Author’s response

The authors would first of all like to thank the reviewers and the editor for their thoughtful and constructive evaluation of our manuscript. Responses to their specific comments are given further below. As the manuscript has undergone substantial revision, we provide an overview of the main changes here.

The goal of our manuscript is to evaluate different methods for merging weather radar rainfall estimates with commercial microwave link (CML) data, using two publicly available datasets (OpenMRG and OpenRainER) that represent contrasting network densities and climatic settings. The analysis is implemented in the open-source Python package mergeplg, which is released alongside the manuscript.

Our main takeaways from the two reviewers were:

- The variogram estimation step of kriging should not be skipped; variograms should be estimated from data and the sensitivity to variogram parameters and neighbourhood size should be investigated.
- The ratio field in multiplicative merging should be log-transformed to make the adjustment symmetric.
- The discussion should be expanded to cover CML rainfall uncertainties and their consequences for merging.

We agree with all three points and have addressed them in the revised manuscript.

Regarding variogram estimation: we have substantially expanded this part of the study. Variograms are now estimated from data using an event-averaging approach adapted from Haberlandt (2007), with parameters averaged over 14-day windows to obtain stable estimates. A dedicated variogram study is included (Fig. 3), showing the temporal evolution of variogram parameters for both datasets and comparing CML and rain gauge observations. A systematic comparison of variogram estimation strategies is presented (Fig. 4), and a neighbourhood size sensitivity analysis is included (Fig. B1). The neighbourhood size has been increased from 12 to 70 nearest CMLs. The discussion includes a new subsection interpreting these results. We note that the editor raised a specific concern about the rigor of the variogram response; we hope the above changes demonstrate that this has been treated as a substantive methodological investigation rather than a sensitivity analysis.

Regarding the log transform: We agree with the reviewers that a log-transformed ratio field provides a more symmetric and robust adjustment. While we had internally investigated log-transforming the ratio field during the early development of the study, it was not included in the original manuscript because those early tests produced larger errors. We have since identified the reason for this: those initial internal experiments were conducted before the range-checking step had been implemented in our workflow. Without range checks, the back-transformation amplified extreme outliers. Following the reviewers' suggestion, we re-evaluated the log transform in combination with the range checks already described in the manuscript. By applying the transform after the range checks, the most extreme outliers are

removed first, yielding significantly improved results. All figures and results for multiplicative merging have been updated to reflect this refined methodology.

During revision we also identified and corrected two technical issues in mergeplg: the sill parameter was incorrectly defined as partial sill rather than absolute sill (in practice this means that the sill was set lower than thought), and merged rainfall estimates were not capped, allowing physically unrealistic values in extreme situations. A cap of 150 mm/h has been introduced for all methods.

The main figure changes are:

- All result figures have been updated due to two technical corrections affecting all methods: the sill definition fix (partial to absolute sill) and the 150 mm/h output cap.
- Multiplicative merging figures have additionally been updated to reflect the log-transformed ratio field.
- Line plot figures (overall metrics, distance, intensity and range check analyses) have been updated to a colorblind-friendly palette with added symbols.
- New figures: variogram study (Fig. 3), variogram sensitivity analysis (Fig. 4), and neighbourhood size sensitivity analysis (Fig. B1).

Please find our point-by-point response below.

Response to reviewer 1:

The paper entitled “Improving weather radar rainfall estimates by merging with commercial microwave link data: a fully reproducible, large-scale method intercomparison” by Erlend Oydvin and co-authors investigates the contribution of commercial microwave link data to mitigate observation errors and biases in weather radar rainfall data through data fusion. The idea is potentially interesting but the current version of the study includes a major methodological error that in my opinion should prevent publication in the present form.

The authors propose to skip the variogram estimation step of Kriging, which is missing the point of Geostatistics that are based on estimating the spatial variability of the process at hand from data through variogram estimation (e.g., Chilès and Delfiner, 2012). To justify their choice the authors mention their own sensitivity analysis without sharing any result, and cite two papers: Haberlandt (2007) and Goudenhoofdt and Delobbe (2009). But I red these papers and they do not support the idea of not estimating the parameters of the variogram from data. In Haberlandt (2007) the hypothesis is tested, but the author clearly states that “the use of an assumed linear variogram is not recommended, because this shows by far the largest RMSE values for both interpolation methods (p 151)”. And in Goudenhoofdt and Delobbe (2009) the parameters of the variogram are estimated from data. In my opinion what one can conclude from Haberlandt (2007) and Goudenhoofdt and Delobbe (2009) is that the choice of a complex model of variogram is unnecessary in Kriging when a local neighborhood is used for interpolation, and that in this setting a linear model of variogram is good enough. But the idea of using a spherical model of variogram with arbitrary parameters is not at all supported by these references.

We agree that the original justification was insufficient. We have expanded the study in three ways:

1. Variograms are now estimated from data using an event-averaging approach adapted from Haberlandt (2007). Rainfall episodes are detected automatically, observations pooled and standardized by their variance, and parameters averaged over 14-day windows to obtain stable estimates. This approach is described in the new methods subsection 3.5.2 "Variogram estimated from data" (lines 195–210). The neighbourhood has also been increased from 12 to 70 nearest CMLs, with the choice justified by a sensitivity analysis (lines 106–109).
2. A dedicated variogram study (Fig. 3) is included, showing the temporal evolution of nugget and range parameters for both datasets, comparing CML and rain gauge observations (Section 4.2, lines 252–262). Results show that CML-estimated ranges in OpenMRG are consistently larger than gauge-estimated ranges, consistent with the spatial extent effect described by Western et al. (1999): rain gauges confined to the city centre cover too small an area to recover the full spatial correlation length, while CMLs spanning the wider network provide a more representative estimate. For OpenRainER, CML- and gauge-estimated ranges are similar, suggesting that path-averaging does not substantially inflate the apparent correlation range.
3. A systematic comparison of variogram estimation strategies is presented (Fig. 4), comparing the fixed reference variogram against data-estimated variograms with free and fixed nugget values (Section 4.3, lines 263–277). Results show that all methods are largely robust to variogram choice provided a non-zero nugget is used, with multiplicative block kriging on sparse networks being the most sensitive case. The fixed reference variogram performs at least as well as any data-estimated strategy across all methods and datasets. A neighbourhood size sensitivity analysis is also included (Fig. B1, Appendix B, lines 621), showing that performance is robust to neighbourhood size.

These results are discussed in Section 5.3 (lines 456–477), covering the robustness of the merging methods to variogram choice (smooth adjustment fields; a non-zero nugget is sufficient), and the competitiveness of the fixed reference variogram (data-estimated parameters offer no consistent advantage, likely due to instability of episodic fits from limited networks).

Another misleading use of a reference is the way the authors cite van de Beek et al. (2012) to justify the choice of the variogram parameters they use. The main point of van de Beek et al. (2012) is to show that variogram parameters vary a lot (in this case in time) depending on rainfall characteristics (cf Fig. 3 in van de Beek et al. (2012)). But in the paper in review the authors pick one single value for the variogram parameters, which is in contradiction with acknowledging the variability of variogram parameters. In addition, the way they derive this value from the work of van de Beek et al. (2012) is totally unknown and not justified. Finally, I would like to point out that van de Beek et al. (2012) focuses on the Netherlands, while the present paper investigates rainfall in Sweeden and Italy, where one can expect that rainfall spatial variability is different.

We agree that the original use of van de Beek et al. (2012) was insufficiently justified. Section 3.5.1 (lines 185-193) now explicitly acknowledges that van de Beek et al. (2012) demonstrate substantial temporal variability in variogram parameters, and clarifies that we retain their summer range of 30 km as a fixed reference baseline specifically to compare against data-estimated variograms, not as the single recommended choice. The justification

for the specific parameter values is now made explicit: the 30 km range is derived from Dutch summer rainfall, which is the closest available reference for summer conditions; the sill is set to 1 consistent with variance-standardised scaling; and the nugget of 0.3 is chosen to be broadly consistent with our own variogram estimates from data (Fig. 3). The limitation that van de Beek et al. (2012) focuses on the Netherlands rather than Sweden or Italy is acknowledged.

The problem of variogram estimation is also likely to impair block Kriging. The authors acknowledge that former studies estimated the point covariance from block averaged data (Graf et al. (2021); Goovaerts (2008)), but they renounce to this approach without any justification and prefer to use an arbitrary model of variogram with arbitrary parameters. This is very surprising since two of the co-authors of the present paper are also part of Graf et al. (2021), in which point variogram parameters are estimated in a sound way.

We have updated the reference to Graf et al. (2021) and Goovaerts (2008) to more clearly reflect how these studies estimate the point variogram (lines 172-176).

Further, the variogram estimation from data now applies to all kriging-based methods including block kriging (Section 3.5.2, lines 195–212). The impact of variogram parameters on interpolation is shown in Fig. 4 (Section 4.3, lines 264–277), which covers both point and block kriging variants for all merging methods. Results show that block kriging performance is similarly robust to variogram choice as point kriging for additive and multiplicative methods, and that both benefit from a non-zero nugget.

Response to reviewer 2

This paper introduces the Python package mergeplg, and demonstrates its value by applying it to two publicly available commercial microwave link (CML) datasets. Several of the merging algorithms in this package are tested, and some sensitivity analyses are carried out. I think the paper is interesting, well-written, and a valuable contribution to the field. The paper needs some clarifications and I think there are some important things missing in the discussion. However, I don’t think any new analyses are needed. I therefore recommend minor revisions. Specific comments are listed below.

On lines 56-57, I think it would be good to explicitly name mergeplg. I think it is a very valuable contribution to this field, and deserves a bit more attention in the paper.

An explicit mention of mergeplg has been added to the introduction (line 55), including a reference to the package and a brief description of its key features (KED, block kriging, uncertainty estimates).

Section 2: are the signals rounded/truncated (e.g. 1 dB r 0.1 dB), or do they have a precision that high enough to not be relevant?

Quantization information has been added to the datasets section: approximately 0.1 dB for OpenMRG (lines 64–65) and approximately 1 dB for OpenRainER (lines 74–75). A note on

the implications for data quality and range check sensitivity has been added to the discussion (lines 522–533).

Line 91: do you mean 30 dB? Please specify the units.

Units (dB) have been added to the quality control threshold (line 92).

Line 92: can "excessively noisy" be quantified?

The noise-based quality control threshold is now quantified with specific rolling standard deviation thresholds (lines 92–95), citing Graf et al. (2020). A note on the tradeoff between removing noisy data and retaining valid observations during intense rainfall has been added to the range check discussion (lines 522–533).

On line 100, radar-CML pairs are introduced. How is the radar rainfall estimate R defined? Is it the path-averaged radar rainfall estimate or is it the rainfall at the pixel over the center of the link? And are these definitions different for KED than for other methods? It would be good to also include a discussion about the implication of this choice on the results.

The methods section now clarifies that R is the path-averaged radar rainfall along the CML using intersection weights, and that this definition is the same for all methods (lines 104–107). A note has been added to the KED discussion explaining that using a path-averaged R introduces a scale inconsistency in the drift condition that may contribute to smoother KED fields, scaling with link length (lines 417–433).

On line 112, the ratio between CML and radar rainfall estimates is introduced. It would seem logical to me to log-transform this variable to make it symmetric (i.e., underestimation by a given factor would deviate as much from 0 as overestimation by the same factor). What is the reason for not using such a transform?

A log transform is now applied to the ratio field after range checks and before interpolation (lines 120–121). The results and discussion have been updated accordingly, and figures have been regenerated using log-transformed ratio data.

On line 115, please include units (I assume them to be mm/h).

Units (mm/h) added to the range check threshold (line 126).

On line 116, why is this condition assymetric? I would expect the lower boundary to be the reciprocal of the upper boundary, which is not the case here (0.1 is not equal to 1/15).

A note has been added that the thresholds are asymmetric by convention, and that other choices are possible (line 125).

On lines 110-120, I suggest using different symbols for the differences and ratios, and for the values of these differences and ratios on the one hand and their interpolated fields on the other. Now all four of these are called Z(v), and this can be confusing.

D(v) and F(v) are introduced as primary symbols for the difference and log-ratio fields in the additive and multiplicative sections respectively (lines 115–121). Z(v) and Z-hat(v) are

retained as generic symbols in the interpolation section only, with an explicit statement that Z(v) refers to either D(v) or F(v).

In Eq. (6) the covariance function C is introduced. However, the same symbol is used for CML rainfall estimates. This is confusing. Please select a different symbol for one of these variables.

CML rainfall is now denoted R throughout, removing the symbol conflict.

On lines 140-141, it is mentioned that the results do not depend too much on the variogram that is used. I think this would only hold if the distance between the points to be interpolated is (much) smaller than the range of the variogram. This could affect the results of the OpenRainER dataset.

A variogram sensitivity analysis has been added (Fig. 4, Section 4.3, lines 264–277) and a neighbourhood size sensitivity analysis (Fig. B1, Appendix B, lines 623). Results show that the merging methods are largely robust to variogram choice. A discussion paragraph has been added interpreting these findings (Section 5.3, lines 461–481). See also the response to Reviewer 1 above for a full account of the variogram investigation.

On lines 141-143, the variogram parameters are introduced. The nugget and sill are expressed in (mm/h)2, so I'm assuming these are the parameters for the additive method. What are the parameter values for the multiplicative method?

A sentence has been added clarifying that the same parameters are used for all methods, justified by the variogram study showing no substantial differences between difference and ratio fields (lines 190–193).

On lines 145-154, KED is explained. It would be good to include here that the CML rainfall estimates are directly interpolated, and not a difference or ratio of CML and radar estimates.

A clarification has been added to the KED section (lines 151–153).

On lines 156-166, it is unclear what the meaning of a block is in this context. Is it the path of a link? If so, it would be good to state that. And with that, it would be good to more precisely define the "average point covariance between blocks" and the "average point covariance within each block".

A definition of block as the CML path has been added to the block kriging section (lines 160–170), with a pointer to Appendix A for explicit definitions (line 170).

In Fig. 2, it is unclear to me what the gray-scale colors in the background of the third columns are. Can this be added to the caption?

The caption now states that the rainfall field from each method is shown as a gray-scale background for spatial context (Fig. 2 caption).

In Fig. 5, I have trouble telling the different lines apart because of slight color-blindness. Is there a way to make the graph easier to read, for example by using different symbols/line types?

Line plot figures have been updated to use a colorblind-friendly palette and symbols have been added to lines (all line plot figures).

On line 259, I think “near rain gauges” should be “near CMLs”.

Fixed in the distance-based analysis section (line 329).

On lines 261-274, the sensitivity to the range checks is discussed. How many samples were removed (percentage of the total amount of samples) by the different range checks? I think it would be valuable information to include in the paper.

Removal percentages have been added (lines 335–340): less than 2% for the additive range check in both datasets (0.0% OpenMRG, 1.4% OpenRainER); 37.4% (OpenMRG) and 53.2% (OpenRainER) for the multiplicative range check under standard settings. The large difference is explained by ratio-based checks being sensitive to near-zero radar values, where even small CML observations produce large ratios.

On lines 261-274, a sensitivity analysis is presented of the range checks. Was it considered to also carry out a sensitivity analysis on the variogram parameters?

See response to reviewer 1 above (Section 4.3, lines 263–277).

On line 264, please include units (mm/h).

Units (mm/h) added to range check thresholds in the range check sensitivity section (line 337).

On line 270, the difference between the two datasets in terms of their response to the range checks is mentioned. What could be the cause of this. Are the CML estimates worse for the OpenRainER dataset?

An explanation has been added: OpenRainER operates within a narrow single-frequency band (24.6-25.6 GHz), where rainfall sensitivity varies with link length, making uniform range check thresholds harder to apply (lines 522-533). OpenRainER also operates with coarser signal quantization which could impact results (lines 522-533).

On lines 297-320, the effects of the dataset characteristics on the results are discussed. What I’m missing here is a discussion of the uncertainties or errors in CML rainfall estimates and how they affect the results. I think this should be included in the paper.

A paragraph has been added covering wet-dry classification, baseline estimation, and WAA correction as sources of CML uncertainty (Section 5.5, lines 534–546).

On lines 340-341 and on line 348, it is mentioned that the path-averaging effect of the CMLs could play a role in dampening rainfall extremes in the interpolated product. If that would be the case I would expect to see a large effect of block Kriging versus point Kriging, but that’s not visible in the results.

A discussion on why block kriging does not improve on point kriging has been added (Section 5.4, lines 488–497).

On line 344 it is mentioned that KED has the tendency to produce smoother rainfall fields. Is this generally the case, or is this specific for these datasets (or only the examples shown in Fig. 2)? I would expect that KED would result in a smoother field only if the radar data are smooth to begin with. Please elaborate on this in the paper. This is important because this apparent smoothing character of KED is repeated throughout the discussion.

The discussion now states that smoother KED fields are a consistent property across both datasets, and that the neighbourhood size sensitivity analysis confirms this is not driven by neighbourhood size, but rather the path averaged drift (Section 5.2, lines 417–433).

On lines 346-348, the potential effect of non-linearity of the relation between CML rainfall and radar rainfall is discussed. What could cause such non-linearity, and how severe could it be? If there's no reason to expect non-linearity, I suggest removing this statement from the paper because it would then not be relevant.

The statement has been removed.

On line 349, it is stated that path-averaging of the CMLs combined with possible nonlinearity of the relation between CML and radar rainfall can lead to a smoothed drift surface. Is this really the case? And if so, please explain more clearly in the paper what the mechanism behind this is.

The sentence has been removed.

On line 353, I think "from rain gauges" should be "from CMLs".

Fixed in the discussion.

On lines 361-366, multiplicative methods are discussed. I think it would be good here to mention that most errors in radar rainfall estimation are multiplicative, and that from that perspective, it would make sense to use a multiplicative method for merging. The fact that the multiplicative methods actually perform worse makes these conclusions even more relevant.

The discussion now notes that radar errors are often multiplicative in nature, making multiplicative adjustment theoretically well-motivated (lines 441–443). This is framed positively given that multiplicative methods achieve the best RMSE and MAE for OpenRainER.

On line 404, I think "performed on pair" should be "performed on par".

Fixed.

On line 406, what is meant by "rainfall simulators"? Does this refer to the correlation or variogram functions?

Replaced with "stochastic spatial reconstruction" (lines 509).

On lines 410-412, and important characteristic of Kriging is discussed that wasn't discussed before: the interpolation uncertainty estimates. I think it would be good to mention this in the

introduction of the paper. And I think that the claim that is made about the improved uncertainty representation of block Kriging cannot be made based on the results presented in this paper (uncertainties are not studied here). So I suggest to reformulate this part.

A mention of KED and block kriging including uncertainty estimates has been added to the introduction (line 55). The conclusions have been reformulated to remove the unsupported claim.

It would be interesting to also discuss the effect of the particluar characteristic of Kriging that it will give more weight to isolated observations. For example, if there are 10 links to the south of the point where I would like to have an estimate and there is only one link to the north, the link to the north will have more weight than, e.g., IDW (where all links would get equal weights if they are all at the same distance from the point of interest). This could be an advantage when dealing with a network that has a very heterogeneous distribution in space like the OpenRainER dataset.

This property is now noted in the interpolation discussion (Section 5.4, lines 498–507).

I'm missing a discussion about the use of the 12 nearest links. For IDW this makes perfect sense. However, for Kriging, you could miss information that would have received a large weight depending on the link topology. Please include a discusison on this in the paper.

A neighbourhood size sensitivity analysis has been added (Fig. B1, Appendix B, lines 624), and the neighbourhood has been increased to 70 nearest CMLs (lines 105–108). A discussion on the advantages of line topology has been added to the discussion (lines 498 - 507). A discussion on the potential stabilizing effect of a local neighbourhood has been added to the merging discussion (lines 394-405).