

We thank both reviewers for their comments on the discussion paper. We have transcribed both reviewers comments into the current file (and reformatted the listing of the second reviewer's specific queries that got corrupted in conversion to the posted review pdf). We respond to each point below in *italics* for ease of interpretation.

Anonymous Referee #1 Received and published: 6 June 2017

This work present methodology for assessing the level of maturity of measurement networks. The authors face a big challenge defining the meaning of “maturity” of a network after providing metrics for assessing it.

We agree that performing such an assessment is far from simple. To date it hasn't been systematically attempted and this current exercise, to our knowledge, is the first attempt to do so.

In sections 2 and 3 the authors define criteria and levels of maturity associating a numerical scale to different common aspects of measurement networks.

From my experience the choice of the criteria is correct. This mean that I would have chosen the same criteria, but maybe I would have added some other aspect like geographical extension and density of stations. This because the same approach could be used for a network made of one station. However, the authors in their results explore the maturity of 54 networks that are clearly groups of systems and not just individual stations.

We agree that geographical coverage is an important aspect of choosing which data to use for a given application. However, we see this as being distinct from an assessment of the measurements themselves. Both are important and various papers in preparation for GAIA-CLIM will address the geographical coverage aspects for the specific use case of satellite characterisation and validation. We prefer to keep the geographical aspects a distinct consideration. We have added a brief section to the discussion (Section 5.3, which is new) to clarify this point.

The authors are aware of “the inevitable and irreducible level of subjectivity” involved in the process of assessing the scores for individual networks. However, for each network more than an evaluation was performed by distinct assessors (the authors say at least 3), providing estimates with a statistical meaning in some extent. The approach used on my opinion is very similar to those used in modern evaluations of services, but the number of samples used in the statistics looks to be poor.

Indeed, the sample size is a potential issue and one that we already stress in the online versions of the assessment presented via the GAIA-CLIM website. The reality is that for any given network there are very few people suitably experienced and qualified to fully complete a network assessment. That would be true both for the assessment approach being proposed herein or any other similarly in-depth consideration. This is thus a challenge with no seemingly obvious solution. We have added some text to the discussion to acknowledge this in section 5.1 through expansion and splitting of the third paragraph.

Despite this objection the assessors are not a “normally” distributed sample, but highly qualified PI, that in principle should behave accordingly to the scientific ethic. This could give to the assessment for individual network a certain degree of objectiveness. The overall

evaluation of network of networks is based on a more robust statistic and gives a good picture of the distribution of reference to baseline measurement networks.

We have also attempted to acknowledge this in the edits to Section 5.1 noted above.

As the authors suggest in their conclusion this work is a good basis for further discussions and refinements of criteria for assessing existing networks. However, they also implicitly define or state the criteria that should be considered designing future networks and stations that aim to be used for reference. Which by the way are already broadly shared between the scientific communities.

We would agree that these are already broadly shared criteria amongst many in the community. We have tried to reflect this point through a slight revision to the opening of Section 5.2 which makes the point now more explicitly.

Anonymous Referee #2 Received and published: 4 August 2017

This is a comprehensive paper and I only have minor comments on it. Whilst the paper is specifically directed towards the validation of satellite observations, the approach outlined could be applied to a wide variety of data types (it would be worth pointing this out in the paper).

In response to this suggestion we have strengthened the commensurate statement in the abstract. We feel that Section 6 also makes this point at the close as it stands in the original version.

A few specific comments on the paper:

1. I have made no attempt myself to independently verify the assessment in Figure 4 (which would be a massive task in itself).

Indeed it would be a very substantial task to do so and, as discussed at various points in the paper, be very hard if not impossible for any single individual to perform given the required depth of knowledge in all the contributing networks which would be required.

2. There's a slight inconsistency between section 3.1 (which says that six categories are mandatory and one optional) and 4.3.6 (where both software and usage are listed as optional). I assume this was because the process started with the intention of usage being a mandatory criterion and this decision was changed during the evaluation – in which case that should be stated.

We have attempted to clarify this issue in minor edits to Section 4.3.6

3. The final network classification (as described in 4.4) is not listed anywhere in the paper. If not done as a standalone table, it could reasonably be done as an additional column in Figure 4.

We have clarified this by describing more comprehensively the approach in edits to the opening paragraph of Section 4.4.

4. The classification of Reference, Baseline and Comprehensive networks ' as being mutually exclusive categories (rather than Reference being a subset of Baseline, and both a subset of Comprehensive) is a little counter-intuitive – especially when looking at Figure 6 where the only 'Comprehensive' networks are tiny. Perhaps the caption of Figure 6 could emphasise their mutually exclusive nature more?

We have added text to the Figure caption to reflect this. We would also note that text to this effect is already present in the original paper at e.g. lines 149/150. Reference data is a little distinct in that it should include uncertainty information and therefore may be available only in delayed mode owing to the additional processing applied. This would be important for certain applications. So, although networks may well be subsets of one another its not quite 100% the case. Hopefully the text in Section 2 is clear in this regard.

5. P29 ' line 18: 'we shall concentrate in future work upon those classified as Reference'. Do Reference-level networks exist for all elements of interest, and are Reference-level networks sufficient for regional-scale validation? (Figure 6 indicates, for example, that for water vapour, Reference-level networks are sparse outside Europe and North America; I imagine this would apply to many variables).

Reviewer #1 raised a similar point and we have addressed both in the newly added Section 5.3

6. In 5.2.1, it could perhaps be noted ' that there is no obvious mechanism for driving the adoption of a consistent nomenclature – WMO and GCOS are perhaps the most central organisations in this context, but many of the data sets under consideration will have limited or no involvement from WMO or its member countries.

We agree and have added a paragraph to the end of Section 5.2.1 to address this point.