Response to Reviewer Comment #1

We would like to thank you for your comments and very constructive suggestions. Your feedback on the manuscript is very valuable, and it will help us to improve the manuscript.
Response to general comment on model comparison is given below under Comment no. 13. Criticism of linguistic quality caused us to repeat and extent our proof-reading procedure. The linguistic revision is not yet completed and will be addressed separately in an additional response. Here we will focus on the content-related comments by the referee.

Response to detailed comments:

1. "Page 3, line *20* ~~30~~: The computational demands of semi-distributed models with polygon based delineation is nowadays often negligible."

   We agree that the computational demand to perform a single model run is often negligible, but this omits the demands for model calibration or extensive sensitivity analysis studies. However, the actual formulation does not address this subjects and has been reformulated as follows:

   "The aim of the proposed algorithm is set to offer a basin partition with a minimum of heterogeneity by a minimum of sub-divisions, i.e. to ~~reduce the computational calibration effort~~ *a unnecessary high number of sub-divisions and, hence, number of parameters* in cases of hydrological modelling."

2. "Page 4 and Fig. 2: Pore volume and available water capacity should be defined. For instance is with "total pore volume" the "porosity" meant? Usually these variables are given in volume percent? Here they are given in mm, why? Does this not require an information about the soil depth?"

   Our data basis provides the opportunity to assess soil properties in different depths. In previous projects, where this data has been applied, we experienced that the majority of soil storage is located in the upper zones of the soil layer. Namely a depth of up to 2 meters covers nearly all available soil storage. Why we applied the *[mm]* data and not the [%] is in this case due availability (data from previous project) than for any particular consideration.
   The employed TPV data is calculated for a soil depth of up to 2 meters, this information has been added to the manuscript, Legend in Figure 2 and Page 4 Line 18-19 have been revised:

   "To characterise the soil characteristics of the German catchments, a gridded soil data map from the German Federal Institute for Geosciences and Natural Resources (BÜK200) and CORINE land coverage data (CLC) (Bossard et al., 2000) were used. Pedo-transfer functions (Sponagel, 2005) were applied to transfer these information into gridded data about (available) water capacities (AWC) and other soil storage parameters as well as hydraulic conductivities, *both for a soil column of up to 2 meters depth*."

3. “Page 5, Lines 24-25: I was trying to picture the basin split into stripes. I found Figure 5 as an visual explanation. However, comparing Fig. 5 with Fig. 3 there should be much less points of the distance-factor function in Fig. 3 (only 5 points). Later I realised that Fig. 3 does not belong to Fig. 5. For demonstration purpose I would suggest providing a pair of Figures with distance classes and corresponding distance factor function using a few classes only. May be Fig. 5 needs to come before Fig. 3.”

   The enhance comprehensibility we follow your recommendation. The distance-factor function of the synthetic catchment has been created (see Figure below) and added to the manuscript as Fig. 4. The synthetic catchment is now introduced on Page 5, Line 30 as follows:

   “Please note that w(x) is the non-normalised value of the area-function (Snell and Sivapalan, 1994). *To visualise the proposed function a simple synthetic basin with its stream network, distance-classes and an arbitrary characteristic are shown in Fig. 3. To keep things simple Fig. 3c shows the unified flow length (comprising x_S and x_H) derived from flow direction data in Fig. 3b. As it can be seen in Fig. 3c the basin has been split into 5 distance classes. When Eq. 1 and 2 are applied to the shown data (Fig. 3a) the average and standard deviation are calculated within these five distance classes. The obtained distance-factor function is shown in Fig. 4.*
   Figure ~~3~~ *4* shows the application of ~~both values (Eq. 1)~~ *to real data in a meso-scale catchment, namely for* AWC in the Mulde catchment.”

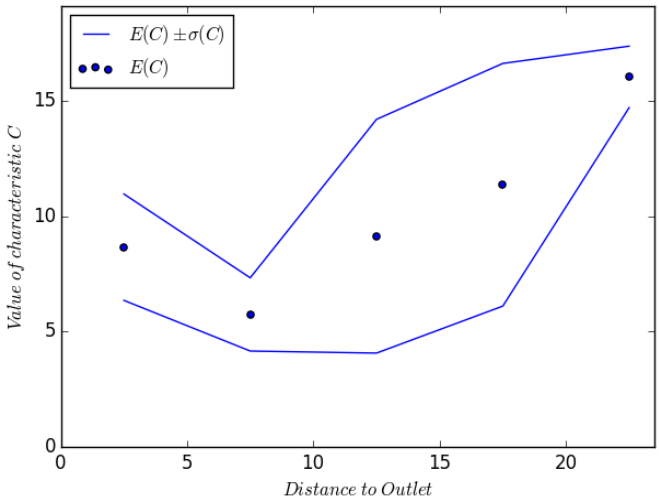

Figure 3: Distance-factor function of sample Data in synthetic catchment

4. “Page 5, line 31: (Eq. 1 and 2)“

   Already addressed in previous point.

5. “Fig. 4: Use the same units on the x-axis as in Fig. 3.”

   Figure 4 (now 5) has been revised.

6. "Fig. 5: Example input data are on the "left"!"

   Subscript of Fig. 5 (now 3) has been changed as noted.

7. "Page 12, Eq. 8: I would suggest to explain this equation in words. I also would suggest to exchange the sides of omega and sigma (even if mathematically not necessary) in both numerator and denominator."

   We suppose the suggestion to switch omega and sigma is indented to clarify the summation. Instead of switching the summands we added parenthesis for clarification. An explanation of Eq. 8 has been added to the manuscript on page 12, line 4:

   "Second measure α_2 is intended to show cases where the total heterogeneity has been lowered significantly, but still remains above the objective. *Threshold Ω is used to formulate this objective. Finally α2 is calculated as the difference between the threshold Ω standard deviation in the separated catchment σS(C), normalised by the difference between Ω standard deviation in the unseparated catchment σU(C)*:

$$\alpha_2 = \frac{\sum_{i \in M(S)} \left(\Omega - \sigma_{S;i}(C)\right)}{\sum_{j \in M(U)} \left(\Omega - \sigma_{U;j}(C)\right)} \qquad (8)"$$

8. "Page 12, line 15: What are cases with "negative outcome"? Do you mean insufficient variance reduction?"

   That is correct. Page 12, line 15 has been revised as suggested:

   "If we focus on the cases with ~~negative outcome~~ *insufficient variance reduction,* we are able to identify some limitations of the algorithm."

9. Fig. 8, 9 and 12: The distance factor functions are hard to read. Use vector graphics and/or large fonts and/or larger figures.

   Figures have been revised. Figures are now embedded in EPS format, fonts have been enlarged.

10. "Fig. 10: In the legend of AWC "max" and "min" need to be exchanged."

    Legend of Fig. 10 has been revised.

11. Fig. 11: When resampling of AWC for the Mulde river basin is shown also the original AWC map of the Mulde basin should be shown for comparisons. Why are there some kind of horizontal stripes in Fig. 11?

    The map of original AWC values in the Mulde basin is not shown in the manuscript, because the spatial arrangement of AWC and the shown TPV is identical (as mentioned on page 13, line 15) due to origin of these data grids. Both values follow the arrangement of the soil map, their only difference is the range of

values (also given in line 15), caused by different calculation formula (Sponagel, 2005). Since the manuscript already has a great number of figures and the information content of the AWC figure low we omitted it in the original manuscript.
Nevertheless, for the purpose of completeness the map of AWC can be added to manuscript if referee and editor insist on this topic.

The stripes in Fig. 11 are caused by the process of quantile exchange. In Fig. 2 it is visible that on a small scale the variance of the characteristic is in comparison to the total variance very low. When these values are transferred to their empirical quantile, their numbers differ only from the fifth or sixth digit. A reduction of digits and the application to another empirical distribution function, with values that are again rounded, resulted in the visible stripes.
We are very thankful for this review comment since it pointed out a problem that got lost along the way of preparing this manuscript. We revised and re-applied the resampling scheme, with the results that we were able to remove the stripes (see revised Fig. 11). Changes in the input to sub-basin ascertainment led to slightly different results (rev. Fig. 12) and performances compared to the initial results. However, altered results do not lead to different conclusions.

Table 1: Normalized reduction of standard deviation for resampled basins

Catchment	Pore Volume		Slope	
	α_1 [%]	α_2 [%]	α_1 [%]	α_2 [%]
Mulde (res)	~~57.9~~ *58.9*	~~12.8~~ *26.2*	~~8.5~~ *8.9*	~~82.6~~ *76.4*
Salzach (res)	~~38.7~~ *39.7*	~~10.5~~ *4.8*	~~14.7~~ *19.7*	~~58.5~~ *37.5*

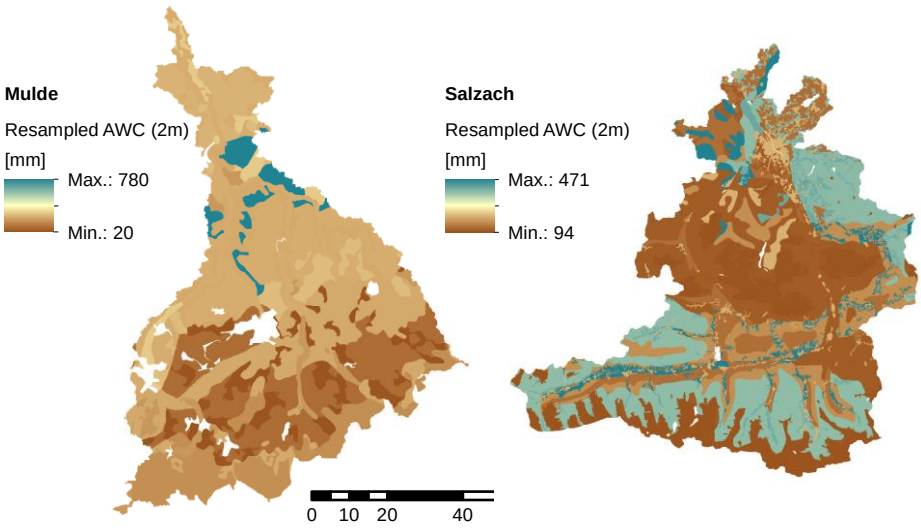

Figure 1: Resampled AWC values for Mulde and Salzach catchment

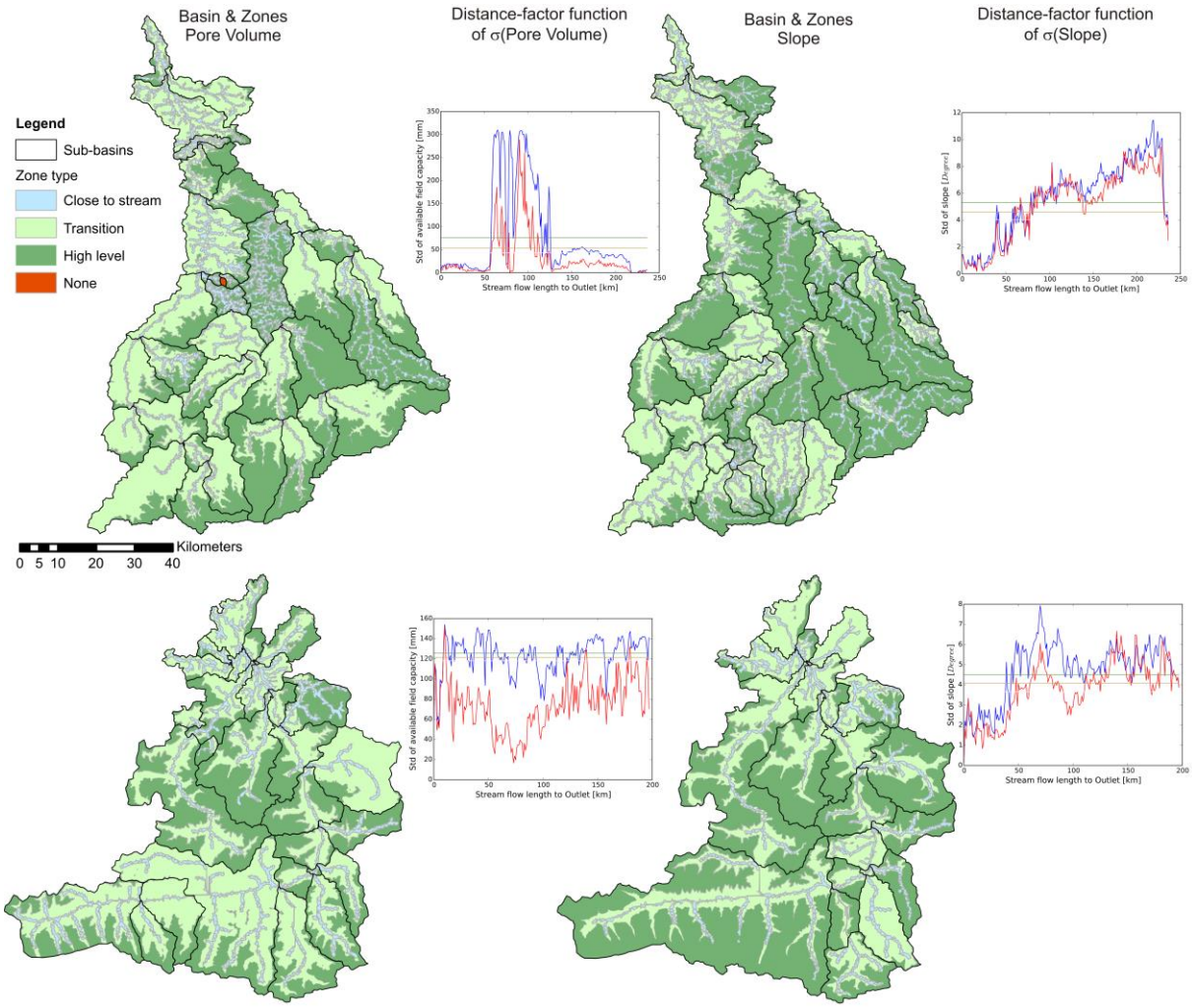

Figure 12: Results of ACS application for resampled catchments of the Mulde and Salzach (from top to bottom), sub-basins based on resampled pore volume (left) and slope (right). Comparison of $\sigma_U(C)$ and $\sigma_S(C)$ for each application (red and blue lines).

Changed results have been transferred to the manuscript as on page 13:

Removed line 28/29.

Added following passage in line 30:

"*We also experience a change in performance for slope. The exchange of heights values creates a lower range of slope values and a lower amount of heterogeneity. This patterns resulted in all other applications to inferior α_2 performances. Still the geomorphologic structure of the basin remains unchanged and heterogeneity can be assessed by the algorithm (visible through unchanged total reduction).*"

12. "Page 16, line 29, Table 4: Where are the 30 different zones per sub-basin coming from; why so many; how are these zones defined?"

In this section the partition by land cover and heights from Sec. 5.1 is converted into a model structure. As indicated, this separation scheme follows the recommendations by Lindström et al. (1997): Land cover to account for soil and heights for precipitation and evaporation correction factors. Land use has been divided

into forest (all types), bare soil/rock and field (all remaining cells that are not water). Threshold for heights partition has been set to 100m. The obtained spatial setup is shown in Fig. 13 & 14.

Choice of 100m as height threshold is due to the usage in the original HBV publication by Bergström (1976) who used a similar heights subdivision (900m / 10 height-zones) and the HBV model of the German Federal Institute of Hydrology which has been applied in a project at our institute (http://doi.bafg.de/BfG/2016/BfG-1877.pdf).

13. "Page 17, Table 5: As already mentioned in the general comments the comparison of the performance for two model versions with such a large difference in number of parameters needs to be given some more thought. Could the reason for the better model performance not be just because of the smaller number of parameters and therefore the smaller complexity and easier calibration. This might be tested by an additional model version using the same small number of parameters as in the new delineation scheme but applied on the old conventional basin separation (using only 3 zone per sub-basin too)?"

The trade-offs between parameters and model performance are a topic for themselves and we intended to address it just as scantly as possible. Therefore, we tried to stay consistent in our modelling choices. We applied the same basin partition schemes that we applied in the previous sections and used the same parameter coupling scheme. Parameter coupling included only 6 zonal parameters, all remaining zonal parameters were due to calibration. Our decision for consistency resulted in the shown number of parameters for the three models.

On the one hand we agree with your statement that more parameters make the model more complex which could lower the model performance. On the other hand, a higher number of free parameters offers more flexibility to fit the observed behaviour. We expected that a model with more parameters would perform better than the parsimonious setup in the calibration period. Due to the high degree of specialisation (fitted model parameters) we expected a greater loss of accuracy for high parametrised model structures than for the parsimonious. Our results so far show that even in the calibration period the performance of the parsimonious setup is superior. Leading to our conclusion that the advantage of more flexibility, even with the same parameter coupling scheme, does not compensate the value of the information we added to the model. We closed our analysis at this point because we defined our benchmark as the "common" scheme for the HBV_{96} model.

However, as your comment pointed out, we did not consider a benchmark model with a comparable number of parameters but with different spatial resolution. We suggest to stick with the benchmark basin partition scheme and extend the coupling scheme. Previously, parameters that could be somehow reasonably be related to storage information (like SM – soil storage, ICMAX – interception storage etc.) were coupled between the zones within a sub-basin. When we loosen this restriction and take all parameters into the coupling scheme we are able to reduce the number of free parameters (see Tab. RS1). This advanced coupling scheme has been applied to the ACS model as well to gain comparability of calibration effort.

Table 4: Parameter quantities

	Benchmark *Free*	Benchmark *Coupled*	Benchmark *Advanced*	ACS *Coupled*	ACS *All coupled*
Sub-basins	38	38	38	44	44
Zones per Sub.	~30	~30	~30	3	3
Parameter (total)	19562	12198	1710	2244	1980
Parameter (per Sub.)	~495	~321	45	51	45

Evaluation of the new calibration schemes gives the following results:

Table 5: Nash-Sutcliffe Efficiencies of Benchmark and ACS-model

Simulation (Start-End)	$NSE_{B;Free}$ [-]	$NSE_{B;Coupled}$ [-]	$NSE_{B;Advanced}$ [-]	NSE_{ACS} [-]	$NSE_{ACS;All}$ [-]
1995 - 2006 (C)	0.678	0.659	0.682	0.792	0.791
2007 - 2011 (V)	0.524	0.570	0.578	0.622	0.647
1984 - 1995 (V)	0.496	0.516	0.525	0.607	0.546
1951 - 1961 (V)	0.433	0.568	0.458	0.660	0.572

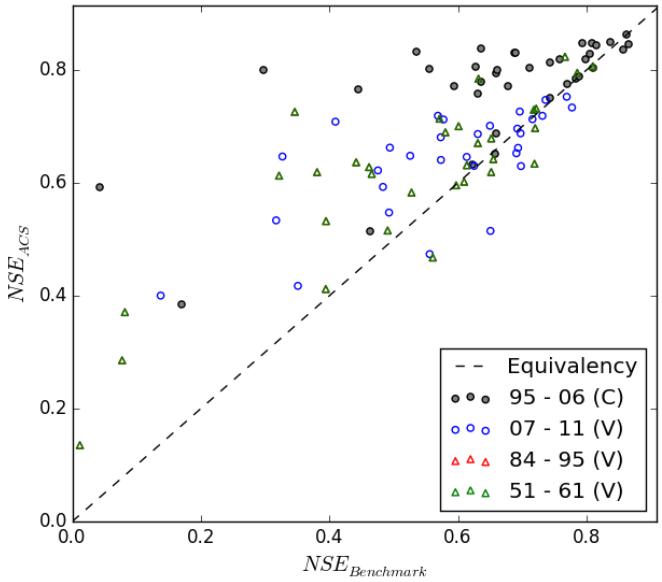

Figure 18: Nash-Sutcliffe Efficiency of ACS- based model and benchmark model, all zonal parameters coupled

Section of model performance has been extended to include the presented results, Page 16, Line 32:

"*A high number of parameters is believed to be offer a model structure more flexibility to fit the observed data, though its higher complexity might lower its performance. To compare our proposed model structure with a benchmark comprising a similar number of parameters we performed added a third calibration strategy. The performed approach coupled all zonal parameters as described above. This lowered the amount of parameters per sub-basin to 45 for both model setups. As it can be seen in Tab. 4 the total amount of parameters in the benchmark partition is higher than in the new ACS-based partition.*

After the calibration (time period 1995-2006) we evaluated model performance in three validation periods. Two in direct (temporal) neighbourhood to the calibration period and the last at the very beginning of the time series. Model performance has been calculated as the average Nash-Sutcliffe-Efficiency (NSE) (Nash and Sutcliffe, 1970) of all gauges and is tabulated in Tab. 5. Results show that ACS-parametrisations are

superior in all cases. Its increase in performance ranges from 17-52% in comparison to the free- , 11-21% *to the 6-parameter*-coupled benchmark *and 5–19% to the all-coupled parametrisation*.

Beside this "lumped" evaluation we compared the performance of the models at each gauge in each period. Comparison of NSE for *6-parameter coupled* models are shown in Fig. 15, for ACS and free-benchmark model in Fig. 16. *Comparison for the all-coupled parametrisation is shown in Fig. 17*. We can see that the individual performances offer the same conclusion as the lumped performance, though some results are better for benchmark models (both parametrisations). To be more precise, in case of *6-parameter*-coupled models 20 points (rep. a single gauge in one of the time periods) are below equivalency (rep. a better performance of the benchmark model), in case of the free-benchmark model 12 *and for the all-coupled benchmark 23* points. Representing 15%, *9% and 20% of* evaluated cases."