Reviewer #2

This study compares 26 precipitation datasets with respect to the Stage-IV product at 0.1 degree resolution over the CONUS and at the daily time scale for the period 2008–2017. The Kling-Gupta efficiency is primarily used to rank the datasets, emphasizing the correlation component. The importance of gauge reporting times in daily gauge corrections is highlighted. Examples comparisons between product versions, satellite versus reanalyzes, deterministic versus ensemble reanalyzes are provided.

The topic fits the scope of the journal as it presents an overview of a selection of available precipitation products. The paper is easy to follow and the methodology is clear. However the limitations of such exercise need to be better highlighted. The paper would be suitable for publication after the following comments are addressed:

We want to thank the reviewer for their thorough assessment of the paper. We appreciate the comment that the paper is “*easy to follow*” as this is exactly what we were aiming for.

1. While the use of a score like KGE is convenient for intercomparison exercises, it must be recalled that such an integrative metric only partially depicts actual performances of precipitation products. This is all the more true since the median KGE values are primarily used in this analysis, which further prevents a detailed assessment. Errors are multi-faceted and scale dependent. As the authors point out KGE in this context emphasizes correlation, which tends to bias the assessment by favoring products designed to correctly capture the timing of daily events such as MSWEP. Other precipitation aspects such as daily totals are of primary importance for hydrological applications. The use of another metric could generate different conclusions. As stated by Gupta et al. (2009) who initially proposed the KGE score, “the primary purpose of this study is not to present an improved measure of model performance”, but “suggest possible ways forward that may move us towards an improved and diagnostically meaningful approach to model performance evaluation and identification”. It is recommended to explicit the relevance of the KGE with respect to the assessment
Purpose.

The KGE is indeed an integrative metric that does not immediately reveal which aspects of the datasets are different. We therefore also show detailed results for the components of the KGE (correlation, bias, variability ratio) for all 26 datasets. See Figure 2 of the main paper and Supplementary Figures S1, S2, and S3.

We agree that “*daily totals are of primary importance for hydrological applications*” and this is exactly what the bias component of the KGE evaluates. Summary results for the bias are presented in Figure 2 of the main paper with detailed maps for all datasets provided in Supplementary Figure S2.

We have chosen the KGE for the present study since it is an objective metric that evaluates the most important aspects of data series, as explained in Section 2.3. For this reason, the metric is becoming more and more popular in different fields of science.

Nevertheless, we are in complete agreement with the reviewer that "*the use of another metric could generate different conclusions*" and we already explicitly state this in the paper in the last paragraph of the Conclusions: "*results may differ when using another performance metric or when evaluating other timescales or aspects of the datasets.*" To emphasize the importance of follow-up research using different performance metrics, we have modified the last sentence in the Conclusions from: *"Similar evaluations should be carried out in other regions with ground radar networks (e.g., Europe) to verify and supplement the present findings*" to *"Similar evaluations should be carried out with other performance metrics and in other regions with ground radar networks (e.g., Australia and Europe) to verify and supplement the present findings.*"

2. Stage IV is not a homogeneous precipitation product over the CONUS. Its generation varies across River Forecast Centers, e.g. it relies more heavily on PRISM in the Western U.S. A homogeneous reference is ideal for such assessment of precipitation products, and the lack thereof should be mentioned.

Thanks for the comment. We have added the following sentence including three references: *"The Stage-IV dataset is a mosaic of regional analyses produced by 12 CONUS River Forecast Centers (RFCs) and is thus subject to the gauge correction and quality control performed at each individual RFC (Westrick et al., 1999; Smalley et al., 2014; Eldardiry et al., 2017)."* The poorer performance of Stage-IV (as well as the evaluated precipitation datasets) over the western CONUS is discussed in Section 3.2 and is attributed to *"the more complex topography and greater spatiotemporal heterogeneity of P in the west"*.

3. No seasonal dependency of the performances is reported in this analysis, although it is an important factor. See e.g. Gebregiorgis et al (2018) for a comparison between TMPA and IMERG over the CONUS. For example performances during the winter season are of significant interest especially for snow conditions. Accurate solid precipitation estimation is of primary importance for applications such as water resources management. Can the authors comment this aspect? Reference: Gebregiorgis et al., 2018: To what extent is the day 1 GPM IMERG satellite precipitation estimate improved as compared to TRMM TMPAâ˘A˘RRT?. Journal of Geophysical Research: Atmospheres, 123, 1694–1707. https://doi.org/10.1002/2017JD027606

We have added a new question to the paper which compares the performance of the datasets in summer versus winter (see Section 3.4), providing some very interesting insights. In addition, we cite Gebregiorgis et al. (2018) in the revised paper. Thank you for pointing us to this very useful study.

4. Other important aspects of precipitation such as occurrence or extremes are not assessed in this study, although they are of primary importance for an array of applications of precipitation products. Can the authors comment this aspect?

We agree that precipitation extremes are important for a host of hydrological applications. However, we do not agree that precipitation extremes have not been assessed in the present study, as the correlation (the first component of the KGE) is primarily sensitive to the dynamics of the largest values (i.e., the extremes). In addition, the variability ratio (the third component of the KGE) reflects the distribution of the daily values and is thus sensitive to the magnitude of extremes.

5. Precipitation (solid, liquid and mixed phase) has a large spatial and temporal variability. This scale dependency limits the representativeness of this intercomparison exercise to the daily time scale. Applications such as flash flood forecasting would require an evaluation at finer time scale. Can the authors comment this aspect?

Many of the precipitation datasets evaluated in this study have a daily temporal resolution and we therefore focused on the daily time scale. We agree that a daily time scale is not suitable for short-range flash flood forecasting. The value of an evaluation at a finer temporal resolution has also been brought up by Reviewer #1. We have already performed an evaluation at the 3-hourly scale using Stage-IV (see Beck et al., 2019), but agree that a more thorough evaluation is needed and emphasize the need for more research in the Conclusions.

6. For the above reasons it cannot be stated as in the abstract and in the conclusion that “Our findings can be used as a guide to choose the most suitable dataset for a particular application”. Applications require a refined and more detailed assessment than the one proposed in this study. Please modify this statement.

This comment is in contradiction with the final comment by Reviewer #1. Nevertheless, we appreciate the suggestion and have modified the statements in both the abstract and the conclusion. The old statement in the Conclusions read: *“Our findings can be used as a guide to choose the most suitable P dataset for a particular application.”* The revised statement in the Conclusions reads: *“Our findings provide some guidance to decide which P dataset should be used for a particular application.”* The statement in the abstract has been revised in a similar fashion.