Discussion comment 2

The manuscript describes a newly developed global precipitation data, that takes advantage of existing products and uses a weighing approach to merge that into on consistent product. The manuscript is well written and I think the authors did a great job in creating this new MSWEP product. Nonetheless, I have some comments that I think should be thought about and or addressed in a revised version of the manuscript before publication.

We thank Dr. Wanders for his positive and useful remarks on the m/s.

Major comments
Figure 1, If I understand correctly, the authors use the monthly performance weights for the daily and 3-hourly merging? It is unclear if the authors first compute 3-day precipitation timeseries and then compute 1 annual correlation values between those 3-day precipitation values and the observations or that the 3-day precipitation timeseries are used to compute an annual estimate of the correlations which is dependent on the day of the year (i.e. varies throughout the year).

This may be a misunderstanding. Figure 1 states that, for the daily time scale, we used the same weights as those used at the monthly time scale. Figure 1 also states that the weights used at the monthly time scale are based on 3-day correlations obtained at surrounding gauges. Thus, 3-day correlations are used for the merging at the monthly, daily, as well as 3-hourly time scales. The weights are constant in time, with two exceptions: (1) the satellite weight is zero when the air temperature is <1°C; and (2) the gauge weight depends on the gauge density which varies from day to day. These details can be found in Section 2.3.

In addition to the previous comment, if annual correlations are used or monthly weights are computed as indicated in Figure 1 this will have an impact on the daily merging. I can imagine that some products better capture the 3-hourly variability than the 3-day precipitation average. For example, some of the latest satellite precipitation products might be in this latter category, while I think a product like WFDEI, does a better job in monthly totals. A product like WFDEI is more heavily bias corrected and other observations are assimilated into the product (e.g. soil moisture). Therefore, WFDEI will get a high weight from the monthly analysis, while on the 3-hourly resolution the performance might not be as good as for a satellite product.

In including Figure 1 we did not mean to suggest that we use annual correlations or monthly weights. Furthermore, although WFDEI is one of the five datasets included in the comparative performance evaluation, it is not incorporated in MSWEP and thus has no actual weight assigned to it.

Why do the authors perform a product validation at 0.5 degree? I understand that the gauge density might be low at 0.25 and that more erroneous observations might be included, however,

that is the resolution that product is going to be used at by users. I think including a validation at that resolution might prove valuable, if not in the main manuscript, maybe in a supplement.

We thank the reviewer for this question. The main reason is that not all products are available at 0.25°, and we wanted to rule out resolution differences as a potential cause for the performance differences. However, motivated by the reviewers comment we re-did the analysis using the native resolution of each product. The results were virtually identical however, and we decided not to present them in the m/s.

Why did the authors select this merging procedure and not a more standard Bayesian methods where the errors between the products are weighted and their crosscorrelation is taken into account. I might have missed it, but reading the manuscript, I come to the impression that the cross-correlation in the errors between the different products is not taken into account. I think that a large part of the errors in most of the product show a strong cross-correlation, which could be exploited in the merging of the products. That could further strengthen the added value of the MSWEP product compared to the existing data.

The employed merging method is essentially just a special case of the Bayesian average with the constant *C* set to 0 (see https://en.wikipedia.org/wiki/Bayesian_average). We do in fact take cross-correlations into account: We assumed that most of the cross-correlation exists among datasets of the same type (satellite or reanalysis), and to account for this cross-correlation we introduced Merging stage 2, in which we first compute a weighted average over all satellite products and over all reanalysis products separately, prior to merging the gauge, satellite, and reanalysis components in Merging stage 3 (see Figure 1).

One thing that I missed after reading the manuscript is the development of an uncertainty product. The authors have P anomalies from all products and they have the weights, so they can indicate an uncertainty on their product. This would significantly strengthen the MSWEP product, especially in terms of ensemble modelling. Many studies just make assumptions on the uncertainty of a precipitation dataset when they use them for their modelling studies, while the authors are in the position to actually provide this valuable information to the reader and data user. I understand this might be some undertaking and too much for this manuscript, but I think it could be a valuable addition in the future.

We agree and intend to explore the estimation of uncertainty in the near future. As the Reviewer correctly states, this is a major endeavor and we are currently pursuing it. However, we want to emphasize that this exercise is not as straightforward as it may seem at first, because for most grid cells there are (1) a limited number of independent estimates and (2) large differences in weights among the estimates. An ungauged arctic grid cell, for example, only has two estimates with any weight assigned to them (JRA-55 and ERA-Interim), which is insufficient to reliably quantify the uncertainty. On the other hand, grid cells containing one or more gauges would have at least three estimates. However, the gauged estimate would have a considerably greater weight, confounding the uncertainty quantification. Although in the tropics there are usually

three estimates (CMORPH, GSMaP, and TMPA 3B42RT) with comparable weights, they are not completely independent and hence would lead to underestimation of the uncertainty.

Minor comments
Table 1 PRISM is missing from the table

Table 1 only lists "(quasi-)global gridded *P* datasets". PRISM is a regional climatic dataset and has therefore been intentionally left out.

I feel it would be good to elaborate a bit more on the merging in the manuscript, the assumptions made here are very important for the final product. Why is this method chosen over others etc.

We agree and have added the following to Section 2: "This method was used since it: (i) is relatively easy to understand and implement; (ii) accommodates the inclusion of datasets with 3-hourly, daily, as well as monthly temporal resolutions; (iii) is largely data-driven (i.e., the weights are based on gauge observations); (iv) accounts for cross-correlation among datasets of the same type (satellite or reanalysis); (v) treats random (i.e., temporally variable) and systematic (i.e., long-term) errors separately; (vi) accounts for gauge under-catch and orographic bias; and (vii) yields reliable estimates (as the comparative performance evaluation described in Section 3 will demonstrate)."

How is the performance of the product over mountainous regions, and more specific the Hindu Kush – Himalaya region? Immerzeel et al. (2015) showed in a recent study that most the annual totals of a selection precipitation datasets does not even match the annual discharge (Q > P), which indicates some severe biases in the products. Is it possible with MSWEP to correct for these biases or would MSWEP suffer from the same problems? No HBV validation has been done in this region, while it is a region of major importance with regard to water demand, availability etc. Why not perform a quick check to see if Q < P (long-term average to excluding changes in storage) for most of the GRDC stations and see if the annual totals could at least account for the observed discharge. For some of the original products, this would definitely not be the case. This makes me curious to see if MSWEP can overcome that problem.

Reference: Immerzeel, W. W., Wanders, N., Lutz, A. F., Shea, J. M., and Bierkens, M. F. P.: Reconciling high-altitude precipitation in the upper Indus basin with glacier mass balances and runoff, Hydrol. Earth Syst. Sci., 19, 4673-4687, doi:10.5194/hess-19-4673-2015, 2015.

MSWEP has indeed been corrected for biases in the Hindu Kush – Himalaya region as well in many other regions with even more severe biases. Figure 2c presents the global map of bias correction factors based on *Q* observations, showing that the bias correction factors for the Hindu Kush – Himalaya region exceed two for most of the area. Thus, we are confident that MSWEP performs better than other *P* products in this regard.