Reply to Referee #4 Anna E. Sikorska-Senoner:

Anna E. Sikorska-Senoner *(AS): Summary and Recommendation:* *This paper proposed an adaptive modelling as an alternative to a distributed model for representing spatial variability of the catchment and forcing input (precipitation). Such an adaptive modelling should be able to run faster than a distributed model but should provide a similar model performance as its fully distributed version. The manuscript is generally well written and it is easy to follow. The idea of a spatially adaptive model that dynamically adjusts its spatial structure during runtime is indeed very interesting and has a potential for being applied in many (hydrologic) modelling approaches. Yet, I have few major issues that should be addressed before considering this manuscript for a publication in HESS. Thus, I recommend a major revision.*

Ralf Loritz (RL): We would like to thank Anna E. Sikorska-Senoner for her comments and the time she invested to review our Manuscript (MS). We hope that after the discussion as well as after we have revisited our MS all open issues she raises can be clarified.

Comments:

1. AS: *The adaptive model (model c) is tested here only on two rainfall events, which I see as the major weakness of this manuscript. As the strength of this approach should lie in the possibility to apply it to a continuous modelling and not to an event-based modelling. Thus, I think it would be important to demonstrate how the model c works on continuous time series. As this is missing in the current manuscript, we still do not know at the end whether it is a good or a bad option to be used.*

RL: *Model a* and *model c* simulate close to identical hydrographs at the end of both rainfall events when *model c* represents the Colpach catchment again by a single hillslope model. This is also true for the soil moisture distributions, which we did not show in the current MS. This means that the information about the spatial organization of a past rainfall event have already been dissipated closely after the spatial adaptive *model c* represents the catchment by a single hillslope. In other words, there is no difference between *model a* and *c* after this point and we would learn not much by letting *model c* run continuously until the next rainfall event.

Furthermore, as rainfall event II is characterized by one of the longest rainfall durations in summer and event I by the highest intensity and third highest spatial variability we see no reason to expect that the spatial adaptive model will fail at other summer rainfall-runoff events. We think that it is not the length of the simulation that matters here but the fraction of the visited state space (or in other words if your training data set is representative). The latter means that we do not assume that the catchment and the model which represents it will function differently at the other untested events. This is underpinned by the fact that also the 42 model elements in the distributed *model b* do not drift apart. The latter reflects the highest complexity *model c* could reach.

However, we agree that we did not well justify the selection of the two events. Following your comment, we will hence plot the soil moisture distribution of *model a* and *c* for event I and II at the time step when the catchment is again represented by a single hillslope. This will show that there is no difference between the *spatially aggregated model a* and *model c* already shortly after the rainfall stopped. Furthermore, will we improve our discussion regarding the choice of our two rainfall-runoff events. Again, we would like to thank AS for her comment.

2. AS: *The performance metrics of the calibrated (tuned) models should be provided so that the model ability to predict rainfall events could be assessed*

RL: The reference model is the only model which was manual tuned to match the seasonal water balance of the Colpach. This procedure is described in detail in *Loritz et al. (2017)* and in the current MS in section 3.1. The KGE value of the reference model is reported in table 1. We will furthermore add the three components of the KGE as discussed with Wouter Knoben to the appendix.

3. AS: *A model set-up between the model a and b could be very didactical, i.e. having a structure as the model b but using the precipitation input as the model a (the same for each grid cell)*

RL: The only difference between *model a* and *b* is the precipitation input. Running *model b* with the input of *model a* would mean to produce the same hydrograph as *model a* 42 times.

4. AS: *It is not quite clear how the switch between different model setups (i.e. the number of model run in the model c) affect the setup of initial conditions for next runs, which is important to be considered for continuous model simulations but also for simulations of events. More details should be provided on that.*

RL: Please see the discussion above.

5. AS: *It would be also very didactical to see the comparison of the precipitation records from the ground station with the precipitation fields obtained from the gridded data. This is never done in the manuscript and no reason for not doing that is given.*

RL: Agreed. We will add the precipitation from the ground station to Fig. 2b.

6. AS: *The (rather) poor model performance of all tested models' set-ups for two selected events requires some discussion. It appears that none of these model can really capture the dynamics of these two events even if using the distributed model and the distributed rainfall information (with KGE<0.3). Hence, it is even more important to verify the model performance (pkt. 2). An addition of other metrics that focus entirely on the flood event such as peak or time to peak could be here very informative. Given a rather poor models' performance, it is difficult to justify the need of developing the adaptive model based on the distributed model if the latter does not provide acceptable simulation results.*

RL: Respectfully, the focus of this MS is not to minimize residuals between an observed quantity and a model simulation. The main goal of this study is to introduce an approach with the goal to setup a spatially adaptive model and equally important underpin this approach with a physical meaning. Furthermore, would we like to highlight that we a) discuss the model performance and how it could be improved on page 21 line 26 to 28 and b) would like to reiterate that the *reference model*, which is the basis of this study, was tested against a series of different variables (sap flow, discharge, water balance, soil moisture, etc.), at different hydrological years, in an additional sub-basins as well as mainly setup based on field observations. We believe that such an evaluation and model-building process underpins the quality as well as the ability of a model to mimic the hydrological dynamic of a landscape sufficiently and maybe even more than adding another performance metric.

As the model was setup to simulate the seasonal water balance we think that the annual performance is quite "good" and we are not surprised that if we zoom into a single event that we loss performance. Furthermore would we like to highlight that the performance metric, which is important here is the KGE between *model b* and *c*, which is 0.98. To improve the interpretability of our model scenarios we will add a second table with the three components of the KGE to the appendix as discussed with Wouter Knoben. Furthermore, will we clearly state that the goal of this study is not to perform a best as possible streamflow simulation.

7. AS: *A fair comparison of all presented models should involve the same metrics, i.e. computation over the same time at a continuous time scale. In this study, different model setups are compared at different scales that makes it difficult to get an overview of their performance.*

RL: We compare and discuss the connection between *model b* and *model c* only for the two events as well as for the corresponding summer period. Respectfully, we do not think that the comparison is unfair.

Detailed comments

1. AS: *Abstract: 'a mesoscale catchment'; a 20-km^2 catchment appears rather small tome than meso-scale.*

RL: Meso-scale: 5 to 1000 km^2, we refer here to the work of *Zehe et al. (2014)* and *Dooge, (1986).*

2. AS: *Abstract: 'three hydrological models', the model is actually the same but different set-ups are used that span from the averaged model until the distributed model. Please clarify that here.*

RL: This depends on your the definition of the term "model". However, I agree and we will use the term model setups here.

3. AS: *L. 20-21 p. 3: It is not always possible and justified to switch from a continuous model to an event based model. Hence, continuous modelling is often required in many applications.*

RL: Agreed. Could you provide a reference here?

4. AS: *L. 19 p. 4: The tested catchment appears rather small to me. How do you define the cut here for a small/meso-scale catchment?*

RL: We refer here to the work of *Zehe et al. (2014)* and *Dooge, (1986)* which is around 5 to 250 km^2. The definition of organized complexity is that such systems are too complex that we can tread them exclusively in a mechanistic manner but too organized that we can represent them in a purely statistical manner.

5. AS: L. 7-9 p. 5: consider restructuring this sentence.

RL: Thank you. We will consider rephrasing it.

6. AS: *L. 11-13 p. 7: could you add the location of these meteorological stations to the map in fig 1?*

RL: The station "Useldange" is too far away to be added to the map. But its location is provided in the corresponding reference. We will add this information.

7. AS: *L. 15 p. 7: it should be 'and measures...'*

RL: Thank you. Changed.

8. AS: *L. 16 p.7: is there any weighting applied here and what kind of?*

RL: No weighting applied.

9. AS: *L. 28 p. 7: could the locations of radars be also placed in the fig. 1?*

RL: No, they are too far away. However, their location is displayed in the reference provided by *Neuper and Ehret, (2019).* We will add this information.

10. AS: *L. 12-14 p. 8: Why do you compare these values with literature and not with the ground station records from your catchment? Is there any reason that you are not using the precipitation records from the ground station?*

RL: These values represent the climatic averages of the area. We have only data for about 10 to 15 years.

11. AS: *Fig. 2: One would expect that the radar values would be compared with the ground station values as a corresponding mean (Fig. 2b). Could you add these values to the figure?*

RL: Agreed. Will be added.

12. AS: *L. 10-17 p. 9 till 21 p. 10: I am not quite sure if this text is really helpful. After reading these lines, we still do not know how the reference model and other models look like. Maybe you could merge these lines with the sections 3.1-3.3.*

RL: We will restructure section 3. Please see the discussion with Daniel Wright.

13. AS: *L. 19-20 p. 10: I would say that the main goal is to test or verify whether similar model performance can be achieved with the adaptive model as compared to the model b. However, by the comparison that you did we still do not know the answer to this question as the comparison is done only based on two pre-selected events both having rather a poor model performance. Please comment on that and also state why these events were chosen for the comparison (and not others)?*

RL: Please, see the discussion above.

14. AS: *L. 21-22 p. 10: Why do you compare the adaptive model with model b using only these two events and not the entire simulation period? In my opinion, the greatest potential of the adaptive modelling lies in continuous modelling and not in the event-based.*

RL: Please, see the discussion above.

15. AS: *L. 31 p. 10 – l. 1 p. 11: why do you test the model only based on the annual assessment and not on hourly simulations? It is quite surprising because you use the model for assessing the model performance at an event-based scale in the second step, i.e. when comparing different models. I think it is important to report here how the model behaves at an hourly time scale so that one knows what can be expected from the model.*

RL: I am not sure if I have understood that comment correctly. But we tested our models by comparing hourly simulations with hourly observations for one hydrological year.

16. AS: *L. 3 p. 11: Which metrics were used here for assessing that the model performance agreed well with the dynamics of observed values? Can you give some more details on that?*

RL: We use the Spearman rank correlation, the Nash-Sutcliff eff. and the KGE. We refer to the study of *Loritz et al., (2017).*

17. AS: *L. 7-8 p. 11: It is not surprising that the model performs poorly at time series scale if it was evaluated only on an annual basis. Some insights should be given here; why was the model tested at an annual basis if its intention is to predict events?*

RL: Respectfully, the goal of this study is not to perform a best possible streamflow simulation with regards to minimize residuals. If this would be the case we would have picked a more data driven approach.

18. AS: *L. 3 p. 12: similar to what?*

RL: They are the same in all models.

19. AS: *L. 12. p. 12: the model analysis should go after the introduction of all models.*

RL: Agreed. We will restructure section 3.

20. AS: *L. 16-19 p. 12: an additional model between the models a and b would be here very useful, i.e. a model that has a structure as the model b but uses precipitation as in the model a (so it uses the same precipitation for each grid cell). This inclusion could nicely show the added value (or no value) of including a spatial distribution of i) the model and ii) of the precipitation input data.*

RL: Please, see the discussion above.

21. AS: *L. 8-10 p. 13: why exactly? In my opinion, the strength of this approach lies in the possibility to apply it to a continuous modelling and not to an event-based modelling. Thus, I think it would be nice to demonstrate how the model works on continuous time series in terms of the model performance and computational efforts. As such a test is missing in the current manuscript, we still do not know at the end whether it is a good or a bad option to use the adaptive modelling approach... Based on the two events selected, we cannot say much about the value of the adaptive approach as the model performance remains poor for these events (as seen from the Table 1 and fig. 7). If the full simulation is not possible(could you give more details why exactly?), already a simple test with shorter but continuous time series of few months or weeks could provide some more insights on how this approach is really working.*

RL: Please, see the discussion above.

22. AS: *Tab. 1: as the initial idea is to improve the model performance for rainfall events, it appears from the table that the model c and model b have still rather poor model performance for the event I and II. In addition, all models perform poor for these events. Yet, an inclusion of the spatial variability does not improve much the model performance that is still not so good. Thus, it calls a question of the need of such an adaptive inclusion to this spatially distributed model which performance is rather low.... Could you comment on that? A decomposition of KGE into its components would bring more insights on the models' behaviour.*

RL: We will add the three components of the KGE to the appendix.

23. AS: *L. 9-10 p. 15: How many grid cells need to have a difference higher than this threshold to use the model c?*

RL: One.

24. AS: *L. 30-31 p. 15 fig.3: is the re-arranged model running with the same initial conditions of the original model or how do you decide on these initial conditions if you want to increase or decrease the number of M in the subsequent time intervals particularly if a continuous simulation is performed?*

RL: We aggregate their states.

25. AS: *L. 4-6 p. 18: for a fair comparison of different models, you should use the same metrics and the same time periods for evaluation. It is not clear why this is not the case here.*

RL: See discussion above.

26. AS: *Fig. 4: could you add simulations with the model a?*

RL: If we add all simulations the figure is hard to read. However, we will consider your comment when we revisit your MS.

27. AS: *L 8. P. 20: The reference Knoben et al. (2019) is missing in the literature list.*

RL: Thank you we will add this reference.

28. AS: *L. 6-7 p. 21: the performance of KGE below 0 is still rather very poor, which re-quires some further explanations. According to Knoben et al. (2019), simulations can be considered as behavioral if KGE>0.3 (with KGE≈−0.41 for a mean flow benchmark).*

RL: See comments above.

29. AS: *Table 1: the performances (KGE) of all models is rather poor for the events here selected (with KGE between -0.41 and 0.29). As already the model b (distributed)cannot simulate the two events in a*

good way (as also seen from the fig.7), why would you spend time on developing the adaptive model based on the model b instead if improving the model b or testing different models here? Could you comment or justify that?

RL: See comments above.

30. AS: *Fig. 7: the simulations with the model a and the reference model should also be added here. Moreover, for both events, all models largely underestimate the events. Could you comment on that?*

RL: In fig. 7 we focus on the comparison of *model b* and *c*.

31. AS: *L 10 p. 26 – l. 2 p. 27: do you have any idea where this large underestimation may come from and how it could be improved?*

RL: Discussed in the MS (page 21 line 26 to 28).

32. AS: *Discussion: I missed some recommendations for other works. When and how would such an adaptive modelling be recommended? How one can set up the adaptive process? And why it is really needed to implement such an adaptive modelling?*

RL: Thank you for this comment. We will revisit the discussion of the MS in this regards.