

Atmos. Meas. Tech. Discuss., author comment AC2
<https://doi.org/10.5194/amt-2021-34-AC2>, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Reply on RC2

Wagner Wolff et al.

Author comment on "Rainfall retrieval algorithm for commercial microwave links:
stochastic calibration" by Wagner Wolff et al., Atmos. Meas. Tech. Discuss.,
<https://doi.org/10.5194/amt-2021-34-AC2>, 2021

Author's reply to the referees comments to manuscript AMT-2021-34

Anonymous referee 2

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

Review summary for "Rainfall retrieval algorithm for commercial microwave links: stochastic calibration"

Summary:

This manuscript presents a new stochastic calibration of the most important parameters of the well established RAINLINK method, which is used for processing CML attenuation data to derive rainfall estimates. Since the RAINLINK method is applied by an increasing number of researchers, a detailed sensitivity analysis and an improved calibration would be an important contribution that could provide guidance for choosing the RAINLINK parameters in future analyses. The manuscript is well written and well structure and would be of interest for readers of AMT. I found several major issues with the analysis, though. Solving these issues will require to redo most parts of the analysis. Hence, I recommend a major revision. I do, however, not see the need to add a comparison of RAINLINK with other methods to this manuscript. The focus on calibration and sensitive analysis of parameters is a reasonable scope for one manuscript.

We gratefully thank the Referee for the constructive comments and recommendations.

General comments and recommendations:

1. Short calibration period with potentially biased fraction of wet and dry periods:

The calibration period is fairly short, only 12 days, and hence might not cover challenging dry periods with strong fluctuations, noise or artifacts. Since these 12 days have been selected from a longer period, I assume that these are all rainy days. If this is the case, this would shift the false-positive and false-negative rates in the validation period compared to the calibration period. As a result the optimal wet-dry parameters from the calibration period might not be optimal for the validation period (see my comment on L104). This can lead to unexpectedly high numbers of false classifications. Based on the result in table 5, I conclude that this is the case here. According to my interpretation, a large number of false-positives contributes to the overall CML rainfall sum, see my comment on L311 for more details. I strongly recommend to, either chose calibration and validation data so that the wet-dry ratios is similar, or to use a performance metric that is more robust to changes in this ratio.

Reply:

We appreciate this comment that would make the paper more complete when implemented. In short: we accepted this suggestion and will employ a more robust performance metric.

Note that this dataset was also used in Overeem et al. (2016) for sensitivity analyses. The reviewer is right that this dataset is likely more rainy than an arbitrary period of 12 days, but this does not imply that it rains all the time. We expect that this 12-day period contains more dry than rainy time intervals. The exact distribution of both dry and rainy intervals for the calibration and validation datasets will be derived from the path-averaged radar data and will be mentioned in our revised manuscript.

Moreover, we will investigate time series of dry intervals from the calibration period to find out how often periods with strong fluctuations, noise or artefacts are present. Percentages of discarded links according to the wet-dry classification from the calibration and validation dataset will be compared.

2. Usage of questionable classification metric, Simple Matching (SM):

The Simple Matching (SM) is chosen as performance metric for the binary classification into wet and dry periods. SM, which is the same as Accuracy (a more common term for this metric for binary classification performance) is very sensitive to the balance of positive and negative samples, see my comment on L187 for an explanation. In general Accuracy is thus not a recommended, but

still widespread, metric. More info can be found e.g. here <https://dx.doi.org/10.1186%2Fs12864-019-6413-7>. This article recommends to use the Matthews correlation coefficient (MCC), which I would also recommend. Other options would be to study the ROC curve, or to be more careful with balancing wet and dry samples in the calibration and validation period. I strongly recommend to redo the optimisation of the wet-dry parameters, taking all this into account.

Reply:

We admit that this issue, which is one of the critical issues that should be discussed, is currently lacking. We will redo the optimization by employing MCC as performance metric and if results improve we will use the corresponding optimized parameter values to rerun our analyses.

3. Unclear method for determining optimal wet-dry parameters:

There is another problem with the optimisation of the wet-dry parameters. The optimal parameters are not those that clearly provided the highest values of SM, see my comment on L225 and on Fig 2a. It is not 100% clear to me how the optimal parameters are derived. If they are derived from the "behavioral" solutions I find this problematic, because these distributions are somewhat arbitrarily selected, see my comment on L225 for a more detailed explanation. I might, however, not have fully understood how the optimal parameters are found. In this case, please explain the method better and also, in particular, explain why not the parameters at the best SM values are chosen. Of course, as stated above, SM is not a good metric for judging wet-dry performance. Hence, in case a different metric is used, things will look differently here anyway.

Reply:

Thanks for the comment and recommendation. We will investigate the usefulness of a different metric (see comment #2). We hope that redoing the wet-dry calibration by using the recommended performance metric (MCC) will lead to satisfactory results, leading to SM being removed from our revised manuscript.

4. Missing validation of wet-dry classification:

The validation section is completely missing a validation of the wet-dry classification. Given the issues with identifying the parameters of the wet-dry classification and its potential impact on rain rate estimation (see my comment on L311), it is crucial to add it here, also including an analysis of its impact on rainfall sums.

Reply:

We agree with this observation, a validation of the wet-dry classification during the validation period would complete the manuscript. Hence, we will add such a validation with metrics such as probability of detection and critical success index in our revised manuscript.

5. Unclear motivation of the proposed calibration:

It should be made clearer why the calibrations that have been done in other RAINLINK publications are not sufficient. Furthermore it should be made clearer why LH-OAT and SPSO have been selected, highlighting and explaining their advantage compared to past calibration efforts. (See my specific comment on L79)

Reply:

We appreciate this observation and we will better explain the added value of our approach compared to those in other RAINLINK publications as well as why LH-OAT and SPSO have been selected. The main idea is to highlight how a stochastic and pinpointed calibration approach can be more parsimonious, reducing computational demand and driving the algorithms to a better performance. An advantage is that all RAINLINK parameters are initially taken into account, whereas previous studies focus on a limited set of parameters. Moreover, the optimization of the wet-dry classification is separated from the rainfall retrieval, i.e., first the wet-dry classification is optimized, next the rainfall retrieval.

Additional note:

In the light of the (according to my interpretation of the presented results) large impact of false-positives on PBIAS, one could (or maybe should), consequently calibrate the rainfall estimation part of the algorithm with taking the wet-dry classification from the reference to avoid an overestimation of wet antenna attenuation that has to compensate the long-term rainfall overestimation from false-positives. This is just an idea that, assuming that large parts of the analysis have to be redone for a revision, could be explored.

Reply:

Good point, we will consider this possibility.

Specific comments:

L22: One has to be careful with the interpretation of the number of stations available in GPCC. Large delays in data delivery and data processing lead to a delayed peak of available stations. From how I interpret the GPCC documentation, this might explain most of the "decline" since the 1980. The GPCC authors write "The decrease of the number of stations from more than 45,000 in 1961-2000 down to 10,000 stations after 2019 is caused by the delay of the data delivery to and by post-processing at GPCC" (Source: https://openda.ta.dwd.de/climate_environment/GPCC/PDF/GPCC_intro_products_v2020.pdf, end of page 9). Hence, this sentence should be reformulated accordingly.

Reply:

Thanks for the observation, we will correct this.

L48: Providing the information about the study area for Chwala et al. (2012) is a bit misleading here, because they did not study spatial rainfall information. Hence, the very low CML density in this study that is listed here, was not a relevant factor.

Reply:

Thanks for your suggestion, which we will implement this.

L60: Since pycomlink contains different algorithms, of which Graf et al (2020) only used a selection, I would write "...rainfall retrieval packages" here.

Reply:

We will modify this accordingly.

L79-L81: I do not understand the argumentation here. If one can get the "most precise path-averaged rainfall intensity estimates" using the optimised parameters from the empirical calibration, why is a new calibration needed. Aren't the old RAINLINK calibration enough? Maybe this should be improved together with the parts around L87. It is not clear what the drawbacks of the "deterministic" calibration of RAINLINK are. Since this is the core motivation of this work, I recommend to make this clearer here.

Reply:

What we would like to highlight is that we have many similar “best” solutions for the optimized parameters (sometimes referred to as equifinality). Employing a stochastic approach allows us to access the uncertainties associated with the best set of parameters. See also our reply to comment #5.

L104: How have the 12-days been selected in this period from June till September 2011? In case you only select rainy days, you skew the average distribution of wet and dry data points. This shifts your false-positive and false-negative rates in the validation period compared to the calibration period. Hence, the optimal wet-dry parameters from the calibration period might not be optimal for the validation period.

Reply:

We selected summer rainy days. We will use a different performance metric for reducing the distribution mismatch problem. See our reply to comment #1.

L155: It would be nice to learn a bit about the computational demand of the sensitivity analysis.

Reply:

We thank you for your interest and will add more information about the computational advantage related to the LH-OAT method.

L165: How is this relative importance related to the parameter range that was selected. Without understanding the details of the LH-OAT method, I can imagine that the parameter range influences the step size and hence the relative impact of each step. Please comment (or just correct my wrong assumptions on how LH-OAT works...).

Reply:

The step size is selected as a fraction of the parameter range. In this manuscript the fraction was 0.1, which is the default of the R package hydroPSO. We will add this information in the text.

L169: Why was SPSO selected? What are the advantages, also compared to other optimisation methods? What are potential disadvantages?

Reply:

The Standard Particle Swarm 2011 (SPSO2011) used in the manuscript is a recent member of the calibration/optimization family in water resources, which is more efficient than DREAM, SCE-UA and other well-known algorithms, as observed in Zambrano-Bigiarini et al. (2013).

SPSO-2011 is a major improvement over previous PSO versions, with an adaptive random topology and rotational invariance constituting the main advancements.

L178: Why was simple matching chosen as metric for the binary classification? It seems to be sensitive to unbalanced distributions of true and false values. E.g., if, in the case of wet-dry classification, the number of dry data points is by far larger than the number of wet data points, very high values of SM can just be reached by setting everything to "dry".

Reply:

We thank you for your constructive view and we will redo the analyses by considering a proper performance metric (MCC). See our reply to comment #1.

L178: Here it sounds as if the modified KGE is used as metric for the wet-dry classification. This should be rephrased.

Reply:

We split the algorithm processing into wet-dry classification and rainfall retrieval itself. Thus, we employed the KGE metric only for the rainfall retrieval process. We will make this clear in our revised manuscript.

L180: I guess the gauge-adjusted radar product comes with 0.01 mm resolution or similar. The path-averaging along the CLM paths results in even smaller values. Wouldn't it makes sense to define a threshold slightly above zero to divide between wet and dry periods because something below 0.1 mm in 15-minutes can hardly be considered rain?

Reply:

Yes, we used a threshold of 1 mm to classify intervals as rainy when above this value.

L182: "...where d is the number of links classified correctly as dry...". I expected that this is done for all data points and not for each link. If this is done for each link, that would mean SM is calculated for each time step. But in the context of this work, it seems to be calculated for all samples for the whole calibration period, correct? Please clarify.

Reply:

Thanks for giving attention to this aspect. Yes, this was calculated for all data points. We will rephrase this part.

L203: It is not clear here if the "main rainfall over the Netherlands" is based on interpolated rainfall maps, or the average of the rainfall values for each CML.

Reply:

Here we mean the average of the rainfall values for each CML, and we will make this clear in our revised manuscript.

L220: What is "behavioral" supposed to mean here?

Reply:

"Behavioral" is the set of solutions which outperforms the solution obtained by using the

default parameters.

L225: I do not understand how the optimal values have been identified. The only metric that is used here is SM. Hence, I expected to find the optimum where the cyan coloured dots (highest SM) in Fig 2a are. The parameters reported in the text are, however, more in the centre of the parameter range, while the highest SM values are at the smallest WD_p4 and highest WD_p1 values. Maybe this has to do with the "Wilcoxon signed rank test" that is mentioned in the sentence before. I could imagine that the derivation of the optimum is somehow based on the distribution of "behavioral" solutions. But, since the distribution of "behavioral" solution heavily depends on the arbitrarily chosen threshold of SM, this is not a reliable procedure. If the SM threshold would be set to e.g. 0.95, the distribution would look very different and for WD_p1 show a clear tendency towards very high values. In conclusion, I find the results very counterintuitive. Please either provide a good explanation for the chosen method or correct your procedure of determining the optimum. Please note that using SM is not a good choice anyway, see my comment on L178. Hence, potentially redoing this step of the calibration should then be done with a different performance metric.

Reply:

Thanks for pointing this out. We considered the optimum value equal to the median obtained by the non-parametric Wilcoxon signed rank test applied to the "behavioral" solutions. Considering that all "behavioral" solutions are similar, we chose the one that represented the median. We will rephrase this part to give attention to the right terminology. Anyway, we will redo the analyses by using another performance metric as recommended (see our reply to comment #1).

L229: What is the point of the 95% confidence interval of the "behavioral" solutions? Or maybe more general, what is the point of the "behavioral" solutions, which have been obtained by arbitrarily selecting solutions with SM larger than 0.90? Why not use $SM > 0.95$ as threshold?

Reply:

We appreciate your observation. The point of the 95% confidence interval is arbitrary, purely a statistical threshold commonly used in the literature. As for the "behavioral solutions", we considered them to correspond to the threshold $SM > 0.90$, because this was the result obtained by using the default parameters.

L232: If I understand the analysis correctly, a SM of 0.9 for the whole calibration does not mean that "90% of the microwave links provide a correct wet-dry

classification considering the entire period of 12 days". I would rather say that 90% of the data points are classified correctly. It is not clear how these correct classifications are distributed between the individual CMLs. Maybe I do not understand how SM is calculated here for the calibration periods (see also my comment on L182). Please clarify.

Reply:

Perfect point. You are right this is related to the data points and not to the CMLs themselves. We will clarify this in the manuscript.

Fig 2a: I find it strange that very high SM values are more or less equally distributed over the full range of WD_p5, but WD_p5 is considered the parameter with the highest relative importance according to Table 3. How can that be explained?

Reply:

Well observed. Likely, the greater number of dry than wet periods leads to high SM values in any case. As observed previously, this sounds related to the performance metric characteristic. We will redo the analyses by using the MCC metric and check if the WD_p5 behavior is still the same.

L260: If the optimisation is done only with rainfall data at the CMLs and not on CML-derived rainfall maps, I do not see how an optimisation of the outlier filter can be done. Assuming that there are a few outstandingly good performing CMLs, all others would be removed in the process, because this would result in the highest average KGE. Please make this clearer in the text.

Reply:

Thanks for giving attention to this aspect. We will make this clearer and add a brief discussion. A sensitivity analysis for the outlier filter can be carried out (Overeem et al, 2016), but it is indeed difficult to take the related threshold parameter into account in the optimization. One way forward could be to include the number of available links in the optimization. The suggested optimization based on rainfall maps, which can be influenced by the underlying CML network density, would also be interesting to add to our discussion.

L269: "...which is in line with what can be seen in Fig. 3.". I find it interesting that this is the case here but not for Fig. 2a. Please explain (which is maybe already done in response to my comment on L225).

Reply:

Exactly, the results in Fig. 2a. are counter-intuitive, we will redo the analyses.

L270 and following: I find it most striking that there seems to be a clear correlation between RR_p4 (wet antenna attenuation) and RR_p5 (alpha). The explanation probably is that a higher alpha leads to higher rain rates, because the weight of the maximum attenuation increases, which has to be compensated by a higher value of wet antenna attenuation correction, decreasing the rain rate estimates. Hence, this two parameters clearly influence each other. This should be mentioned in this section.

Reply:

We appreciate your recommendation and will mention this likely relation between the parameters.

L285: The validation of the wet-dry classification seems to be missing completely here. I strongly suggest to include it, in particular because I expect the results to be very different from the calibration period because of the different ratio between wet and dry data points in the two periods and because SM is not robust to changes in this ratio.

Reply:

Yes, we will add here the wet-dry classification for the validation period. See our reply to comment #4.

L303: It would be nice to see a figure similar to Fig. 4 also for the 15-minute data. I am aware that similar plots have been shown in several RAINLINK publications, but, it would be interesting to see the differences between default and calibrated processing not only for the daily data.

Reply:

We appreciate your suggestion. We already present the metrics for the 15-minute data in Table 5, but we are willing to add the associated scatter density plots for path averages to our revised manuscript, also because most RAINLINK publications compare rainfall maps rather than path averages.

L304: "For a complete evaluation we use different rainfall thresholds." It took me some time to understand this sentence. If the reader does not already know the details of Table 5, it is not clear what the "complete evaluation" is and what the "different rainfall thresholds" are used for.

Reply:

Well observed, we will rephrase this part.

L311: My explanation for the strong influence of the threshold "Reference > 0" on PBIAS is the following. There is most likely a large number of false-positives. These false-positives contribute significantly to the overall CML-rainfall estimates and result in a positive PBIAS. This impact of false-positives on the CML rainfall estimation is nicely shown in Fig 9. in Polz et al. (2020, <https://doi.org/10.5194/amt-13-3835-2020>). If the false-positives are removed, which is what the threshold "Reference > 0" does, the resulting CML-rainfall estimates are missing this large amount of "false-positive" rainfall. As a consequences, PBIAS shows a strong underestimation of CML rainfall estimates. This effect also explains the other observations, made in the sentences before. The fact that PIBAS is is "better" for the calibrated parameters turns into a disadvantage when applying "Reference > 0", because the shift of PBIAS towards underestimation seems to be similar for calibrated and default parameters (explaining the observations in L307). The reason why the effect on PBIAS does not appear when applying a threshold like "Reference OR RAINLINK > 0" is that this threshold does not remove the false-positives, because if RAINLINK > 0 and Reference = 0, the data point is kept in the dataset. Your sentence in L309 "This underestimation is not observed if both RAINLINK and the reference are above the threshold" is not correct, because you apply an OR not an AND for these threshold. As stated above, I strongly recommend to include an analysis of the wet-dry classification for the validation data. Furthermore, as stated in my comments on the calibration of the wet-dry classification, the choice of parameters might not be optimal for the calibration period. Hence, there might also be less impact of false-positives, if another "optimal" parameter set is found.

Reply:

Thanks for the constructive observation. It seems that this issue is related to the wet-dry classification methodology, which we need to redo with the MCC metric. In addition, we will also provide a validation of the wet-dry classification. After that, we hope to have gained more insights about the obtained results and a conclusive answer. We will modify L309 accordingly.

L317: I guess you are referring to Table A1 in de Vos et al. (2019). There seems to be a typo, either in this table or in the sentence here, because in the Table A1 the Pearson correlation for the revaluation is 0.27 and not 0.52 as written here.

Reply:

Well observed, it is a typo, which we will correct.

L320: Since the reevaluation covers winter months and since this is know to introduce overestimation of CML rainfall estimates, I would have guessed that de Vos et al (2019) have a high bias in their analysis, which apparently is not the case. Please explain a bit more detailed where this difference in PBIAS could stem from, because I do not understand how "different periods, with different durations" lead to the high PBIAS in this study compared to de Vos et al (2019).

Reply:

Excellent observation. It seems again that the wet-dry classification methodology can be the reason for this difference. Even with false-positive observations the SM metric yields good results, although it leads to poor PBIAS in the rainfall retrieval. We will investigate if this difference still exists after rerunning the wet-dry optimization.

L327: As explained in my comment on L311, I assume that false-positives play an important role for the overestimation of CML rainfall.

Reply:

We agree. Using a different performance metric is the solution. We should redo the analyses.

L334: Why is this not done with rainfall maps, which are also easily produced with RAINLINK? That would be a more relevant basis for doing an analysis "over the Netherlands".

Reply:

Thanks for giving attention to this aspect. Actually, we intend to evaluate the path-averaged rainfall when distributed over an area and the associated error behavior. Evaluating a spatially interpolated map, the error of interpolation process would be added in the analyses, which is beyond the scope of this work.

L337: I do not understand how the area plays a role here. You average the data from the individual CMLs, not taking into account how they are distributed over this area. The effect on PBIAS and beta has nothing to do with the fact that the CMLs are within a certain area.

Reply:

Yes, we will remove or rephrase this phrase because this is indeed not related to the area.

L339: I would not call this an "areal time series". One could maybe argue that an sensor-average from a fairly homogeneously distributed rain gauges network is representative of certain area, but not an average of a very heterogeneous sensor network like the one of the CMLs here.

Reply:

Interesting observation. The CML coverage over the Netherlands is not homogeneous indeed and those urban areas with high network density will have a larger weight in the computation of the areal rainfall. On the other hand, the number of CMLs is much higher than those from official rain gauge networks, i.e. they could provide a better spatial average than gauge data. Hence, we think that calling this "areal time series" is justified, but we will add the above mentioned limitation to our revised manuscript. Studying the calibration for different classes of CML features would be an excellent topic for a future researcher.

L347: Shouldn't one reason for the differences between calibrated and default

parameters be that the calibration here is done with a more sophisticated, presumably better, method?

Reply:

Yes, this is likely one of the reasons and we will add this to our revised manuscript.

L349: I would add WD_p1 and WD_p4 here, because Fig 2a shows that the highest values for SM are reached at the end of their parameter range. Hence, it can be expected that SM could further increase beyond the current parameter range if it would be extended. So the questions is, why was this not done.

Reply:

We need to redo the wet-dry classification anyway by using a different performance metric and will likely also extend the parameter range for WD_p1 and WD_p4.

L372: I can not follow this argumentation. While I agree that "hydrological and meteorological scales of application are defined over areas", I would say that these scales, in particular in hydrology, are much smaller than the Netherlands for which the positive effect of aggregation over an area is found in this manuscript.

Reply:

The reviewer is right that hydrological scales in the Netherlands are much smaller than the country, so we will remove this argumentation.

L389: Just a comment. Yes, comparing to gauges avoids the impact of radar errors, but the path-averaged nature of CMLs has to be considered when comparing to rainfall data from point observations. Furthermore, since the gauges would have to be fairly close (maybe less than 2km) to be able to assure comparability with CMLs on 15-minute or 1h time scales, this would limit the number of CMLs that can be analysed.

Reply:

Yes, excellent point, a perfect reference to evaluate CML rainfall estimates is a challenge yet. Also note that representativeness errors in radar data are a limitation when providing path-average reference data. We will add your point of view to our revised manuscript.

Editorial comments:

L131: Maybe write "summarises" instead of "highlights" here.

Reply: We will replace "highlights" by "summarises"

Final remarks:

We thank the referee for the kind comments about the contribution to the field and the paper being well organized and interesting to read. We accepted all the referee's contributions and hope to reformulate the manuscript following the same quality, care and effort employed by the anonymous referee.

References

Overeem, A., Leijnse, H., and Uijlenhoet, R.: Retrieval algorithm for rainfall mapping from microwave links in a cellular communication network, *Atmos. Meas. Tech.*, 9, 2425–2444, <https://doi.org/10.5194/amt-9-2425-2016>, 2016.

Zambrano-Bigiarini, Mauricio, and Rodrigo Rojas. "A model-independent Particle Swarm Optimisation software for model calibration." *Environmental Modelling & Software* 43 (2013): 5-25, <https://doi.org/10.1016/j.envsoft.2013.01.004>.