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 #1 (L.A. Melsen)

Jehn et al. provide a case-study of incremental model-breakdown; starting off with a (benchmark) model including a high number of processes (and parameters), the model is compared to models where fewer processes are explicitly represented. Finally, based on this information, a simplified model is presented with a higher model performance than the benchmark model. The manuscript is well written and well-structured, and the figures and tables are to the point. I liked the fluxogram.
We created a citable repository for the fluxogram and its code which is now referenced in the paper (http://doi.org/10.5281/zenodo.1137703).

There are, however, some questions, especially about the rationale, that I think need to be addressed, and the results and discussion sections are limited. This can improve with a more in-depth analysis of the results, for which I provide a (first) suggestion. About the rationale: In the introduction and the conclusion the ‘incremental model-breakdown’ is presented as an alternative next to step-wise model building and comparison of pre-defined structures.
1) My intuition would be to conduct a sensitivity analysis, and based on that determine which processes are relevant and which are not. What is the advantage of doing the incremental model-breakdown rather than a sensitivity analysis? (except that the parameter is completely removed from the model rather than fixed at some point).
“Incremental model-breakdown” has the same aims as a sensitivity analysis. The main difference is, as you state, that a process (together with its parameters) is removed completely and does not remain in the model anymore within the incremental model-breakdown. We see this as advantageous, as it reduces the structural complexity of the model. To make this clearer we added the following sentence to the discussion: *“In this regard, the incremental model breakdown is different to methods like sensitivity analysis where the model structure is untouched, as we reduce the structural model complexity.”*

2) Another alternative, besides the incremental model-breakdown, step-wise building, and pre-defined structures, is to replace formulations of certain processes with alternative formulations, for example the SUMMA framework which you cite (Clark et al, 2015ab). How does the incremental model breakdown compare to this approach?
Step-wise model building has the same goal as the incremental model breakdown: Finding the right model. However, the focus differs. The SUMMA approach (which is entirely possible using CMF as a base framework) deals with the question: “What is the best formulation of a process in a given model?”, while the question for the incremental model breakdown is: “What is the best overall structure for a model in a given catchment?”. In future studies it would be worthwhile to combine both approaches, to get an even more thoroughly exploration of the catchment. To clarify this, we added the following sentence to the introduction: *Clark et al (2015ab) propose with the SUMMA concept another approach to test multiple hypotheses. Their question is: do we use the right formulation for this process? This study asks instead: Is the process relevant for this catchment at all?*

3) What is the added value of the incremental model-breakdown compared to all the alternatives? p.2,l.28 states that only a minor quantity of the vast space of possible model structures is explored, but isn’t this also true for the incremental model-breakdown as presented in the manuscript, since only a single ‘complex’ model was employed?
We agree that this was ambiguously worded. It is clear that our approach will not be able to sample the space of possible model structures exhaustively. Nevertheless, we think that incremental model-breakdown samples a larger part of the potentially available model

structure space than most other approaches, as we start with a very complex model structure (complex in the realm of lumped models), containing all processes which seem to be important for the catchment and trim it down sequentially. This way, more processes might be considered, as when starting with a simple model structure, adding pieces and settle for a structure once a sufficient value of the objective function is reached. To clarify, we added the following line to the conclusions: *From the surface, the water is either directly routed to the river or enters three serial soil/groundwater layers, which in turn would allow a completely exhaustive exploration of the space of possible model structure*. And the following sentence to the introduction: *While still not being able to sample the entire space of possible model structures, this approach might find some model structures which are likely missed with other methods.*

Main points:
The model was run with a daily time step for a catchment in the order of 3000 km2. As becomes clear later on (section 2.5), the response time of the catchment is less than a day. How do you expect this influences your results?
Obviously, this temporal resolution is not sufficient to capture the dynamics of the catchment. (follow up on that; It is unclear to me why you had to move the time-series; the river-part could easily be implemented as a routing with a time delay rather than a storage-system, which is more common for rainfall-runoff models).
We agree that a daily time step is insufficient to model all subdaily dynamics in mesoscale catchments. However, most of the public hydrological data are only available on a daily time step. Therefore, modellers have to cope with it. 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.

The discussion of equifinality in the manuscript seems incosistent. Generally, the risk on equifinality is higher with more degrees of freedom (more parameters compared to the information in the available data for calibration). But on page 13, l.3 is written: ‘[incremental model-breakdown]..have a positive impact on model performance, given the increased number of behavioural runs’. So; more behavioural runs is positive? But also an implication of equifinality? On p.14, l.11 it states ‘[incremental model-breakdown]. is a good way to improve model performance and reduce equifinality’. Please clarify. This relates to my next point: is it a fair comparison to take the mean of the behavioural runs? I have not figured it out myself completely yet, but I don’t see why a particular model should be ‘punished’ for having more (or less) behavioral parameter sets, see e.g. p.16, l.3-7. Perhaps consider another metric to compare the models.
Our use of the term “behavioural” was indeed inconsistent. We deleted this section and rephrased all other sections where the term “behavioural” was used in this way. We further do not use the number of behavioural runs as a performance indicator anymore. To further increase the quality of the evaluation we now included the NSE as fourth objective function particularly focusing on model performance for higher flows.

p.9, l.20; for every model, a LHS of 300.000 is taken, despite the number of parameters. So, for models with fewer parameters, each parameter is sampled more often. This could explain why the more frugal models (fewer parameters) have more behavioural runs. Do you think this is the case?

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, our procedure might allow models with less parameters to get more behavioural runs. Therefore we now excluded the number of behavioural runs as a performance indicator.

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.

Please add a motivation why you chose these three objective functions. None of the objective functions focusses on high flows, but still peaks and high flows are continuously discussed in the results and discussion section (e.g. p.13,l.17), while low flows are not discussed at all.

We agree and see this shortcoming. Therefore we now included the NSE in our multi objective calibration approach as a fourth objective function. Accordingly, we added respective parts in the methods, results and discussion sections.

Can you provide an order of magnitude for the drinking water abstraction? The process is included in the model because water is abstracted for 80,000 inhabitants (p.4,l.2) but turns out to be unimportant, possible because of low population (159 persons per km2, p.15,l.10). In other words: where did you base the min and max parameter boundaries for drinking water extraction on? (Table 3)

As we did not find reliable data to quantify the influence of the drinking water abstraction, we decided to include a subjective estimation, to test whether there is a potential influence on the water flux estimation and if so, how large this influence is. In the revised version of the manuscript, we state this up-front: "*As the influence of the drinking water abstraction is not known, the amount of water abstracted is calibrated*". As it turns out, drinking water abstraction is of marginal influence in the catchment for the annual water balance.

In general, please provide references or motivation how and why you defined these boundaries for your parameters (Table 3).

We included the following section in the calibration and validation section to explain the parameters and also present units for all parameters in Table 3:

The lower and upper bounds for V0_soil and ETV1 were taken from Blume et al. (2016) for typical field capacities reported for German soils in the range of 20 to 300. Canopy parameters are in line with values provided by Breuer et al. (2003). Groundwater transit times are roughly corresponding with Wittmann (2002) and Wendland et al. (2011). For all other parameters we could not find reliable data and thus estimated them subjectively. The parameters use a wide range intentionally to allow the parameters to adapt to the very different model structures.

Blume, H.-P., Brümmer, G. W., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., Wilke, B.-M., Scheffer, F. and Schachtschabel, P.: Kapitel 9: Böden als Pflanzenstandorte, in Scheffer/Schachtschabel Lehrbuch der Bodenkunde, Springer Spektrum, Berlin Heidelberg., 2016.

Breuer, L., Eckhardt, K. and Frede, H.-G.: Plant parameter values for models in temperate climates, Ecol. Model., 169(2–3), 237–293, doi:10.1016/S0304-3800(03)00274-6, 2003.
Wittmann, S.: Tritiumgestützte Wasserbilanzierung im Einzugsgebiet von Fulda und Werra, <http://www.hydrology.uni-freiburg.de/abschluss/Wittmann_S_2002_DA.pdf>, Diploma-Thesis at the Institut for Hydrology, Albert-Ludwigs-University Freiburg, 2002.
Wendland, F., Berthold, G., Fritsche, J.-G., Herrmann, F., Kunkel, R., Voigt, H.-J. and Vereecken, H.: Konzeptionelles hydrogeologisches Modell zur Analyse und Bewertung von Verweilzeiten in Hessen, Grundwasser, 16(3), 163–176, doi:10.1007/s00767-011-0169-6, 2011.

To continue on that, I would also like to suggest for further analysis; why not showing the distribution of the parameters for the different model formulations? I would be interested to see if any of the parameters is taking over the job of one of the parameters that has been left out. This would result in a shifted parameter distribution. If I may undisclosed refer to my own work; see figure 8 in https://doi.org/10.5194/hess-20-2207-2016

This is a good addition to the paper. Therefore, we now created a parameter distribution plot for the parameters shared by Model 1 and Model 15, to enable a more thoroughly comparison. To explain this plot, we added the following sentences to the results:

The remaining ten parameters in Model 15 behave different from the same ones in Model 1 (Figure 5). Some parameters like tr_soil_GW and fEVT0 have almost the same density distribution. Still, there are several parameters like tr_soil_river and ETV1 whose density is much more focused around a specific value for Model 1 than for Model 15.

Regarding the discussion:

Model 1 has less equifinality in some parameters, compared to Model 15 (Figure 5). E.g. the parameter ETV1 has two very distinct peaks for Model 1, while for Model 15 the distribution for this parameter is widely spread. The behavior of ETV1 might also be linked to the rightward shift of the parameter beta_soil_GW. This parameter controls the speed in which water leaves the soil in the direction of the groundwater. The increase in its value lets the water stay longer in the soil storage, allowing more evapotranspiration, which in turn allows the parameter ETV1 be handled more flexible by the model.

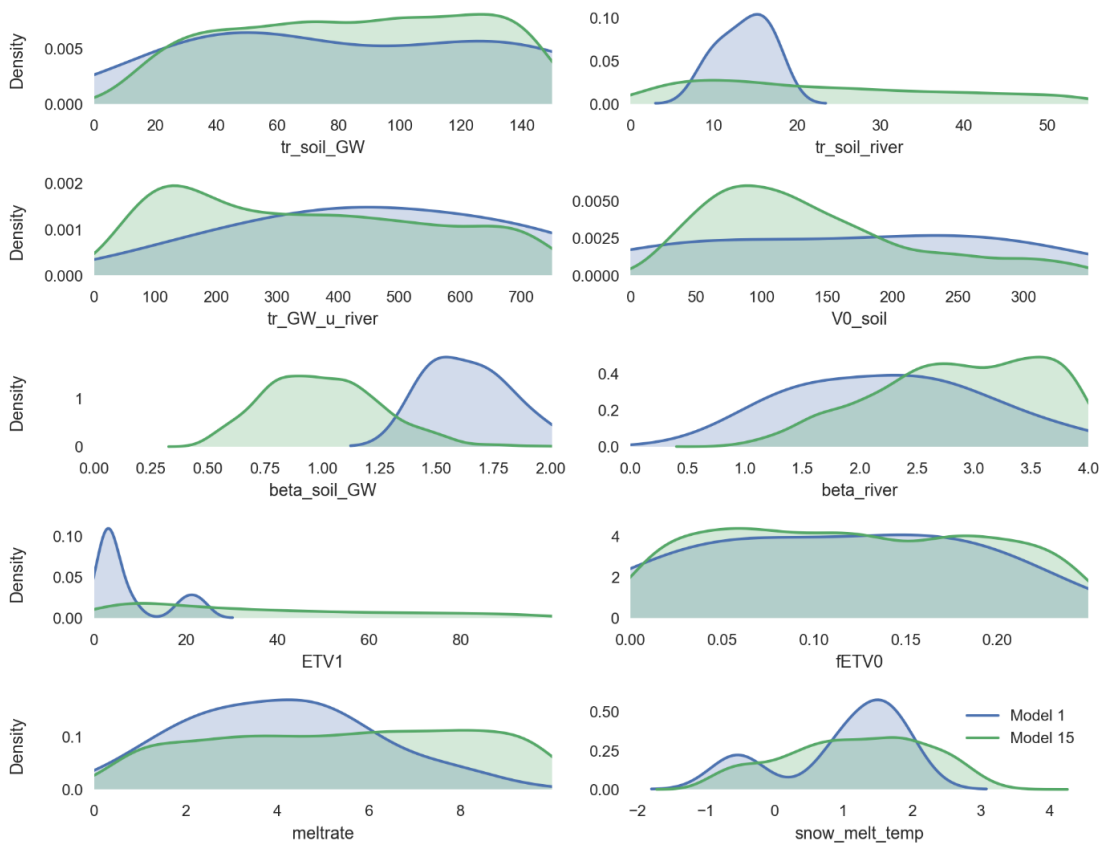

Figure 1: Distribution of all parameters shared by Model 1 (blue) and Model 15 (green), fitted with kernel density.

Then, it would be interesting to see which parameters compensate for which processes. The manuscript lacks a discussion of how the calibration period relates to the validation period. More parameters could fit better in the calibration period but can be flawed in the validation, which is something that should be discussed in relation to model complexity and number of parameters (see Kirchners paper on being right for the right reasons).

How do models compensate lack of realism is indeed a highly interesting question. We think this is better to show exemplary for selected processes in a smaller catchment were the relevant processes are known and merit a study on its own.

The main criteria to determine which processes were important, was the ability of model to have any behavioural model runs at all. To understand the influence of the model structure on the parameters we included Figure 1. We did this only for Model 1 and Model 15, as those are the most important models in the study. To make it more clear on what the process selection was based we added the following sentence to the Material and Methods section:

The main criteria to determine the value of a process was the ability of the model to produce behavioral runs in the calibration period at all.

To better explain the differences between the calibration and the validation period we added a cumulative sum plot for the precipitation and discharge (Figure 2). With this it is more clear that both periods are different. In addition to the figure, we added the following text to the Material and Methods section:

The model time step and temporal resolution of the data are both daily. Both the validation and the calibration period behave differently in regard of their patterns of precipitation and discharge (Figure 1). The calibration period is wetter and contains six of the seven large rainfall events (>30 mm d^{-1}). In addition, in both periods there is one year representing contrasting extreme weather conditions. In 1985, during the calibration period, very little discharge is observed with at the same time high precipitation, while in 1988 during the validation period high discharge was recorded at comparingly low precipitation.

Also we added the following sentences to the end of the discussion:

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.

All in all, Incremental model breakdown and inspection of parameter distribution, as well as comparison with already established models and the flowpath in a model with a fluxogram might help determine if models do the right things for the right reasons.

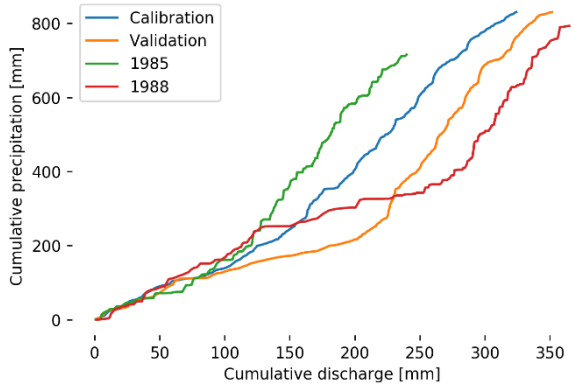

Figure 2: Cumulative discharge plotted against cumulative precipitation for the calibration and validation period and two years with extreme behavior. For the calibration and validation period, the cumulative discharge and precipitation are the average of the corresponding years.

Other points:
Please mention the model time step and the temporal resolution of the input data in the 'model input and validation data' section.
Added as proposed.

Please check the units of Eq.1. V0 is a volume (p.4,l.24) but has the units of a rate. What are the units of V and Q?
We changed the section to make it more clear. The description of the kinematic wave in CMF is now more exactly described by discarding *tr* and introducing *Q0*, which is the flux in m³ per day when the volume of the storage equals the parameter V_0.

Calculating the mean for a NSE is tricky since the NSE is highly non-symmetrical (+1 to minus infinity). Consider using the median.
Changed as proposed. Boxplots use the median now.

Figure 3, caption. I think the word 'uncertainty' in 'uncertainty of the behavioural model runs' is not in place here. All you look at is the spread in your behavioural runs, which is certainly different from uncertainty.
Changed "uncertainty" to" range".

The same holds true for p. 17, l. 1, 'less uncertain'
We realized that this wording is confusing and deleted it.

p.2, l.21 comparability -> comparison
Changed as proposed.

p.4, l.6 unnecessary brackets around CMF, 2017
Changed as proposed.

p. 10, table 3; caption; ‘indented’, in the table: ‘intendent’ -> intended
Changed as proposed.

p. 16, l.8-13 repetition of p. 15, l.26
Deleted all repeated sentences.