Reply to Anonymous Referee #2:

Anonymous Referee #2 ***(AR2): Summary and Recommendation:*** *The manuscript introduces an adaptive spatial clustering of hydrologic response units (HRU) to cope with the dynamics of the intermittent rainfall by keeping the model as parameter parsimonious (=model states) as possible in terms of reduction of similar-reacting HRUs. The manuscript is well-written and I enjoyed reading it. The introduced clustering is innovative from and fits into the scope of the journal. I have a few moderate and a number of minor comments, which are stated below. My overall recommendation would be a moderate revision to give the authors enough time to solve the open issues. Since I can only choose between minor and major revision, major revision it is.*

Ralf Loritz (RL): We would like to thank the second referee for the time and the effort she/he put into his review. The points she/he raises are relevant and addressing them will help to improve our manuscript. We hope that after this discussion (as well as after we revised our manuscript) all issues she/he raises can be clarified.

Moderate comments:

1. AR2: *The manuscript is about the reduction of the spatial model resolution based on the variety of precipitation as input signal. I'm wondering if there is not an adaption of the temporal resolution required as well since scales in space and time are not independent of each other (see Melsen et al. (2015) and references therein)? Maybe it's not an issue for the small catchment studied here …*

RL: Important comment. The results of the study of *Zhu et al. (2018;* recommended by the first reviewer) highlight that the timing of the precipitation is more important in smaller catchments while it is the spatial pattern in larger catchments. We will discuss this in a revisited MS and carefully read the study of *Melsen et al. (2015).*

AR2: *… but for larger catchments with a small hydrologic variability the numeric stability can be questioned due to the large spatial discretization and the high temporal resolution (e.g. in terms of the Courant-Friedrichs-Lewy condition, Courant et al., 1928). The authors should proof this condition for their model setup and discuss possible issues in the manuscript. An alternative would be to reduce the temporal resolution as well, which would lead to an additional reduction of parameters/computational costs.*

RL: CATFLOW uses an adaptive time-stepping, which means that time steps can be reduced down to seconds depending on the numerical solver. In the presented study the Darcy Richards equation is solved implicitly while the surface runoff is solved explicitly (for more details see also *Zehe et al., 2001*). As the horizontal grid resolution of the CATFLOW hillslope (reference model) is below 1 m, the vertical below 10 cm and time steps are small we have no issue to fulfill the Courant criteria in our model.

Nevertheless, you mention an important point here and the fulfillment of physical and numerical constraints were the main motivations of our former "representative hillslope" study (*Loritz et al. 2017*).

CATFLOW hillslopes are typically interconnected by a river network and runoff is routed downstream with a diffusion wave approach (explicitly solved) assuming a prismatic river cross-section and roughness that changes with changing Strahler order. However, the combination of a river network with the raster layout of our adaptive model is not straightforward (although not impossible, for instance by linking the centroid of a raster cell to the closest node of the river network). To make things not more complicated as necessary we decided to use a lag function in this study. This lag function is not solved numerically but shifts the simulated hydrographs in time by a constant velocity. Again we have no issue with the Courant criteria here. The latter is different if we would have used an adaptive mesh approach where the numerical grid is changed during runtime. Here we carefully need to check the courant criteria when we increase the size of the grids. We will discuss this in a revisited MS.

2. AR2: *The authors have selected two events to show the ability of adaptive clustering. The choice of both events seems to be very arbitrary. From Fig. 4 it seems that the resulting runoff peaks are not representative for runoff mechanisms of the catchment. As far as I understand from P13 l8-10 the clustering is carried out manually and not automatically so far, which is the reason why the authors decided for two small events covering only a few time steps. However, I disagree with the hypothesis that "a test on a longer timescale...would provide only little more scientific inside" (P13 l9-10), which is also not proven by the authors.*

RL: Respectfully, we do not agree with the assessment of the reviewer regarding the selection of our rainfall events. We chose event I as it has the highest intensity and third-highest spatial variability in the chosen period. We chose event II because we wanted to test our adaptive model at a rainfall event with a longer duration and lower intensity). From examining the rainfall-runoff events in summer we believe that both events represent the runoff generation in summer well as long as subsurface storm flow is dominant. We agree with the reviewer that our statement is a bit misleading and we will explain better what we mean here. Please see also the discussion with the third and fourth reviewers.

AR2: *I rather expect that the reduction of model parameters due to the adaptive spatial resolution is reduced significantly for long-lasting rainfall events causing a direct runoff response over several days as e.g. in Nov 2014, Jan 2014-Mar2014 and Aug 2014.*

RL: You are correct. The needed spatial model resolution in winter is much lower compared to the chosen summer rainfall-runoff events. This is indicated by the fact that the distributed *model b* and the *reference model* perform almost identical with respect to simulate the observed discharge in the winter season. We would argue that is is difficult to justify the use of a spatially distributed model over a

spatially aggregated model if they perform similarly as long as the focus is on an integral response of a system.

AR2: *Another point that can be questioned is snow, which does not cause runoff immediately, but when snow melt begins. How will this be affected/can be incorporated by the adaptive clustering? The impact of more complex events than those analysed in the current study has at least to be discussed sufficiently in the manuscript, although an analysis of more events is encouraged to represent the effect of the adaptive clustering on the variety of runoff responses*.

RL: Interesting point. Snow is rare in the Colpach catchment which is fortunate as CATFLOW has no internal snow routine. However, let's assume we would have used a model with a snow routine in an area where snow is a dominant control on the runoff generation. In this specific scenario, we would indeed have to adapt our definition of similarity between the model states. In other words, instead of using the only slope of the simulated hydrograph alone to define similarity, we would also need to check the snow cover before we would group models as functional similar based on their state. Two similar hillslopes would then have the "same" snow cover (given a threshold) as well as the same slope of the hydrograph. We very much like the idea of testing the approach in an area where snow is an important factor. However, for now, we will discuss the limits of choosing a single variable to group model states in a revisited MS.

3. AR2*: The model states are identified by the slope of the resulting runoff curve. However, the slope can be more or less identical for one time step independent of the current runoff situation, e.g. if runoff is reduced in one tile from 25mm to 20mm and in another tile from 10mm to 5mm (which could be the case in a stratiform event with a convective cell inside), the resulting slope is the same, right? So the soil moisture and other storage elements is then "averaged" due to the same model state of both tiles, although both tiles are in completely different hydrologic situations. It would be useful if the authors would comment on that issue or, if I understood it not correctly, clarify the part where I got lost.*

RL: You are correct. In an earlier version of our spatial adaptive model, we used the absolute discharge to identify similar model states. The issue here was that two models could produce the same discharge at a given time step but one model would simulate a rising hydrograph while the other a declining. We hence decided to take the slope of the hydrograph assuming that the model differences would be small given the size of the Colpach catchment, the focus on the summer season and because we only simulate shallow subsurface stormflow. In the case of a stratiform event with a convective cell inside or if we have snow in a catchment our assumption might be violated. Thank you for raising this issue and along your lines, we will add another criterion to our spatially adaptive model. In a revisited MS only model elements which share a similar dQ dt^{-1} (0.05 mm hr^{-1}) as well as Q (0.05 mm hr^{-1}) will be grouped together.

Specific comments:

AR2: *P4 l5-8 The difference is not clear formulated at this point. It becomes clearer while reading the manuscript, but should be communicated concisely at this point.*

RL: We will rephrase this sentence.

AR2: *P7 l27 Where are the disdrometers located? Can they be used to improve the rainfall input for the reference model to achieve a more realistic uniform areal rainfall? If not, could be an increase of rainfall amounts with altitude improve the areal rainfall estimate? The Roodt station is situated in the raster field with the lowest rainfall amounts (Fig 2) and not representable for the catchment. So any correction has to be done to enable a fair comparison between reference model and model a.*

RL: When we were setting up the reference model for our proceeding study (*Loritz et al., 2017*) the only rainfall measurement available at that time was the ground station in "Roodt". As we are aware that the comparison between the *reference model* and *model b* (the distributed model) is not entirely fair as they used different rainfall data we added *model a* to the model ensemble. In a restructure section 3 we will clarify this as well as add the location of the distrometers to the appendix A1.

AR2: *Fig 2 Please add rain gauge data to Fig 2b) to enable a comparison of all rainfall inputs.*

RL: We will add the rainfall data from "Roodt" to Fig2b.

AR2: *P10 l2, p11 l26 area-weighted -> As I understand the areal mean is estimated by the arithmetic mean of the satellite data. How do weights for different areas affect this estimation? This is not clear for me, please rephrase/add the explanation.*

RL: As not all of the 42 raster cells of the distributed rainfall data are entirely within the borders of the Colpach catchment their weight was reduced when we calculated the average precipitation for *model a*. We will rephrase this sentence accordingly.

AR2: *P11 l2 sap flow -> Do the authors mean by sap flow the flow in plants? I can't imagine at this point how the authors applied observations like that in the current study. If so, please describe a bit more detailed, since it is not a conservative measure for model validation and hence of great interest for the community.*

RL: By sap flow we indeed mean sap flow in plants. In our proceeding study (*Loritz et al. 2017*) we compared normalized sap flow velocities against normalized transpiration simulations of CATFLOW to evaluate the transpiration simulation. Sap flow measurements where thereby one of the keys for a successful simulation in the Colpach catchment as they helped us to identify the onset of the vegetation (when the trees started to transpire). The comparison is described in detail in *Loritz et al. (2017*; Figure

12). Although we agree that this might be of great interest for the community we would like to avoid discussing this once more to keep the MS as focused as possible.

AR2: *P12 l23 "average distance of each grid cell to the outlet" -> Should it not be the distance along the flow path/flow direction? So it would be possible that the runoff is assumed to stream upwards in some areas of the catchment- Please rephrase or reconsider.*

RL: For each grid cell of the precipitation field we calculated the average flow length along the surface topography to the outlet of the catchment using an underlying DEM with a 10 m resolution. We used the averaged flow distances in our lag function. We will explain this in more detail in a revised MS.

AR2: *P12 l30 "3.2.1 to 3.2.3" -> "3.3.1 to 3.3.3"*

RL: Following the discussion with the first reviewer Daniel Wright we will restructure section 3 and remove all "subsubsections".

AR2: *P13 l2 "wetness state" Please define this term. It sounds as only soil moisture is included without any additional information, but there is more included, right? If not, why not using the term soil moisture? Section 3.3 and 3.3.1 There are repetitions among the paragraphs, please remove them.*

RL: We will remove the term "wetness state" with the term "catchment state" to make clear that we also mean the shallow groundwater table, soil moisture, etc. We will furthermore restructure section 3 and remove the repetitions.

AR2: *P15 l4 & 21 Both thresholds are catchment size-dependent (as the authors state also later).For other applications it would be useful to introduce a catchment size-dependency to derive these thresholds. This is beyond the scope of the study since it demands a multi-catchment analysis, but the authors should add a small sensitivity analysis by e.g. using ΔP >{0.5, 1, 2} mm/hr as thresholds. This is along with a comment I have for the results section later, but I want to state it already here. In the results discussion it is often mentioned, that the number of parameters is reduced between model b and c, there is no figure illustrating it, although I would imagine it would bean impressive plot with y as KGE over x as the summarized number of model parameters per time step (or on average) for one event. Model reference, a, b, c (ΔP>1mm), c(ΔP>0.5mm) and c(ΔP>2mm) would be the points to show in the diagram. I assume model c would represent a break in the curve (KGE not increasing, while number of model parameters do) and the different thresholds would represent the uncertainty of this approach.*

RL: Interesting comment. Using a typically physically-based model (CATFLOW) and specifically the setup of our model based on field measurements it is kind of difficult to estimate the number of model parameters in our study. However, we will add a plot with the distributed precipitation binned into different thresholds (0.1, 0.5, 1, 2, 5 mm hr^{-1}) to the appendix. Based on this plot we will discuss how the binning will most likely affect our spatial adaptive model. Furthermore, will we discuss that a

sensitivity analysis with different thresholds is needed along your line of arguments. Again thank you for this comment.

AR2: *Table 1: As far as I understood the calibration was done only for the reference model, right? Although that seems to be done in a former publication, a brief information about calibration and validation period, objective function and so on is required to interpret the table. For model a, b and c no parameters were changed, so the same parameter set was used throughout the study to enable comparisons? If there was a re-calibration for model c, the reference model and models a and b should be re-calibrated for the events only as well to enable a fair comparison*

RL: Exactly the calibration was done in a former publication exclusively for the reference model. All model parameters remain the same. The only differences between the models are the rainfall data which we use to drive the models as well as their spatial resolution. We will add more details about the calibration in a restructured section 3.

AR2: *Fig. 5: I'm a bit confused here. The authors state P=12 for t=2, but from counting it is P=13 – please double-check (also the number of entries in the following text refer ring to t=2). Additionally, for t=4 M=3 results from P=2 and S=1 – from my understanding the maximum of model states is max(M)=2 in this case, please double-check.*

RL: You are correct. We will check the figure as well as the corresponding text passages. Thank you for checking the figures so carefully.

AR2: *P27 l4-22 This paragraph provides already a good overview of related references. However, from my understanding the reference of Nicotina et al. (2008) concluded that spatial patterns of rainfall are only important for large catchments (8000km2 in their study) for hourly time steps, the correct estimation of areal rainfall is sufficient for smaller catchments. The authors should review this reference again and check their implementation in the current manuscript.*

RL: Thank you for this comment. We were referring to the following section in *Nicotina et al. (2008)*: "*As noted in section 4, this is because the spatial scales of variability of rainfall are very often much larger than the typical hillslope scale. Whenever infiltration excess mechanisms are important, the spatial distribution of areas of intense rainfall may be an important factor in determining the hydrologic response, …* ". In the current MS the use of this reference is indeed misleading and a leftover from an earlier version. We will revisit the corresponding sentence.

AR2: *Also, Ogden and Julien (1993) state that only for rainfall with durations shorter than the concentration time of a catchment the spatial distribution of the rainfall matters, for longer rainfall events only the temporal distribution matters. To highlight the importance of distributed models the authors could also look at Krajewski et al. (1991), Bardossy & Das (2008) or Müller-Thomy et al. (2018)*

RL: Thank you very much for pointing us to these publications we will read them carefully and see if they can help us to improve our argumentation.