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

R. Cornes (Referee)

r.cornes@uea.ac.uk

Received and published: 21 July 2014

This paper provides a generally thorough survey of the benchmarking concepts required for homogenizing the global ISTI dataset. The main comment I have for improving the paper concerns the need for a better demarcation of the two aims of the paper: the background to the benchmarking (the "concepts" part) and the summary of the processes that will actually be undertaken to construct the benchmark data. At times it is not clear which concepts will actually be used in the benchmarking, and which are general considerations. I understand that the authors are conscious of not wishing to influence the blind-benchmarking process, but I feel that more information could be given without adversely affecting this process. This is most apparent when describing the construction of the "analog-clean" dataset (see comment 2 below).

Thank you for taking the time to read this paper thoroughly and provide useful comments. We have tried to address all points and hope that our responses are satisfactory.

In terms of the aims of the paper (concepts vs methodological description) we have done some work to make this more clearly a concepts paper. We wish to keep this as a concepts paper for two reasons. Firstly, this is the first time we have a global scale internationally agreed framework with which to assess homogenisation and so we feel that it is important to lay down exactly what we have decided are the important features of a homogenisation benchmark system and how it would work. This is both so that it is clear for us to refer back to and also that it is a starting point to either build on or challenge and modify. Obviously this isn’t going as far as making it an ISO standard but it is a first step potentially. No doubt ideas will evolve over time though. Secondly, we feel that each of the three components of the benchmarks (clean world creation, development of error worlds, assessment scheme) warrants its own follow-on technical paper. This will allow much more detailed description than we could have here but the approach absolutely requires first that the general concepts be laid out and peer reviewed so we see the current concepts paper as a truly critical first step in the whole endeavour..

We recognise that the paper as it stands is not clear enough in its aim and so we have revised it reasonably substantially to make it much clearer that this is a concepts paper only – and why we think that is necessary. We have also removed many of the figures as we thought that a line of text served just as well (figures 2, 3, 5, 6, 7) and these were probably just wasting space.

SPECIFIC COMMENTS

(1) PAGE 240, LINES 10-11.

The statement "This is especially problematic for large datasets ..." is the key to the entire paper but I feel this message gets lost here. Consider including this statement earlier in the paper and in the abstract. In addition, it could be worth pointing out that although there are many more stations in the ISTI database than previous databases, these are largely in areas where there was previously a high density, i.e. the remote global areas generally remain under-represented.

We have added a sentence into the second paragraph to make this point about automation/lack of metadata to guide/validation adjustments clearer:

"In most cases, these methods are automated, given the large number of stations, and purely statistical due to very poor metadata availability."

We have decided not to discuss the coverage of the ISTI databank as that is described fully in the Rennie et al, 2014 paper (referenced) and clear from Figure 1.

(2) PAGE 240, LINES 23-27.

A clearer description is required when you refer to the benchmark data as representing "truth" (albeit with known inhomogeneities introduced).

We have moved the Introduction around a little bit to make this clearer and expanded in the third paragraph:

"In particular, the ISTI focuses on homogenisation algorithm skill. This can be tested using a set of synthetic temperature records, analogous to real station networks but where inhomogeneities have deliberately been added. As such, the 'truth' about where and what errors exist is known a priori. The ability of the algorithm to locate the changepoints and adjust for the inhomogeneity, ideally returning the 'truth', can then be measured. This community-based validation on a realistic problem is referred to as benchmarking henceforth. "

The papers you describe in the literature review on page 241 (lines 9-29) use either purely GCM data (e.g. Williams et al. 2012) or purely station data that are considered to be homogeneous (e.g. Venema et al. 2012) to generate the benchmark datasets. The process you seem to be suggesting in Section 2 is a combination of GCM-derived components (possibly the l and V components from EQ. 1) and station-derived components (c and possibly the m & v components). If correct, this difference needs to be stated as it marks a deviation from previous benchmarking studies.

The main difference from the other studies is that the ISTI benchmarks will be global in scale, part of an internationally developed framework and hopefully, a closer representation of real station characteristics. We have discussed potential methods to help outline what we are attempting to create but in reality there are various *different methods that could be chosen. These may not be our final choice of methods. For these reasons we prefer not to add further text about GCM or obs only verses GCM+obs and rather clarify the main advantages of the ISTI benchmarks over what currently exists (Introduction, paragraph 10):*

"The ISTI benchmarks should lead to significant advancement over what is currently available. They will be global in scale, a better representation of real world complexity both in terms of station characteristics and inhomogeneity characteristics and provide a repeatable internationally standardised assessment system."

Clearly, releasing detailed information about the construction of the "analog-clean" data at this time may jeopardise the blind-benchmarking process, but the danger in the current draft of the paper is that it is uncertain the degree to which the benchmark data could be affected by inhomogeneities in the ISTI data, if indeed any components are derived from these data.

This is a good point. We do not wish to specify our methods precisely in this paper – rather keep it as an overview. We would rather have three later technical papers which build upon the concepts outlined here and are focussed on each part of the benchmarking (clean worlds, error worlds, assessment) where there would be space to go into sufficient detail and show a number of figures. We have added some text to make the point that whatever methods are chosen, inherent errors in the real data should not be transferred (Section 2, paragraph 3):

"If real world data are used to formulate all or part of a model to synthetically recreate station data, we need to be sure that errors within the real data (random or systematic) are not characterised and reproduced by the model."

(3) PAGE 246, LINE 20:
Consider stating that in the case of the ISTI data at level 3, this QC process is already undertaken.

Actually, no QC has been performed on the ISTI stage 3 data unless a station has been QC'd at source by the owning country or host databank. This is likely to be the case for a number of stations but the intention is that stage 3 data are raw and stage 4 data are QC'd. We recommend that users of the ISTI databank perform their own QC on the data prior to homogenisation (p 246, line 24-26).

(4) PAGE 247, LINES 17-26:
It could be worth explaining some of the problems that may occur from homogenizing to the final time-series segment, such as short time periods or the potential for non-representative data.

We have added the following to that paragraph:

"This creates issues for a user interested in the actual temperature because for any one station the period of highest accuracy may not be the most recent period.

However, it is not really possible to detect which period is the most accurate for each station and having multiple reference periods in the benchmarks would make assessment far more complex and less useful."

(5) PAGE 251, LINES 16-29. The false-detection-rate described relates to detections within time series. It may be useful to consider including some information about the relationship of this false-detection to the number of repeated hypothesis tests that arise from the number of stations involved.

This is a problem when different groups decide to homogenise different subsets of the ISTI databank rather than the entire thing. We have added a note about this in Section 4, paragraph 10:

"However, it is more likely that different groups will select different stations based on their desired end-product. These may be limited to long stations only or limited to specific regions. ***This could be problematic for contingency table assessment given the inherent tendency for false alarm rates/miss rates to grow with increasing numbers of test events (i.e., number of stations)."***

(6) FIGURES 1 & 2: It needs to be stated in the captions to these figures that these are reproduced (at least in part) from Figures 8 and 2 in Rennie et al. (2014).

Good point, done.