Reply to Referee #3 Wouter Knoben:

Wouter Knoben *(WK): Summary and Recommendation:* *The authors develop and test a hydrological model that is able to change its spatial complexity in time. In its most simple state, the model represents the Colpach catchment in Luxembourg as a single representative hillslope. In its most complex state, the model would be able to use 42 hillslope elements to simulate the catchment's response to extremely spatially variable rainfall inputs. The model adds hillslope elements based on the spatial complexity of incoming precipitation and removes hillslope elements based on the change of runoff over time. Both processes use a threshold to decide when upscaling or downscaling the model is needed or possible. The authors show that the adaptive model reaches the same KGE scores as a fully distributed model that uses 42 hillslope elements all the time, while the adaptive model needs 16 representative hillslopes at most. This is shown for two short-duration event that occurred during summer.*

I have read this paper with much interest and found it generally easy to read and understand. As models grow more complex, computation times go up and studies such as this could open up great opportunities to reduce computation costs by avoiding redundancy in model calculations. However, I have some questions about the tests and metric the authors use to show that the adaptive model is as good as the fully distributed one. These are outlined below. I've provided additional requests for clarification in the line-by-line comments in the hopes that these are helpful.

Ralf Loritz (RL): We would like to thank Wouter Knoben for the interesting discussion on our Manuscript (MS). WK raises a couple of important and well-thought comments and we hope that after this discussion as well as after we have revisited our MS all open issues can be clarified.

Comments:

1. WK: *My main concern is the choice of using dQ/dt to reduce the number of model elements. Using the change in discharge over time to measure similarity of states can only work if there is a unique relationship between model state and dQ/dt. Given the equifinality in the fluxes-discharge relation that's typically visible in hydrological models (see e.g. Khatami et al., 2020), I think the section that introduces this concept (P16, l17) is not quite clear about why this dQ/dt assumption can be used together with CATFLOW.*

RL: Important comment and also the second reviewer had similar concerns. We will hence add Q as a second variable to group and ungroup model states and improve the discussion on how using a single variable to define similarity between model states will always lead to errors in certain scenarios and following that these variables need to be picked carefully.

WK: *Reading further, the authors address this concern to some extent in section 4.4 (P23, l18). This section however seems to show that CATFLOW does not exhibit such a unique relationship and the model reduces the number of model elements before the groundwater states reach similarity. This does* *apparently not affect the quality of the simulations much, because the KGE scores in Table 1 seem to indicate the adaptive model is as good as the fully distributed model for the two testing events.*

RL: In the current MS we did not mention that CATFLOW simulates only shallow subsurface stormflow in the entire summer period. This means that we do have a rather unique relationship between our model states and their function. Furthermore, the example in Fig.8 shows two extreme cases where one model receives much more precipitation than the other exactly intending to show the limits of our approach (section 5.3). As discussed in more detail with the fourth reviewer we will show that there is no difference between *model a* and *model c* (also concerning soil moisture in both depths) already shortly after the rainfall stops and when *model c* represents the entire catchment with a single hillslope. Furthermore, by calculating the Shannon entropy of the 42 hydrographs simulated by the spatially distributed *model b* we can see that there is no reason to assume that two models drift apart in the selected time frame. We will disscuss this in a revisted MS.

WK: *Fig. 4 shows that both testing events are selected in the middle of summer, when presumably the catchment is in quite a dry state (catchment state is not mentioned when selection of the two events is discussed on P18, l20 to P19, l6).*

RL: We mention the catchment states for event I and II on page 18 line 26 to 28 and page 19 line 5 to 6. Furthermore, do we refer to our former study where we showed 38 time series of soil moisture in the Colpach catchment in various depths and locations for the same hydrological year.

WK: *The fact that both events are selected during the dry summer could mean that the model can reset itself to mostly empty between the events and as such the long term (seasonal) impacts of not keeping the groundwater states separate cannot be investigated with the current two testing events.*

Equally, the events concern high flows so the impact of differences in slow ground-water states probably do not register in the dQ/dt values during the falling limb of the hydrograph (and thus the adaptive model simplifies itself).

There is the compounding issue that the KGE scores used to calculate the performance of model c are only calculated during the high flow event and that metrics such as KGE are typically not very sensitive to errors in low flows. This means that the parts of the simulation time series where the differences in groundwater states could be seen are both not included in calculation of the KGE score of model c and if they were, the KGE metric might not be able to pick up on any differences.

RL: Again an important point. As already mentioned above we discussed in section 5.3. "*While this finding is surely constrained by the chosen threshold, the picture is nevertheless quite different in deeper soil layers where the diversity of the rainfall forcing leads even after 24 hrs to increasing differences between the "driest" and "wettest" models. A part of the information about the different meteorological forcings between the two models is hence still stored in the model state after 24* *hrs and has not yet been dissipated. The importance of those differences likely depends on the dominant runoff generation process. In the present case, they have a minor impact as model ...*"

In a revisited MS we will underpin once more that our approach with the current definition of similarity regarding the model states can have significant impacts for long-term simulations however that there is no reason to expect that in our specific case.

WK: *Summarizing the above, I'm not sure that the dQ/dt criterion is entirely appropriate to determine when the adaptive model can reduce its complexity, and I'm equally unsure if the current two testing events would be able to show if the dQ/dt criterion is or is not appropriate. The straightforward solution would be to run model c for the year, add these results to Table 1 and briefly investigate for example the relative contributions of different fluxes to the overall water balance and the model's response to a few precipitation events during winter. Given that the adaptive model should be faster than the fully distributed one, this should not be a large computational burden and it will provide a much more complete impression of the capabilities of the adaptive model.*

RL: We hope that an improved discussion of the limits of the dQ/dt criterion (or any other criterion) as well as the addition of a second similarity measure (Q) will clarify the issues WK raises. We would also like to stress once more that we focus exclusively on the summer season as the distributed *model b* outperforms the *reference model* only in this period and because the meteorological boundary conditions change between the winter (frontal) and summer seasons (convective; Fig. 4 b). Furthermore, did we chose two rainfall-runoff events instead of the entire period as it allows us to analyzes the events in great detail (Event I: Fig. 5, 6, 7) and as our main focus in this study is on the rainfall-runoff interaction and not on low flow conditions. We selected event I because it has the highest rainfall intensity and third-highest spatial variability (highest Shannon entropy) in the selected period and event II because we wanted to test our spatial adaptive model at a summer rainfall event with longer duration. We believe that both events represent the state space of the runoff formation of the Colpach in summer well and see no reason to assume that the *model c* would fail at another rainfall event. A test of the spatial adaptive model for the entire hydrological year (or even for a longer period), in a different environment, with more variables and different thresholds to group and ungroup the model states and maybe even with a different type of model, is indeed desirable. However, to keep the already quite elaborated MS as focused as possible we will focus on improving the discussion with respect to the limits of our approach, add Q as second criteria to define similar model states, add a new figure to the appendix where we show how the thresholds impact the number of precipitation groups (please see the discussion with reviewer 2) and finally plot the soil moisture of *model a* and *c* at the end of event I and II to highlight that both models are in a similar state also with respect to their soil moisture.

Line-by-line comments

WK: *P5, l5. This question seems quite strongly related to the contrasting results in the literature that the authors discuss in the first and second paragraph of the introduction, where they conclude that the impact of using a distributed model and/or distributed forcing data is conditional on the catchment under investigation. This research question seems a bit generic in that light, given that only a single model and catchment are being investigated in this work. As is, question 1 seems more like a formality to me (it must be answered with "yes" before Q2 can be investigated) and the main focus of the manuscript seems to be on Q2. Perhaps the manuscript can gain a bit in focus if only the current research question 2 is specified, and the work done to answer the current Q1 is presented as a prerequisite to address the current question 2. For example, "We test this hypothesis by first showing that the model CATFLOW applied to the 19.4 km2Colpach catchment using a gridded radar-based quantitative rainfall estimate improves in performance when it is distributed in space and driven by distributed rainfall. We then address the following research question: "Can adaptive clustering be used to distribute a bottom-up model in space that it is capable to represent relevant spatial differences in the system state and precipitation forcing at the least sufficient resolution to avoid being highly redundant as a fully distributed model?"*

RL: Good idea. We will consider rephrasing this section following your lines.

WK: *P5, l14. Assuming that "> 1 m" refers to soil depth, should it be "< 1 m"?*

RL: No, soils are rather deep in this area and vary between 1 to 2.7 m according to several drillings and electrical resistivity tomography (ERT) measurements.

WK: *P7, l9. Which numerical scheme is used by CATFLOW?*

RL: Darcy-Richards: implicitly solved by a mass conservative modified Picard iteration scheme (Celia et al. 1990); Surface runoff (1d St. Verdant eq.) explicit Euler forward. We will add this information to the MS.

WK: *P7, l20. If possible without using too much space, it might be helpful to the reader to briefly summarize the main findings of Loritz et al. (2017).*

RL: The main findings of this study are summarized on page 10 section 3.1.

WK: P7, l21. What are the outcomes of this quality control?

RL: Manually quality checked by the Luxembourg, Institute of Science and Technology (LIST; no negative values, etc). We will remove the term "quality checked" as it is not necessary here.

***WK:** P7, l28. I'm not quite sure I understand why these distances are given as a range if only a single station is concerned. Does this indicate minimum and maximum distance of the catchment bounds to each radar station.*

RL: Exactly. These are the distance to the boundaries of the Attert catchment in which the Colpach is located. We will add this information in a revisited MS.

***WK:** P9, l13. I find this sentence a bit hard to follow. Is the part from "apart from..." onwards necessary here? This is already discussed in the introduction.*

RL: We will remove this part.

***WK:** P10, l21. Why is the model tested during two events instead of over the full year? How were these events selected?*

RL: Please see the discussion above and the discussion with the second and fourth reviewer.

***WK:** P11, l14. The conclusion that a distributed model is needed to account for runoff driven by convective precipitation would be stronger if the authors can (briefly) list which processes are represented at too coarse a scale in the reference model for it to properly deal with convective precipitation.*

***WK:** P11, l14. I believe this sentence would be more complete if it also explicitly mentioned that distributed instead of catchment-averaged precipitation data is needed to properly simulate the result of convective precipitation events.*

RL: We wrote on page 11 line 14: *"In other words, this entails that a hydrological model, distributed at a sufficiently high spatial resolution, is required to capture the spatial variability of the precipitation field to satisfactorily simulate the runoff generation of the Colpach"*. We believe that our argumentation is well justified here.

***WK:** P11, l27. It would be helpful for the reader to repeat that the only difference between reference model and model a is the choice of precipitation data.*

RL: We wrote in the sentence before the sentence you mention: "*Model a is identical to the reference model, however, driven by the area-weighted mean of the spatially resolved precipitation data described in section 2.4 (Fig. 2 b).*"

***WK:** P12, l3. Are these variables similar or identical to those used in the reference model?*

RL: Identical. We will change the word accordingly.

WK: *P12, l4. To clarify, does this mean that model b is run in a gridded fashion with the catchment divided into 42 grids (matching the precipitation grid)? If not, it would be good to clarify this in the text and mention the number of model elements that the precip field similarity approach gives. Line 18 on this page could benefit from a similar clarification.*

RL: Yes, this means that model b is "*divided into 42 grids (matching the precipitation grid)".* We will consider rephrasing the corresponding sentences.

WK: *P12, l23. Are there some observations that could help support the choice for 1 m/s?*

RL: We will add the reference of *Leopold, (1953)***.** Fig. 1 in this reference shows an average relation of flow velocities and discharge in rivers. Correspondingly we picked an average value of 1 m s^{-1} (2 to 3 feet per second).

WK: *P14, l29. It might be good to extend this line of reasoning to soil types and vegetation cover, as these are commonly used as model inputs/parameters.*

RL: Agreed. We will rephrase the corresponding sentence.

WK: *P15, l7. This sentence is quite general (referring to humid environments) and could use a reference. However, if the authors chose 1 mm hr-1 based on their expertise and knowledge about this catchment, then I think it's more accurate (and in no way worse) to phrase this decision along those lines, e.g.: "We chose this threshold as a reasonable value upon which we expect differences in hydrologic behavior, based on our collective understanding of the Colpach catchment."*

RL: Valuable point, we will rephrase this sentence.

WK: *P17, l10. I think it's import to repeat the similarity condition of dQ/dt here, because for a model that has no unique relation between model state and dQ/dt values this method cannot be applied without accounting for this difference.*

RL: Please see the discussion above and in section 5.3 in our MS.

WK: *P20, l6. The authors use KGE values in this section and Table 1. I'm not sure to what extent the aggregated value is a useful metric for events that last only a handful of time step. It would be good to at least disaggregate the KGE into its correlation, variability and bias components (e.g. quantify what can be qualitatively estimated from Figure 7) to see if the total KGE scores of the individual models are generated by (roughly) the same types of errors in the simulations.*

RL: Good point. We will add the three components of the KGE in the appendix for each model.

WK: *P21, l25. “acceptable” is somewhat subjective because no standard of acceptability has been defined. It might be cleaner to simply report the correlation component of the KGE to quantify to what extent the hydrograph shape is simulated.*

RL: Agreed. We will rephrase this term.

WK: *P21, l26. This trial of a direct runoff component seems somewhat ad-hoc to me. I don’t think this adds anything to the manuscript and that it will take more space than is available to properly justify this change. I suggest to remove these sentences.*

RL: Thank you. We will consider removing this sentence.

WK: *P30, l4-24. These sentences seem as if they would be better placed in the introduction or methodology sections.*

RL: We will rephrase some of these sentences. Please see the discussion with reviewer #1 (Daniel Wright).