Clim. Past Discuss.,
https://doi.org/10.5194/cp-2017-64-AC2, 2017

Interactive comment on "Effects of undetected data quality issues on climatological analyses" *by* Stefan Hunziker et al.

Stefan Hunziker et al.

stefan.hunziker@giub.unibe.ch

Received and published: 9 August 2017

We want to thank the anonymous referee #2 for the important and helpful comments and suggestions.

Anonymous Referee #2

General comments

This paper uses data from the central Andes to illustrate the impact of systematic data issues on climate analyses. By applying standard and enhanced quality control procedures before data homogenisation, the study shows that systematic biases – which might not be detected using standard QC methods – can have a big impact on the analysis of temperature and rainfall trends.

This paper is a nice contribution to the field of climate data quality, and raises some important considerations that are often overlooked when large-scale climate trend analysis is conducted. The study is succinct and well written, and is a good complement to this group's previous paper (Hunziker et al. 2017).

Some suggestions for possible improvements to the paper are given below for the authors' consideration.

Specific comments

1. The main concern I have about the results from the study is how much data are removed in the enhanced QC procedure, and what impact that would have on the final trends. Around 40% is a significant amount of data, and you'd expect that removing that much information would have an effect on the homogenisation and analyses whether or not the data have issues. It'd be great if the authors could repeat their analysis using a synthetic dataset, or a dataset that does not have systematic data issues and see the impact of removing 40% of the data. That way they could ascribe some statistical significance to their findings. If this is not plausible, then at least a discussion on the role of missing data in the results, OR more detail on where and when the removed data are.

Response 1: We agree with the referee that 40% of the data is a relatively significant amount of information. The effect of removing 40% of the data, however, will strongly differ between different datasets. Furthermore, the results may differ significantly depending on which time series segments are (randomly) excluded. In the authors' opinion, the best approach would be to resample all the available station data quality controlled with the standard QC approach by removing 40% of the data randomly. Repeating this process many times, the resulting climatological analyses from these random groups may provide information about the statistical significance of the results of the present paper. However, the process from the dataset to the climatological analyses is time consuming (clustering, data homogenization, data analyses).

Therefore, such an approach is hardly feasible in frame of the present study. Two main evidences make it very unlikely that the differences found in the present study are just an effect of removing 40% of the data: 1) the enhanced QC removes obvious observation errors that cannot be corrected by data homogenization and consequently affect climatological analyses, and 2) the results of the dataset quality controlled with the enhanced approach are highly consistent (e.g., apparently better performance of data homogenization approach, more consistent station trends, low trend spread of the diurnal temperature range which indicates that the individual data homogenization of TX and TN are in accordance). We will add more information on the availability of time series for the different analyses (for the standard and the enhanced QC approach), and the frequency and temporality of the occurrence of the different data quality issues.

2. In section 3.1.2, I think a little more information is needed about the QC issues mentioned. I know it's in Hunziker et al. 2017, but the two papers are independent and should be understandable on their own.

Response 2: We agree with the referee. We will add information on the data quality issues in form of a table to make the paper fully understandable without reading the IJC publication of Hunziker et al. (2017).

Technical corrections

Introduction line 2: replace "most" with "many national"

Response 3: Will be adapted.

Intro lines 20–30: This part confused me a little. Are you saying that trends actually do vary between neighbouring stations (due to climate factors), or that issues with data quality cause the differences?

Response 4: We will reformulate this paragraph to make it clearer. The observation of varying trends between neighboring stations was a motivation to investigate if UDQI can cause or increase such inconsistencies. Later in our paper, we show that UDQI

indeed may increase spatial trend inconsistencies (Figure 6 in the paper).

Page 3, line 1: as "the" enhanced approach

Response 5: Will be corrected.

Page 3, lines 5–9: The Central Andean region is also a good case study because of its complex terrain, as quality issues/homogenisation is notoriously difficult in such conditions.

Response 6: We will include this aspect.

Page 3, lines 9–12: Change chapter to sections

Response 7: Will be adapted.

Page 3, line 20: add "network" after weather observation

Response 8: Will be corrected.

Page 4, line 5: data "were" quality-controlled

Response 9: Will be corrected.

Page 4, lines 10–15: I'd add that the Durre et al. tests include spatial inconsistency tests.

Response 10: Will be included.

Page 4, lines 27–29: Why doesn't the GHCN-Daily QC approach flag these clearly erroneous values?

Response 11: The reason for the problem is that some extremely high and low numbers did not fit in the maximum allowable number of digits in the GHCN-Daily data format. We will better explain the problem in the paragraph.

Page 4, line 30: add % after 0.35

Response 12: Will be corrected.

Page 5, line 6: on "an" annual time scale

Response 13: Will be corrected.

Page 5, line 26: add % after 0.26

Response 14: Will be corrected.

Page 5, lines 28 and 30 [and throughout the manuscript]: I think you need "a" before monthly and daily time scale.

Response 15: Will be corrected.

Page 6, line 5–6: I suggest you reword to: Note that Hunziker et al. (2017) further suggest the inclusion of additional information derived from 5 metadata into the QC process. This allows the removal station records that were generated under inappropriate conditions such as poor station siting or severe. . .

Response 16: Will be adapted.

Page 6, line 15: one day "a week"

Response 17: Will be corrected.

Page 6, line 30: I'd add reference to Peterson et al. 1998 when discussing using firstdifferences: Peterson, T. C., T. R. Karl, P. F. Jamason, R. Knight, and D. R. Easterling (1998), First difference method: Maximizing station density for the calculation of long-term global temperature change, J. Geophys. Res., 103(D20), 25967–25974, doi:10.1029/98JD01168.

Response 18: We will include the suggested reference.

Page 6, line 31: Spearman (with a capital S)

Response 19: Will be corrected.

Page 7, line 1: add "the" before monthly time scale

Response 20: Will be corrected.

Section 3.4.1. Is there a way to graphically show the different clusters? Perhaps some additional panels in Figure 1 using polygons?

Response 21: The clusters differ for each parameter and for the different QC approaches. Therefore, they cannot be marked in Figure 1. We will include a summarizing figure showing the different clusters in the supplementary material.

Section 3.4.2. It sounds like you've used ACMANT3 because it is fully automated. Perhaps mention this earlier in the section to make that point stronger.

Response 22: We may adapt the paragraph to make the point stronger as suggested.

Page 8, line 14: remove "of"

Response 23: Will be corrected.

Page 8, line 15: remove the hyphen between DECADE and dataset.

Response 24: Will be corrected.

Page 8, line 16: add "and" after requirements

Response 25: Will be corrected.

Page 8, lines 20–24: You need an extra line or two here to define/explain the Theil-Sen estimator, and how you have pre-whitened the data.

Response 26: We will include the required information.

Page 9, lines 5-15: I understand what you're trying to say here, but it's a bit confusing. Consider revising the text.

Response 27: We will revise the paragraph.

Page 13, line 13: "and" Switzerland

Response 28: Will be corrected.

Page 13, line 14: Why are you hypothesising this? Why would the tropics have more UDQI? Do you also mean that there are more likely to be UDQI in developing countries?

Response 29: Weather observation in less developed countries is at a different stage than weather observation in industrialized countries. Therefore, it is likely that more data quality issues occur in data from less developed countries that are particularly located in the tropics. We will revise these sentences.

Page 15, line 5: add "a" before few climate change indices

Response 30: Will be corrected.

Table 1 caption: were, not where

Response 31: Will be corrected.

Table 2: can you add any information about the spread of the data in this table?

Response 32: We may include the standard deviation to each median.

Figure 2 caption: I'd add the word "show" after the words green triangles and red triangles.

Response 33: Will be corrected.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2017-64, 2017.