Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-414-RC1, 2022

Comment on hess-2021-414

Anonymous Referee #1

Referee comment on "Signature and sensitivity-based comparison of conceptual and process oriented models, GR4H, MARINE and SMASH, on French Mediterranean flash floods" by Abubakar Haruna et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-414-RC1, 2022

General Comments:

Abubakar et al. are presenting their work on comparing three divers hydrological models regarding their capability to model flash floods. The comparison of the lumped, conceptual GR4H model, the distributed, conceptual SMASH model and the process-oriented, event-based MARINE model is based on a sensitivity analysis for each model, a performance comparison and a soil moisture comparison of the model states with the modelled SAFRAN-ISBA-MODCOU (SIM) soil moisture predictions.

In my perception, the manuscript is currently lacking a concise story line and presentation of the study goals and relevant outcomes. There is a lot of information the authors try to convey which makes it hard to grasp the core of the study. A few more tables should help to achieve a clearer presentation of all the very individual characteristics of the compared models and their setups. This will be necessary to understand and conduct a final evaluation of their acquired results. I feel that not all information presented are valuable for the goal of the paper and could be moved to the appendix to make the manuscript an easier read. While the results are elaborately presented, I am missing some depth in the analysis/discussion and a substantial conclusion. I find this study very specific (case study) with only limited insights into the question they set out to answer (which model brings which benefits/challenges when modelling flash floods?). I even feel the authors are missing a chance to generalize their results to offer some insights or advice on modelling flash floods with one or all of the three used models.

Thus, I advise on a thorough and extensive review with several iterations between the co-authors (1st author & supervisors) before resubmitting this work in a format that makes it easier for the reader to identify the goal and relevant outcomes of this study.

Specific Comments:

Structure: The formulated goal of the paper ("the objective was to understand how [the 3 models] simulate catchment's hydrological behavior, the differences in terms of their simulated discharge, the soil moisture, and how these can help to improve the relevance of the models") should be matched with the specific analyses that were conducted to answer these questions and with the conclusions that result out of these analyses. These connections should then be formulated very clearly in the paper and parts that don't contribute to answering the questions should be removed, or the part they play be made very clear.

Introduction: I would recommend adding a paragraph on previous model comparison studies and their challenges in order to better evaluate how comparable the results acquired in this study actually are or at least what the challenges might be. Especially since the authors refer to their work as an intercomparison study (Line 158). Some references to start out with might be:

Refsgaard, J. C., & Knudsen, J. (1996). Operational Validation and Intercomparison of Different Types of Hydrological Models. Water Resources Research, 32(7), 2189–2202. https://doi.org/10.1029/96WR00896

Butts, M. B., Payne, J. T., Kristensen, M., & Madsen, H. (2004). An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. Journal of Hydrology, 298(1–4), 242–266. https://doi.org/10.1016/j.jhydrol.2004.03.042

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research, 47(9), 1–16. https://doi.org/10.1029/2010WR009827

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resources Research, 47(11), 1–13. https://doi.org/10.1029/2010WR010174

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., & Zappa, M. (2015). Does model performance improve with complexity? A case study with three hydrological models. Journal of Hydrology, 523, 147–159. https://doi.org/10.1016/j.jhydrol.2015.01.044

Models, Tools, and Data:

Regarding the SIM model please describe why it is feasible to use this as a benchmark for the modelled soil moisture (Is model VS model really a good idea? Why not satellite data as benchmark?) Please describe the differences between SIM1 and SIM2 and why you chose SIM 1 to initialize MARINE and SIM2 as the benchmark. Why is it okay to compare to the daily SIM product when the models run on hourly time steps?

The authors compare very diverse model structures and I believe it would be helpful to the reader to see a somewhat aggregated version of the main differences of the model setups and their data requirements. E.g. it might be helpful to see the differences in the input data in a table format. As there is a large difference between the input for the lumped GR4H (P, PET, Q) and the distributed, process-based MARINE (landuse, soilmaps etc). The table could also be used to show the main differences of the models itself (lumped vs process-based etc., delta t and delta x etc.) in one glance. Please include delta t, delta x and also the number of calibrated parameters for all models of the study. This is a long paper and you want to make it as easy as possible for the reader so they don't lose interest.

Why is delta x and delta t different for MARINE and how do you think this influences your results? Please justify why different delta x and delta t are used for the models.

Please also provide a table for a quick comparison of the 2 catchments. It should include the same information for both catchments. This is currently not the case in the text. E.g. average slope is only given for Gardon. Also "a lot of intense rainfalls" could be backed up by some climate data statistics in the table. At least mention the average precipitation input for each catchment.

Input Data: Is there any information on the quality of the radar observation reanalysis data? Radar data is known to underestimate especially heavy rainfalls of short duration which are important for flash flood generation. Could this be a problem for the current study? Also it seems the precipitation input of the SIM model is different to the tested models. How could this influence the results, especially as it is used as a benchmark for soil moisture comparison. What are the implications?

Sensitivity Analysis:

While I agree the sensitivity analysis is necessary for better understanding the model results later on I currently fail to see how we can learn anything from the comparison of the model parameters here as they are different for each model (are differently implemented in the model structures). You might want to state which parameters are comparable and why you believe they are comparable. Please clarify in the manuscript.

The tables with the sensitivity ranks of the parameters (Table 6 + 7) are currently given

without much description or discussion before the section starts. They should be included in the subsections when they are referred to or removed to the Appendix.

If a comparison of the models is the goal I find it essential to analyze the event based sensitivity for SMASH and GR4J as was done for MARINE as well. Otherwise, it's really hard to attempt a fair comparison here.

Table 3 - Why is the number of classes so much higher for Gardon than for Ardeche?

Line 285 – Why only 5000 runs if the other models had 10000 for their SA? Please justify.

Currently the SA seems to tell us which parameters are sensitive in which model/catchment and we see some expected differences. Due to the small sample we can't conclude anything general though and thus I currently fail to see the benefit for comparing the models with regard to flash flood modelling. Especially since the comparability of the sensitivities lacks some justification. As the paper is very long, I would advise to move the SA to the Appendix and focus on the actual comparison of the models. Section 4.1.4 may be kept as the main outcome of the SA but doesn't require all the plots and descriptions in the main text. Unless the authors can clarify the benefit and insights we gain from the results.

Calibration:

The section on the response surface and functioning points is interesting but does not seem to immediately add to the point of the paper. Consider moving to the appendix.

How were the events separated into 2 periods and how is this justified? There are only 2 events in one period one for Gardone? Why?

Is Table 8 showing the mean from the unmasked calibration and then the STD from the masked calibration? Please clarify. Are we only looking at the masked calibration results for the rest of the paper? Please indicate more precisely.

Figure 11 does not seem to be mentioned or described anywhere even though it clearly indicates that GR4H and SMASH have a problem in modelling the high discharges for Ardeche, which is quite relevant for this study. Please comment on this in the manuscript.

Section 4.4.3 I don't understand what was done here and what Figure 18 is supposed to tell us. What is the y-axis "change in available storage"? Is this in percent? What does Figure 11 tell us and why is it relevant?

I feel it is a rather large drawback of this study that all models have a rather different calibration routine. What's the take of the authors on this?

Conclusions:

There are no general conclusions after the description of the results they acquired. Are there cases when SMASH is the better choice or when MARINE is? What do we actually learn from these results that is of relevance to people attempting future work with these models in a flash flood context?

The authors conclude from their results "The difference in the model performances could stem from differences in the levels of complexity of the models, the processes described and the constrains of the models, and thus highlights the need for future improvements in the models and calibration methods." – 1.) It is very much expected to have differences in performance when testing 3 so very divers models – so what did your study contribute? Please describe either HOW they perform differently, WHY they perform differently or what would be the new insight on that they actually perform differently. 2.) Why does that highlight the need for improvement of the models? Which weaknesses were identified that need to be improved? Should a lumped conceptual and distributed process-based model perform identical?

The authors state in line 465f that "MARINE has its efficiency in validation decreased by around 25%, while SMASH and GR4H have a decrease of 5.2% and 4.8% respectively." The conclusions read rather as if MARINE comes off pretty well. I feel that the results need to be contextualized a little more and general advice be given. What did you learn from your model comparison study that may be of benefit for a reader?

There are a lot of unnecessary relative terms in this paper. Try to be more precise! (e.g. line 405 " "somehow similar conclusions" or line 440 "relative robustness")

Minor/Technical Comments:

There were a few terms that felt unfamiliar to me. E.G "flow operator" (especially used in 2.1) for process description/algorithm etc. I would advise to use more commonly used terms in the literature such as process description/algorithm. Also I wouldn't use "numerical experiments" for a normal methodology consisting of calibration-validation and

model comparison.

Line 9f - If the catchment names are used it should be mentioned that these are catchment names somehow. Otherwise, it's a little confusing.

Line 29 – potentially use "internal states and fluxes"

Line 41 – another suitable citation in that context would be Bouaziz et al. (2021)

Line 85 – potentially use "as well as" instead of "and next"

Line 91 – please add reference where in the paper this is analyzed

Line 122 – is there a reference for the "Michel calibration algorithm" you can add?

Line 125 – it looks as if the authors missed to describe which warm up period they are using to avoid the effects on the results they mentioned

Line 164 – is detailed after – please refer to the location of the paper where this is detailed

Line 166 – MARINE is an event-based, physically based, […]; add based

Line 184 - is detailed after – please refer to the location of the paper where this is detailed

Line 259 – KS test: abbreviation is only defined in appendix, so please do so here as well

Line 268 – Section B should be Appendix B

Line 292 – as shown after – please refer to the location of the paper where this is detailed

Line 294 – “dividing the data into two”: two what? Please specify by adding time periods or similar.

Line 295 – The sentence is hard to read. Maybe change to something along the lines of: A “Time series of 13 years at hourly time step is considered and “ divided into “ two sub-periods of 7 years each for calibration and validation. The Period 1 is defined from …

Line 296 – Why are Period 1 and 2 overlapping by a year?

Line 300 – does this mean you concatenated the single events to form a consecutive time period of events which you then split for calibration? If so, this is not yet clear from your text. Please clarify.

Line 304f – “These experiments are designed to compare 3 models at flash flood modelling […]” should appear much more prominently and earlier in this paper!

Line 310 – which aggregation are you comparing? The Catchment size aggregation? Please clearify. Also this should already be mentioned around line 210, so the questions about the different resolutions and their comparability doesn’t arise.

Line 315 – how do the events compare relating to their number of peaks, gradients in limbs and precipitation patterns? It’s not specified in table 5. Please specify!

Line 337 – gives the delay in hours? Please specify.

Line 338 - why is this more rigorous in terms of safety? Please elaborate.

Line 421 – “a few parameters are stuck at the bound”. For SMASH it seem to be most parameters that are stuck at the bound. What does this tell us?

Line 483 – For which time period are the signatures calculated? Only for the event time period? Please specify.

Figure 2 – I find the map could use a scale and north arrow to really allow for the term map.

Figure 3 – Maybe add a line for your divide between behavioral and non-behavioral runs at NSE=0.7 in 1st row of plots

Figure 14 – the black cross is missing in the legend. Why does the header for CR look different?