incoming solar radiation, windspeed? This wouldn't be public info but could help with 'realistic' error input.

Notes for Team Corruption:

Need to add in realistic inhomogeneities that do not reward specific algorithms by being too obvious/exaggerated. For example, having an over exaggerated seasonally dependent shift will penalise algorithms with a flat detection/adjustment more than necessary and reward algorithms detecting/adjusting based on strong seasonal shifts. This is a difficult balance to achieve but having final errors added by those not building the algorithms and keeping the benchmarks blind will help.

Specific types of inhomogeneity:

- Add in station moves by cutting a pasting a nearby station series. May have to tweak a little to avoid exact duplication though – could create 'duplicate' stations by using the average of 2-3 neighbouring stations to downscale the GCM gridbox therefore creating a unique but realistic station. These 'duplicate' stations will differ slightly and can be substituted for part of a station series to mimic a station move.
- Instrument change/calibration error – this could be a flatter change but could also be a change to the variance on hourly timescales (not necessarily monthly). Instrument sensitivity may change.
- Shelter change – cotton region to stevenson screen – would be a seasonally varying change
- Manual to automated – more missing data, more repeated data (QC), fewer outliers? (QC), more or less sensitivity?
- Changes in observation times – how will this be manifested in monthly data?
- Significant changes to network density – a very real problem that may be reflected in the

analogs anyway as they follow the real station drop-in/out – although do we want 100+ years of benchmarks? If we're shortening the record we need to ensure a similar station fall out in at least one of the worlds. When validating we need to be clear on the reasons why algorithms are failing if possible. 1972 seems to be an important year in ISD (NCDC's global sub-daily data) where vast numbers of digitised records drop out and then come back in in 1973.
- Changes in observation frequency and reporting resolution. Increases in reporting frequency from 6 hourly to hourly may mean that lower minimums/higher maximums are now recorded – and vice versa. Rounding procedures may lead to changes from resolution changes – do they truncate or round?

Have a few established break characteristics to input but make them not too predictable or people will know what to look for.

Reference period – this should be the most recent homogeneous subperiod. This is a problem, especially for algorithms doing seasonal shifts, when the last breakpoint is very close to the end of the record. However, this could be a real break location and so should not deliberately be avoided. Assessment should be aware of this though – algorithms could be penalised by this because they would not be able to model the seasonality effectively but assessments may look like the algorithm is failing because of the types of breaks or another complicating feature that was added – importance of useful assessment.

Future benchmarks:
- should be realistic
- Correlations in perturbations within a network – geographical clusters
- study seasonal cycle
- Provide metadata – some good, some bad, some incomplete, some negligible

Include other key climate features – solar radiation/sunshine duration affects the break characteristics, wind, ENSO etc. Largest effects in clear skies – full solar radiation. This info can be stored from the climate model data when creating the analog-known-worlds for later use by team creation.

Be realistic but also have ability to isolate certain break types/questions to make analysis useful – need for a series of worlds with well posed questions.

Regional knowledge is valuable – how to obtain this?
- Norway: Most breaks due to relocation (55%), screen changes (14%), instrument change (15%), other (15%) - very little effect of changing observer – NOT QUITE SURE HOW THAT ADDS UP TO 100%? SIMULTANEOUS CHANGES?
- France/Germany found most changes due to changes in shelters. Norway may have less changes with shelters because of radiation? Or many changes happen at the same time so difficult to distinguish.
- Norwegian data are composites of multiple nearby stations – not official station moves but later station mergers! Similarly in Czech Republic.

Proportion of known to unknown breaks – I would expect that for most countries there are more 'unknown' breaks than 'known' breaks – Czech has 50% backed up by metadata.

Some algorithms are trying to adjust more than just the mean, some of the higher order moments. Do we know enough to be able to add in errors in this way? Can we assess this fairly?

Notes for Team Validation:

Use the existing benchmarks and validation to look at which methods are more or less useful. Importance of looking at both ability to recreate the 'truth' and to detect the different types of breaks so that algorithm creators can get something positive about this – what exactly is causing problems for the algorithms? (station density, break frequency, break magnitude, background trend, seasonal cycle, natural variability, missing data, breaks near end-points etc.).

Reference period – this should be the most recent homogeneous subperiod. This is a problem, especially for algorithms doing seasonal shifts, when the last breakpoint is very close to the end of the record. However, this could be a real break location and so should not deliberately be avoided. Assessment should be aware of this though – algorithms could be penalised by this because they would not be able to model the seasonality effectively but assessments may look like the algorithm is failing because of the types of breaks or another complicating feature that was added – importance of useful assessment.

I think that false alarm rates are very important. I would rather a conservative and low false alarm rate than one that gets a higher number of breaks but adds a lot of error too. The false alarm rate should take into account the impact of incorrectly detecting a break given the adjustment applied. A detected break with a negligible adjustment applied is not so bad.

RMSE error seems to be a simple and useful metric – to root or not to root though?
analog-error-worlds minus analog-known-worlds = FULLRMSE
adjusted analog-error-worlds minus analog-known-worlds = REDUCEDRMSE
REDUCEDRMSE should be less than FULLRMSE if the algorithm is improving the network

Watch out for temporal variation in contingency scores – fewer breaks detected near the end of series?

Validating on annual verses monthly (or daily) – should validate on the highest resolution that will be used – so monthly I would say. This will penalise against flat adjustments but we know that inhomogeneities are not flat changes – this means that the errors added MUST be as realistic as possible.

How to calculate True negatives for the contingency scores?

Ensemble approach – hopefully this approach will be growing in popularity and so we need to be able to cope with this. An argument for keeping validation simple.

Some algorithms are trying to adjust more than just the mean, some of the higher order moments. Do we know enough to be able to add in errors in this way? Can we assess this fairly?

Benchmarking Programme:

Growing the benchmark network and user community is essential – as COST HOME showed, it is the only way to compare and validate algorithms

Need to be clear about things from the start of the cycle:
Reference period is the most recent homogeneous subperiod
Assessment at monthly or annual level?

How do we publish actual break detection and adjustment info from the real data alongside datasets? Like we do for metadata. How do we want break information for the adjusted break data to be provided? I gave a stab at a standard ascii file in the white paper. This could get very large and unmanageable with 60000 stations though. We need to think about this.

Value in comparing nationally homogenised datasets with globally homogenised stations for those countries. We should expect differences, especially where station density differs.

Ensemble approach – hopefully this approach will be growing in popularity and so we need to be able to cope with this. An argument for keeping validation simple.

COST ACTION funds training schools – this is what we need to get people using the benchmarks – any way of getting some one-off funding?

EGU session on homogenisation and benchmarking – get an ISTI slot in there.

Daily benchmarks:

Moving to daily scales means that we must deal with weather and not climate. Much greater noise to deal with. The way that inhomogeneities are manifested may be very different on daily and subdaily scales to monthly.

Changes in observations times, frequency and resolution – some breaks will be easier to do at the daily scale than monthly.

At some point people will want to use multiple variables from a station to identify breaks.

COST proposal on Daily benchmarking is a possibility and I have funding for a PhD student to work on daily data. Great value from regional efforts/expertise at reproducing changes at specific locations, between specific instruments and practices etc. on daily scales – the working group can learn from these as we go along. We make a first stab at global benchmarks while encouraging other research outside the working group which can then feed into future benchmark cycles. Having some agreement on starting points (creating the analog-known-worlds) for all these efforts would be helpful though – downscaled model data? End product would be a set of networks for as many regions of the globe as possible – sourced from the same source data and methods but with regionally specific changes applied.

This really needs people from outside Europe: China, Japan, India, Malaysia, Australia,, Africa, USA, Canada, Caribbean, South America, Australia, New Zealand etc.
What happens at different latitudes/elevations/climate zones/locations when a station is moved, an instrument is changed or begins to drift, practice changes (resolution, reporting time) – need to reproduce known dates and locations and changes i.e. USA automation mid-1990s.

Validation methods – these may differ to monthly

Seemless daily to monthly benchmarks?

Tools to aid use of benchmarks for algorithm development and dataset uncertainty – aligning with other communities working on uncertainty in climate data

Need funding for training schools – creating benchmarks, using benchmarks, validating benchmarks