We would like to thank the reviewer for the second revision of our manuscript and the helpful comments that further improved the manuscript. We agree with most of the raised points and questions and clarified or rephrased the manuscript accordingly. We additionally changed the color scheme of Fig. 6 and 8 to be more comprehensible to people with color vision deficiencies.
In the following the reviewers' comments are in black, our answers in blue, and quotation marks additionally mark the text changes in the manuscript.

REVIEW OF THE PAPER “Improved rain event detection in Commercial Microwave Link timeseries via combination with MSG SEVIRI data”, AMT 2023-175 (round II)

General comment

The authors did a huge effort in editing the manuscript according to reviewers comments. I think that this new version has improved a lot and it is much clearer than the original one. I also would like to thank the authors for their detailed replies to my questions. There are still a few points, which, in my opinion, are due a minor revision.

- Sampling and resampling of SEVIRI and CML data: for wet /dry classification (ADB methods), SEVIRI has been resampled to 1-min, i.e. the same as CML data. However, comparison against radar data is carried out resorting CML to 15-min sampling. Why you did not use 15-min all the way, just resampling CML data?

There are good reasons why CML data is processed at a 1-minute resolution and evaluated at a 15-minute resolution:
The processing at a 1-minute resolution is due to compatibility with established processing methods. The reference methods (CNN, q80), but also the wet antenna attenuation and baseline estimation were either developed, or calibrated, and tested for one-minute instantaneous CML data (Graf et al. 2020, Polz et al. 2020). They would have to be adapted to a lower resolution, which is not our aim here. We aimed to see if SEVIRI data, with its 15-minute resolution, can be used for rain event detection with the CML data we have available. It was straightforward to downsample the SEVIRI wet-dry indicator to a 1-minute resolution leaving the rest of the processing unaltered and we could show that this approach leads to quite satisfactory results.
To evaluate the binary wet-dry information from SEVIRI (with its 15-minute resolution) we relied on the radar reference which already has a lower temporal resolution (5-minute) than the CML data and, therefore, we decided to further resample to a 15-minute resolution. To improve the comprehensibility of the study we compared all results at the same resolution.
To improve the clarity on this we added an explanation to the last section of 3.1.1 Individual methods for rain event detection:
“We forward-filled the 15-minute classification to a 1-minute resolution in the CML processing described in Sec. 3.2. This temporal resolution is necessary for the two TSB methods and other CML processing methods such as WAA compensation and baseline estimation as they were developed and tested for this resolution (Graf et al. 2020).”

- About my comment on performance indicators in the first round of review. Specifically about MCC. It is not a matter of being familiar or not with it. I think that several readers would not be able to rate the statement in the abstract “Compared to basic and advanced TSB methods, these combinations improved the Matthews Correlation Coefficient of the rain event detection from 0.49 (0.51 resp.) to 0.59 during the day and from 0.41 (0.50 resp.) to 0.55 during the night". Is it a significant/huge improvement or not? Are 0.49 or 0.41 acceptable values for the MCC? To help understanding MCC, if we compare (6) and (1), we get that MCC is basically PCC for binary data. That would be a synthetic and easy explanation of MCC. When I see it from the reader's perspective, it would be more effective to summarize in the abstract the improvement brought by combinations using TPR and FPR, or just writing a simple statement as the one on p. 19 lines 420-21.

On the one hand, the actual value of MCC (as PCC or many other metrics) cannot easily be put into categories of good or bad because they depend on the use case. In some cases, a MCC of 0.8 might be really bad while for other use cases, a MCC of 0.2 might already show a success. Some rain event detection method studies already used the MCC, allowing for a comparison of the values. On the other hand, we agree that for a reader it is easier to understand a qualitative statement. Therefore, we added one sentence to the abstract with a qualitative explanation of the results of the manuscript:

"Additionally, these combinations increased the number of true positive classifications, especially for light rainfall compared to the TSB methods, and reduced the number of false negatives while only leading to a slight increase in false positive classifications."

- About my general comment on wet/dry rationales for radar data: the authors explain how radar pixels are combined (p. 5, line 125) to overlap a CML as the comparison with SEVIRI is done over CML paths. The radar-based precipitation value is derived by a weighted average in space (according to the fraction of the path overlapping the radar pixel) and an arithmetic average in time (from 1 to 15-min). However, it is not well explained how radar time series (i.e. rainfall rate estimates) were reduced to wet/dry time series for the validation of SEVIRI products in Fig. 4. It is stated that wet/dry threshold on radar data is 0.1 mm/h (p. 4 line 130). Hence, I guess the authors first calculated the radar-based rainfall estimate over the CML path and then they thresholded it at 0.1 mm/h. I think that this procedure should be explained in the text for instance on p.5 after line 125.

We agree with the reviewer and added a sentence explaining the application of the threshold at the suggested location:

"For the usage as a binary wet-dry reference, we used a rainfall intensity threshold of 0.1 mm/h at the 15-minute resolution. All values below 0.1 mm/h were considered dry."

- Threshold on SEVIRI probability of rain. As the authors said in Sec. 5.1, it is surprising that PC and PC-Ph work at their best with such a low probability of rain. This looks even more surprising considering that it is calculated over a rather large pixel (table 1). Can the author provide any information about how these precipitation products were extracted from SEVIRI measurements to justify these outcomes?

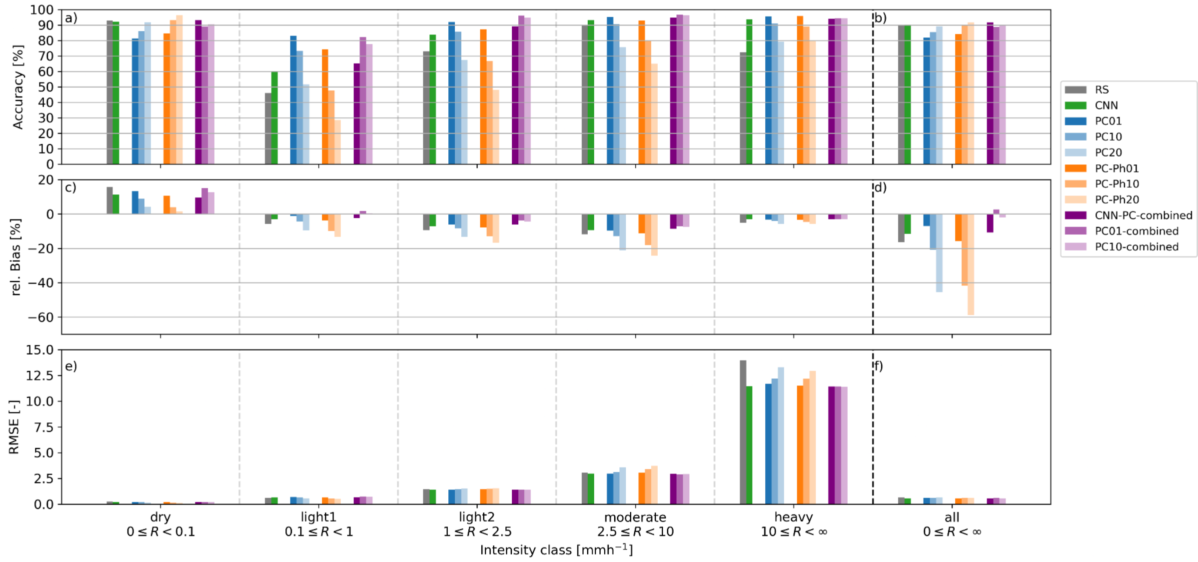
As the SEVIRI products are indirect rainfall indicators based on optical and infrared measurements an estimate is based on observed cloud mask, cloud optical thickness, cloud water path, cloud top temperature, etc. As a result, they have large uncertainties in detecting light rainfall and the exact outline of precipitation fields. However, for Europe the cloud mask product has a very high rate of detection of dry pixels, rejecting most dry areas for rainfall

anyway (leading to a small number of FP in the final rain event detection). For lower thresholds of PC and PC-Ph, the increase in TPR mostly equals the increase in FPR as shown in Fig. 4 with P10 as the best product.
"For the SEVIRI products, this does not seem to be the case. One reason might be that the cloud mask, which both products use, has good performance over Europe and therefore has a high probability of true dry detection, resulting in a low number of FPs in general and thus also for low thresholds."

- Figure 7: I cannot get why the relative bias of each class is normalized to the average rainfall intensity over all classes (denominator of Eqn. 2) if I got it, which I am not sure. If we assume an order of magnitude 1000 mm of rainfall per year, the average rainfall intensity would be around 0.12 mm/h. Is it correct? In my view, for class, say, light 2, the denominator should be the average of the occurrences of radar-based rainfall rates between 1 and 2.5 mm/h. The important information I retain from Fig. 3c is the sign of the relative bias. The height of bars is about the balance between negative and positive errors. In the evaluation of CML rain estimates time vs radar, using different wet/dry classifiers, I think a useful indicator is missing, that is the RMSE, which would help in assessing also the performance over individual classes.

About the relative bias and the normalization:
Eqn. 2 was slightly wrong because it should be total error divided by total precipitation (we corrected this now). We hope this resolves the understandable confusion.
The main purpose of Fig. 7 is to analyze the "Performance of rain event detection methods for different rain intensity classes". The relative bias shown in this figure is the mean error with respect to the radar reference in that class divided by the total radar rainfall intensity of all classes. This way, we can see how this overall bias for each method (Fig 7d) is a sum of the bias each method has in the individual classes. Therefore, we can also see where improvements are more important. If we normalized each class by the "average of the occurrences of radar-based rainfall rates between 1 and 2.5 mm/h" we would not gain additional knowledge, but we would lose an overview of where the overall bias comes from.
The updated Figure caption is:
"The relative bias in each class is the sum of all errors in one class as a percentage of the total rainfall of the full reference data. This way, the five leftmost classes add up to the 'all' class on the right."

About the use of RMSE:
The Figure below shows the draft of an updated Fig 7 that includes the RMSE.



In general, the only reason not to include this information is the added length of the manuscript and the resulting difficulty in extracting useful information from the large number of indicators presented. As you can see, an RMSE comparison between intensity classes is not possible because of the scaling of the RMSE with the rainfall intensity. Additionally, the different magnitudes of the RMSE in the different classes make a comparison of methods in one class harder. Therefore, we decided not to add the RMSE to the analysis.

- One thing I guess remains without an explanation is why all methods underestimate precipitation with respect to the radar reference whatever the rainfall intensity class, while they overestimate dry periods (relative bias in Fig.7). I think a comment in the manuscript is due. Even if this fact is explained somewhere else, please not only add only the reference, but at least one explanatory statement.

Thank you for pointing this out. There are good reasons for the observed biases in the different classes which we will explain in the following:
Let's first assume that derived rainfall rates during correctly detected rain events perfectly fit to the rainfall reference. Then the effect of a negative bias would still be visible because of false negatives since the rainfall intensity classes only include TP and FN periods. A part of this negative bias is compensated by FPs in the dry class where only a positive bias is possible due to the radar reference rainfall being 0. With this assumption laid aside, a bias for estimated rainfall rates is also present which may be due to WAA compensation. The two combined products have almost no overall bias compared to the reference. All others show a negative bias, that is true. This fact can potentially be attributed to WAA compensation. We therefore added a statement to Section 4.3:
"The contribution of the different intensity classes to the overall relative bias in Fig. 7 shows that the overestimation due to FPs in the dry class is smaller than the underestimation in the positive intensity classes where FNs are one major factor for missed rainfall. This way, all methods except PC01-combined show an overall negative bias. One additional explanation for an overall negative bias could be a too-large compensation for WAA. Tiede et al. (2023) described strong WAA fluctuations during long-lasting rain events, a fact that is not considered in the WAA compensation method we chose (or any other available method). Therefore, there is some uncertainty in WAA compensation which influences the bias."

Specific comments

- P.4, line 144: “it is calculated by a regression of IR and Water Vapour channels (WV).” What you mean by WV channel? Channels are identified by a frequency.

While this is technically true, the two IR channels at 6.2 and 7.3 µm from MSG SEVIRI are commonly called water vapor channels see e.g. https://www.nwcsaf.org/pc-ph_description or https://www-cdn.eumetsat.int/files/2020-04/pdf_conf_p46_s3_04_georgiev_v.pdf

- P. 7, line 171: “We computed RS and CNN on a 1-minute basis”. Do you mean in previous paper?

We clarified this point by changing the sentence to:
“Identically to Polz et al. (2020), we computed RS and CNN based on 1-minute TL data.”

- P. 7, line 175 what you mean by “we forward filled?” You just classified all minutes within a 15-min SEVIRI wet slot as wet? It could be a problem when it starts/stops raining or during intermittent rain. This point has to with the first bullet in my general comments.

We think we answered this point already below the first bullet point in the general comments.

- P. 10 Eqn. (2) I think the terms on the numerator should be switched, as all methods underestimate rain intensity as the authors state several times, it means that RB<0

Indeed, we switched r_ref and r_cml in the numerator.

- P. 12 lines 299-301 and P. 19 line 406-408: the authors do not bring a physical evidence that dew formation is the responsible for such a drop of the MCC for RS from day to night. They just say that the difference between RS and CNN performance suggests this conclusion. I suggest to smooth the statement on p. 19 which sounds like an harsh statement. (also I think the word “assumption” is not correct in this context)

We have worked extensively with CML data and have often seen dew formation and the described behavior of the RS and CNN rain event detection methods. We have also investigated these events using temperature, humidity, and dew point temperature from ERA5 (Polz et al. 2023). While the RS often classifies dew events as rainy because it cannot distinguish between high-frequency (rain-induced attenuation) and low-frequency changes of the signal (caused by dew), CNN was able to learn that they are not. However, we will weaken the statement in the manuscript, as this is not a very important point to the overall conclusions of the manuscript. We rephrased the section to:
“While CNN was able to perform equally well during nighttime, the other TSB method RS showed a decreased performance. One possible explanation could be the formation of dew on the antennas during nighttime that can regularly be observed in CML time series as a slowly increasing (after sunset) and decreasing (after sunrise) attenuation. The more sophisticated pattern recognition algorithm of the CNN method seems to be able to correctly classify these periods as dry.”

- P. 13 Figure 4 caption: better to add that TPR, FPR, MCC refer to wet/dry classification performance while PCC is for rainfall intensity estimate.

Good idea, we rephrased this caption to:

“Performance metrics of the binary rain event classification (TPR, FPR, and MCC) and the rainfall rates (PCC) of the PC (blue) and PC-Ph (orange) products compared to the radar reference. The results of each score are presented for different thresholds (x-axes) and split into day (light colors) and night (dark colors).”

- P. 13 line 316-319: from Figs. 5 and 7, as far as I see it, the three combined methods shown perform the same except for the bias. Moreover, the 10.8% bias is attributed to PC1 combined, while from Fig. 7 it seems the one of CNN-combined.

For the MCC (Fig 5) there are only small differences, but the accuracy and bias for different rainfall intensities (Fig 7) show larger differences, we therefore kept the three shown combinations. Regarding the description of the rel. bias in Fig. 7 we accidently mixed the rel. bias values for PC01-combined and CNN-combined up. We now placed the correct value and product together:
“PC10-combined also showed the lowest relative bias of -2.1% compared to PC01 combined with 2.6% and CNN combined with -10.8% (see Fig. 4d).”

Technical corrections

- p. 6 line 163, “adapted”

Corrected this mistake.

- p. 7, line 185: I guess it is “Fig. 2”

Corrected this mistake.

- p. 10, line 219 “For comparison against the benchmark” sounds better

Since we have not used the word “benchmark” anywhere else in the manuscript, we do not want to use it here.

- P. 12, line 2929: “one” instead of “on”

Corrected this mistake.

- P. 14, line 326-27: I guess you are referring to Fig 7 c) and d)

Corrected this mistake.

References

Graf, M., Chwala, C., Polz, J., & Kunstmann, H. (2020). Rainfall estimation from a German-wide commercial microwave link network: Optimized processing and validation for 1 year of data. *Hydrology and Earth System Sciences*, *24*(6), 2931–2950. https://doi.org/10.5194/hess-24-2931-2020

Polz, J., Chwala, C., Graf, M., & Kunstmann, H. (2020). Rain event detection in commercial microwave link attenuation data using convolutional neural networks. *Atmospheric Measurement Techniques*, *13*(7), 3835–3853. https://doi.org/10.5194/amt-13-3835-2020

Polz, J., Glawion, L., Graf, M., Blettner, N., Lasota, E., Schmidt, L., Kunstmann, H., & Chwala, C. (2023). Expert Flagging of Commercial Microwave Link Signal Anomalies: Effect on Rainfall Estimation and Ambiguity of Flagging. *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 1–5. https://doi.org/10.1109/ICASSPW59220.2023.10193654

Tiede, J., Chwala, C., & Siart, U. (2023). New Insights Into the Dynamics of Wet Antenna Attenuation Based on In Situ Estimations Provided by the Dedicated Field Experiment

ATTRRA2. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5.
https://doi.org/10.1109/LGRS.2023.3320755