How uncertain are precipitation and peakflow estimates for the July 2021 flooding event?

by

Mohamed Saadi, Carina Furusho-Percot, Alexandre Belleflamme, Ju-Yu Chen, Silke Trömel,
and Stefan Kollet

Submitted to *Natural Hazards and Earth System Sciences*

Manuscript ID: NHESS-2022-111

Answer to Anonymous Referees

1 Summary of main changes

We gratefully acknowledge the valuable suggestions made by the two Anonymous Referees and we would like to thank them for their time and evaluation. We addressed all comments in detail (Sections 2 and 3). In summary, the major changes in the data and methods applied in the study are:

- We included new radar-based, quantitative precipitation estimates (QPE) that better account for the vertical gradients of radar variables (and hence of precipitation rates). Compared to state-of-the-art QPE products (Chen et al., 2021), these new products (with VPC in their names, for Vertical Profile Correction) exploit measurements of Micro Rain Radars (MRR) that helped characterize the precipitation rates below the height monitored by the C-band radars of the DWD (*Deutscher Wetterdienst,* German Weather Service). In addition, a vertical profile correction was applied to horizontal reflectivity *Z* and specific differential phase K_{DP} following an approach by Chen et al. (2020). These new products significantly improved the radar-based QPE with respect to estimates from rain gauges.
- We removed the QPE product based on specific attenuation at vertical polarization (A_V) and K_{DP} (RAVKDP in the original manuscript) as it yielded similar results to RAHKDP, the one based on specific attenuation at horizontal polarization (A_H). Hence, the number of radar-based QPE products is now RADOLAN + six other products (RZ, RZKDP and RAKDP, in addition to the version with corrected vertical profiles RZ-VPC, RZKDP-VPC, and RAKDP-VPC)
- We added a new simulation of ParFlowCLM with distributed Manning's coefficient assigned based on land cover.
- The number of rain gauges used for comparison with the radar-based products was reduced from 67 to 63 after re-checking the gaps.

The conclusions of the paper have slightly changed. Namely, the new products with vertical profile correction improved the estimates of event precipitation with respect to rain gauges. The point-scale evaluation and catchment-scale evaluation led to similar ranking of the different QPE products with respect to RADOLAN. Finally, the probabilities of exceeding the historical peakflow were highly sensitive to QPE for all catchments.

Below we provide a detailed reply to the comments of Referee #1 and Referee #2.

2 Response to comments of Anonymous Referee #1

General comment: ***"**The authors present a modelling study targeted at evaluating different rainfall products and two hydrological models to simulate the flood event of 2021 in West Germany, in order to sow the uncertainties in quantitative precipitation estimates (QPE) and the modelling. In general, the study provides insights in the usefulness and weaknesses of the different radar-based rainfall products and their use in simulating extreme flood events. The manuscripts is overall well written and structured.*

However, I have some reservations to the conclusions drawn, mainly because of the study design, in particular the hydrological modelling part. My concerns are as follows:"

Comment 1: *"Different model parameterizations (ParFlowCLM) and calibrations (GR4H) were derived and later used without any differentiation of their performance simulating the historic period. This is actually hindering a proper evaluation of the QPEs, because poor performing hydrological models might be (or are) used to simulate the flood in 2021 with the different QPEs. I strongly recommend to list the performance of the different model parameterizations/calibrations and sort out poor performing ones. In any case the model performances should be provided by the Nash-Sutcliffe and Kling-Gupta performance measures, because these were already calculated. The claim of the authors that the model parameterization/calibration has a larger impact than the QPEs is not that surprising, considering the sensitivity-analysis-like selection of the parameters and calibration routines. The conclusions towards selecting a particular QPE would be more meaningful, if only well performing models for flood events (high discharge) during the calibration period would be used."*

Authors' response: From a hydrological perspective, what is actually obstructing a proper evaluation of the QPEs is the absence of streamflow observations including error estimates for the event. Therefore, we performed a sensitivity study to understand how the different QPEs and model parameterizations impacted the peakflow estimates for the July 2021 event. In addition, we studied how QPE impacted the model estimation of the severity of the event (quantified by the chances of exceeding the historical peakflow). In the new version of the manuscript, we also added the estimates of peakflow based on water level (Mohr et al., 2022), which showed the ability of the models (especially the uncalibrated ParFlowCLM) of reproducing these estimates when RAKDP-VPC is used as QPE input. Acknowledging that this is not enough to discriminate the QPEs, we included an evaluation with respect to measurements from rain gauges, the results of which can help select a particular QPE in a meaningful way (in this respect, RAKDP-VPC seems to be the best one).

We would also like to stress that relying on the performances of hydrological models on historical events does not guarantee a proper evaluation of the QPEs for the event under consideration, because the conditions of the event we are simulating are unprecedented. In addition, the criteria under which one can consider a model to be well-performing are not well defined. We illustrated this issue using GR4H and adopting several calibration objectives (i.e., with respect to reproducing the whole historical period, or with a focus on the top 10% of the discharge data, or with focus on the top 1% of the data), as explained in Lines 110-119 of the original manuscript. As can be seen from the results, this does not overcome the high uncertainties in the simulated hydrographs by GR4H (Figures 5 and 6 of the original manuscript). As suggested, Figure R1 shows the calibration performances for each of the 12 sets of parameters and for each catchment. Most of the obtained calibration scores are better than the climatology model (i.e., NSE > 0 and KGE > - 0.41), except for four parameter sets for the Erft at Neubrueck, and one parameter set for each of the Erft at Bliesheim and the Kyll at Kordel, for which the calibration scores (in this case, NSE) were less than the chosen threshold (i.e., NSE = 0).

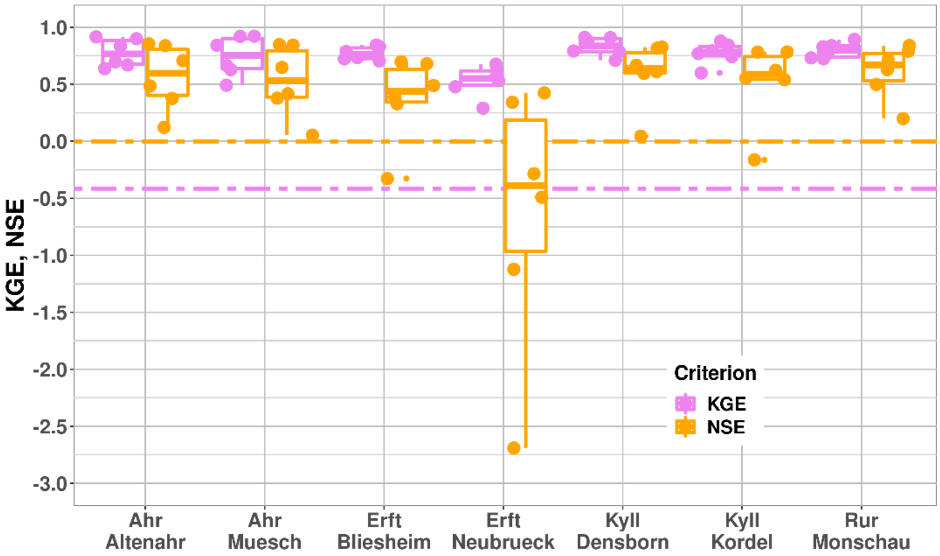

Figure R1 : Calibration performances of GR4H for the study catchments. Orange dashed line indicates the threshold NSE = 0, in which the hydrological model is as good as the climatology (i.e., mean observed flow). Pink dashed line indicates the KGE score for the same benchmark.

Removing these parameter sets had little effect on the conclusions. Specifically, the probabilities of exceeding the highest (measured) peakflow for each catchment were not impacted by excluding these ill-performing parameters, as can be seen in Figure R2.

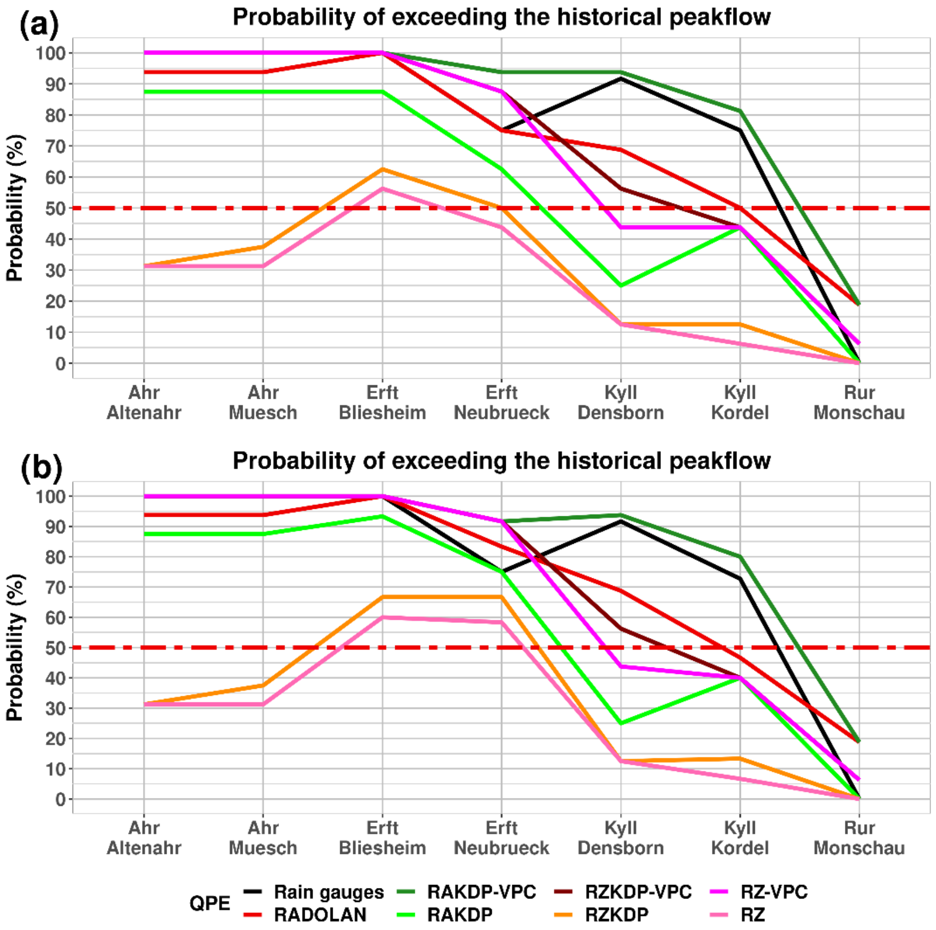

Figure R2 : Effect of different QPE on the probability that the simulated peakflows by GR4H and ParFlowCLM exceed the historical peakflow for each catchment, with (a) all GR4H parameters included, and (b) with only well-performing parameters (i.e., with NSE > 0 and KGE > -0.41).

To conclude, even if it is not possible to rank the tested QPEs with hydrological simulations because of the absence of measured peakflows for the event, these results show that it is possible to compare QPEs from the flood-forecasting perspective by looking at their ability to detect historical peakflow exceedance probability. The comparison with rain gauges can identify the best QPE product with respect to rain gauges, but this does not necessarily inform us of their utility for flood forecasting. In addition, we kept the different parameterizations to consider the uncertainties related to the different modelling approaches and calibration options. The fact that the modeling approaches are contrasting may explain the dominant effect of model parameterizations compared to that of QPEs. Moreover, the level of uncertainty in peakflow estimates is in line with levels of uncertainty reported by Kreienkamp et al. (2021) for the Ahr at Altenahr (see their Table 2, p. 8). We added this specification in the revised manuscript, by modifying Lines 267-270 to:

“The sensitivity of model simulations confirms the dominant impact of QPE on hydrological model performances (Braud et al., 2010; Oudin et al., 2006), underlining the need for reliable precipitation estimates especially for extreme flooding events. However, the effect of QPE seemed relatively smaller (but still important) than that of model parameterizations (Fig. 8), and it was variable from one catchment to another for the 14 July event (Fig. 6-7). The large differences between model estimates for a single QPE input reflect how uncertain peakflow estimates can be for such an extreme event (see Table 2 for the Ahr at Altenahr in Kreienkamp et al., 2021). The stronger effect of model parameterizations with respect to QPE may be due to the inclusiveness of our approach that did not exclude ill-performing parameterizations, especially in the case

of ParFlowCLM. Removing these would lead to lower differences due to hydrological models, but this removal needs streamflow measurements for the event, which are unavailable or highly uncertain for our catchment set”

Comment 2: *“The parameterization of ParFlowCLM with uniformly distributed roughness values is very unrealistic for these catchments with diverse land uses, i.e. land surface properties. I am surprised that such a simplistic approach is used for such a sophisticated, physically based and spatially distributed model. Thus I strongly recommend to re-run the simulation with distributed roughness values estimated based on land use and standard roughness values, as mentioned in the outlook. This would give the ParFlowCLM simulation much more credibility.”*

Authors’ response: We agree with the Referee and added new runs in which the roughness values were distributed based on land use types. For all catchments, this yielded hydrographs that were bracketed by the uniform simulations of ParFlowCLM using the median roughness case (MMann, in which the parameter was set to 0.1 s/m$^{1/3}$) and the low roughness case (LMann, where the parameter was set to 0.03 s/m$^{1/3}$ in the new manuscript version). We added the following paragraph to the revised manuscript:

“To account for the uncertainty in Manning’s roughness coefficient, which highly impacts the peakflow simulations (Lumbroso and Gaume, 2012), different scenario simulations with spatially homogeneous and distributed roughness values were performed. In total, three spatially homogeneous values were tested for the whole domain: a default value of 0.2 s m$^{-1/3}$ (HMann, i.e. high roughness, from Schalge et al., 2019), and two additional values of 0.1 s m$^{-1/3}$ (MMann, medium roughness) and 0.03 s m$^{-1/3}$ (LMann, i.e. low roughness). These three values cover the whole range of Manning’s coefficient values reported by Lumbroso and Gaume (2012), but adopting a uniform spatial distribution (although simple to implement and to interpret) is unrealistic given the differences in roughness values between land-cover types. Therefore, a fourth simulation was performed using distributed Manning’s coefficients (DMann) based on land cover types (and following Table 2 in Asante et al., 2008), with low values for water bodies (0.02 s m$^{-1/3}$) and urban and barren surfaces (0.03 s m$^{-1/3}$), mild values for croplands (0.033 s m$^{-1/3}$), natural vegetation mosaics (0.037 s m$^{-1/3}$), shrublands, grasslands, snow/ice, and permanent wetlands (0.05 s m$^{-1/3}$), and high values for forests (0.1-0.12 s m$^{-1/3}$).”

The results of the new simulation for distributed Manning’s coefficients were accounted for in the remainder of the paper.

Comment 3: *“For the GR4H model I find using the calibration not focussing on extremes for the analysis of the QPEs not convincing, because a conceptual model calibrated on mean flow is unlikely to get the peak discharges of floods right, and should thus not be used for evaluating the QPEs. You might prove me wrong listing the performance values.”*

Authors’ response: Some of the 12 parameter sets of the GR4H simulations (actually 2/3 of them) were obtained with focus on the top 10% and 1% of the discharge values (see Lines 109-119, where $Q_{obs,th}$ was changed to the discharge values with frequency of non-exceedance of 10% and 1%). The corresponding performances are shown in Figure R1. Most of the parameters obtained good NSE and KGE scores. However, we do keep in mind that the use of daily discharge values limit the information content of the estimated parameters, as mentioned in Section 5.3 regarding study limitations. But this can be less detrimental given the size of the catchments, for which the daily time step is somewhat reasonable.

Comment 4: *"Furthermore, some of the comparisons/evaluations of the QPEs and simulations are based on comparison with uncertain or unknown quantities. The missing flood hydrographs are a major obstacle here. Meanwhile reconstructed flood hydrographs are available at least for the catchments in Rhineland-Palatine by the Landesamt fuer Umwelt (LfU). Similar data should be available from the authorities in Northrhine-Westphalia. These hydrographs can be seen as the best estimate of the actual flood event. I strongly recommend to obtain these data sets. This would increase the impact of the evaluation in terms of ability to simulate the flood 2021 significantly."*

Authors' response: We were in close contact with the Environment office of North Rhine-Westphalia (Mr. Martin Brinkmann, martin.brinkmann@lanuv.nrw.de), and he told us that (5th of May) that unfortunately their analysis of the event is still unavailable. We sent another email recently and we're waiting for a response. We also sent requests to the Rhineland-Palatinate office of Environment to get their analyses, with no response. Fortunately, a transdisciplinary study in review by Mohr et al. (2022) reported some of these estimates, which were obtained with hydraulic approaches based on relationships between water level and discharge or using hydraulic models. We added these estimates to Figure 6 of the manuscript, and we compared our model estimates with them. We found that ParFlowCLM estimates bracketed well these estimates when RAKDP-VPC (the best product relative to rain gauges) is used. In the new version of the paper, we added these results as follows:

"Overall, the ranking of QPE products with respect to the total precipitation depth for the 14 July event was preserved by model simulations for all catchments, as shown in Fig. 6. Model simulations with RADOLAN as input barely reached reported estimates by Mohr et al. (2022) based on relationships between water level and streamflow (red dashed lines in Fig. 6). Using RAKDP-VPC as input, simulations of ParFlowCLM bracketed well the estimates based on hydraulic approaches, with the best estimates obtained with median or distributed Manning's coefficient (MMann and DMann). GR4H also succeeded in bracketing these estimates except for the Erft at Bliesheim, but most of GR4H peakflow estimates for this catchment were lower than the ones based on hydraulic approaches."

Comment 5: *"Another point: the comparison of the catchment average precipitation used the Thiessen polygons as reference, but these values are also very uncertain. Thus, the general statement that some of the QPEs outperform RADOLAN in catchment average is actually not supported. You only show that these products are closer to the uncertain catchment average based on rain gauges. Which of the QPEs is actually closer to reality cannot be derived form this comparison. This should be mentioned."*

Authors' response: We agree with the Referee's comment. We changed the sentence in question to underline that some QPEs outperform RADOLAN *with respect to reproducing the estimates from rain gauges*, which themselves (we admit) are also uncertain. We specified this in the revised manuscript:

"Conclusions about the agreement between QPE products and rain gauges are similar when we look at the catchment-scale evaluation. Specifically, QPE based on specific attenuation (A) with corrected vertical profiles for K_{DP} (RAKDP-VPC) outperformed RADOLAN in reproducing estimates from rain gauges (using Thiessen polygons) across the seven catchments (Fig. 4), and reduced relative error from a median of -18 % for RADOLAN to +2 %."

We also added estimates from the REGNIE product, which uses a better interpolation method of precipitation fields from rain gauges (but available at the daily time step) to the new manuscript version. REGNIE estimates were similar to those we obtained from Thiessen polygons except for the Erft at Bliesheim (Figure R3).

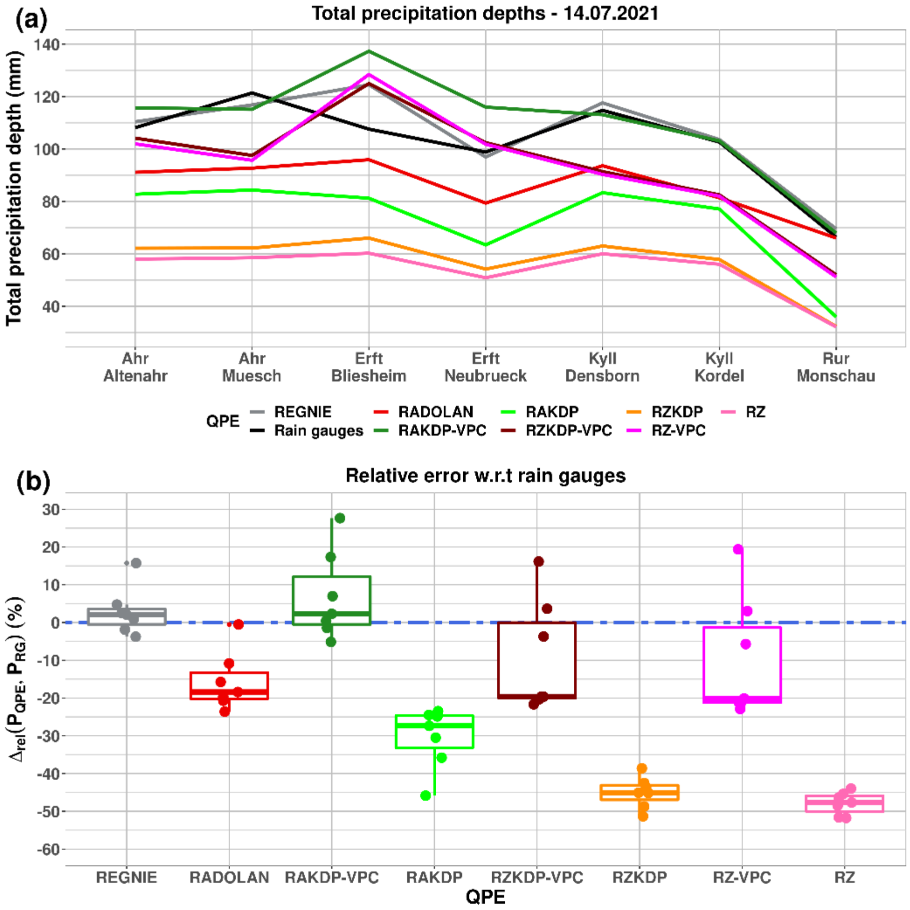

Figure R3 : (a) Total precipitation depths for the 14 July 2021 estimated by rain gauges, REGNIE and radar-based QPE products. (b) Relative errors in REGNIE and radar-based QPE with respect to estimates from rain gauges using Thiessen polygons of the total catchment-scale precipitation depth for the 14 July 2021.

Comment 6: *"I am also missing the discussion of hydrologic processes that might become relevant or only occur during extreme floods. This is a generally ongoing discussion in hydrology, but for this particular event the increase interflow and thus runoff generation by field drainage pipes or the creation of additional drainage channels by erosion has been reported. Unfortunately, this is not published yet, thus you cannot cite it, but there should be reports in newspapers or by the authorities available."*

Authors' response: We focused on using available hydrological tools and informing them with different precipitation estimates for the event, and then analyze how the peakflow estimates varied. The main message is to show that for an extreme event, such as the July 2021 event, uncertainties in peakflow estimates can be very high due to high uncertainties in precipitation estimation and hydrological modeling. Other studies provide this discussion by focusing on the description of the event, such as the recent one by Mohr et al. (2022).

Nevertheless, we included a paragraph in the discussion section advocating for a coupling of hydrological and hydromorphological models to account for the crucial interactions between hydrology and river morphology in the context of anthropogenic influence. We added the following lines to the revised manuscript:

"Accounting for the 3D, soil and subsoil heterogeneities in the representation of hydrological processes allows for ParFlowCLM to well represent the runoff generation by overland flow and increased interflow in the upstream steep part of the study catchments, but it would be improved by including anthropogenic effects on

hydrological processes that had a large impact on the flood generation mechanisms for this event (Mohr et al., 2022). The structure of ParFlowCLM allows for coupling the complex hydrological and morphodynamical processes (sediment and debris transport, bank erosion, and developing landslides) that non-linearly interacted with the flood propagation and river morphology increasing the destructiveness of the event.”

Comment 7: *“The role of the antecedent soil moisture has been briefly discussed in the manuscript, but studies for its impact on flood generation has been given as an outlook only. I wonder about two aspects: First, the used initial soil moisture for the simulation of the flood 2021: what initial soil moisture was assumed? Was it assumed dry, a guess of some wetness, or maybe based on satellite observation? Or did you use the hydrological simulations until the event to prime the model for the flood simulation? In the latter case the antecedent soil moisture should be realistic to some extent. If assumed, some justification or at least explanation has to be given. Second, an interesting aspect would be if the flood would have been different if the soil was in different state (drier, wetter) than in reality. You mentioned this in the outlook, and this is surely worth investigating, as the role of antecedent soil moisture is likely to differ in different flood/rainfall situations. If you have any capacities, I recommend to include this aspect, and drop the discussion of the simulation results of poor performing models.”*

Authors’ response: For the estimation of the antecedent soil moisture conditions, we used hydrological simulations to initialize the models for event simulations. Both GR4H and ParFlowCLM were run continuously starting from 2006-2007 for all catchments. This allowed for exploiting the whole record period to yield the best estimate of model initial conditions prior to the event. We now mentioned this in the revised manuscript when we present how QPEs are evaluated using hydrological models. The following statement was added to Section 3.4:

“Second, we examined the effect of QPE on the frequency of exceeding the highest historically observed peakflow for each catchment (Table 1) by simulated peakflows. Both GR4H and ParFlowCLM were initialized using a long spin-up period starting from 2006 for GR4H and 2007 for ParFlowCLM. This allowed for exploiting the whole available record period of climatic forcing to yield the best estimates of antecedent soil moisture conditions. Then, each radar-based QPE was used as input to both models to obtain twelve peakflow simulations from GR4H and four peakflow simulations from ParFlowCLM. These peakflows are compared with the highest historically measured peakflow.”

For the question of the impact of uncertain antecedent soil moisture conditions, we agree that it is an important aspect to look at, but our aim is to focus on the quality and uncertainties of QPE products for the event and not their quality and uncertainties in front of initial conditions. Furthermore, we are limited by the computational costs for ParFlowCLM: the model is actually implemented at the scale of Central Europe ($4*10^6$ grid cells times 15 soil and subsoil layers), and analyzing other scenarios of antecedent soil moisture will require significant amount of computational resources. In addition, the current paper has already significant results with respect to the effect of QPE or modelling approaches, and adding another aspect (in this case, the effect of antecedent soil moisture) would occult the main messages.

Comment 8: *“In addition to these general comments, I have some more specific comments in the annotated manuscript.”*

Authors’ response: We accounted for the specific comments in the annotated manuscript as follows:

Specific comment	Answer
“the actual damages were even much larger due to non-insured damages in infrastructure. might be worth mentioning.”	Based on the paper by Mohr et al. (2022), we updated the line in question: ***“The flooding events of July 2021 in Europe resulted in more than 220 deaths (Deutsche Welle, 2021), large-scale damages to infrastructure (Koks et al., 2021) and costs of up to €8.2 billion in insured losses (GDV, 2021) and up to €32.05 billion in total losses in Germany alone (BMI, 2022), making them the most severe natural disaster caused by heavy rain and flooding in Germany (Mohr et al., 2022).”***
“some information about major geological (underlying rock formations) and soil properties would also be helpful.”	We computed the catchment average silt, sand, and clay content (%) of the catchments from the European Soil Database and added these to the paragraph describing the study region: ***“The region is characterized by sedimentary rocks interbedded with volcanic rocks, with relatively shallow soils characterized by low water-holding capacity (Kreienkamp et al., 2021) and dominated by sand (catchment averages: 34%-41%) and silt (catchment averages 29%-38%; Panagos, 2006).”***
*“this means (calibration with either period 1 or 2) * (NS or Kling-Gupta) * (three thresholds)? The first two terms are not that clear from the description. should be improved. And why calibrating on two different performance values separately, and not combine them in one performance measure to make the best out of both performance measures? This would ease the selection of a selection of models for the evaluation of the QPEs.”*	Yes. The equation is now clarified according to the Referee’s suggestion. We chose not to combine the measures in order to see how the use of each of the objective functions impacts the estimated parameters. Of course combining both measures would lead to better constraining the parameter sets, but this would not help elucidate the impact of choosing one or the other criterion.
“the definition of a uniform roughness over the whole domain must be justified. It makes much more sense to differentiate the hydraulic roughness according to landuse/landcover. The high roughness of 0.2 is likely appropriate for forest, but not for the river course of build-up areas. The opposite holds true for the low roughness. From a hydraulic perspective the selection of uniform roughness for simulating overland flow is highly questionable, thus the approach of making a sensitivity	We added a simulation using distributed Manning’s roughness coefficients. See our response above to Comment 2, where we acknowledged that using uniform roughness is unrealistic. However, this sets a comparison framework using simple benchmark parameterizations to evaluate the added value of using a distributed Manning’s roughness.

analysis with uniformly distributed roughness values needs to be justified."	
"why averaged over the 8 neighbouring cells, i.e. over an are of 9 km2, and not only the radar pixel? This implies the assumption that the rain gauge is representative for an area of 9 km2, which might be appropriate, but also not, depending on the topography and the rainfall spatial distribution. Thus an explanation is required."	We chose to average over the 8 neighboring cells to account for small differences in location between radar cells and gauges observations due to motion and vertical variability, as done by Dai and Han (2014) and Schleiss et al. (2020). Using only one radar cell without averaging had little effect on the estimated criteria. We added a justification for this choice to the revised manuscript.
"derived by the Thiessen polygons, I assume?"	Yes, we specified this in the revised manuscript.
"this is not that surprising, considering that RADOLAN is adjusted to rain gauge records, isn't it?! might be worth mentioning, as international readers will not understand the meaning of the acronym RADOLAN"	This statement is no longer valid when adding the radar-based QPE with vertical profile correction (VPC). The sentence in question now reads: ***"At the point scale, the comparison with N = 63 rain gauges over the region shows that the radar-based QPE with vertical profile correction and gap-filling are the ones that agreed most with the rain gauges (Fig. 3)."*** The meaning of the acronym RADOLAN is already mentioned in Table 2.
"If I am not mistaken, the reference catchment scale precipitation is obtained by Thiessen polygons based on the rain gauges. This means that also the reference is very uncertain, thus I have reservations about the conclusions. You can state that the QPE under-/overestimate the catchment rainfall in relation to the uncertain gauge-based catchment rainfall, but not to the actual catchment rainfall, which is in fact unknown. It might also be argued, that the spatial distribution of rainfall is much better represented by the radar QPEs that be the interpolated rain gauges. *This is a dilemma, unfortunately, but you should take this into consideration when interpreting your results."*	See our response to Comment 5 above.
"which spatially uniform distribution is unrealistic!"	See our response to Comment 2 above, where we acknowledged that the uniform distribution is unrealistic. In addition, we modified the part in question to: ***"Both the choices of GR4H calibration and Manning's coefficient for ParFlowCLM led to high uncertainty of***

	peakflow simulations. With a high Manning's coefficient, ParFlowCLM succeeded in estimating both the timing and the magnitude of the last recorded peakflow at the catchment outlet (~330 m^3 s^{-1} at ~19:00 of the 14 July), whereas the median simulation of GR4H was quite delayed with respect to simulated hydrographs by ParFlowCLM. Using a distributed Manning's coefficient (DMann) led to similar ParFlowCLM simulation as when using uniformly distributed, median Manning's value (MMann) for the Ahr at Altenahr. Finally, all model simulations with both RADOLAN and RAKDP-VPC illustrate how the heavy precipitation event resulted in a record-breaking flood for the Ahr at Altenahr."
"what causes the spread of the GR4H simulations? the different calibration periods, the performance measure of the different thresholds for peak flows? *This is an interesting information for the interpretation of the results and should be mentioned."*	Looking at Figure R4 below, Most of the spread is caused by the period of calibration, with systematically higher peakflows obtained when calibrating on P2 then when calibrating on P1. Simulated peakflows based on KGE led to higher peakflows on average compared to NSE. However, the effect of thresholds is variable from one catchment to another. For some catchments, focusing on the high flows in the calibration led to higher estimated peakflows, whereas for others (especially the Erft at Neubrueck, albeit with some ill-performing parameters), it led to the same or lower peakflows compared to the default calibration (i.e., no specific focus on high flows). Although this is an interesting result, we chose not to add another figure to the manuscript. Instead, we added in the Results section in the part related to interpreting Figure 6 the following: ***"For GR4H, analyzing the effect of calibration choices (not shown here) showed that the choice of the calibration period had the greatest impact on the simulated peakflows across the catchments, with higher peakflows obtained when the latest period in time is used for calibration."*** When discussing the GR4H simulations (second paragraph of Section 5.2), we added the following:

	"Finally, the analysis of the effect of the calibration choices on GR4H simulations (not shown here) highlighted the effect of the hydroclimatic specificities of the calibration period on the model simulations for an unprecedented or future events (Brigode et al., 2013)."
"antecedent soil moisture surely plays a role in flood generation. it's impact depends, however, on the rainfall intensities. It might play a large role particularly on floods on a alarge scale like in the flood in 2013 in Germany (Schröter et al. 2015), but its role can also be negligible in case of heavy convective rainfall and flash floods, as for e.g. in case of the flood in Braunsbach in 2016. It would be interesting to know and find out by you modeling concept, if antecedent soil moisture played a role, or could have played a role in the 2021 flood (i.e. would the flood have been different with wetter or drier catchments, or the same). I suggest to includde this in the manuscript. *Schröter, K., M. Kunz, F. Elmer, B. Mühr, and B. Merz (2015), What made the June 2013 flood in Germany an exceptional event? A hydro-meteorological evaluation, Hydrol. Earth Syst. Sci., 19, 309-327, doi: 10.5194/hessd-11-8125-2014."*	If we would like to assess the effect of soil moisture, we need historical events that are similar in terms of precipitation amount to the July 2021 event but with different antecedent soil moisture conditions, as done by Schröter et al. (2015). However, this is beyond the scope of our study (see our response to Comment 7 above). We nevertheless mentioned in the new manuscript how antecedent soil moisture plays a role in flood generation especially in extreme flooding events (including the proposed reference by the Referee) as follows: ***"High enough antecedent soil moisture conditions can indeed lead to extreme flooding events even when the precipitation amount is not relatively extreme (with respect to historical events), as shown by Schröter et al. (2015) for the exceptional June 2013 flooding event in Germany."***
"how is this translated into antecedent soil moisture at the onset of the event?"	We did not use those estimates to initialize the event, but we run the models continuously starting from 2006-2007 to estimate the initial conditions for the event. See our response to Comment 7 above.
"what about the large mining pits? They should, and as far as I am informed, indeed had a significant impact on the flood generation, as thy stored a lot of water."	We added the large mining pits as a possible factor for differences between GR4H and ParFlowCLM.
"The Landesamt für Umwelt in Rhineland-Palatinate reconstructed the water levels and discharge of the event. These might (or should?) be acquired to evaluate the models. *Likely also the authorities in Northrhine-Westphalia have similar information."*	See our response above to Comment 4. In addition, these are also highly uncertain estimates, but are complementary with ours: theirs are based on hydraulic approaches, whereas ours is based on hydrological considerations.
"I strongly suggest to include these aspects in the study to increase its impact and reduce questionable assumptions like the uniform Manning roughness."	We added a simulation with distributed Manning's roughness, see our response to Comment 2 above. For the effect of soil moisture, see our response to Comment 7 above.

“explain the green shaded area. I assume that it is the min/max range of the simulation with different parameter sets for GR4H, but this needs to be explained.”	We added an explanation of the green shaded area in Figure 5 of the manuscript: ***“The green shaded area is delimited by the minimum and maximum values of estimated discharge by GR4H for each time step.”***

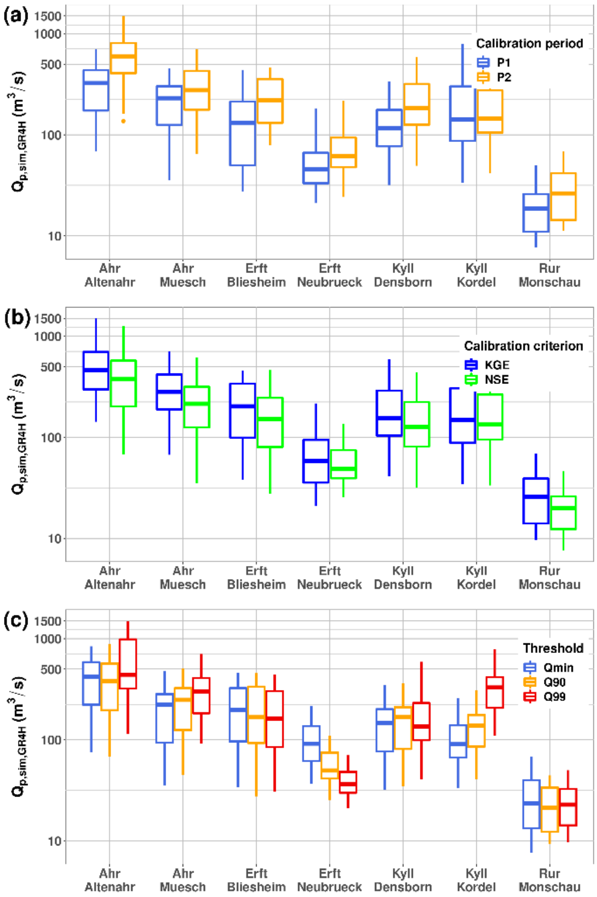

Figure R4 : Differences between simulated GR4H peakflows due to (a) calibration period, (b) calibration criterion (Nash-Sutcliffe Efficiency – NSE or Kling-Gupta Efficiency – KGE), and (c) the threshold defining the range on which model calibration is focused (the whole range for Qmin, the top 10% of the discharge data for Q90, and the top 1% of the discharge data for Q99).

3 Response to comments of Anonymous Referee #2

General comment: *This work aims to investigate the influence of using a set of different radar-based QPE and different hydrological models on the uncertainties in simulating the record-breaking July 2021 flood event in Germany. Given the lack of peak flow information (the flood partly destroyed the monitoring systems), the analysis is focused on the probability that the simulated peakflow exceeds the highest historically observed peakflow before the flood. This is a very interesting point of view, given the challenges offered by the prediction of a record breaking flood to both precipitation estimation and hydrological prediction. The work is appropriate for NHESS and its readership.*
The manuscript is broadly well written and well structured. However, there are some specific issues listed below that should be considered before acceptance.

Comment 1: *"Better identifying the main focus of the work. The July 2021 flood in Germany is not only a record-breaking flood. It is a flood that far exceeded previously observed records (the authors could report existing post flood estimates that shows how far the estimated July 2021 peak exceeded the previous records). Of course, existing methods and models for flood forecasting cannot predict these floods well because flood generation processes of large extremes differ from those of smaller, more frequently observed events. Therefore, research aiming precisely to this issue by considering these kind of megafloods is timely and helpful. However, this point is completely ignored in the abstract, and it is elaborated relatively late in the introduction."*

Authors' response: We agree that identifying the main focus of the work is essential. The first two sentences of the abstract were meant to convey this idea. Following the Referee's suggestion, we reinforced the main idea as follows:

"The disastrous July 2021 flooding events made us question the ability of current hydrometeorological tools in providing timely and reliable flood forecasts for unprecedented events. This is an urgent concern since extreme events are increasing due to global warming, and existing methods are usually limited to more frequently observed events with usual flood generation processes."

We would like to stress that our aim is not to provide an exhaustive analysis of the flooding event (such in Mohr et al., 2022), but to focus on how precipitation estimates are uncertain for this event, and how this uncertainty in precipitation estimates compares to that of hydrological models to impact peakflow estimates. Since there are no measurements for the event, we proposed to focus on the probability of exceeding the highest measured peakflow, which is itself a novel way of circumventing this problem. The first paragraph of Section 1.3 of the manuscript identifies the focus of the work and its novelties:

"This study investigated the influence of improved QPE and different representations of hydrological processes on the uncertainties in simulating extreme flooding events. The novelties of our study consist in: (1) using new QPE products from vertical-profile corrected, phase-based observables of C-band and X-band radars, (2) contrasting hydrological modeling approaches (conceptual vs. partial differential equations (PDE)-based model), and (3) proposing an evaluation framework of the hydrometeorological prediction chain for unprecedented extreme events with unavailable discharge measurements. Since no peakflow measurements are available (partly due to destroyed monitoring systems), our analysis focused on the probability that the simulated peakflow exceeds the highest historically observed peakflow. This is relevant because hydrological models are often evaluated based on their ability to detect the probability of flows exceeding catchment-specific, critical thresholds for flood warning applications (Anctil and Ramos, 2017)."

Comment 2: *"The point (L205-2010) made on the different results obtained based on considering raingauges and raingauge-based catchment-scale precipitation estimates is someway misleading. First, it totally ignores the uncertainty in the catchment-scale estimates based on raingauges (and here I urge the authors to consider techniques better than Thiessen for this). Second, this conclusion obviously depends on the set of raingauges considered. If the reference raingauges are those considered for estimating the catchment-scale precipitation, I doubt outcomes may be different. By the way, this conclusion is missed in the conclusion section."*

Authors' response: We acknowledge that our choice of Thiessen polygons to compute the catchment-scale precipitation is subjected to uncertainties and the density of the rain gauge network. For this reason, we included the daily estimates from REGNIE (Rauthe et al., 2013), which is a gridded, high-resolution product that accounts for several attributes of rain gauges in the interpolation process. Note that this product covers only 50% of the catchment area for the Rur at Monschau. In the new version of the manuscript, we added the following:

“Acknowledging the uncertainties that may arise from using Thiessen polygons to compute catchment-scale precipitation depths, we compared these to catchment-scale precipitation estimates from the daily gridded product REGNIE (1-km resolution), which accounts for the position, the height, the exposition and the slope of the gauge stations in the interpolation of the precipitation fields from rain gauges (Rauthe et al., 2013).”

Figure R3 above (Figure 4 of the revised manuscript) shows that the estimates from rain gauges using Thiessen polygons are similar to REGNIE’s, except for the Erft at Bliesheim, where Thiessen method underestimated the total precipitation depth for the 14 July 2021. We can conclude that the Thiessen polygons give reasonable results for our case study.

For the second point considering the conclusion in L205-210 of the original manuscript, we agree with the Referee that this depends on the set of rain gauges considered, but also on the spatial variability of precipitation fields. If the network of rain gauges missed the spatial variability, then the catchment-scale evaluation can be strongly different from the point-scale evaluation. In the revised manuscript, considering new precipitation products with correction of vertical profiles, the conclusions at the point-scale and the catchment-scale were quite similar with respect to the ranking of the different radar-based QPE. Therefore, we changed the lines 205-210 of the original manuscript to:

“Conclusions about the agreement between QPE products and rain gauges are similar when we look at the catchment-scale evaluation. Specifically, QPE based on specific attenuation (A) with corrected vertical profiles for K_{DP} (RAKDP-VPC) outperformed RADOLAN in reproducing estimates from rain gauges (using Thiessen polygons) across the seven catchments (Fig. 4), and reduced relative error from a median of -18 % for RADOLAN to +2 %. With the exception of RAKDP-VPC, radar-based QPE products tended to underestimate catchment-scale precipitation with respect to rain gauges in most cases, confirming the point-scale results (see NMB scores in Fig. 3). However, this comparison underlines the fact that the assessment of QPE products is catchment-dependent. RAKDP-VPC outperformed RADOLAN (with respect to rain gauges) for the catchments drained by the Ahr and the Kyll, whereas they both agreed for the Rur at Monschau. For the catchments drained by the Erft, RAKDP-VPC overestimated precipitation depths with respect to rain gauges, whereas RADOLAN underestimated the total precipitation depth. Finally, using the Thiessen polygon method led to similar catchment-scale precipitation depths compared to the regionalized REGNIE product, except for the Erft at Bliesheim where the Thiessen polygon method underestimated the total precipitation depth with respect to REGNIE.”

In the revised manuscript, we stated that both the point-scale and the catchment-scale evaluations led to similar results, i.e. improved precipitation estimates thanks to better characterization of the vertical profile of radar variables:

“Better characterization of the vertical profiles of radar variables led to significant improvements of radar-based QPE for the extreme event of 14 July 2021 with respect to rain gauges. These improvements were confirmed at both the point scale and the catchment scale.”

Comment 3: *“The point (L254-256) about the causes leading to the strong underestimation (For the 14 July 2021 event, this underestimation may be explained by intense collision-coalescence processes taking place close to the surface..) lacks any ground. I mean: it is likely that collision-coalescence processes may cause those underestimation, but this attribution needs a far better explanation.”*

Authors' response: The use of Micro Rain Radar (MRR) observations showed that all radar variables increased towards the ground which clearly suggests the dominance of collision-coalescence processes below the melting layer for this event (Figure R5a-b). Although collision-coalescence processes alone do not change the precipitation flux within a column, the retrieved rain rate still increases toward the surface (Figure R5c). By examining the contributions of drizzle (D < 0.5 mm) and raindrops (2 mm < D < 4 mm) to the DSDs, the former shows a secondary peak at 1 km height followed by a rapid decrease downwards, while the number of raindrops constantly increases toward the ground below the ML (Figure R6). Accordingly, the increasing rain towards the surface can be explained by the transformation of water vapor into droplets above 1 km height and its transformation into rainwater via warm-rain processes below (Chen et al., Submitted). In addition, accounting for these measurements and correcting the vertical profiles of Z and K_{DP} reduced the errors of the radar-based QPE with respect to rain gauges, as can be seen in Figure R3.

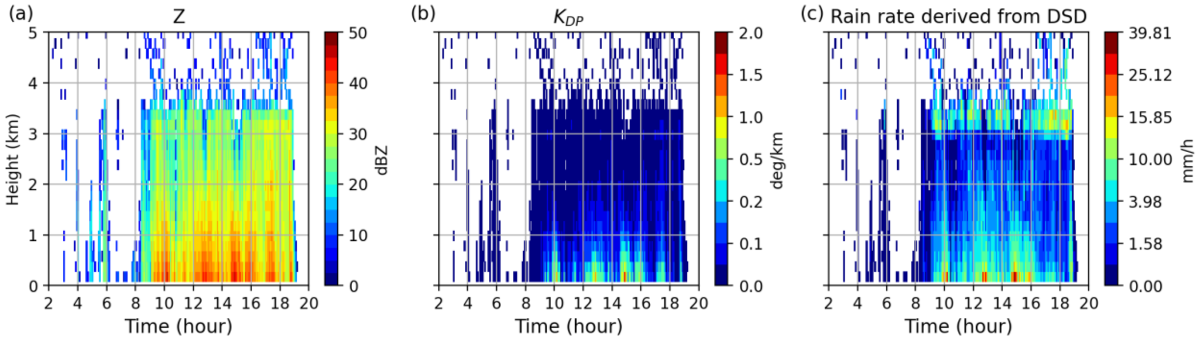

Figure R5: Retrieved profiles based on the MRR measurements from University of Bonn on 14 July 2021, including (a) reflectivity Z, (b) specific differential phase KDP, and (c) rain rates derived from the retrieved raindrop size distributions (DSD).

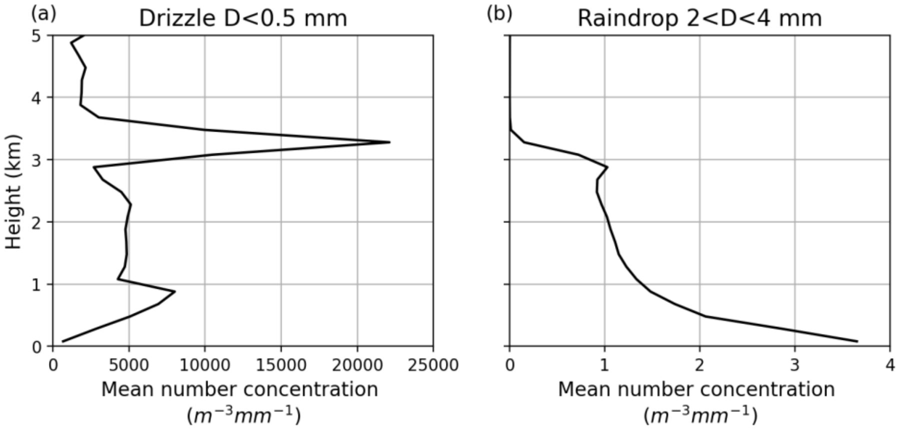

Figure R6: Mean number concentration profiles of (a) drizzle with diameter D < 0.5 mm, and (b) raindrops with 2 mm < D < 4 mm calculated from the DSDs retrieved from the observations of the MRR located at University of Bonn.

Comment 4: *"Information on how antecedent conditions were computed, and about the accuracy of these estimates, is missing, in spite of the critical role this information have on the sensitivity of the model to QPE error."*

Authors' response: Prior to 14 July 2021, both GR4H and ParFlowCLM were run continuously starting from 2006-2007 for all catchments. This allowed for exploiting all the record periods to yield the best estimate of model initial conditions prior to the event. We now mentioned this in the revised manuscript when we present how QPEs are evaluated using hydrological models. The following statement was added to Section 3.4:

"Second, we examined the effect of QPE on the frequency of exceeding the highest historically observed peakflow for each catchment (Table 1) by simulated peakflows. Both GR4H and ParFlowCLM were initialized using a long spin-up period starting from 2006 for GR4H and 2007 for ParFlowCLM. This allowed for exploiting the whole available record period of climatic forcing to yield the best estimates of antecedent soil moisture conditions. Then, each radar-based QPE was used as input to both models to obtain twelve peakflow simulations from GR4H and four peakflow simulations from ParFlowCLM. These peakflows are compared with the highest historically measured peakflow."

Comment 5: *"The parameter uncertainty of ParFlowCLM is strongly underestimated when focusing only on Manning values, as the authors did. At least they should do a better job considering uncertainty in the information about soil properties (lets only think to soil depth)."*

Authors' response: We agree with the Referee that the uncertainty of ParFlowCLM is underestimated without looking at other parameters, such as soil properties. We stated this in the Discussion section as one of the limitations of our study:

"The large uncertainty due to the Manning's coefficient is perhaps accentuated by the nature of the relationship between the coefficient and the discharge, but it is still here a lower bound since uncertainty to other parameters (hydraulic conductivity, van Genuchten parameters) was not included."

and

"Fourth, the accuracy of the parameter estimation in our study could be improved by investigating the uncertainty related to other distributed parameters (such as hydraulic conductivity; Poméon et al., 2020), or using hourly discharge streamflows for the GR4H calibration."

However, our objective was not to give an exhaustive quantification of the effect of parameter uncertainty on ParFlowCLM simulations. The large uncertainties caused by Manning's coefficient and QPE inputs illustrate how peakflow simulations are uncertain, let alone the contribution of other parameters. In addition, there are some computational limitations for us to do such an exercise. With the objective of having a regional scale model for flood forecasting, ParFlowCLM is currently implemented at the scale of Central Europe with $4*10^6$ grids and 15 soil layers, yielding a total of $6*10^7$ grids. We chose Manning coefficient as the peakflows are highly sensitive to this parameter and it is usually the focus in extreme flooding events studies (Lumbroso and Gaume, 2012).

In a very similar study with more focus on parameter uncertainty, Poméon et al. (2020) included uncertainty in the hydraulic conductivity on the simulations of ParFlow for flash floods events in several German catchments. However, they only adopted a uniformly distributed values of each parameter.

Comment 6: *"The use of English in the paper, while of a reasonably high standard, contains many idiosyncrasies, like the sentence: "The QPE impacted both GR4H and ParFlowCLM simulations", where 'Errors in the QPE impacted both…' is more likely."*

Authors' response: We corrected the sentence in question and checked for other idiosyncrasies in the revised manuscript.

4 Cited References

Anctil, F., Ramos, M.-H., 2017. Verification Metrics for Hydrological Ensemble Forecasts, in: Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H.L., Schaake, J.C. (Eds.), Handbook of Hydrometeorological Ensemble Forecasting. Springer, Berlin, Heidelberg, pp. 1–30. https://doi.org/10.1007/978-3-642-40457-3_3-1

Asante, K.O., Artan, G.A., Pervez, M.S., Bandaragoda, C., Verdin, J.P., 2008. Technical Manual for the Geospatial Stream Flow Model (GeoSFM) (USGS Numbered Series No. 2007–1441), Technical Manual for the Geospatial Stream Flow Model (GeoSFM), Open-File Report. Geological Survey (U.S.). https://doi.org/10.3133/ofr20071441

BMI, 2022. Bericht zur Hochwasserkatastrophe 2021: Katastrophenhilfe, Wiederaufbau und Evaluierungsprozesse, Bundesministerium des Innern und für Heimat, Berlin, Germany.

Braud, I., Roux, H., Anquetin, S., Maubourguet, M.-M., Manus, C., Viallet, P., Dartus, D., 2010. The use of distributed hydrological models for the Gard 2002 flash flood event: Analysis of associated hydrological processes. J. Hydrol., Flash Floods: Observations and Analysis of Hydrometeorological Controls 394, 162–181. https://doi.org/10.1016/j.jhydrol.2010.03.033

Chen, H., Cifelli, R., White, A., 2020. Improving Operational Radar Rainfall Estimates Using Profiler Observations Over Complex Terrain in Northern California. IEEE Trans. Geosci. Remote Sens. 58, 1821–1832. https://doi.org/10.1109/TGRS.2019.2949214

Chen, J.-Y., Trömel, S., Ryzhkov, A., Simmer, C., 2021. Assessing the Benefits of Specific Attenuation for Quantitative Precipitation Estimation with a C-Band Radar Network. J. Hydrometeorol. 22, 2617–2631. https://doi.org/10.1175/JHM-D-20-0299.1

Chen, J.-Y., Reinoso-Rondinel, R., Trömel, S., Simmer, C., and Ryzhkov, A., Submitted. A radar-based quantitative precipitation estimation algorithm to overcome the impact of vertical gradients of warm-rain precipitation: the flood in western Germany on 14 July 2021.

Dai, Q., Han, D., 2014. Exploration of discrepancy between radar and gauge rainfall estimates driven by wind fields. Water Resour. Res. 50, 8571–8588. https://doi.org/10.1002/2014WR015794

Deutsche Welle, 2021. German floods: Climate change made heavy rains in Europe more likely. Available online: https://www.dw.com/en/german-floods-climate-change/a-58959677 (accessed 2.16.22).

GDV, 2021. 2021 teuerstes Naturgefahrenjahr für die Versicherer, Gesamtverband der Deutschen Versicherungswirtschaft (GDV), Berlin, Germany. Available online: https://www.gdv.de/de/medien/aktuell/2021-teuerstes-naturgefahrenjahr-fuer-die-versicherer-74092 (accessed 7.5.22).

Koks, E., Van Ginkel, K., Van Marle, M., Lemnitzer, A., 2021. Brief Communication: Critical Infrastructure impacts of the 2021 mid-July western European flood event. Nat. Hazards Earth Syst. Sci. Discuss. 1–11. https://doi.org/10.5194/nhess-2021-394

Kreienkamp, F., Philip, S.Y., Tradowsky, J.S., Kew, S.F., Lorenz, P., Arrighi, J., Belleflamme, A., Bettmann, T., Caluwaerts, S., Chan, S.C., Ciavarella, A., De Cruz, L., de Vries, H., Demuth, N., Ferrone, A., Fischer, E.M., Fowler, H.J., Goergen, K., Heinrich, D., Henrichs, Y., Lenderink, G., Kaspar, F., Nilson, E., Otto, F.E.L., Ragone, F., Seneviratne, S.I., Singh, R.K., Skålevåg, A., Termonia, P., Thalheimer, L., van Aalst, M., Van den Bergh, J., Van de Vyver, H., Vannitsem, S., van Oldenborgh, G.J., Van Schaeybroeck, B., Vautard, R., Vonk, D., Wanders, N., 2021. Rapid attribution of heavy rainfall events leading to the severe flooding in Western Europe during July 2021. World Weather Attribution (WWA).

Lumbroso, D., Gaume, E., 2012. Reducing the uncertainty in indirect estimates of extreme flash flood discharges. J. Hydrol. 414–415, 16–30. https://doi.org/10.1016/j.jhydrol.2011.08.048

Mohr, S., Ehret, U., Kunz, M., Ludwig, P., Caldas-Alvarez, A., Daniell, J.E., Ehmele, F., Feldmann, H., Franca, M.J., Gattke, C., Hundhausen, M., Knippertz, P., Küpfer, K., Mühr, B., Pinto, J.G., Quinting, J., Schäfer, A.M., Scheibel, M., Seidel, F., Wisotzky, C., 2022. A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe. Part 1: Event description and analysis. Nat. Hazards Earth Syst. Sci. Discuss. 1–44. https://doi.org/10.5194/nhess-2022-137

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V., Michel, C., 2006. Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. J. Hydrol., The model parameter estimation experiment 320, 62–83. https://doi.org/10.1016/j.jhydrol.2005.07.016

Poméon, T., Wagner, N., Furusho, C., Kollet, S., Reinoso-Rondinel, R., 2020. Performance of a PDE-Based Hydrologic Model in a Flash Flood Modeling Framework in Sparsely-Gauged Catchments. Water 12, 2157. https://doi.org/10.3390/w12082157

Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., Gratzki, A., 2013. A Central European precipitation climatology – Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS). Meteorol. Z. 235–256. https://doi.org/10.1127/0941-2948/2013/0436

Schalge, B., Haefliger, V., Kollet, S., Simmer, C., 2019. Improvement of surface run-off in the hydrological model ParFlow by a scale-consistent river parameterization. Hydrol. Process. 33, 2006–2019. https://doi.org/10.1002/hyp.13448

Schleiss, M., Olsson, J., Berg, P., Niemi, T., Kokkonen, T., Thorndahl, S., Nielsen, R., Ellerbæk Nielsen, J., Bozhinova, D., Pulkkinen, S., 2020. The accuracy of weather radar in heavy rain: a comparative study for Denmark, the Netherlands, Finland and Sweden. Hydrol. Earth Syst. Sci. 24, 3157–3188. https://doi.org/10.5194/hess-24-3157-2020

Schröter, K., Kunz, M., Elmer, F., Mühr, B., Merz, B., 2015. What made the June 2013 flood in Germany an exceptional event? A hydro-meteorological evaluation. Hydrol. Earth Syst. Sci. 19, 309–327. https://doi.org/10.5194/hess-19-309-2015