Author's replies to the referee's comments to manuscript AMT-2021-34

Please see the file with track changes to check the edited lines.

The original referee's comments are written in bold and the author's replies are written in regular font.

Anonymous Referee #1, revision 2

The authors addressed the two main concerns I have previously raised. Indeed, I would still like to see a more thorough discussion referencing the relevance of the power-law coefficient calibration as a complement method with respect to the presented approach (e.g., there have been studies presented different ways to calibrate the power-law coefficients for a specific climate zone using the CML's standard min/max measurements and a rain-gauge, and not only disdrometer-based calibration). However, since this issue is now somewhat mentioned in the paper, it is not a major issue anymore.

Dear Referee #1 , we appreciate our revision and wrote the following sentence related to power law coefficients.

329-335: "…A physically-based approach which derives these coefficients from drop size distribution observations and scattering computations is preferred compared to optimizing these coefficients in a statistical manner. Especially, for frequencies higher than 35 GHz. The drop size distribution dependence of the k-R relation in the frequency range of approximately 20-35 GHz is considered small compared to errors from wet antenna attenuation or erroneous wet-dry classification. Although a physically-based approach is considered better, a calibration of power-law coefficients may be a way forward for regions which lack disdrometer data (Ostrometzky and Messer, 2020)…"

Anonymous Referee #2, revision 2

The authors have provided a substantial major revision. They have redone their analysis according to my suggestions and have updated the manuscript. The updated analysis did not change the results much, but the methodology is now much more sound. My only complain is that the newly added text is often not of good quality and hence should be revised carefully. In summary, I only have some minor and specific comments that can be addressed in a minor revision.

We gratefully thank the Referee for the constructive comments and recommendations. We carefully checked the quality of the added text.

Minor comments:

1. The impact of false-positives and false-negatives that results from the wet-dry classification could be discussed in more detail. Besides providing optimized RAINLINK parameters, this manuscript also shows the sensitivity of the wet-dry classification to its parameters. Most important, it also shows the challenge of wet-dry classification and its limitations, i.e. there is still a significant number of false positives and false negatives even after optimisation. This is important information and hence should be pointed out more clearly. See also my comment on section 3.2.1.

Reply:

We rewrote this section giving attention to the limitations and the impact of false-positives on rainfall estimates. Please see your comment on section 3.2.1.

2. I am missing a table with the optimised parameters so that they are quickly to grasp. Maybe this info could just be added to table 1 and table 2. Or the optimised parameters could be shown together with their performance metrics in comparison to the default values and those from other RAINLINK calibrations. The later option might be hard to put into one table without making it confusing, though.

Reply:

We added a Table in the conclusions section. Now the values of parameters can be checked easily.

Specific comments (line numbers refer to the diff version):

L19-23: This still sounds as if there is an ongoing decline of rain gauges that is evident from the data availability plot of GPCC. Writing that the GPCC database "underwent a decline" sounds as if GPCC would get less and less data. From how I understand the last GPCC report that I referenced in my last review, this is not true. There is a constant increase of data. The largest portion arrives at irregular intervals and with large delay, though. There might be a global decrease of rain gauges which are in operation, but
the GPCC data availability plot cannot be used to deduce such a trend. I suggest to reformulate this section once more.

Reply:

We reformulated this section given attention to the constant increase of the GPCC database. We added the following sentences:

L19-26: "... Another issue is the data availability of ground-based measurements. For instance, the largest worldwide rain gauge database, maintained by the Global Precipitation Climatology Centre (GPCC) had 45,000 rain gauges in 1961-2000 and down to 10,000 after 2016. This decrease was caused a delay in data delivery and by post-processing at GPCC (Schneider et al., 2021). Although, decreasing in the past due to quality control, the GPCC database has been increasing in recent years as a result of delivery of updates as well as supplements with additional stations and long time-series of data (Schneider et al., 2021)..."

L62: I would not say that "data-driven solutions are not feasible for places or countries without sufficient reference data". The training can, of course, only be carried out in regions with sufficient reference data. But the trained methods, like the ones in the references that you cite, can potentially be used with data from any region. Transferability can be questioned, though. But this is also true for most other CML processing methods which are typically developed with data from only one climatological region. The big disadvantage I see with data-driven solution is that you cannot readjust them to a new dataset just be tuning two or three parameters. I suggest to slightly rephrase this new section.

Reply:

We rewrote:

L63-69: "...These data-driven solutions also hold a promise for ungauged areas, but it will not be feasible for places or countries without sufficient reference data to train the machine learning algorithms. That is, data-driven models require a huge number of observations to learn and detect the whole behavior of the phenomenon to be modeled. For other algorithms, such as RAINLINK, it may still be feasible to at least tune a few parameters, for instance, by employing drop size distribution observations (from a region with a similar climate) to obtain more appropriate coefficients of the relationship between specific attenuation and rain rate..."

L85: This new sentence is hard to understand. Please reformulate and/or split into two sentences.

Reply:

We rewrote and tried to simplify the sentence as:

L89-91: “…In fact, many optimum solutions can occur, corresponding to different parameters sets (a phenomenon known as equifinality)…”

L94: the part "…, also we…" does not sound like correct English to me. Anyway, a new sentence could be started here.

Reply:

We modified this as follows:

L100-101: “…Moreover, we optimize for the first time the main RAINLINK processes, i.e., wet-dry classification and rainfall retrieval, separately…”

L176: My question from the last review is still not answered: "How is this relative importance related to the parameter range that was selected?" Let's say, you select a too small parameter range because you do not yet know the sensitivity. Then the "relative partial effect", which as far as I understand, will depend on the absolute step size, which will be very small for the too small parameter range. So my question is not, what is the relative step size, but how the parameter range, which influences the absolute steps size, could impact the importance of a parameter in this analysis.

Reply:

This is an important point, however, we have not tested different parameter ranges to obtain such insight. We delimited the ranges based on expert judgment and trial runs and believe that the employed ranges are quite wide. We can consider this question as a research gap to be analyzed in the future.

We wrote in the conclusion section:

L492-495: “…Further research can be conducted to test how the parameter range affects the importance of parameters in this approach. Specifically, even wider parameter ranges could be tested. Moreover, a longer calibration period could be analyzed to make the optimized parameters more generally applicable to other data from other periods. This especially holds for the wet-dry classification process.

L241-244: Since WD_p2 is now by far the most important parameter, this should be explained in the text. I would also like to understand why WD_p2 suddenly is so much more important than before.

Reply:

Well observed. Actually, it surprises us a bit that WD_p2 is now the most important parameter in the sensitivity analysis. We find it difficult to explain this. We wrote:

L244-247: "...The highest importance reached by the WD_{p2} parameter highlights the rain-induced attenuation temporal correlation. Since, this parameter represents the number of previous hours over which the maximum value of the minimum received power (P_{min}) is computed, it governs the wet-dry classification process by influencing on the attenuation (median(ΔP)) and specific attenuation (median(ΔP_L)) computation..."

L249: Remove "the" before "all solutions"

Reply:

We wrote:

L253: "The distributions are obtained for all solutions..."

Fig 2.: It is hard to see relations between the different WD parameters. I suggest to rearrange the individual plots to a scatter plot matrix, e.g. using https://ggobi.github.io/ggally/reference/ggpairs.html, because this way all relations and potential correlations would be visible. The distributions that are now shown on the bottom, would also fit on the diagonal in the scatter plot matrix.

Reply:

Excellent point. We rearranged this Figure by using the ggpairs function.

L271: "Due to the similar value of WDp1..." I do not understand this sentence. Why are data excluded due to similarity of WDp1 values?

Reply:

We wrote:

L269-273: "...This parameter has a direct relation with data availability, since it determines the minimum number of hours needed to compute max(P_{min}). Note that max(Pmin) is only computed if at least a minimum number of hours of data are available; otherwise it is not computed and no rainfall intensities will be retrieved (Overeem et al., 2016b)."

Table 4: Similar to table 3, what is the reason that the order of relative importance changed and that there is clear leader, RR_p5 here?

Reply:

We wrote:

Well observed, we identify that all ranks changed in our analyses. We presume that the change of the cost function (MCC now) is the reason for that. Although MCC does not increase the calibration performance, this metric seems to be a better choice, because of the imbalanced feature observed in this dataset. A possible explanation is that the overall effect of the attenuation correction, 1.74 dB, is rather small compared to the attenuation due to rain. Hence, the actual derivation of mean rainfall intensity from minimum and maximum rainfall intensity may be dominant.

L309: I do not understand what "bears to similarity" means.

Reply:

We wrote:

L306-307: "...the optimum values, 1.7 and 0.23 are almost identical with the median values for the "behavioral" solutions, 1.74 and 0.24..."

L320: Here you probably mean RR_p4 and not RR_p5.

Reply:

Yes, you are right and we modified this.

L330-336: It could be noted here that the k-R relation is not very sensitive to DSD variations for frequencies in the range of approx. 20-35 GHz. Compared to errors from wet antenna or wrong wet-dry classification, the DSD dependence of the k-R relation in this frequency range can be considered to be small.

Reply:

We wrote:

L329-335: "...A physically-based approach which derives these coefficients from drop size distribution observations and scattering computations is preferred compared to optimizing these coefficients in a statistical manner. Especially, for frequencies higher than 35 GHz. The drop size distribution dependence of the k-R relation in the fre-

quency range of approximately 20-35 GHz is considered small compared to errors from wet antenna attenuation or erroneous wet-dry classification. Although a physically-based approach is considered better, a calibration of power-law coefficients may be a way forward for regions which lack disdrometer data (Ostrometzky and Messer, 2020)…"

L346: The MCC of 0.4 for the validation is significantly smaller than the minimum MCC of 0.53 of all "behavioral" solutions from which the mean parameters where taken. What is an explanation for this strong decrease in performance?

Reply:

We wrote:

L347-353: "…We find a MCC value of 0.4 for the validation dataset, being smaller than the MCC threshold for "behavioral" solutions, i.e., 0.53. This occurred because the calibration did not generalize at all the wet-dry classification process. It was focused on the calibration dataset, capturing many details and noise, and subsequently failed to capture a different trend from another dataset, i.e., became an overfitted model. Thus, the performance for the validation dataset was worse, because the calibration dataset will not be entirely representative for other periods. A solution could be to increase the size of the calibration dataset, encompassing more characteristics and trends about the phenomenon…"

L353: Maybe write "wet-dry observations of the reference" to make it clear that these labels are derived from the reference.

Reply:

We wrote:

L361: "…According to the wet-dry observations of the reference during the validation period…"

L353-355: While 97% (number of dry data points in validation period) and 93% (number of dry periods in calibration period) are numbers which seem close to each other, I want to point out that the relative number of wet periods is more than twice as high in the calibration period (7%) compared to the validation period (%3). I am, however, not sure about the exact impact on the results, e.g. the significantly decreased MCC in the validation periods. You might want to think about this issue and add a comment to the text.

Reply:

We wrote:

L361-366: "...According to the wet-dry observations of the reference during the validation period, we observed that 97% of the data points represent non-rainy intervals. Being just four percentage points higher than the calibration period (93%), the fraction of dry periods can be considered comparable to each other. However, the fraction of rainy periods for the calibration period (7%) is more than twice as high as for the validation period (3%). This implies that the calibration dataset is at least different with respect to the validation dataset concerning the percentage of rainy periods, which may have resulted in a lower MCC value for validation..."

Section 3.2.1: In my opinion it should be pointed out here that the absolute number of false-positives is higher than the number of true positives. This is important for the interpretation of CML rainfall estimates because it means that more than 50% of the data points where CMLs estimate rainfall can be considered artifacts. As can be see in Polz et al. 2020 (https://doi.org/10.5194/amt-13-3835-2020) in Fig 9 this is not uncommon. The impact of the false-positives on the resulting rainfall amount is, however, smaller than the count of the false-positives suggest, as can be seen in in Polz et al. 2020 Fig 9d and 9f. In your case the impact of the false-positives on the rainfall amount might be different, though. Given the impact of false-positives on PBIAS in your analysis, the false-positive rain rates might play a larger role here. This should be discussed in more detail, maybe also in the conclusion section because the frequent occurrence and impact of false-positives seems to be a peculiar characteristic of CML rainfall estimates that all potential users or producers of CML QPE should be aware of.

Reply:

We wrote:

L357-360: "...Approximately 50% of the rainy events are classified as dry, both for the calibrated and default parameter sets. Similar results were reached by Polz et al., (2020), however, the impact of false rain detection on the resulting rainfall amounts was found to be smaller than the relatively poor wet period classification suggested...".

L374-377: "...Due to overestimation observed by the PBIAS values, we can conclude that the significant number of false-positives (i.e., erroneous rainfall detection), plays an important role here. Polz et al. (2020) observed a different behavior, in the sense that even having a large number of false-positives was not translated into such an overestimation of the rainfall amounts...".

Moreover, Table 5 gives total interpretation of the false-positives impact on our analyses.

Fig 5: Why did the results for the default parameters change compared to the same plot, Fig 4, in the initial submission? E.g. KGE is now 0.37 for the default parameters. It was 0.45 in the initial submission for the default parameters.

Reply:

We redid the evaluation by using data.table package (https://github.com/Rdatatable/data.table) syntax, instead of pure R. Perhaps, this was the reason of difference in KGE values. The time aggregation performed by data.table works with a syntax near to SQL and proper for “big data” databases.