Geosci. Instrum. Method. Data Syst. Discuss., 4, C85–C87, 2014
www.geosci-instrum-method-data-syst-discuss.net/4/C85/2014/

Geoscientific Instrumentation Methods and Data Systems Discussions

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).

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.

(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). 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 I 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. 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.

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

(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.

(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.

(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).

Interactive comment on Geosci. Instrum. Method. Data Syst. Discuss., 4, 235, 2014.