Interactive comment

J. ten Harkel

Note to the editor and authors: As part of an introductory course to the Master programme Earth & Environment at Wageningen University, students get the assignment to review a scientific paper. Since several years, students have been reviewing papers that are in open online discussion for HESS or BGS, and they have been asked to submit their reports to the discussion in order to help the review process. While these reports are written in the form of official (invited) reviews, they were not requested for by the editor, and we leave it up to the editor and authors to use these reports to their advantage. While several students were often asked to review the same paper, this was not done with the aim to provide the authors with much extra work. We hope that these reports will positively contribute to the scientific discussion and to the quality of papers published in HESS. This report/review was supervised by dr. Ryan Teuling (teacher within the ITEE course at Wageningen University and also associated editor with HESS).

We would like to congratulate Wageningen University for this initiative — peer review is a fundamental aspect of science and scientific publishing and adding it to the curriculum is an excellent idea.

The article of Beck et al., 2018 compares 26 different precipitation datasets and compare these datasets to one another by analysing the Kling-Gupta efficiency score (KGE score). The authors show what the limitations are of the current research performed and explains the added benefit of their research to the science community by highlighting characteristics such as the number of datasets used and the size of the geographical area (the conterminous US). Furthermore, the authors present a clear overview of the performance of these 26 datasets using a gridded KGE score for the period 2008-2017. As a reference to compare these 26 data sets to they used a radar-gauge product (Stage-IV) which has been resampled to 0.1◦. They reduced systematic bias using PRISM data by matching Stage-IV long term mean to the long term mean of PRISM.

The article by Beck et al., 2018 fits the scope of the HESS Journal well. Especially the following line from the scope of HESS: “the study of the spatial and temporal characteristics of the global water resources (solid, liquid, and vapour)”. It provides the reader with a helpful guide in choosing which spatiotemporal precipitation dataset they can use for specific research questions, therefore helping others in their modelling efforts. The research by Beck et al., 2018 also highlights the benefit of the newly updated precipitation datasets, showing the evolution of precipitation monitoring over the years. The manuscript provides a good overview and evaluation of current precipitation datasets. The text is generally well-structured and concise. The conclusion of the article is in line with the evidence provided. Although the manuscript shows only limiting reasons for the performance of individual precipitation datasets, it links very well to other studies performed in this area. It can become an important reference paper for future research that uses gridded precipitation datasets. My recommendation would therefore be to publish the article after some relatively minor issues have been addressed.

Thank you for the thorough review which has helped us to improve the paper.

[minor issue 1] The first paragraph of the chapter 3 Results and Discussion gives the overall performance of all precipitation datasets by calculating the mean median KGE score and the KGE score components for all datasets. I wonder how useful these calculations are. In the next paragraphs the authors show how the datasets are different, so showing a mean median and making such a generalisation to start with is not useful in my opinion. I like the thought of an analysis to find the most important factor determining a high KGE score, however I wonder if for different datasets the results might be different and what the benefit is of using the KGE over normal correlation is correlation seems to be the most important factor. I would recommend leaving this paragraph out of the manuscript or clarify my concerns above. Especially clarifying the choice for KGE.

We calculate mean scores for the KGE and its components to demonstrate that, among the three components, the correlation is on average the "worst" and therefore exerts the dominant influence on the final KGE scores. The calculation of the mean scores is necessary to make this point. We hope this clarifies the issue.

A further recommendation to analyse and assess general performance would be to include an analysis on the error associated to each dataset. Figure 2 does show box-and-whisker plots; however, no further detail is given on the underlying reasons for sometimes large whiskers. I would advise the authors to analysis this spread, instead of only focussing on the median KGE score. Analysis of this spread may prove useful in determining if specific geographic areas are underperforming compared to the median of each dataset.

The data underlying the box-and-whisker plots shown in Figure 2 are presented in Figure 1. It is difficult to ascertain the reasons why a certain dataset performs as it does, but we have made every effort to do so.

[minor issue 2] As a reference to the precipitation datasets the authors used the Stage-IV dataset, which is a combination of radar and rain gauge data, they state that the dataset provides high accurate precipitation estimates. However, the authors introduce PRISM as a correction to the used Stage-IV dataset to correct for long-term mean. Again, they state that this the most accurate monthly dataset. I would like to see a better explanation of why Stage-IV is not sufficient, and the claim of the most accurate monthly dataset should be backed up with at least a reference. Plus, there should be a number showing the difference in long-term mean because at the moment it is not possible to see the difference an assess the necessity of this correction.

We have added a sentence stating that the PRISM dataset has been used as reference in several precipitation dataset evaluation studies and provide three examples: *"It is generally considered the most accurate monthly P dataset available for the US and has been used as reference in numerous studies (e.g., Mizukami et al., 2012; Prat and Nelson, 2015; Liu et al., 2017)."*

[minor issue 3] Why is the WRF dataset included according to table 2, it stopped producing data in 2013, this conflicts with the goal of the mauscript to provide a guide for the reader to choose a dataset that can be used in further research. Also, it is a mismatch to the described analysis period in paragraph 2.3, where the authors state they analyse the period 2008-2017. There are more products that mismatch this analysis period.

The fact that a precipitation dataset is only available for the past does not mean that the dataset cannot be used for research. The WRF dataset, for example, can be used to study the impact of climate change on precipitation patterns in the US.

I would recommend that the authors explain this mismatch between available data en the chosen analysis period. Including an explanation on how this might affect the KGE scores for these specific datasets.

This is indeed a (minor) drawback of the study but one that is impossible to overcome due to the different start and end times of the datasets. This is however unlikely to influence the results since even the shortest period of record of four years (for IMERGHHE V05) is still more than sufficient to calculate robust performance statistics.

[minor issue 4] There are 26 data products mentioned, why is there only special focus on the dataset that have a corrected and uncorrected version in the second part of the article? Please elaborate the choice for these dataset in the introduction.

We do not fully understand the question. However, the distinction between uncorrected and corrected datasets was necessary to avoid unfair comparisons.

[minor issue 5] Paragraph 3.2 lines 24-31: The product SM2RAIN CCI V2 is a possible option for evaluation and correction of other datasets however the KGE of SM2RAIN CCI V2 is only 0.28, in my opinion this conflict one-another, I would like to see this further explained or removed.

We appreciate the suggestion but the two studies we cite clearly demonstrate that SM2RAIN makes it possible to evaluate and correct other precipitation datasets. We refer to those studies for more information.

[minor issue 6] In the introduction, the division between the research questions 1-4 and 5-9 should become clearer, indicating that he second set of research questions is to evaluate the evolution of precipitation datasets.

Questions 7 (now 8) and 9 (now 10) are not related to the evolution of precipitation datasets. The proposed distinction is thus not entirely valid.

Paragraph 2.3 lines 25-26 is already mentioned in on page 3 line 25.

Good suggestion. We have corrected this.

Paragraph 3.7 line 27-18: A product MSWEP is mentioned which is completely new and doesn't add anything to the paragraph before.

We agree and have removed the sentence in question.

In chapter 4 conclusions page 15 line 28, new things are introduced such a rain gauge density as a possible explanation, why?

This statement serves to bring two additional factors to light that should be considered when attempting to generalize the results of gauge-corrected precipitation datasets to other regions.

In the conclusion the actual goal of the manuscript becomes clear, should be clear from the start.

The last paragraph of the Introduction clearly lists the objective of the paper: *"We shed light on the strengths and weaknesses of different P datasets and on the merit of different technological and methodological innovations by addressing ten pertinent questions."*

Page 10 lines 9-11 You state that a bias is expected but this ended up not being the case, please elaborate on the expectation and on which data this expectation and conclusion are based.

We expected a bias to be present *"since PRISM, the dataset used to correct systematic biases in Stage-IV (see Section 2.2), lacks explicit gauge undercatch corrections (Daly et al., 2008)."*

Page 10 line 28, already a conclusion, can be left out here

The statement that *"the results found here for the CONUS do not necessarily directly generalize to other regions"* is important and does not feel out of place to us.

Page 10 line 32 "suggest that its gauge-correction methodology requires re-evaluation", based on what is this statement included, please elaborate or include a reference backing up this statement.

The fact that *"GSMaP-Std Gauge V7 shows a large positive bias in the west (Supplement Figure S2)"* suggests that its algorithm exhibits serious issues. We are not aware of other studies reporting about this issue and hence cannot provide a reference.

Paragraph 3.5 mentions that IMERGHHE V05 performs better than TMPA-3B42RT V7 based on KGE scores, however figure 3a shows that in the west there are significant areas where TMPA-3B42RT V7 performs better, please indicate this in paragraph 3.5

Thanks for the comment. We have added the following: *"In the west, however, there are still some small regions over which TMPA-3B42RT V7 performs better (Figure 4a)."*

Page 14 line 5, reference to a figure form Beck at al., 2017b), would be helpful if the figure is included in the article as a back-up to statements made in paragraph 3.8

We are not in favor of repeating the results of previous studies as this would make the paper significantly less concise.