We would like to thank the reviewers for their highly constructive comments on the manuscript “Incremental model breakdown to assess the multi-hypotheses problem”

(comments of the referees are printed in blue, responses of authors are held in black, added text to the manuscript is in italic)

Response letter to Reviewer #3 (Anonymus)

The article presents an incremental model breakdown approach to determine an optimal hydrological model structure for rainfall-runoff modeling. The hypothesis of the authors is that one should start from a model structure that includes all possible processes and that this structure should then be incrementally simplified by successively removing the unimportant processes, i.e. those for which the model performance is not degraded or even improved when they are removed from the structure. The approach is demonstrated on a catchment in Germany. Though the approach is interesting, I have several concerns about the way it is applied and demonstrated:

- I think that the “one-at-a-time sensitivity analysis” approach that is applied makes the hypothesis that all processes are independent from each other in the model structure. However, this is probably not the case and it is most likely that there are interactions and compensations between model components. Therefore, I find it is difficult to conclude on the individual value of each component based on these tests only. There is no guarantee that the model structure selected at the end is optimal, since only a very limited number of structures among all the possible ones have been tested. It is likely that there are many options which are close to each other in terms of performance.

We thank the reviewer for the valuable comments on the paper. However, we would like to point out that this paper is meant to introduce a new concept to explore the space of possible model structures. We realize that the method would be validated in a more profound way if

- several more catchments had been used
- the validation and calibration time period had been swapped
- several different time series from the same catchment had been compared
- the incremental model breakdown had been iterated several times, and
- the comparison with other approaches like step-wise model modelling had been more in depth.

We think though that all these suggestions sufficient to fill several papers and would bloat the current paper that is mainly meant for an introduction of a new concept.

We further agree that it is correct to criticize that even in our study only a limited number of models is used. Still, 15 (or 16 with HBV-Light) different models is a large set of different model structures that only few other studies have compared. We therefore think that our work as a good starting point for future, even more comprehensive applications of incremental model breakdown.

We also agree that we cannot ensure independency of all processes from each other and we do not have a guarantee that this process is successful in the end. Failures of this approach would lead either to a model with lower performance than the original in the calibration period or to an overfitted model. While the first type of failure is obvious (we would not submit such a result for publication), the problem of overfitting has been not sufficient discussed in the original manuscript. We added a discussion of overfitting to chapter 4.3. Secondly, the connection of processes is shown in a shift of the parameter space, which we have shown exemplary between model 1 and 15 in Fig. 1. We do not claim , that our method leads to a single optimum model, but we explore a new path to model structure improvement. To clarify this we have extended the last sentence of the introduction with: *…obvious, even if a theoretical optimal model structure is still unknown.*

- The parameter sampling approach, drawing 300,000 parameter sets for each structure, makes that the parameter space will be much more densely scrutinized in the case of a model with 10 parameters than in the case of a model with 19 parameters. This means that the chance of getting behavioral parameter sets is much more limited in the second case

than in the first case. This may induce a bias in the way the models are compared when using the GLUE approach. This should at least be discussed or ideally further tested.
It is true that the parameter space of the models with less parameters is sampled more exhaustively. Nevertheless, LHS is a robust enough method to counter this. As the parameter space is sampled very uniformly when using LHS, a smaller number of runs is needed, as in comparison with e.g. Monte Carlo Algorithms.
The LHS allows to calculate how many runs are needed for good sampling of the parameter space (see McKay et al. (1979)) and this threshold (n=262,144 for 19 parameters) is achieved for all models. This is also in line with our personal experience when using LHS. Usually, models reach good values for the objective functions in the first few hundred runs (even when they are complex), and all following runs are adding only small increments in performance. Still, it might allow models with less parameters to get more behavioural runs, but as we now excluded the behavioural runs as a performance indicator, this does not change the observed performance of the models.

McKay, M. D., Beckman, R. J. and Conover, W. J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics, 21(2), 239, doi:10.2307/1268522, 1979.

- The way the structures versions are selected is unclear. Is this based on results in calibration or in validation? Actually these two options should be tested and discussed.
To avoid a selection bias, we reserved strictly a period of the dataset that is never used in any selection process. The validation period that is only used as a final check for overfitting. As for any model calibration study, the division of calibration and validation period is in the end arbitrary. We have considered to use multiple calibration periods, but rejected this approach in favour of longer time series. To clarify the meaning of the validation period we included as the second sentence in 2.5: *The validation period is strictly not used in any selection process to avoid overfitting and only used in the last validation step of the overall method*. And further in the chapter we extend: *We used the Generalized Likelihood Uncertainty Estimation (GLUE) methodology (Beven and Binley, 1992) to find behavioral parameters sets for the calibration period.*

Furthermore, how a model structure is judged to be significantly better than another? Is there any threshold in model improvement or statistical test associated?
We realize that is not completely made clear how the model structures were selected. We have not made clear enough that it is not the “good” model we need for the decision, but to collect the reject models. Following other comments related to this, we emphasized this process more in the Introduction:
A subprocess is marked as necessary, when models lacking it are rejected. On this base, a subsequent model is constructed which uses only meaningful subprocesses. Incremental model breakdown is therefore a rejectionist approach, built on the learning from failure and not an optimization process. Beven (2005) assumed that a rejectionist approach is generally better suited to gain insight about process hypotheses.
and Material and Methods:
A model is rejected, when it is not able to produce runs of acceptable performance for all parameters. And rejected means in this study, the model is missing a process to important to ignore. If a model lacking a certain subprocess is able to produce behavioural runs that subprocess is irrelevant for this application.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320(1–2), 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

The robustness of the structure selection should be discussed. The model structure is selected based on the use of the first period as calibration and the second as validation. I think the authors should at least test the procedure by inverting the role of the two periods. It is likely that the structure selection may end up (maybe not on this catchment but there are probably cases where it may happen) with different model structures in the two cases.

We decided to reserve a strict validation period, never used in any model selection. By switching the periods around the independency of the validation period is lost. However, we discussed the role of the validation period more in chapter 4.3: *Incremental model breakdown bears, as any model intercomparison study of calibrated models, a risk of overfitting. In the context of this study, overfitting would results in the acceptance of a process that seems only by chance relevant in the calibration period, but has only weak predictive power. Another overfitting effect would be a preference of parameter rich models. An indicator for overfitting are great results in the calibration period but flawed results in validation. This shows the importance of a validation period that is never used in any selection process, neither for structure nor for parameters. In this study, the performance of the models during validation generally exceeded the performance of the calibration period, despite the different characteristics of those periods.*

This raises the problem of equifinality in the choice of model structures, and may be a limit of the proposed approach. The selected structure may be overspecialized for the selection period and not really transposable on periods with other conditions. This is what can be observed in the case of model parameters and it is probably also the case in terms of structures. This is probably even a larger problem for periods with much contrasted characteristics.

From the original manuscript it might not be clear, that the validation period is differs slightly from the calibration period. We add, also in reponse to reviewer #1 a new fig. 2 to show the differences between the periods. All models with behavioural runs produce accepted results in the validation period and rejected structures are rejected in the validation period also. We see this as a strong indicator for transferability between time periods, and hope that we clarified this issue with the discussion section above. The GLUE answer to the problem of equifinality is to drop the search for the best model parameterization to accept and instead search for parameterizations to reject. In this study, we transfer this approach to model structures and gain information from the model structures that fail and not from the models that work. The rejectionist approach has been clarified in Material and Methods as given above. Also we are discussing this now in chapter 4.3:

A second effect when both structural and parameter uncertainty are to be compared, we are not only facing an equifinality of parameter sets but add equifinality of structures. We based the recognition of relevant processes on the rejection and not the optimization of certain model structures, as suggested by Beven (2005) to gain a robust method.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320(1–2), 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

- The authors did not really discuss the respective roles of structural and parametric complexity in the results. At the end, they have a much more simple structure than at the beginning but which still has ten parameters, which may appear as overparameterized at the daily time step. It may be interesting to have even more simple model structures, to see how the further simplification possibly leads to degradation in the modeling.

We agree, but see this suggestion as part of future work. This paper is mainly meant to introduce the idea of model breakdown and not to find the “best” model possible. In further studies, it could be tested, to which results it would lead to make several iterations of the incremental model breakdown approach. Testing which processes are the most important for the model, reducing the structure to those processes, define a harder boundary for behavioural runs and repeat the process.

- The authors criticize the usual approach which takes existing models, with interesting arguments. To further demonstrate the value of their approach compared to the classical one, they could test an existing model (e.g. HBV or another model of this type) as a benchmark, to explain the added value of their approach compared to the case when one simply take an existing model.

We now included HBV-Light as a benchmark model and added several sections to explain it.

Material and Methods: *As there have not been many studies regarding the construction of models via modelling frameworks, this study uses HBV-Light as a benchmark to make results more comparable with non-framework studies and to allow a more precise evaluation of the performance of the proposed incremental model breakdown method. HBV-Light is a widely used model, which has proven its functionality in very diverse catchments [Seibert and Vis, 2012]. It is a lumped, parsimonious model. We used the simplest setup of HBV-Light with a single soil storage and no lapse rate. As HBV-Light has no internal way to calculate potential evapotranspiration, we used the same approach by Samani [2000] as for all other models.*
Results: *HBV-Light performs best of all models in this study. Its performance increases from the calibration to the validation period, especially in regard of the maximal values of the objective functions (Table 3, Figure 4). The largest differences manifest in the values for the RSR and the NSE between HBV-Light and the other models. However, HBV-Light seems to have problems in simulating the base flow of the Fulda catchment, resulting in a worse value for the logNSE in comparison to the other models. Here the performance is similar to Model 15. Also, HBV-Light has a very wide range for the values of the objective functions in the validation period, hinting to a large parameter equifinality.*
Discussion: *All three models show a distinct behavior (Figure 5), with HBV-Light and Model 15 behaving rather similar. The main differences between the models are the ability to predict the peaks, an over/underestimation of base flow and the shape of the hydrograph in general. Model 1 captures the shape of the low base flow best, while HBV excels at simulating the peaks. Model 15 is somewhere in between. Those differences are probably caused by the number of storages in the models and processes that mimic saturation excess.*
Model 15 and HBV-Light are quite similar with regard to their model structure and the considered hydrological processes. The main differences in model performances is the way the mathematical process descriptions are implemented. HBV-Light has a maximal value for percolation and the triangular weighting function that changes the shape of the flow curve (Seibert and Vis, 2012) .With the maximal value for percolation, additional water is forced to become discharge, as there is no other way it could go. This allows HBV-Light to forecast the peaks better, but also might make the model react too quickly. This behavior though is counteracted by the triangular weighting function of HBV-Light. In contrast, Model 15 predicts the peaks correct only during times of low evapotranspiration. Another main difference exists for the simulation of base flow. Model 1 depicts a highly correlated base flow to the observed one, but the model is overestimating the total amount. Model 15 and HBV-Light mimic the shape and timing of the low flow worse, but predict the amounts better. One reasons for this behaviour might be that a model needs a good representation of the groundwater to simulate discharge minima (Plesca et al., 2012). This is the case for Model 1, but only to a lesser extent for HBV-Light and Model 15.

- Last, I find that making the test on at least a second catchment with contrasted characteristics may strengthen the conclusions. Here the results may be obtained only by chance. There is no guarantee that the results are general outside this case study.

We agree that an additional catchment might yield some interesting results as well. However, as stated in several studies, lumped models need to be tailored for every catchment separately (see e.g. Kavetski and Fenicia (2011) and Fenicia et al. (2014). Therefore, it is to be expected that different catchment would lead to different model structures. Still, we think that this is a worthwhile endeavour for future studies, but would go beyond what can be done for the work presented her.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L. and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrol. Process., 28(4), 2451–2467, doi:10.1002/hyp.9726, 2014.

Kavetski, D. and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights: Flexible framework

for hydrological modeling, 2, Water Resour. Res., 47(11), doi:10.1029/2011WR010748, 2011.

I also have other comments detailed below. In summary, I think there is valuable material in the article, but that the methodology should be further tested and more thoroughly evaluated to provide a more convincing demonstration of its usefulness. I suggest major revision.

Detailed comments

1. P2,L28: This is probably true for all modelling approaches!
This is true and the sentence is removed.

2. Section 2.1: Say in which country the basin is located. Maybe a location map could be added. Is catchment size actually 2.977 or 2,977 km²?
We apologize for this error. The Fulda Catchment is almost 3,000 km² large. As proposed, we added a map to make this more clear (see response to reviewer #2).

3. P4,L10: I find that the definition of a process in the structure should be given. When a process is removed, what happens in the connections in the structure, especially when there are several branches coming to/departing from this process?
To better explain the removal of processes we added the following sentences to the Material and Methods section: *When a process was removed, the connections leading to it were then connected to the next nearby storage. For example: If the surface storage was removed, the Canopy and the Snow Storage were connected to the soil. Or if the river was removed, all connections leading to it were directly connected to the outlet.*

4. P6,L18-19: As mentioned in the major comments above, I think that it should be explained how a version is considered to be significantly better or worse than another.
This comment is dealt with in the answer to major comment.

5. P9,L3-8: Why there is not a pure time-delay parameter (possibly non integer) in the model that would be added in the model structure to account for this time shift and to make it more generally applicable?
One approach is the routing with a time delay, which we considered in model 1 where we included a “river” storage which simulated a behaviour with a retention time. But this showed to be less appropriate, as Model 1 was not being able to produce behavioural runs with this process included. Our approach of shifting the time series by one day is another viable option, see Bosch et al. (2004) or Asadzadeh et al. (2016).
We would like to stress that routing or shifting does not affect the idea of our paper, which is presenting an alternative blueprint for hydrological model set up rather than a case study and best model practice for the Fulda river.

Asadzadeh, M., Leon, L., Yang, W. and Bosch, D.: One-day offset in daily hydrologic modeling: An exploration of the issue in automatic model calibration, J. Hydrol., 534, 164–177, doi:10.1016/j.jhydrol.2015.12.056, 2016.
Bosch, D. D., Sheridan, J. M., Batten, H. L. and Arnold, J. G.: Evaluation of the SWAT model on a coastal plain agricultural watershed, Trans. ASAE, 47(5), 1493–1506, doi:10.13031/2013.17629, 2004.

6. P9,L12-14: Please remind in brackets for each criterion the optimal value and range of variations, to avoid misunderstanding in the interpretation of results for readers not fully familiar with these criteria.
Added as proposed.

7. Table 3: Please add a column for units. Maybe also add a column to remind in which structural element (as defined in Table 1) each parameter is included. In the caption: “all model parameters”

We included a column for units as proposed.

8. P11,L11-12: Is not that expected by construction that all model structures have less parameters than the original one?
This is true. We deleted the sentence.

9. P16,L8-15: This seems to repeat the last paragraph of the previous page.
Deleted all repeated sentences.