Benchmarking and Assessment Working Group Call #21

Tuesday 12th at 2pm GMT (3pm BST, 4pm CET(summer), 10am EST,(summer), 5am AETZ)

Attending: Peter (PT, will need to leave about 1600 BST to make train), Lucie Vincent, Kate Willett (KW), Robert Dunn (RD), MM (Matt Menne), Victor Venema (VV), Rachel Warren (RW), ian jolliffe(IJ), Renate Auchmann (RA), Claude Williams (CW), Robert Lund (RL), Lisa Alexander (LA)

Not Attending: Enric Aguilar (somewhere in NW Spain)

##########

AIMS:

1) Catch up - where have we got to
2) What next and how for each section
3) Another time line
4) Everyone happy with review responses for Concepts paper

ACTIONS FROM LAST MEETING:

ACTION KW: Work with Richard Chandler and Thordis to get this (gibbs sampling/factorisation) working over at least a single region.
DONE

ACTION: CW, LV, MM, VV - ultimately to decide/discuss on what we want to do. Include jumpy graduals? Do not include jumpy graduals? Explore in World 8 (exotic IH) only?
VV: The analysis performed in Boulder, whether a breaks or trend model fits best may be a way to decide. First results were in favor of jumpy graduals, but we would need to apply it to more cases.
MM: I can work up a couple more cases. Then I would propose that we use the PHA/BFA detection results to help inform how we seed the benchmarks with shifts. I think that would help build realistic scenarios as long as we also "fill in the missing middle" of the distribution

ACTION: TT to send around an email to help to explain this.
No longer so essential

ACTIONS FROM THIS MEETING:

ACTION KW: Work on getting these values from the real data.
ACTION KW and IJ: Chat about Matern for future values possibly. Peter Challenor at Exeter may be able to help too?
ACTION KW: Try a better GCM with higher filtering
ACTION KW: Just try it assuming MVN for stations we cannot model properly.
ACTION MM: Look at a few more series and then make a decision
ACTION MM: Give VV and RA the BFA/PHA results.
ACTION IJ: Send newest version around - next week or two.
ACTION KW: Next call on this - 3-4 weeks.
ACTION KW: Update acknowledgements section and author names where necessary

##########

AGENDA:

1) Where are we now (25mins) / next step:

1a) SAMSI/IMAGe workshop overview (Kate)

I did not get the clean worlds completed in time.

I used the workshop to get the Clean World code running which was successful.

It was an excellent opportunity to tell potential users about the benchmarks:

- Spatial team can use clean worlds to test interpolation algorithms
- All teams can use error worlds to test detection and adjustment methods
- Talk now online: http://video.ucar.edu/mms/image/samsi2014_kate_willet.mp4
- initial presentation/wrap up now on our website:

https://sites.google.com/a/surfacetemperatures.org/home/benchmarking-and-assessment-working-group#Conferences%20and%20Workshops

IJ: Live streams of the talks worked well

- Important improvements for Team Creation in terms of interpolating GCM data, correcting errors in VAR code and testing whether benchmark data are representative enough of real world data characteristics (st dev and autocorrelation of difference series, cross correlation and autocorrelation of station series).
- Debate over gradual inhomogeneities - smooth or jumpy?
- Victor to visit Met Office for a week to work on the error worlds and to search for historical parallel data from outside the mid-latititudes in the archives.
- Claude to visit Met Office for a week to work on many things
- Kate to visit Uni Bern (Renate and Stefan) to talk about the Benchmarks

PT: I hope that a full report will be available soon. The draft is with the two funders for comments before a participant review so timeline is out of my hands.

1b) Team Creation (Kate)

Richard Chandler came to visit. We worked on globalising the whole method and producing spatially correlated shock terms for the globe.

- for each point in time, reproduce globe station by station, each time using the candidate stations neighbour network to run the VAR (neighbour disconnect method) and then added the spatially correlated shock term for that time point
- Richard wrote some R code that can produce global fields of shock terms that are spatially correlated (given a desired covariance matrix) using either a gibbs sampler or factorisation. The factorisation method is preferable. This code works on the scale we require in sub-30 minutes.

I now have R code that can run on 20000+ stations and reproduce synthetic stations with some level of spatial and temporal correlation.

- uses a distance function to approximate spatial correlation (I tried using real correlations but the data are too noisy - results in ridiculous simulated values)
- uses 40 highest correlating stations (greater than 0.37)
- builds a covariance matrix at lag 0 and lag 1 for each station and its neighbours based on simple distance function:

  corr_lag0 = 0.91 * exp(-0.0004*distance) (force diagonals to be 1.0)

  corr_lag1 = 0.41 * exp(-0.0004*distance) (force diagonals to be 0.45)

RL: Why 0.45? USA That is probably ~0.3

ACTION KW: Work on getting these values from the real data.

KW: How do we bring elevation into that?
RL: Matern functions may be able to bring in the elevation as a parameter
ACTION KW and IJ: Chat about Matern for future values possibly. Peter Challenor at Exeter may be able to help too?
- creates the autoregressive parameters for the VAR for each station+40neighbours (Φ1 = Γ(1) %*% Γ(0)INV)
- runs the VAR in reverse using real data to get the real residual shock terms (Zt = At − Φ1 %*% At-1)
- use factorisation to provide global spatially correlated fields of residual shock terms

- need to divide world into networks+surrounding neighbours (slide 3 of wrap up presentation)
- need covariance matrices for the residual shocks for each network+surrounding neighbours (Σz = Γ(0) – Φ1 %*% Γ(1)T)
- transform simulated residual shocks (MVN - slide 4 of wrap up presentation) for each station using real residual shock distribution, mean and standard deviation
- use neighbour disconnect method (slide 5 of wrap up presentation) to propogate VAR model for each station, for each time step
- multiply by real station standard deviation, add long-term trend from interpolated GCM gridbox, add real station seasonal cycle (slide 6 wrap up presentation)

But is it good enough?
We've tried a few different distance functions - little change.

Key measures are:

- station autocorrelation of climate anomly series (slide 7 wrap up presentation) too high - IMPROVE DISTANCE FUNCTION AT LAG1
- station cross correlation of climate anomaly series (slide 8 wrap up presentation) too high - IMPROVE DISTANCE FUNCTION AT LAG0
- st dev of station minus neighbour difference series (slide 9 top wrap up presentation) too low compared to USHCN (st dev ~1.0)
- autocorrelation of station minus neighbour difference series (slide 9 bottom wrap up presentation) ok - higher than we had previously thought.

KW: Anything else we can look at to see if the clean worlds are good enough?
CW: lag0 of 0.91 is probably too low for eastern US 0.96-0.97 not uncommon in the eastern US.
KW: Need to be careful that the statistics won't throw a wobbly - problem with cholesky decompositions.

RL: Need to be very careful about expanding the lags beyond lag1 to get low frequency AC
VV: Need to use a different GCM
PT: Long-term memory generally has very large spatial scales - way beyond the 40 stations we've been using as neighbours.
ACTION KW: Try a better GCM with higher filtering

PT: Open issue is whether we need the clean worlds to exactly mimic the databank and if so how we bring in the 10,000 short / sparse stations that Kate is unable to calculate a VAR functional basis for. My very initial impression is that if a group may use these additional records it will be hard to fairly

compare its performance to those that don't. But, equally, I have no idea how within the framework we might add these extra locations and series without mucking stuff up. Maybe we just can't do it and we are better running with what we have?
VV: I thought we now have the statistics as a function of distance, then it should not be too hard to also use the short pieces. I find it very important that the benchmark mimicks the ISTI dataset itself.
VV: If it is really not possible to make a benchmark from the 32k+ dataset, the I think we should ask everyone to also homogenize a 20k+ real data dataset. Only for this dataset we could then compute the uncertainties.
KW: The problem lies with replicating the residual shocks. At present I derive the autoregressive parameters using the distance function (which can be done for any station) but there needs to be enough data to back out the actual residual shock terms from the real data because I use the mean, standard deviation and distribution to transform the simulated residual shock terms. It may be sufficient to assume a Gaussian distribution, mean zero, standard devation of one for all residual shocks that cannot be characterised in reality.
PT: If this is computationally possible its probably worth doing this just to see what it looks like?
VV: Gaussian, mean zero sounds fine. Is there some way to estimate the standard deviation? Interpolating the values from nearby longer stations?
PT: Or a zonally based estimate from the 20K you do have?
RW: Would it be possible to use some kind of sub-sampling method on these stations to make their records complete enough to get the shock terms?
KW: Quite possibly as its only very basic characteristics of the distribution of the shock terms that we need.
ACTION KW: Just try it assuming MVN for stations we cannot model properly.

Next steps:
Use improved GCM interpolation and play with smoothing level (a way to introduce more low frequency variability)
Improve distance function
Try for all stations

Ideally:
Distance function should take into account elevation, aspect, land use etc but this is becoming a complex statistical model
Improve the variability - not convinced we really are getting ENSO type variability in the series
PT: Don't you import that from the GCM? Maybe you need to use something a tad newer than HadCM3 ;-p
KW: At present the loess smoothed function from the GCM is too smooth to pick out that level of variability. This is because I wanted the smoothing level to match that which was used for the real stations to remove the trend/underlying major inhomogeneities. If we use a narrower filter on the real data then we remove too much of the autocorrelation so its not really simulatable using the VAR. I agree we should try a different (modern) GCM and play around with the smoothing levels a little more - especially now we're using the distance function rather than modelling from the stations directly.

1c) Team Corruption (Victor)

VV: Except for the jumpy or smooth gradual inhomogeneities, I have the feeling that we have debated the main points, we "just" need to implement the inhomogeneities.
VV: One new point is that Ronan Connolly suggested to implement gradual inhomogeneities preferentially in stations that are known to be urban. That may be a bit different from implementing them in random stations, because cities tend to be clustered. Kate gave me a dataset (from Dave Parker?) with the urbanisation of all the ISTI stations estimated from night light. Does anyone know how good this is and what alternative we would have?
PT: The night light data is used by GISS and would be reasonable in the developed and EIT worlds but in at least parts of the developing world (where electrical supply can be very limited) is nightlight a reasonable proxy for urbanisation? Not sure.
CW: The nightlight proxy seems to be too sensitive for urbanization, unless there is a method to scale the signal. Even attempting to cluster the pixel may not help if they are more dispersed than the nightlight dataset makes them appear.... We cannot wait, but the new Carbon Sensing satellite will help us the next round...

PT: What is a realistic timeline now for release of the error worlds? Are we happy to push them out prior to peer review of the underlying methods papers or concurrent with their OA review as 'beta'? How do the group want to handle the release?
VV: Hard to say. I can do a little here in Bern, especially implement the biased inhomogeneities based on Auchmann and Brönnimann and World 8 (exotic IH). How far i will get will depend on how fast I learn R. RA: (I can help with R) and finish the biased inhomogeneities (Auchmann and Brönnimann) with you/soon after you leave.
PT: Are you willing to push the benchmarks out before a paper describing them is accepted?
VV: Had not thought about that. At least I would prefer to submit the paper after the work is finished. While implementing the inhomogeneities we will probably find some things that need to change. Would there be a disadvantage to submitting a paper after publishing the data? Except on getting valuable feedback from reviewers? I think we should send the description of what we did to the homogenization list and ask for feedback, that goes faster than an official reviewer.
PT: Sure. Main thing is whether happy to announce the release before the paper is submitted and / or accepted or whether we stay the release until such a time.

MM: Colin G tried to fit different models to difference series (at workshop) - stair steps vs smooth linear trend, accounting for autocorrelation - step model was at least as good or better. Only two case studies though. Thinking about physics behind - may expect a jumpy series more than a smooth. Next step to look at a few series outside of USA.
ACTION MM: Look at a few more series and then make a decision

RL: Are we confusing seasonal shifts introduced by a changepoint with this stepiness?
LV: Difficult to solve

MM: Seasonal expression of the breaks - difference between Reno and its neighbours has a strong seasonality.

RL: Going to write a paper on properties of target minus reference series.

MM: Not enough to deseasonalise target minus neighbour first - difference series definitely still contains a seasonal cycle.
KW: Is that from the inhomogeneity?
VV: Yes and that is something we want to model.

MM: What about using breakpoint results from BFA and PHA? Concerned that our error scenarios are too easy. Can look at frequency of shifts for different regions. In low density areas and for the missing middle we have to make some inferences. Can also get at the jumpiness from here too.
VV: Ideally need some studies looking at these results for countries other than USA and Europe.
ACTION MM:Give VV and RA the BFA/PHA results

CW: For adding in urban IH - should independently generate the timings/character of these so that they are not too similar to each other when they are close - some will max out - some with start/finish at different times, warm at different rates.

1d) Team Validation (Ian)
IJ: A document was prepared and circulated early last year. I've revised it a bit a few times since then in the light of comments and further thoughts of my own. It can form the basis of discussion when we have a call devoted to Validation. Validation methods/measures chosen will depend to some extent on exactly how the inhomogeneities are created, but we should be able to decide on the main methods/measures we wish to use without knowing full details of the inhomogeneities. We need to have a good idea of what/how we validate when the analogue-error-worlds are released in order to be clear in what we ask the algorithm developers to provide. However, we shouldn't rule out looking at extra aspects that appear interesting once we've seen the 'homogenised' data, not necessarily to report, but at least to inform the next cycle.
ACTION IJ: Send newest version around - next week or two.
ACTION KW: Next call on this - 3-4 weeks.

2) GIMDS Concepts Paper (deadline September 2nd) (25 mins)
Paper:
http://www.geosci-instrum-method-data-syst-discuss.net/4/235/2014/gid-4-235-2014.html
Reviews:
http://www.geosci-instrum-method-data-syst-discuss.net/4/235/2014/gid-4-235-2014-discussion.html

2a) Response to reviewer 1
Generally happy (Blair is generally a happy guy :->)

Spatial cross correlations depend on distance, elevation, aspect, land use, proximity to the coast. In the paper we say it would be good to replicate these exactly but in reality I think we will struggle to do much more than a distance/elevation function this time around. Is this good enough?
Part of this is covered in the GCM if we can get the smoothing level correct.

Inhomogeneities preceded by an increase in random error/variability. - will we include these?
VV: Not so sure about this at the monthly scale - possibly too complicated for now.

I have updated figures 6 - changepoint/inhomogeneity statistics for homogenisers to return
Figure 6:
STN_ID = station ID
STN_Long = station longitude
STN_Lat = station latitude
STDATE = start month and year
EDDATE = end month and year.
STSIZE = difference in median relative to reference period at subperiod start.
EDSIZE = difference in median relative to reference period at subperiod end. (NB: These will be identical for abrupt inhomogeneities and different for gradual inhomogeneities)
Chg_Rate = monthly rate of change for gradual inhomogeneities. (NB: 999.999 identifies a non-linear rate which should be recorded in CHARACTER)
JanSz to DecSz = difference between median January (February, March etc.) over subperiod relative to the median of all months in subperiod for seasonally varying inhomogeneities. CHARACTER = type of inhomogeneity and non-linear equations. (NB: A = abrupt, G = gradual, S = seasonal variation, L = linear, N = non-linear. N should be accompanied by a function where Y = temperature and Xih = month within subperiod)
VV: Why median? Don't people just implement one break size? Does that description come from how the PHA computes its breaks?
KW: Median more robust to outliers but mean may be a more 'normal' thing to provide.
KW: We would expect one size for one period unless its gradual and there may be a seasonal component
VV: More in general, didn't we say that we would only ask people for the dates of the breaks and not for the size of the adjustment. Because that is very error prone (in HOME even just the date was often wrong) and because that can get mightily complicated and anyway hardly machine readable as the nonlinear equation in the Figure 6 already suggests.
KW: I think it might be useful although we would have to be very clear about what the size was relative too - the reference period ideally but that could vary between products even if we specify that it should be the most recent period - it could be 1 to N years long depending on where the last breaks are found. Certainly better to ask for the adjustment size/shape that was applied than try to estimate what was applied ourselves I think.

KW: Compulsory date, optional characteristics.

IJ: Could just focus on changepoint locations easier

CW: Good to know size of adjustment because someone may make one large adjustment for multiple small ones.

IJ: Having a minimum gap between changepoints is moving away from reality.

2b) Response to reviewer 2
Main point queries value of concepts verses description of what we're planning. I've tried to make it very clear that this is concepts only and the value of it being concepts only.

2c) Response to reviewer 3

Main point (as with Reviewer 2) queries value of concepts verses description of what we're planning. I've tried to make it very clear that this is concepts only and the value of it being concepts only.

Clean Worlds - currently no relationship between long-term trends and seasonality (e.g., winters warming faster than summers in some locations). Do we need to worry about this level of variability?

Preventing overtuning to our style of benchmarks. I don't think we can do more than have an evolving cycle and have both blind and open worlds. We do need some open worlds.
VV; Could make the open worlds more simple.

Analog/analogue/pseudo?

VV: Reviewer 3 had a number of questions on Section 2, the equation describing the climate signal. His main problem was that the additive terms also interact with each other.
Explaining this has made this section even longer. I have the feeling that we do not need this equation, we do not use it later on, but the reader (reviewer) put in a lot of effort to try to understand it. Which phyiscal term goes into which variable and now how the interactions are.
In the end we only use X=GCM+Downscaling_increments+Inhomogeneities. The paper would be a lot cleaner and easy to understand if we would rewrite the beginning of Section 2 and simply state that the GCM takes care of the complicated climate signal with trends, seasonal cycles, climate modes, auto- and cross-correlations (and their interactions).
KW: I think I prefer keeping the climate elements though as we just use the GCM example as something that could be used to provide those features but someone could come along with an all whistles and bells statistical model that works. We want to lay out what components need to be included rather than how it should be built in this paper.

2d) Other comments related to the paper

RW: If this is going to cause too much hassle don't worry - but if I can be changed to R.E. Warren instead of R. Warren that would be good as there is already a Rachel Warren who publishes in the area of climate and we need people to be able to distinguish us
PT: Its always possible at this stage. Does anyone else need initials adding?
IJ: Most of my publications have my middle initial (T) but I don't have a strong preference
KW: I can make all of these changes. I also spelled Stefan's surname wrong.
VV: If the journal writes out my first name, I prefer my name without the K.C. middle initials, like it is in the manuscript.
LA: I'll change my initials to L.V.

PT: Maybe its cos after 1000 reviewer comments on a section of an IPCC report to respond to a review has never held the same fear for me again, but I think the reviews are easy to respond to.

PT: The one slight concern I would have would likely be with the editor. I think that in the responses and the redraft you need to make crystal clear that this concepts paper is a key first step to 1.

allowing a framework for more technical papers to follow, 2. building community acceptance of the benchmarking exercise and 3. spreading best practices more broadly to other sub-genres of geosciences. I made some edits to this end in what I sent Kate but I suspect these aspects could be hammed up somewhat more and would help the editor in deciding it is truly journal relevant. A few minutes rereading the journal purpose and playing buzzword bingo may be advisable.

PT: Were any other changes necessary as a result of the discussion that went on in the realclimate thread on the databank that veered off into a discussion of benchmarking?
KW: I think the ISO standard comment was of interest. We are far far far away from that but essentially we are trying to create an international standard for benchmarking. I added something in about this.
KW: I don't think the main commenter really understood what we were doing with the GCM/VAR clean world development which suggests it wasn't well enough described in the paper. I will work on this.
PT: Okay, we should add an acknowledgement to the folks who commented at RC on the paper in the redrafted acknowledgements section as well as the reviewers (and editor?)
KW: Good point
ACTION KW: Update acknowledgements section and author names where necessary

PT: Unless Kate wishes another check I am happy to trust her judgement and do not feel the need for another round of review. So long as any subsequent new changes are minor in nature without changing the core structure or methods I will be okay with resubmission.
IJ: I agree.
LA: Kate I've made some amendments to reviewer 3 and sent them to Kate (sorry I have to sign off now as it's bedtime! Don't want to interrupt discussion)

3) Time line for progress:
New Year 2015
Paper to describe the benchmarks aimed for similar time.

4) AOB (5 mins)
PT: I will advise on the time of the all hands call later in the week. Please sign up at the doodle if you haven't already done so. http://doodle.com/bz7qk6z3znsh2b5b That call will expect some update from this WG.

PT: WG progress report due in October. Doesn't need to be long. Concentrate on doing but please provide some update we can post.

5) Next Call Date (1 min)
Early September

##########
NOTES: