Interactive comment on “Concepts for benchmarking of homogenisation algorithm performance on the global scale” by K. Willett et al.

Anonymous Referee #3

Received and published: 27 July 2014

First of all, I would like to congratulate the ISTI initiative for their efforts, and for their intention to take methods for homogenisation into account in such a systematic manner.

As the authors state themselves, the proposed framework will not change the global picture of surface temperature increases, but might substantially improve the regional representation of temperature evolution.

Thank you very much for your very thorough review. You have made some excellent points which have really made us think about the text and try to improve it, and the benchmarks in general. We have tried to address all of your comments or explain why we prefer not to in a few cases.

In addition to responding to your comments we have taken the opportunity to improve the text where we felt necessary. We have removed a number of the figures (2, 3, 5, 6 and 7) as we felt that these took up unnecessary space and are easily described in words. We have also added a paragraph at the beginning of section 5 to clarify how the benchmarks will be made available as we did not think that this was very clear:

“To ensure that homogenisation benchmarking achieves its full potential in terms of usefulness, the benchmarks need to be easily accessible and the assessment process timely with results that are easy to use. We envisage making the benchmark worlds available alongside the ISTI databank in identical format (Fig. 3) such that data-product creators can easily process them in addition to the real data.”

Still, before publication, I would like to address some major issues:

1. What is the intention of the paper, and how far has the benchmarking working group come to set up their framework? For a mere layout of intended work, the paper is by far too long, yet for a full description of the proposed experiment, it is maybe too short and unspecific. Currently, descriptions throughout the manuscript are still rather vague.

The concept of benchmarking in this context is very new. We feel that the level of detail is warranted to describe what we believe to be the key issues and essential components. We have added text in the introduction (paragraph 12) to make it clear that this is a concepts paper only and why we think it is important to have such a paper:

"The objective of this paper is to lay out the basic ***framework*** *for developing* ***the first*** *comprehensive benchmarking system for homogenisation of land surface air temperature records* ***on the global scale. By defining what we believe a global homogenisation benchmarking system should look like, this paper is intended to serve multiple aims. Firstly it provides an opportunity for the global community to provide critical feedback. Secondly, the document serves as a reference for our own purposes and others wishing to develop benchmarking systems for other parameters of problems of a similar nature. Finally, it constitutes a basis for further improvement down the line as knowledge improves. Future papers will provide detailed methodologies for the various components of the benchmarking system described herein.****"*

I would expect the authors have -planned the specific setup of analog worlds (based on GCMs, based on real world data surrogates...) -worked on a selection of validation measures for different aspects -discussed potential experiments with different types of benchmark data sets (e.g., open, blind, different scenarios, regional aspects,...) I would like to see at least some more detail on these issues in such a paper. Currently it seems a bit, well, premature. Of course not all detail can be given, but direct links to pages with descriptions of the experimental design should be given. In line with this comment, the authors might think of changing the papers title. Instead of concepts, one might use the term "a framework" or "an experiment".

We strongly wish to refrain from describing the exact methods we plan to use for round one of the benchmarks. This is because not only are these methods work in progress, they will also evolve over time with future versions/cycles of the benchmarks. We feel that it is important to lay out the concepts/framework for what we (an international community of experts) think a global benchmarking system should look like and what are the essential components. We hope that this will form the basis of future developments of benchmarking. While we doubt this will ever reach the level of ISO standards we hope that this is a small step in the direction of something standardised internationally.

We agree that 'A framework' is better than 'Concepts' so have changed the title. We hope that you will be satisfied with this explanation. Two papers are currently in progress detailing the analogue-clean-world methods and there will likely be at least one for the error-worlds and one for the assessment part.

2. The authors should clearly state the assumptions and resulting limitations of their benchmark set given by Eq. 1. (=the linear decomposition): -there is no interaction between the annual cycle and internal climate variability. This is a serious limitation, e.g., in the tropics. It is known that El Nino is coupled to the annual cycle. Such an example should be given. -there is no interaction between long term trends and the annual cycle - also this might be questionable (e.g., winters might warm faster than summers) -there is no interaction between long term trends and internal climate variability. ENSO will basically stay the same, as other modes will do. Currently, this information is missing. It might be that I have misread part of the

equation. E.g., c is given a time index. As it is called a climatology, I presume that t refers only to time of the year. There might be further limitations resulting from Eq1 - If aware of any, please state them explicitly!

This is a very good point and should certainly be mentioned in section 2. We have now done this (bold text below) and also tried to make it clear that this is a conceptual as opposed to a strict mathematical decomposition and that while the aim is to be as realistic as possible, there are methodological and computational limitations when it comes to global scale simulation:

Section 2 paragraph 2
"

- ***I** represents any long-term trend (not necessarily linear, **with possible seasonally varying components)** that is experienced by the site due to climatic fluctuations such as in response to external forcings of the global climate system.*

...

*These terms are assumed to be additive in this conceptual framework. **This equation should not be considered to be a formal mathematical representation. All four components are deemed** necessary to be able to subsequently build realistic series of $\boldsymbol{x}_{t,s}$ on a network wide basis that retain plausible station series, neighbour series and regional series. Below, a discursive description of the necessary steps and building blocks envisaged is given. **A variety of methodological choices could be made when building the analogue-clean-worlds. It is envisaged that the sophistication of methods will develop over time, improving the real-world representativeness of the benchmarks periodically.***

... (paragraph 5)

Note that background trends may be seasonally variant, further complicating seasonally varying inhomogeneity detection. Such characteristics may be obtainable from a GCM.

... (paragraph 6+7)

Balancing sophistication of methods with automation and capacity to run on ~32000 stations is key. Ensuring spatial consistency across large distances (100s of km) necessitates high-dimensional matrix computations or robust overlapping window techniques.

The key measures of whether benchmark clean worlds are good enough are as follows:

- ***station to neighbour cross-correlation***
- ***standard deviation and autocorrelation of station minus neighbour difference series (of climate anomalies)***

- ***station autocorrelation***

These measures should be compared between real networks that we know to be high quality (relatively free from random and systematic error) such as NOAA's USCRN (United States Climate Reference Network; http://www.ncdc.noaa.gov/crn/) and the collocated analogue-clean-world stations. "

3. What is actually the mandate of ISTI? On your webpage you state that a proposal had been submitted to the WMO. What was the outcome of this submission? Anyway, you should mention the status here, as an official mandate would give your initiative much more weight as community effort.

ISTI's mandate is to improve our ability to robustly understand historical land surface air temperature change at all scales. This is stated in the opening (was in second) paragraph. After the proposal to WMO CCl the go ahead was given to set up the ISTI. This was done at a workshop at the UK Met Office in September 2010. Since then, recognition has been obtained from WMO GCOS, BIPM and ISI TIES. However, there is no formal funding stream and so the ISTI is a voluntary community effort. We hope that the provision of the website and overarching ISTI paper gives sufficient information to the reader without having to go into more detail here.

Further Issues:

Is it useful to call the surrogates analogs? I don't have a strong opinion, but analog sounds very much like a downscaling method, and in similar contexts, the term "pseudo" has been used, e.g., pseudo proxy, pseudo reality.

We would prefer to keep the term analogue mostly because we have used it in presentations and the website and other publications but also the dictionary definition seems to fit. The definition for pseudo does not seem to fit so well – the benchmarks are fake data but we are trying to make them as similar to the real world as possible.

Abstract, line 3: "at all scales." But this is not what you do.

It is what ISTI aims to do so we feel that the context here is accurate.

p238, l22ff: sections don't discuss etc.

Corrected:

*"**The creation of** spatio-temporally realistic analogue station data **is discussed in Section 2**. **Development of** realistic but optimally assessable error models **is discussed in Section 3**. **A**n assessment system that meets both the needs of algorithm developers and data-product users **is explored in Section 4. A** proposed*

benchmarking cycle to serve the needs of science and policy ***is described in Section 5.*** *Section 6* ***contains concluding remarks****."*

In your list of homogeneities, you might consider to add that inhomogeneities, at least at short time scales, might be weather dependent. Although I am not sure whether this is of relevance or just an academic question (e.g., a building might act as a wind shield from a particular direction only.

We have a couple of sentences on this that I hope might cover your comment (Section 3, paragraph 7):

*"Ideally, the effect on temperature****, and hence d,*** *from the change in* ***weather*** *(e.g., radiation, windspeed, rainfall and humidity****) should be taken into account*** *if possible. Given the current state of knowledge this will in many respects be an assumption based on expert judgement."*

p241, l1. Add "most" before previous. You discuss the home COST Action afterwards, were a different approach has been taken.

We think this is ok as it stands because the COST Action created small networks of synthetic data with added inhomogeneities.

p243 explanation of v: you should explicitly mention that different modes live on different scales (space and time)

We think this is ok as it stands because we discuss interannual, interdecadal and multidecadal variability with regionally distinct patterns. We have added the following though in Section 2 paragraph 2:

- ***v*** *represents region-wide climate variability* ***at a range of scales (space and time).*** *That is to say interannual and interdecadal variability due to El Niño and La Niña events, annular modes (AO and AAO), or multidecadal variations such as the Pacific Decadal Oscillation or Atlantic Multidecadal Oscillation. Such modes have regionally distinct patterns of surface temperature response e.g. a positive AO yields warm winters over Northern Europe.*

p244 first sentence: "these terms are ASSUMED..." See also my detailed comment above.

Good point, added assumed.

p244/245, l13ff: these paragraphs are a bit too vague. If you only want to present a concept, well. But wouldn't it be much better if you had agreed upon how to construct your analogs? How do you estimate m and v without a GCM? How do you separate forced signal from large scale from local variability? Also you should be careful not to construct different error analogs from the same clean analog. Just by

having the combined information available one might estimate the underlying clean analog much better than in a real world situation where one observes only one realisation of inhomogeneities.

We prefer to keep this discussion peripheral and save detailed methodology for a separate paper for reasons discussed above, and now added into the text. In reality, we won't need to separate the forced signal (l – long term trend) from large scale variability (v). We can choose GCMs that represent natural variability well enough and also have a range of forcing scenarios available to prescribe the long-term trend behaviour. It may be that we can downscale a GCM sufficiently such that the VAR statistical model mentioned in the paper may not be required. Certainly later versions of the benchmarks will hopefully use more sophisticated methods than the first round.

We have now stated explicitly that the underlying clean world should differ to some extent for each error world in Section 3 paragraph 1:

"Ideally each analogue-error-world would be based on a different analogue-clean-world to prevent prior knowledge of the 'truth' (analogue-clean-world)."

l 8: stations are not long but their series.

Good point, changed to station records.

l 13ff: work out clearly that an analog needs to be plausible, but not perfect. Otherwise lots of people will criticise the approach.

This is a very good point. We had not been explicit enough here. Now we have added the following:

"The key measures of whether benchmark clean worlds are good enough are as follows:

- *station autocorrelation*
- *station to neighbour cross-correlation*
- standard deviation and autocorrelation of station minus neighbour difference series (of climate anomalies)

These measures should be compared between real networks that we know to be high quality (relatively free from random and systematic error) such as NOAA's USCRN (United States Climate Reference Network; http://www.ncdc.noaa.gov/crn/) and the collocated analogue-clean-world stations. "

p247 l15 replace "stations" by "records"

Changed to station records.

p247 l25 "return" sounds a bit strange. Adjust?

Agreed, changed.

p248 please be more precise in explaining the covariate effect. The examples you present do not clearly make the case for an inhomogeneity - for me this sounds like a real climate response. Maybe you imply that d is amplified when the variability is amplified, but this is not stated explicitly.

By this we mean how an inhomogeneity behaves given different states of the weather, albeit averaged to monthly which we felt justified use of the term climate in this context. We agree that it is clearer to use 'weather' and modified the sentence to Section 3 paragraph 7 to make it easier to understand the precise meaning:

*"Ideally, the effect on temperature**, and hence d,** from the change in **weather** (e.g., rainfall, humidity, radiation, windspeed and direction**) should be taken into account** if possible."*

in your bullet points you should add "shift of" for each item.

Done – used 'change from' instead of 'shift of'.

p249 l13/14: the sentence is not quite logical. Maybe it is enough to delete the "digitally" in the second half, but it depends one what you really want to say.

We agree, removed 'digitally'.

p250 how do you ensure blindness? If your analogs are based on GCMs, it might be easy to detect some of the inhomogenities by just comparing the signal with the CMIP5 data base.

We won't tell anyone which GCMs are used or which actual time period. There is only so much we can do here. We're more concerned about accidental overtuning than deliberate cheating.

p250 l17: shorter compared to what?

By shorter we mean low frequency variability – so interannual to multidecadal. We have revised this sentence to clarify.

p251 l 1 "why e.g.? Is there more possible?

In this context we probably should use i.e. – we have modified the text.

p251 l 2 "across a range of space and time scales" very unspecific

Actually, given the rest of the paragraph we think that this bit should be removed. The paragraph now reads as follows:

*"For Level 1 assessment of large scale features (i.e., **c, l** and **v** in Eq. 1), a perfect algorithm would return the analogue-clean-world features. Algorithms should, ideally, at least make the analogue-error-worlds more similar to their analogue-clean-worlds. **Climatology, variability and long-term trends** can be calculated for stations, regional averages or global averages from each adjusted analogue-error-world. Similarity can be measured in terms of proximity in ^{o} C for the climatology and linear trend approximations and standard deviation as a measure of variability."*

p251 l4 "This information" more precise!

See above.

p251 l17 add "detection" after "location"

Done.

p251 l18 sliding scale? does this term exist?

We believe so.

p251/252: at no place you discuss level 4! As you do it with all others, you should also do it with level 4.

Very good point. We have added the following in section 4, paragraph 8:

"Level 4 assessment should help inform us which analogue-error-world is most similar to reality (if any) in terms of detected changepoints for each algorithm. This is useful for two reasons. Firstly, assuming the error-structure is realistic, it may help to tell us something about uncertainty due to inhomogeneities remaining in the data. Secondly, it helps improve later versions of benchmarks in terms of developing realistic error models."

p252 l4 is there a better term than "correct misses"? "correct homogeneous"?

We think that this term, although ugly, is clear. We have modified it to 'correct non-detections' as we think this is more accurate although probably still ugly.

p253 l9 but then you should separately evaluate these?

This could be done if a list of used stations is provided. This has been noted in the text.

p254 l5-10 this explanation is not quite clear, at least not how the last sentence relates to the previous. Wouldn't the overtuning occur if the clean analog would be released?

It is true that by having open worlds that have been developed in the same manner as the blind worlds we risk algorithms being tuned to our version of reality – hence the need to update with some regularity. However, in this case we mean that there is the potential for algorithm developers to learn the underlying trend signal and tune to that, or some other feature of either the clean worlds or the error worlds.

p254 l16 here you mention for the first time which time scales you want to consider. This should be much earlier in the manuscript!

We have now added this information to the introduction paragraph 13.

"Here, the focus is solely monthly mean temperatures. These concepts broadly apply to daily or sub-daily scales and additional variables (e.g., maximum temperature, minimum temperature, diurnal temperature range). However, both development of synthetic data and implementation of realistic inhomogeneities, while maintaining physical consistency across different variables simultaneously requires significantly increased levels of complexity. "