Final response in the interactive discussion

Dear Referees,

We would like to thank you for your comments to our manuscript entitled *"Enhancing the consistency of spaceborne and ground-based radar comparisons by using quality filters*" (amt-2018-101). In this document, we would like to provide our responses to the comments of each of the three referees in one single document.

The referee comments turned out to be very helpful. Based on these comments, we suggest several changes to the manuscript which we will outline in detail on the following pages.

For that purpose, we will show the referee comments in black font, and our responses in blue. For the sake of clarity, we have also not reproduced some introductory parts of the referee comments in this comment. Parts that were not reproduced, are marked as *[...]*. Furthermore, we have assigned numbers to all comments to enable cross-referencing between comments. Finally, please find all referenced literature at the end of the response.

We hope that the suggested changes sufficiently address the referees' concerns, so that we can, given the approval of the editor, finalize the revision of our manuscript.

Sincerely,
Irene (on behalf of the author team)

Short Comment # 1 (by Daniel Michelson)

[...]

1) How to characterize data quality? Significant effort has been devoted to systematic representation of weather radar data quality in Europe, through COST Actions (717, 731) and EUMETNET OPERA, giving framework approaches for dealing with data quality. The authors have followed Zhang et al. (2011) for representing data quality resulting from beam blockage. It would be useful to have some context in the paper acknowledging previous work and including a rationale for choosing the Zhang et al. approach.

We now refer to the framework for quality representation in OPERA, and briefly compare it to the approach of Zhang et al. (2011). It should be noted, though, that our general approach is open to other definitions of overall data quality, as long as such an overall value can be used to compute weights. Furthermore, we used only one single quality variable (beam blockage fraction). Hence, the challenge to aggregate several variables to a single index is not prominent in our study.

2) What advantages does this work offer when it comes to addressing topographical beam blockage with GR compared to previous work? The paper references Bech et al. (2003) which is a benchmark paper. There are other implementations of the same approach that use other DEM data, e.g. GTOPO30. The authors' use of high-resolution SRTM data is interesting, but are the results better than using ~1 km GTOPO30 data?

Generally, any increase in DEM resolution should be expected to increase the accuracy of our estimate of the beam blockage fraction (see e.g. Kucera et al. 2004). That effect could be particularly prominent in the near range of the radar as has been shown e.g. for high-resolution airborne laser DEMs by Cremonini et al. (2016). Yet, there is no reference (truth) that could be used to actually verify beam blockage estimates for any underlying DEM. What could be done, however, is to repeat the analysis with GTOPO30 data (i.e. 1 km resolution), in order to investigate the sensitivity of results, or whether our estimate of calibration bias becomes more or less consistent using GTOPO30. Yet, we already consider the paper as quite long, especially considering the requested changes in the course of this review process. We are thus hesitant to include such additional analysis that is not expected to add substantial new insight. Instead, we will add a brief discussion on the potential effects of DEM resolution on the quantification of partial beam blockage.

3) GR calibration. There is one sentence in 3.2(2) indirectly indicating that the Subic radar may have been calibrated during the time period covered by the study. This needs to be clarified. Was the radar calibrated during this time? More than once? Are the results available? Any other maintenance that could have impacted calibration levels? This is very important to understand results like those presented in Figure 8. Also, the methods presented in this paper have been applied to data from one GR, yet they would be much more valuable if also applied to data from a second GR. Doing so would reveal which radar is "hot" and which is "cold" and whether there are any other systematic differences that are unique to each GR.

We entirely agree that it would be interesting to apply the methodology to another or even several other radars in order to investigate specific characteristics of individual radars, or in other words, differences and similarities. Yet, we consider this study as a proof-of-concept in which we present the underlying methodology, and show that it adds value in estimating calibration bias for a single radar. An inter-comparison with several radars would be an additional study. Such a study should not only compare the behaviour of different radars independent of each other. It should also investigate the effect of "recalibration" on the consistency between two or more radars in regions of overlap (c.f. Warren et al. 2018). However, we consider such an analysis beyond the scope of the present paper.

Unfortunately, we were unable, despite repeated attempts, to retrieve detailed information on maintenance operations from the radar operator. The only information provided by PAGASA engineers was that in 2012 and 2013, there were problems with the transmitter which caused the output power to be very low (~80 dBm as opposed to the required 89.2 dBm). The modulator was replaced in 2014. In 2015, the magnetron was replaced. In October 2016, the supplier was changed to SELEX. We also have

to assume that after performing major changes in the radar hardware, the radar engineers carried out a calibration.

Technical "corrections"

4) Including information throughout the paper on what software calls are available and have been used is irrelevant and should be avoided. Instead, I recommend a small section following the recommendations given by Irving: https://doi.org/10.1175/BAMS-D-15-00010.1

In the revised version of the paper, we will discuss the context of Irving (2016). In fact, our paper moves beyond the "minimum standard" suggested by Irving, since we not only provide the code (a doi pointing to the wradlib package version will be generated for the final version of the paper), but we combine the enhancement of an existing software package (wradlib, extensively documented) with a fully documented application context (a jupyter notebook), combined with the data, so that we do not have to provide a log file, as suggested by Irving, but directly allow the user to run the analysis, and modify it in order to adapt to different application contexts. However, we agree that it makes sense to remove the references to explicit function calls from the main text. In order to conform with Irving's suggestions, we also add a brief section with the description of the underlying software, its dependencies, and the notebook to reproduce the results.

5) References to Morris and Schwaller and Schwaller and Morris are inconsistent. In the list of references, both are given from 2011, but the paper references one from 2009. The Morris and Schwaller reference appears to be incomplete in the list of references.
The Schwaller and Morris citations will be updated to consistently refer to Schwaller and Morris (2011). The Morris and Schwaller reference will be updated to show complete information.

6) 2.1.2 Version 6 of the GPM 2AKu products is stated, but is not the current version 05B? Reference(s) to product documentation are needed.

The version will be corrected to 5A instead of 6. The latest version of the product used in this paper is 5A (downloaded February 2018). The references to the product documentation will be added in the paper.

7) 2.3 Where does the information on bright-band height and width come from? Also precipitation type and rain indicators? Please add.

The parameters are extracted from the TRMM 2A23/2A25 and GPM 2AKu products themselves. In the paper we refer to Table 3 of Warren et al. (2018) for the exact list of parameters. We will also clarify that in the text as:"*Several meta-data parameters were extracted from the TRMM 2A23 and GPM 2AKu products for each SR gate, such as [...]"*

The following table describes which parameters were extracted for each product, and how they are used in the analysis. This table will be added to the documentation in the code repository, and also to a supplementary to the paper.

Table A. Parameters extracted from TRMM 2A23 and GPM 2AKu products and the derived variables used in 3D matching

2A23 (TRMM)	2AKU (GPM)	Derived variable
rainFlag	flagPrecip	Rain/no-rain indicator
rainType	typePrecip	Precipitation type
status	landSurfaceType	Surface type
HBB	heightBB	Brightband height
BBwidth	widthBB	Brightband width
dataQuality	dataQuality qualityBB qualityTypePrecip	Overall data quality
correctZFactor	zFactorCorrected	Attenuation-corrected reflectivity
sclocalZenith	localZenithAngle	Zenith angle
-	binClutterFreeBottom	Range bin number for clutter free bottom

8) 2.3 GR data are acquired every 9 minutes, but matched within a 5-min window. How is this done?

With GR sweeps being repeated every 9 minutes, the maximum time difference between overpass and the closest GR sweep would be 4.5 minutes. With a buffer of 30 seconds, we set five minutes as the search window *before* and *after* the overpass. We will clarify that issue in the paper.

9) Figures 5-7. Sub-plot (e) is a great way to visualize this kind of result, but clearer colours are needed. I'm suspecting that light-gray points are covered by dark grey points. A colour table might be a better approach, perhaps combined with slightly smaller point sizes.

We agree that the very light colors for small quality values are difficult to interpret. Then again, we were, after some experiments, unable to adequately convey the visual message of weighting low quality samples less than high quality samples by using two different colors at both ends of the colormap. A color lookup table could not resolve the issue. Instead, we decided to start the "left" end of the colormap with a darker color, so that low quality samples become more visible. We will also implement the referee comment to decrease the point size in order to minimize overlaps.

10) Figure 5 caption: Replace ZPR with ZSR

Z_{PR} will be replaced with Z_{SR} in the caption

11) Figure 8. Might want to clarify in the caption that data from Jan-Apr are not used because this is the dry season.

The suggestion will be implemented.

12) Just a thought: what impact can radome wetting have on the results? Radome wetting is still an issue even if you exclude data near the radar. But is it an issue at all at S band?

In their review paper on sources of uncertainty, Villarini and Krajewski (2010) quoted Austin (1987) in that *"the radome attenuation is significant only for wavelengths smaller than or equal to 5 cm and negligible for wavelengths as long as 10 cm"* (i.e. S-band). Merceret and Ward (2000) reported wet radome attenuation for S-band to remain below 1 dB for rainfall intensities up to 100 mm/h. In summary, we expect wet radome attenuation to be a negligible effect in our study, and we will reference Austin (1987), Merceret and Ward (2002), and Villarini and Krajewski (2010) in order to support that assumption.

Referee Comment # 1 (by Marco Gabella)

[...]

0) I would also suggest the make the final part of the title more specific, for instance " … by using a quality filter based on beam blockage fraction"
The title will be updated to "Enhancing the consistency of spaceborne and ground-based radar comparisons by using beam blockage fraction as a quality filter".

1) I hope the authors can agree with the following three considerations:

a) Quantitative interpretation of radar measurements are based on A MODEL of the backscattering targets.
b) Such A MODEL is an approximation of a very complex reality (Nature).
c) There is never sufficient information in radar measurements to resolve such complexity.

Having said that, I think I can now recommend more emphasis in the text related to the very different wavelengths and sampling volumes (for instance, you may want to have a look at the figures in the paper by Joss et al., 2006) characterizing GR (3 GHz) vs SPR (14 GHz, attenuating frequency!). Yes, one can try to correct for attenuation (e.g., Iguchi et al., 2000), he can even try to convert Z from 10 to 2 cm, but the uncertainties affecting the retrieved quantities are large! (See a) b) c) above …)

By the way, when introducing Eq. (2), you mention Cao et al. (2013) and coefficients in Table 1 for dry snow and hail … I have just quickly opened the pdf and saw that Table 1 lists (retrieval/ simulated) BIAS and RMSE?!?

I am confident that after (re-)considering the above mentioned issues, after thinking of the (necessarily) simplifying approach for beam occultation correction1 (Gaussian shape for the main lobe of the antenna radiation pattern, instead of the simple and practical linear approach proposed by Bech et al., which is an unrealistic "top-hat" radiation pattern), …, the authors will feel more comfortable with what they call "short term variability" at page 15, line 7; furthermore, they will not list "short term variability" at the first place, rather … at the last one!

First, we would like to thank the referee for spotting the mix-up in references: In fact, the conversion coefficients were reported in another paper of Cao et al. (2013), and we will correct the reference accordingly:

Qing Cao, Yang Hong, Youcun Qi, Yixin Wen, Jian Zhang, Jonathan J. Gourley, Liang Liao (2013): Empirical conversion of the vertical profile of reflectivity from Ku-band to S-band frequency, J. Geophys. Res. Atm., 118, 1814-1825, doi:10.1002/jgrd.50138.

Second, we appreciate very much that the referee puts our discussion of "short term variability" into perspective. In order to avoid misunderstandings, though, we would like to emphasize that our notion of "short term variability" does *not* necessarily imply short term variability of ground radar (mis-)calibration. In the paragraph on "short term variability", we reiterate the obvious result that our bias estimates vary substantially between overpasses. We then enumerate potential causes for that variability. Admittedly, putting "hardware instability" first in that enumeration might not be justified if we intended to list the causes ordered by their relevance to explain variability. However, the present study does not provide the scope to further disentangle and rank the different sources of uncertainty, or, in other words, the different sources of ΔZ-variability between overpasses. Yet, we agree with the referee that the role of different wavelengths and sampling volumes, the role of attenuation correction at Ku band, the limitations of assuming a Gaussian antenna pattern, as well as the general limitation of the underlying backscattering model have not been sufficiently highlighted in our list of potential causes of variability. We will add the corresponding discussion and related references to the paragraph.

We will also implement the referee's suggestion to revise the order of points discussed on page 15 of the original manuscript. We will also expand the label of the paragraph from "short term variability" to "Short term variability of bias estimates between overpasses" in order to avoid misunderstandings. In line with comment #15 of this referee, the order will be changed to: 1) Effect of quality weighting on bias estimation, 2) GPM and TRMM radars are consistent, 3) Change of bias over time, 4) Short term variability of bias estimates between overpasses.

2) What happened to the GR in 2014? (lines 20-22, page 15 and Figure 8): +1.4 dB overestimation, after two year of clear and "heavy" under-estimation! (–4.1 dB and –2.5 dB, respectively). What a jump! Was it hardware related? Software? Both? From a weather service viewpoint, it is interesting that this paper bring in the important concept of GR calibration and monitoring, see e.g. the recent successful workshop (https://www.dwd.de/EN/specialusers/research_education/seminar/2017/wxrcalmon2017/wxrcalmon_en_node.html) However, if the authors provided possible explanations of what happened, the paper would become even more interesting and valuable. If you are interested in knowing more regarding monitoring and calibration of modern radar, you may find recent paper regarding: the Transmitter chain (e.g., Reimann et al., 2016); the Receiver chain (for instance, using the Sun: Gabella et al., 2016; Hubbert, 2017,) both Transmitter and Receiver chains, using a 24 GHz vertically pointing radar and disdrometers (Frech et al., 2017).

- Gabella, M.; Boscacci, M.; Sartori, M.; Germann, U. Calibration accuracy of the dual-polarization receivers of the C-band Swiss weather radar network. Atmosphere, 2016, 7, 76.
- Reimann, J.; Hagen, M. Antenna pattern measurements of weather radars using the Sun and a point source. J. Atmos. Ocean. Technol. 2016, 33, 891–898.
- Frech, M.; Hagen, M.; Mammen, T. Monitoring the absolute calibration of a polarimetric weather radar. J. Atmos. Ocean. Technol. 2017, 34, 599–615.

However, maybe, using a more robust definition for the SPR-GR reflectivity Bias, it will come out that the jump is smaller than 3.9 dB; for what concerns the assessment of the Mean Field Bias and its statistical evaluation, please have a look at following point #3)

As of today, we were unable, despite repeated attempts, to retrieve detailed information on maintenance operations from the radar operator. The only information provided by PAGASA engineers was that hardware replacements happened in 2014 and 2015, and the supplier changed in 2016, as already elaborated in our response to Short Comment #1 (comment no 3). As much as we appreciate the recommendations and background information on transmitter and receiver chains provided by the referee, we hope he agrees that, based on the level of information provided by the operator, any discussion of specific causes for the jumps would remain speculatory. We agree with the referee that this lack of information - that should exist somewhere - leaves us somehow dissatisfied.

3) From my viewpoint, the study is a bit limited in the definition of Bias assessment and the corresponding statistical metrics for the evaluation. For instance, it is going to be straightforward for the authors to derive other statistical parameters and present them in a summarizing Table that can complement the nice and informative figure 8. First of all, in addition to the annual mean of {ΔZdB*} (lines 20-22, page 15) also the standard deviation of {ΔZdB*}. Then, I would suggest a more robust definition for the Bias: instead of using dBZ, you use Z values in linear units: Z=10(dBZ/10). Then you derive a weighted average for the numerator (denominator) using linear Z of the GR (SPR). Finally, you compute 10 Log of such ratio (dB). This annual Log_of_the_MFB is more resilient than the Mean_of_the_Log presented in the paper. To avoid weighted-average or in a probability matching scheme, you may want to consider only bins with QBBF larger than say, 0.9 (or larger). After having done this selection, consider the difference (BIASxx) between different quantiles (probability matching): xx=50

(median), 75, 84, 90, 95, 99 percentiles. Maybe, BIASxx is not constant, rather it depends on the percentile? Finally, for these QBBF-selected bins, you may explore the value of the average bias E{ΔZdB} as a function of the intensity of the echo of the GR (using for instance intervals of 3 or 5 dBZ; obviously, you will have less and less samples for larger values of dBZGR). Does this Mean_of_the_Log Bias remain more or less constant? Or do you see a trend? (Maybe, SPR has residual attenuation for large reflectivity values?). Interesting, is not it?

We would like to thank the referee for sharing these ideas. We will implement the suggestions as follows:

- We will add a visual representation of the standard deviation of the annual mean {ΔZdB*} in figure 8;
- We will recompute the bias estimates based on the referee's suggestion to first convert reflectivity to linear units before computing the weighted average;
- We will analyse the sensitivity of results in case we replace the weighted average by a simple quality threshold below which the samples will be discarded in the computation of calibration bias; however, we have the feeling that the paper is already very long, so we suggest to put the results of that analysis in a supplementary and only briefly refer to that in the main paper. Of course, using only partial beam blocking as a quality variable has very specific implications as to the effect of thresholding: any additional sample that exhibits a higher degree of partial beam blockage, and that we include in our computation of average reflectivity, will lead to a lower estimate of average ground radar reflectivity in the sample. Then again, reducing sample size through excessive filtering increases the standard deviation. That problem cannot really be resolved, but using the weighted average appears to us as the least arbitrary solution;
- Finally, we will add an analysis in which we investigate the dependency of our bias estimate on the intensity of the ground radar echo. Again, we suggest to present the results of that analysis in a supplementary, and only briefly discuss them in the main text.

4) Finally, the last issue is related to literature: while several TRMM PR vs GR papers are listed, there is a lack of DPR-related studies and DPR technical literature. regarding the latter, I have suggested at the end (GPM related references), three papers published in 2014 and 2015. Regarding the former, I have listed our recent DPR-related studies in the complex terrain of Switzerland; I am confident the authors will be able to find additional GPM papers also in other parts of the world. Furthermore, is Cao et al. double citation (at page 6) correct? Does Morris and Schwaller (2009) exist? (line 7, page 1 citation) Regarding GPM, please, do not forget to mention that your analysis neglect Ka-band observations (please briefly discuss the reason of such a choice).

Indeed the references were incomplete, missing DPR technical literature. References will be added accordingly, including intercomparisons between ground radar and the GPM DPR.
We will remove one of the references to Cao et al. (2013) on page 6, and correct the actual reference (as already pointed out above). The Morris and Schwaller (2009) citation mistake will also be corrected.

We will add a brief note that GPM Ka-band observations have not been considered in the present study, reasons being Ka band being more prone to attenuation, and limited validity of the Rayleigh scattering hypothesis in a substantial portion of rainfall cases (see e.g. Baldini et al. 2012).

[...]

5) Introduction

- Line 4: … to monitor the bias of the gauge adjustment factor to be applied to precipitation estimates of the GR.
- Lines 6: … to quantify the GR reflectivity bias with respect to the reference (namely, SR reflectivity value after conversion from Ku-band to S-band).

The above mentioned lines in the abstract will be revised accordingly.

6) In fact, I would propose the following terminology:

- Setting the Bias as close as possible to 0 dB between radar QPE and in situ measurements: *gauge-adjustment*
- Assessing the Bias between reflectivity of two radars: *relative calibration*
- Forcing to 0 dB the Bias in measured Power (dBm) between an external or internal reference Noise Source and the radar at hand: *absolute calibration*

We agree with the suggestion and will revise the manuscript accordingly, introducing the labels “gauge adjustment”, “relative calibration”, and “absolute calibration” in the first section.

7) Section 2.1.2

The fact that only 283 overpasses were within the selected, reasonable 120 km range should be mentioned here. There is no reason to wait until former(see **) sec. 3.1.1. Similarly, you can at least anticipate that the number will considerably decrease upon conditional requirements such as min. # of “wet” pixels, time difference, min. # of bins above both GR and SPR sensitivity.

We agree that it might be confusing to state numbers of overpasses or valid samples without already anticipating the effect of spatial limitations or additional filter requirements. We will revise the manuscript accordingly by stating these effects early in the paper.

8) Question: have you only used only months from June to November? Not clear from the text. Please rephrase. In fig. 8, I see two overpasses in December (2012 and 2014). By the way, in Dec. 2012 E{ΔZdB*} is almost 5! dB, while the annual average is -4.1 dB?!? (see my previous points 2) and 3))

We used only the months from June to December, which coincides with the rainy season in the area. We will clarify that in the text. As for the case of December 2012, upon checking the particular GR-SR match (December 5, 2012), the value of ΔZ^* is indeed very high (4.2 dB) compared to the average. Looking at the GR and SR data, the number of samples seems sufficient (n=382), and the GR overestimation is

consistent for the different elevation angles. As a result, we cannot provide a consistent explanation for this outlier.

9) By the way, I would propose the following structure for Sections and Subsections
2. Data
2.1 Spaceborne Precipitation Radar (SPR)
2.2 GR

3. Method
3.1 Partial beam shielding and quality index based on beam blockage fraction
3.2 SPR-GR volume matching
3.3 Assessment of the average reflectivity Bias

Section 3.1: I would move it (including former fig. 4) inside the new Section 3.1 (former Sec. 2.2)

4. Results and Discussion
4.1 Single event comparison
4.1.1 Case1
4.1.2 Case2
4.2 Overall June-November comparison during the 5-year observation period

We would like to thank the referee for this suggestion. We agree with the proposed structure, and will update the paper accordingly.

10) Page 5, Line 4-5: Please delete the sentence, the reader is able to read the simple algebra in eq. (1).
The sentence has been deleted.

11) Page 10, Lines 19-25: misleading. I cannot possibly agree. On the contrary, my interpretation is that partial beam blockage plays approximately the same role (0.7 dB difference between the silly estimate that include blockage and the conservative one that exclude all cases where BBF > 0.5). Please rephrase.

We agree that our interpretation was hard to follow, and the corresponding paragraph kind of confusing. That was also pointed out in comment #11 of referee #2. We will rephrase that part of the paper, and hope that our point will become clearer - because we still think it is quite an important one! At a low elevation angle, substantial parts of the sweep are affected by **total** beam blockage. The affected bins are either below the detection limit, or they do not exceed the GR threshold specified in Table 2 of the manuscript. As a consequence, these bins will not be considered in our matched samples, and will thus not influence our bias estimate - irrespective of using partial beam blockage as a quality filter. At a higher elevation angle, though, the same bins might not be affected by **total** beam blockage, but by **partial** beam blockage, as also becomes obvious from Fig. 4 of the manuscript. If we consider these bins in our matched samples, they will cause a systematic error in our estimate of calibration bias,

unless we use the partial beam blockage fraction as a quality filter by computing a quality-weighted average of reflectivity. As a consequence, the effect of quality-weighted averaging (with partial beam blockage fraction as a quality variable) can be most pronounced at "intermediate" elevation angles, depending of course on the specific topography and the relative position of the ground radar. We had referred to that effect as "counterintuitive" since one might naively expect that the detrimental effects of beam blockage on our estimate of calibration bias would *generally* decrease with increasing elevation angle.

12) Page 13: Would you please add a complementary figure at ELEV= 1.5 for the 1.10.2015 overpass? Just like you did for the 8.11.2013 overpass.

We thank the referee for the suggestion, however, we are hesitant to add the additional figure as it does not provide additional insight as compared to the comparison of two sweeps for 2013-11-08, while adding to the length of the manuscript. As a compromise, we suggest to add the additional figure to the supplementary material.

13) Page 14 and Figure 8. Some journals ask for a graphical abstract as a self-explanatory image to appear alongside with the abstract. I think Fig. 8 would be perfect for such scope. It is nice and rich of information. Suggestion: could you please use color. For instance, the 1.10.2015 and 8.11.2013 overpasses could be in color. By the way, the 8.11 circle in picture a) seems to be very close to 0 dB, while in Fig. 7 it is written that E{∆ZdB*} is -1.1 dB. Am I missing something? Is it related to what you wrote in lines 3-6? These sentences are not clear to me, could you rephrase, please? Furthermore, regarding picture b), do not forget to emphasize that if the QBFF works properly then: E{∆ZdB*}– E{∆ZdB} should be negative in 2012 and 2013 (almost all the point in a) are below the 0 dB dotted line), positive in 2014 (almost all the point in a) above the 0 dB dotted line …).

We thank the referee for the suggestion to highlight the two case studies in Figure 8 by color, and we will implement the suggestion accordingly. We are also grateful for suggesting a potential error, however, in this case, we do not agree: the triangle for Nov 8, 2013, represents correctly the bias estimate on that date, as an average over samples from all sweeps (-3.7dB). Apart from that, Fig. 7 refers to the overpass on October 1, 2015.

We also thank the referee for pointing out the issue of negative differences E{∆Z*} - E{∆Z} in Fig. 8b which we missed to discuss sufficiently in the manuscript. First, we would like to clarify that if the QBBF works properly, the difference E{∆Z*} - E{∆Z} should be positive - the areas suffering from partial beam blockage registers weaker signals (i.e. lower reflectivity) than expected producing the "old" lower mean bias, and giving them low weights in the calculation of mean bias brings the "new" (quality-weighted) mean bias up. In the same vein, the difference in standard deviation should be negative - the "new" standard deviation that considers quality is lower than the "old" standard deviation that does not consider quality, so that the difference between "new" and "old" standard deviation is negative. The negative differences E{∆Z*} - E{∆Z} are therefore inconsistencies, caused by the effect of filtering in the case of very small sample sizes. We will include this clarification in the revision.

14) Page 15. I would change the order of your points and list your point (1) at the end, as # (4) [see my comment 1) at page 1)]. I would start from (3), which is the scope of this paper: indeed an intelligent weighted-average based on QBFF shows a better standard deviation of ΔZdB*. By the way, I recommend adding a table and/or a figure (histogram) that summarizes the statistical properties of σ*{ΔZdB*} and σ{ΔZdB}. Then, I would introduce the important result regarding the consistency of GPM and TRMM radars, followed by the changes of the bias in time

As already pointed out in our response to comment #1 of the referee, we will change the order of points as suggested. However, we decided not to introduce additional figures in terms of histograms of bias, differences in bias, or standard deviations. These histograms would have to be provided separately for each year, because it is obvious from the time series that they would represent different populations. Apart from avoiding to introduce many new figures, the informative value of these histograms is not too high due to the limited number of samples. Instead, we will implement the referee's suggestion from comment #3 by including the standard deviation of the annual mean {ΔZdB*} in Fig. 8a.

15) Page 16.
Line 5, delete coherent.
The word "coherent" has been deleted.

16) Line 14-16. Sorry, you cannot summarize the (mis-) calibration of the GR by simply going from 2012 (–4.1 dB) to 2016 (+0.6 dB) and omit, for instance, the +1.4 jump in 2014. [see my comment 2) at page 2)].

We will revise the manuscript accordingly by providing a more complete and coherent summary of the temporal changes of our bias estimate.

17) Line 17-19. Pleonastic. I would delete it.

We would like to refer to our response to the referee's comment #11: we hope that we were able to clarify a misunderstanding there. Given that the referee agrees with our clarification, we think that lines 17-19 on page 16 are not pleonastic, but rather an important note to emphasize that moving to higher elevation angles does not necessarily help to avoid the problems introduced by beam blockage in the specific case of comparing GR and SR observations. Nevertheless, we will also revise the corresponding paragraph in the conclusions section in order to make it more comprehensible.

18) Line 26. Why do you discuss C-band radar technology ?

Lines 24-28 on page 16 of the original manuscript were intended to provide a brief perspective for future studies, in which we mention that for C-band radars, it would be important to include path-integrated attenuation as a quality variable. In the revised version, we will clarify that point.

19) Minor points

My proposal for radar acronyms: 2-character for ground, namely GR; 3-character for satellite radar. Would you please use TPR for TRMM, DPR for GPM and SPR in those cases where you refer to both, independently of the platform

We appreciate the suggestion. Yet, we think that distinguishing the different spaceborne platforms via acronyms might cause more confusion than clarification, in particular since we rarely address the different platforms separately in the main text. We would thus prefer to stick with GR vs. SR in general.

Referee Comment #2 (Anonymous)

[...]

Comments on other sources of uncertainty in calibration assessment:

1) Attenuation at Ku-band:
The authors should address the uncertainties with attenuation correction at Ku-band. The attenuation correction tech. used for just Ku-band is the HB-SRT method (Seto and Iguchi 2015). It is known that using the HB method alone does not work well in higher rain rates (> 20 mm hr-1, Seto and Iguci 2011, but as low as 12 mm hr-1 Rose and Chandrasekar 2005). Furthermore, the SRT method is more uncertain over land (larger standard deviation of the surface backscatter cross-section, Meneghini et al. 2000). It is anticipated that since the radar is located in the tropics both of the issues above could occur (more likely in convective precipitation). Please discuss these uncertainties and how they could impact your results of the bias correction.
It is mentioned in the conclusions that for C-band attenuation correction is vital, but GPM and TRMM are Ku-band, thus isn't it vital as well?

We agree that attenuation correction is vital for both GPM and TRMM at Ku-band, and there is certainly a large body of literature concerned with the related effects, including the effects of nonuniform beam filling (NUBF) on the attenuation correction procedure. In the present study, we have only used the attenuation-corrected reflectivity values without considering the uncertainty associated with the correction procedure. In the revised manuscript, we will explicitly refer to the uncertainty introduced by attenuation correction. We will also, in the conclusions, provide an outlook on including the spaceborne reflectivity observations in the framework of quality-weighted averaging, just as we suggested for the ground radar observations. That would imply to use the estimates of PIA which are provided through the SR meta-data as a quality variable and thus to consider it in the quality-weighted average of SR reflectivity in the matched samples.

2) Ground Clutter for the SR:
In radar gates near the surface, with respect to the SR, ground clutter is a problem. How are the authors dealing with ground clutter from the SR? Are they using gates below the lowest clutter free bin estimate (included in the GPM file)? If so, is the lowest clutter free gate being

assigned to all the gate below it? If you plot it out, a lot of times that's what is done. Essentially the data looks smeared from the lowest clutter free bin to the surface, which isn't to realistic and it is suggested to just not consider these gates. Please comment on this, potentially in Section 2.3. If you are including these interpolations, you may wish to not (it will introduce error).

Thanks for pointing out this issue which has not yet been sufficiently clarified in the original manuscript. While TRMM 2A25 contains a clutter flag for the variable "Corrected Z-factor" (-8888 indicates ground clutter), the GPM 2AKu product contains a variable "binClutterFreeBottom" to indicate the lowest clutter free bin in a ray. In both cases, TRMM and GPM, we use the SR clutter information to discard the affected bins. We will clarify that point in the revised manuscript, using both table 2 (filtering criteria), and the new table with metadata variables that we introduced as a response to comment #7 of SC1 (as part of the the supplementary).

3) NUBF:
Please also include some discussion of the potential impacts of non-uniform beam filling (NUBF) on your analysis. Edges of large systems, individual cumulus showers could result in NUBF in SR because of the quasi-large footprint. Lowering the reflectivity value in the gate.

We agree that non-uniform beam filling can cause errors in particular for the SR platform which might become more pronounced in case of path-integrated attenuation is present and being corrected for. Durden et al. (1998) provided an excellent discussion of potential effects. Han et al. (2018) attempted to consider the effect in case GR and SR observations are matched, by using the - comparatively highly resolved - GR observations in order to compute the standard deviation of reflectivity in an SR footprint as a measure of NUBF. From the literature, it is hard to tell how much systematic error is introduced in SR measurements by the effects of NUBF. However, the three comments of this referee (reg. attenuation, clutter, NUBF) were very helpful for us to understand the necessity of extending the framework of quality-weighted averaging to the SR, too. So while we consider our present manuscript as a proof-of-concept in the consideration of quality, follow up studies should attempt to achieve a more general implementation that not only includes additional quality variables for the GR data, but that also applies these to the SR observation which already come with extremely rich and helpful meta-data to support such attempts. While our study tries to minimize the effects of NUBF (by setting a minimum fraction of GR bins within the SR footprint to exceed a minimum reflectivity threshold, see table 2 of the original manuscript), a future framework for SR quality might rather consider the variability of GR bins in the SR footprint, as suggested by Han et al. (2018).

Specific Comments:

4) Page 2, line 5: Please add the Kummerow et al. (1998) paper for TRMM, and the Hou et al. (2014) for the GPM reference (page 2, line 6). This will help readers who are not entirely familiar with both platforms.

The Kummerow et al. (1998) and Hou et al. (2014) citations and references have been added.

5) Page 6, line 3: “The gates below and above the brightband were considered in the comparison”. Please provide a brief reason why this is done. I do not want to assume the author's reasoning.

According to Warren et al. (2018), the frequency-corrected reflectivities within the melting layer (bright band) appear underestimated compared to the ones below and above the melting layer. In addition, while usually the samples above the brightband are used in GPM validation, there are significantly more samples below the melting layer, especially in a tropical environment such as the Philippines.

6) Figure 4 & Section 3.1: It is not clear what you are plotting. The figure titles state the quality index but the figure caption and text states beam blockage fraction. Please clarify.

The caption has been updated to match the figures: *Quality index map of the beam blockage fraction for the Subic radar at (a) 0.0° (b) 0.5° (c) 1.0° and (d) ° elevation angles.*

7) Section 3.1.1: Why are the number of overpasses here different than when they were listed earlier (section 2.1.2)? I am referring to the numbers before applying the criteria in Table 2.

Applying the criterion of “Minimum number of pixels tagged as rain = 100” eliminates several overpasses. Only this criteria affects the number of overpasses, not the others listed in Table 2. We will clarify this in the paper.

8) Case studies (Section 3.1.2 and 3.1.3): Could you include the mean BB level height? You can add it to the bottom right with the other statistics. Also comment on fraction of stratiform vs convective. These two will help readers assess the amount of attenuation and NUBF that could be involved (e.g. uncertainty in the SR measurements).

The mean BB level height will be added to the figure as suggested. While stratiform rain dominates the precipitation type for most cases, convective rain is significantly represented, hence we decided to keep both rain types in the analysis.

9) Figure 5 + 6 + 7 a and b: Suggestion. Consider changing the colorscale to one that is perceptually uniform and color-deficient friendly. For example, try the HomeyerRainbow or the LangRainbow included in Pyart (https://github.com/ARM-DOE/pyart)

We thank the referee for the suggestion. Upon trying the different colormaps proposed, we decided that we will go with the HomeyerRainbow colormap. The figures will be updated to reflect the new colormap.

10) Page 10, Line 12: “Major parts of that sector did not receive any signal due to total beam blockage”. Where is this occurring? The reader can refer back to Figure 4, but it might be

helpful to outline the circles with a thin black line in Figure 5d where there is SR data, but no GR data. That way the readers would see where there is 100% beam blockage and thus no signal from the GR, but also gain insight of size of the precipitating system.

The figures for the case studies show only the matched bins, but the referee is right, information such as location of bins where there is SR signal but no GR signal and the size of precipitating system are not conveyed. We will address this by showing all the available SR bins for the first panel and outlining the circles with SR data but no GR data in black, as suggested.

11) Page 10, Line 24-25: "That might be considered counterintuitive, as one might expect the blockage to disappear with higher elevations". Please provide some discussion explaining why this is the case.

We thank the referee for pointing out the lack of adequate explanation. As can be seen also from the comments of referee #1, this paragraph appears to be confusing in the original manuscript. We will revise the paragraph accordingly in order to make our point clearer. Please also refer to our response to the comment #11 of referee #1.

12) Page 16, Lines 13 – 16. 'We could' and 'we could also' imply that you did not conduct this analysis when it seems you have. I suggest to change these phrases to be definitive. 'We showed that…' 'we also demonstrated that…'
The sentences will be updated as suggested.

Technical corrections:

13) Page 3, line 20: The most current GPM version is version 5, version 6 is not released yet.
The version will be corrected (version 5A instead of 6).

14) Page 18, line 18: Reference Cao et al. 2013 is incorrect. It should be:
Empirical conversion of the vertical profile of reflectivity from Ku-band to S-band frequency
We apologize for the mixup. The citation and reference will be corrected to refer to
Cao, Qing, Yang Hong, Youcun Qi, Yixin Wen, Jian Zhang, Jonathan J. Gourley, and Liang Liao. 2013. "Empirical Conversion of the Vertical Profile of Reflectivity from Ku-Band to S-Band Frequency." *Journal of Geophysical Research: Atmospheres* 118 (4): 1814–25. https://doi.org/10.1002/jgrd.50138.

15) The reference Warren et al. should be 2018, published Feb 2018 in J. Atmo. + Ocean. Tech.. Page 2, line 8;Page 3,line 25;Page 5,line 11;Page 15,line 14
The citations and reference will be corrected.

16) Figure 4: Missing y-ticks and tick labels on bottom left subplot

Axis labels will be restored in Figure 4. The color scheme has been changed so that the lightest color is made a bit darker for better visibility in Figures 5-7 subplots d and e, following the suggestion of another reviewer.

17) Page 8, line 5-6. No need for new paragraph. You can combine the two.
The paragraphs will be combined as suggested.

18) Figure 5: Figure caption has Zpr instead of Zsr
Z_{PR} will be replaced with Z_{SR} in the caption

References

Austin, P.M. (1987): Relation between measured radar reflectivity and surface rainfall. Mon Weather Rev., 115, 1053-1071.

Baldini, L., V. Chandrasekar, D. Moisseev (2012): Microwave radar signatures of precipitation from S band to Ka band: application to GPM mission, European Journal of Remote Sensing, 45:1, 75-88, DOI: 10.5721/EuJRS20124508.

Biswas, S. K. (2017): Cross Validation of Observations from GPM Dual-Frequency Precipitation Radar with S-Band Ground Radar Measurements over the Dallas — Fort Worth Region. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Fort Worth, TX, USA: IEEE. https://doi.org/10.1109/IGARSS.2017.8127393.

Cremonini, R., Moisseev, D., and Chandrasekar, V. (2016): Airborne laser scan data: a valuable tool with which to infer weather radar partial beam blockage in urban environments, Atmos. Meas. Tech., 9, 5063-5075.

Durden SL, Haddad ZS, Kitiyakara A, Li FK (1998): Effects of nonuniform beam filling on rainfall retrieval for the TRMM precipitation radar. J. Atmos. Oceanic Technol. 15: 635.

Han, J., Z. Chu, Z. Wang, D. Xu, N. Li, L. Kou, F. Xu, Y. Zhu (2018): The establishment of optimal ground-based radar datasets by comparison and correlation analyses with space-borne radar data, Meteorol. Appl. 25, 161-170.

Kucera, P. A., W. F. Krajewski, and C. B. Young (2004): Radar Beam Occultation Studies Using GIS and DEM Technology: An Example Study of Guam. Journal of Atmospheric and Oceanic Technology 21 (7): 995–1006.

Merceret, F. J., J. G. Ward (2002): Attenuation of Weather Radar Signals Due to Wetting of the Radome by Rainwater or Incomplete Filling of the Beam Volume, Technical Report NASA/TM-2002-211171, NAS 1.15:211171, 20 p., URL: https://ntrs.nasa.gov/search.jsp?R=20020043890

Villarini, G., W. F. Krajewski (2010): Review of the Different Sources of Uncertainty in Single Polarization Radar-Based Estimates of Rainfall, Surv Geophys (2010) 31:107–129