The authors have responded well to the four reviews. I am particularly happy with the response to the similarity metrics used to determine when model states are similar and separate model elements can be merged. The authors have updated their choice of similarity metrics and gone to great lengths (e.g. section 4.1, page 23, section 6.3) to impress on the reader that their choice of metrics is catchment and study-dependent, which is a welcome addition. The experiment still relies on just two rainfall events instead of a continuous simulation for a longer period. I understand the authors' justifications for this but I think the paper can be improved further by expanding on this choice a bit more (qualitatively). Details below.

1. Reliance on two rainfall events

The experiment remains constrained to only two rainfall events during summer. The authors have clarified their reasoning for this (there is no benefit from using a spatially distributed model in winter conditions anyway, and the two events are the most extreme conditions available during summer). I find this a reasonable response although I think further testing over longer time scales would still be good to build confidence in this approach. It may be worthwhile to slightly reframe the paper and call this a proof of concept, until a run for a whole year or multiple years is feasible.

I also get the impression that there was a practical element to this choice in the sense that creating an automated version of the spatially adaptive model is "technically difficult" (p13, l10 in the original manuscript) and a "full automation of the proposed adaptive clustering approach [..] is beyond the scope of this study" (p13, l8 in the revised manuscript). I don't think the paper mentions this in so many words, but I get the impression that upscaling and downscaling the number of model elements was a (mostly) manual activity in this experiment.

The cases where this approach would be most useful (long simulations with many catchments) are equally the cases where manually upscaling or downscaling the model setup is least feasible. I think it would be helpful to the community if the authors can provide some context for the decision not to automate this process and describe the practical limitations that make automating this approach difficult. Is there something specific to CATFLOW that makes automation difficult or do the authors foresee similar difficulties for models that are regularly ran across large domains, such as land-surface models? Will there be additional difficulties to overcome, such as issues with routing in larger catchments, where also the location of the rainfall event will start to play a role and a hillslope far away from the outlet cannot be easily merged with one near the outlet on account of their different travel times?

A note on the additional computations needed (binning rainfall field on every time step, binning model states and averaging them in case the number of model elements can be reduced, etc) could be helpful too to determine potential gains in computational efficiency but I would understand that such numbers may currently be difficult to estimate.

A discussion such as this could help the community decide whether this approach is something that can actively be pursued with current models or whether this remains a largely academic exercise until a new generation of models can be developed that are specifically designed to take advantage of spatially adaptive setups. Section 6.2 may be a good place for this.

Line-by-line comments

P1, l1. Given the scope of the test cases in this manuscript (only 2 events during 1 year are assessed) and the overall message from the four reviews, one might say that this paper is more about providing a proof of concept for a spatially adaptive model and less about investigating the "role and value if distributed precipitation data for hydrological models". I would suggest that the authors rephrase their title to either emphasize that this paper shows a novel model approach or to be more specific about the limited scope of the investigation of role and value of distributed precip data for modelling.

P3, l22. It's not fully clear to me why moving to the event scale instead of running continuous simulations is one way to achieve dynamical allocation of the model space. Can the authors clarify this by adding the reason(s) why such a shift enables this new type of modelling and/or why this type of modelling is not possible for continuous simulations?

P5, l18. Following up on my earlier comment, I asked whether "> 1 m" should be "< 1 m" because the soils are described as "shallow" in this same sentence. Given the authors response: "No, soils are rather deep in this area and […]" perhaps this description needs to change. This of course doesn't impact the paper's results in any way, but it might avoid some confusion for other readers.

P9, l8. The manuscript mentions four model setups here (reference, a, b and c) but the abstract (p1, l9) mentions only three.

P10, l25. "One goal of … *rather than a result of important structural difference [..] within the Colpach catchment*." Technically, this second part is not tested as such in the paper. The results indicate that with a spatial variable precipitation field modelling results improve but this does not imply that model structure issues do not play a role. It might be good to clarify this sentence.

P21, 27 and further. I still think this addition of a direct runoff component is somewhat ad-hoc and contributes little to the paper. If the main goal is not to accurately match observations in this catchment, this semi-section can be removed to streamline the message in the results section.

P28, l6.It may be more accurate to change "summer season." to "summer season in this catchment."

P28, l22. Should these words be underlined?

P32, l2. Is "constellations" the right word here? Should this be "occurrences" or something similar?

Editorial

P3, l22. "one-way" > "one way"

P13, l4. "The main goal of the model testing of the spatially adaptive …" > "The main goal of testing the spatially adaptive …"?

P14, l7. "are so large, that" > "are so large that"

P14, l11. "elements by" > "elements is by"

P14, l11. “Hydrology” > “hydrology”

P14, l28. “similar” > “similarly”

P15, l18; also l19. “upon we” > “upon which we”

P15, l20. “in humid catchment” > “in this humid catchment”?

P16, l8. “Section 3.3.2” > this needs updating but I’m unsure to what. Section 4.1 maybe?

P17, l3. “models two” > “models to”

P21, l25. “acceptable” > “acceptably”

P23, l29. “is” > “are”

P27, l8. “simulate” > “simulating”

P32, l5. “equifinality” > “as equifinality”

Misc

I’ll also takes this opportunity to briefly respond to comment #28 by reviewer 4, to clarify a potential misunderstanding. Part of comment #28 states:

“*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)”.*

While we do use KGE = 0.3 in our paper, this is only to illustrate that if arbitrary thresholds are used to determine which models are behavioural, NSE and KGE do not produce consistent results. In fact, later in the paper we try to warn the reader against using such thresholds in an arbitrary manner:

“Regardless of whether KGE or some other metric is used, the final step in any modelling exercise would be comparing the obtained efficiency score against a certain benchmark that dictates which kind of model performance might be expected (e.g. Seibert et al., 2018) and decide whether the model is truly skillful. These benchmarks should not be specified in an ad hoc manner (e.g. our earlier example where the thresholds are arbitrarily set at NSE=0.5 and KGE=0.3 is decidedly poor practice) but should be based on hydrologically meaningful considerations.”

I see that the authors also argue against using this threshold in their manuscript and I think is an appropriate response.