Benchmarking Working Group Call #7

Friday 10th May 4pm BST (GMT+1), 3pm GMT,

Attending: Kate Willett (KW), Victor Venema (VV), Ian Jolliffe (IJ), Robert Lund (RL), Steve Easterbrook (SE), Lucie Vincent (LV), Peter Thorne (PT)

Not attending: Claude Williams (has gone down sick and is heading out the door, sends profuse apologies)

ACTIONS FROM PREVIOUS MEETING:
ACTION: KW to test different methods:
- Improve missing data handling – DONE – INTERPOLATE SMALL AMOUNTS OF MISSING DATA AND TEST MISSING DATA ALLOWANCE
- play with and without loess – VARIED LOESS FITTING
- Different GCMs – atmosphere only – NOT DONE
- compare variance and autocorrelation of difference series – DONE – SEE COMPARISON FIGURES EMAILED 'Kate_Simulation_May2013.ppt'

ACTION VV: Lets try and get big Qs sorted by Friday next week. - STARTED
ACTION VV/IJ: VV pass on HOME documentation to Ian. IJ have a look. - DONE

ACTIONS FROM THIS MEETING:
ACTION: RL Needs 100 stations - KW to send MCDW data to Robert.
ACTION: KW to set up googledocs/okfn? wiki? blogpost?

NEXT MEETING: Wednesday May 29th 4pm BST/GMT+1, 11am Eastern USA, 5pm Euro time.

AGENDA:
10 mins: Update on Team Creation KW ('Kate_Simulation_Statistics_May2013.ppt') and RL ('Robert_Simulation_May2013.ppt')

RL: Vector Autoregression (VAR) order 1 - Uses standardised anomalies - subtract monthly mean, divide by monthly standard deviation, vector autoregression to match space time correlation.
KW: no trend removed before modelling? RL: No, not in this version
Aurocorrelation at lag 1 0.16 is pretty low - very variable weather. This is the spatial correlation at lag 1.
RL: Optimal order is 8 for these data. We're assuming order 1 here. Argument for temporal and spatial memory in there.
Bartlett's formula for testing significance of autocorrelation values - are they significantly different from zero?
ITSM package to test for optimal order - KW used BIC/AIC R package.
Sample variance, phi and other VAR matrices should match if this is a good simulation.
Simulated sample variances likely to be greater than 1 because a random draw from a normalised distribution will not be exact.

0.95 comes down to 0.88 when scaled to convert to correlations.
PT: Pulling back to a 30K feet view if we are creating multiple perfect worlds can we not do it several ways and are we not going to drive the corruption models off several initial cases because the phasing and nature of variability are confounders to homogenization? Indeed we might want to put EXACTLY the same error structure on more than one distinct perfect world starting series to assess this?
KW: so I think we have a working method but now need some far field correlation structure that ties the globe.

ACTION: RL: Needs 100 stations - KW to send MCDW data to Robert.

PT: First order Q here (see word doc I'll send): Is the plan to look at Tavg only or Tx and Tn in addition and does this drive how we look at corrupting and evaluating the algorithms that may participate?
KW: Tavg for now
PT: Does this yield issues when the databank itself is multi-elemental and many groups may look to analyse in a multi-elemental manner?
KW: Possibly but I think we need to start here.

Team Corruption 'Questions_error_models.doc'
15 mins: Questions to answer with Team Corruption's error worlds

Claude says he has some input he will get to us next week about apparent seasonality artefacts arising from NCDC's PHA. In a nutshell he says that either seasonality is less of an issue than some suspect or that we need a stronger test than the current PHA permits to find it.

PT: Another 1st order Q: How many analogs can we produce? Very many! BUT ... at what point do we scare folks off? We should alight on a total number of analogs we will produce that we think is reasonable and that then by necessity drives the experimental design. I think if we go over 5 open and 10 blind we will lose groups. Even that may be too many?
KW: 10 blind maximum.

Questions about uncertainty remaining in climate observations:
i. Do homogenisation algorithms 'improve' trends/variability?
ii. What is the uncertainty in trends/variability of homogenized data due to common weaknesses in homogenisation algorithms?
    e.g. do they generally under/over adjust?
    e.g. do they generally not detect gradual changepoints?
    e.g. do they generally distort the seasonal cycle?
    e.g. do they generally perform worse in the Tropics/High lats/Data sparse regions etc.?
iii. What is the false alarm rate/chance of homogenisation make trends/variability worse?
iv. How much do we need to worry about difficult to detect inhomogeneities? Do more difficult to detect inhomogeneities cause greater errors in the data than easy to detect inhomogeneites?

Questions about homogenisation algorithm strengths and weaknesses:
i. How well do homogenisation algorithms perform in the presence of missing data?
ii. Can homogenisation algorithms detect seasonally varying changepoints which can be in the same direction or both positive and negative at different times of the year?
iii. Can homogenisation detect changes in the variance where there is no change in the mean?
iv. Are break points and gradual inhomogeneities more or less additive or do they influence each other?
v. Are small breaks more difficult to detect purely because they are smaller (signal to noise ratio)

or is it because they are more frequent or both? In Williams et al. more small breaks were found to be more difficult as a realistic setting. We could investigate whether this is because of the larger number or due to the smaller size. In other words, we could vary the size and the number separately. (In an upcoming idealised validation study of the HPA and BEST it is already studied whether the number of breaks is important (while keeping the size of the individual breaks the same), but this study is just for the USA.)

vi. Can homogenisation algorithms detect biased changepoints that result in a trend in the data?

vii. Does it make a difference whether the trend bias is caused by many biased breaks or just a few? (Biased breaks are breaks with a nonzero average that have the potential to bias the trends of raw data.)

viii. How well can be find and correct breaks that occur simultaneously/within a short period of time across a network/region/country?

ix. Can homogenisation algorithms detect and adjust for gradual inhomogeneities?

x. What type of inhomogeneities are easiest to detect?

xi. What type of inhomogeneities are hardest to detect?

xii. Are there regions/climatic zones where it is easier to detect homogeneities? Is this linked to station density or background climate (climate variability)?

KW: Cannot easily pick 1-2 Qs to answer. Most Error worlds will go some way to addressing all. Keep these Qs in mind for Team Validation.

15 mins Types of worlds needed to answer those questions

Realistic Worlds:

– try to vary one thing at a time where possible? All 'errors' to be physically based from GCM or real station properties. All to have seasonal cycle. All to have some spatial correlation in places. All to have a spread of inhomogeneities over time (end points and middle). All to include a range in size and direction of inhomogeneities.

A. Best guess world/like HOME

A world that includes a mix of abrupt, gradual, isolated, grouped, changepoint free, various magnitudes, biased and non-biased, ISTI matched missing data, in the presence of 20th Century climate forcing (background climate change) for all regions of the globe.

B. Seasonal Cycle exploration

Bi – no seasonal cycle – change in mean only

Bii – change in seasonal cycle – no change in mean

Biii – small seasonal cycle, same direction

Biv – large seasonal cycle, same direction

Bv – small seasonal cycle, both directions

Bvi – large seasonal cycle, both directions

NB some changes may only effect summer time or only effect winter time.

C. Changpoints direction bias

Ci very few large abrupt changepoints throughout series – including some close to end points – both directions

Cii many small abrupt changepoints throughout series – including some close to end points – both directions

Ciii very few large abrupt changpoints throughout series, biased to one direction

Civ many small abrupt changpoints throughout series, biased to one direction

D Missing data exploration

Di Best guess with no missing data

Dii Best guess with real matched missing data (same as A)

E Gradual Trend Exploration

Ei few large and small gradual inhomogeneities only – no climate change

Eii few large and small gradual and abrupt inhomogeneities – no climate change
Eiii few large and small gradual and abrupt inhomogeneities - in presence of climate change
Eiii Sawtooth mix of multiple large and small gradual and abrupt inhomogeneities – in presence of climate change

KW: Other option - but keeping things a 'real' as possible
Continuum (assume most are seasonally varying breaks? - in all cases test the station/region/global trend, station/region/global month/year/decade variance and total hits/misses (weighted for location, size and duration) for regions/globe):

Regional inhomogeneity characteristics are needed. PT: put a google docs up and ask people outside to contribute. Good to know anything - timing, cause, numbers if possible.

KW ACTION - set up googledocs/okfn? wiki? blogpost?

When - month, year, spread of years?
Where - whole country, region?
What happened - shelter change, automation, time of observation, move to airports etc.?
Quantify effect - change in mean and/or variance, any quantitative value?

Idealised Worlds:
A. Small breaks. It is typical to introduce breaks from a normal distribution with a standard deviation of 0.8°C and we are quite sure that this number is realistic. This size has been found for the breaks in metadata and the sizes of the detected breaks in the HOME benchmark data (which used 0.8°C) and in real dataset is very similar. For understanding the algorithms, it may be an idea to reduce the size in a few steps: 0.8, 0.4, 0.2°C.
B. Many breaks. The typical length of homogeneous subperiods is 15 to 20 years. We could vary the frequency of breaks between 20 and 2 per century, while keeping the size fixed.
C. No breaks. Study false alarm rate.
D. Best Guess ScenarioShall we also have one realistic scenario in the open for testing the algorithms?
E. Strong spatial dependence. One biased break at the same date per country. Or would that be too extreme?
F. A lot of missing data periods.
G. For the question whether breaks and local trends are additive, we could generate one world with only abrupt shift, one with only gradual shifts and one in which the perturbations of these two worlds are added together.
H. Most breaks are likely unbiased (have a zero mean), some may be biased (especially in case of technology changes or systematic relocations, for example to airports). Technology changes would be the transition from North wall to garden screens, from open garden screens to Stevenson screens and from Stevenson screen to automatic weather stations. Thus in a 150 year time series, I would expect 2 to 3 breaks to have a bias, the rest to be unbiased. We could vary this from inserting one break per century with a large bias, to (almost) all breaks having a small bias, to investigate whether this makes a difference. (The rest of the breaks would be unbiased)

10 mins Make some decisions

10 mins Discuss Team Validation if time - see Ian's circulated document 'ValidationCommentsMay13.doc'

AOB

Just to note that there is a planned joint Steering committee / benchmarking / databank call posited for 8am EDT (12Z) on Tuesday June 4th. I'll send details nearer the time. By then Kate, Jay and I should have a new shiny Implementation Plan for initiative as a whole for folks to get their teeth into.
Matt Menne advises that NESDIS still are pending decision on the proposed workshop. He will advise as soon as he hears.

NEXT MEETING: Wednesday May 29th 4pm BST/GMT+1, 11am Eastern USA, 5pm Euro time.

Minutes sign off:

NOTES: