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 #2 (F. Anctil)

In this paper, submitted by Jehn et al., a breakdown approach is proposed in order to simplify a complex model into a structure with “improved model performance, less uncertainty and higher model efficiency” (line 17, page 1). The method is validated on a 3-year time series from a single gauging station in Germany.

General comment:

The main argument in favour of experimenting with the proposed incremental model breakdown is that it may lead to a better model than the more common stepwise bottom-up approaches, arguing that “there is a chance that they have missed an even better model performance by including further modifications” (line 28, page 2). Yet no comparison with a stepwise model building is presented, providing no evidence that a breakdown approach is superior.

We do not see the incremental model breakdown as being superior to the other approaches, but more like another way to explore possible model structures. The main difference is that incremental model breakdown tries to explore the model space another way by turning the stepwise process upside down.

A direct comparison of both approaches by the same set of authors would not work, as experience from one approach will inevitably influence the decision of model building during the other approach. For future work, it might be a worthwhile idea to give two separate research groups the same information about a catchment and let them built a model: One group using incremental model breakdown and one group using stepwise model building. Finally, both resulting model structures are compared in their performance and structure. However, for the current work presented here, which focuses on the general idea of incremental model breakdown, such a comparison would go beyond the scope of the paper.

Major comments:

There is possibly some confusion on the size of the watershed, which drains only about 3 km2 according to line 14, page 3. It is more likely that the size be 2977 km2 and not 2.977 km2, in order to accommodate 108 meteorological stations and an altitudinal range from 150 to 950 m a.s.l. A map of the watershed would have allowed to clarify this issue. It is recommended to add one.

This was a typo. The Fulda Catchment is 2977 km2 in size. We now added a map (Figure 1).

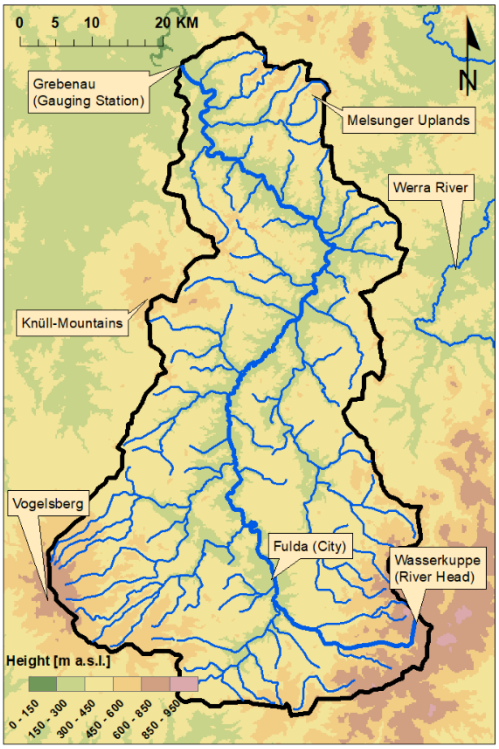

Figure 1: Relief map of the Fulda catchment for the gauging station Grebenau (black border).

Lumped hydrological models often need shorter time series for calibration than distributed ones. But in the context of a research on the selection of structural components, I am surprised that only 6 years of data was selected for calibration and only 3 more for validation (line 3, page 9). This needs to be justified. Longer series offer the advantage of stabilizing the results in regards to climatological variability. Were there no data available after 1988? At least, the authors need to inform on the climatology of the calibration and validation datasets in regard to the general, say, 30-year climatology. For instance, models usually work much better in wet years than in dry years. Was it the reason for selecting observations from the 80's?

We thank the reviewer for this comment, but we have a different opinion on this point. Indeed, a longer time series contains more climatic variability. However, a good model should be able to cope with climatic variability, as its inner structure should resemble the real processes in the catchment. This viewpoint is also shared for example by Kirchner (2006) or Klemeš (1986).
Uncertainty about rainfall is one of the major sources of model uncertainty. To reduce this uncertainty, we selected the time period with the greatest number of rainfall stations without missing data relevant for our study area. Any longer or later time series would result in a strongly reduced number of stations. To better describe the data we used, a figure on cumulative discharge and precipitation is now included (see also response to reviewer #1). We also added the following sentence: *Still, the precipitation stays in the long term range for this catchment for all years (Fink and Koch, 2010).*
Finally, we would like to add that the objective of this paper is not to find the "best" model for the Fulda catchment in the sense of a case study, but present a new way of model building using a rejectionist approach.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42(3), doi:10.1029/2005WR004362, 2006.
Klemeš, V.: Operational testing of hydrological simulation models, Hydrol. Sci. J., 31(1), 13–24, doi:10.1080/02626668609491024, 1986.

The authors should also avoid vague statements like “climatic conditions during the calibration (1980-1985) and validation period (1986-1988) were rather similar” (line 31, page 3). Chances are that they are not so similar at least in terms of low flows, otherwise how can one interpret the raise in validation logNS values in Table 2, in comparison to their calibration counterpart.

We deleted the statement and added additional data and Figure 2 (in response to reviewer #1) to communicate the climatic conditions more clearly. See also replies to the comment above.

The issue of shifting the simulated discharge one day into the future to improve overall performance (line 8, page 9), thus simulating Q(t+1) instead of Q(t), typically falls from some failure in the routing components of the model and sounds more like fudging than modelling. What is the operational consequence of that trick? The argument that rainfalls occur in the “later time of a day” is weak and needs to be substantiated. This information should be included in Figures 1 and 4.

We still think this is a valid method. We now included further information in respective figure captions.

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.

GLUE is a convenient tool to assess the level of the parameter uncertainty of a model and to identify a number of equifinal (behavioural) parameter sets. Its use here as a calibration tool needs to be better justified (line 13, page 9), for example in comparison to more operational calibration schemes.

We used GLUE as it is widely recognized in the hydrological community and gives a clear statement in regard of the model’s capabilities to accomplish predefined criteria. As we were not aiming to find a single best parameterization of our models but rather scrutinize the associated parameter space of our models, we still think that GLUE is an appropriate tool for this question. To make this clear, we added the following sentences to the calibration and validation section:

It should be noted, that other calibration schemes, objective functions and parameter ranges might have lead to different results. However, we are not striving to find the best performing parameter set. Instead, we uses GLUE for the identification of behavioral model runs to evaluate the various model structures.

Here, models variants are essentially compared in Table 2 on the basis of their number of behavioural runs that surpass three thresholds advocated by Moraisi et al. (2007), while parameter uncertainty is not explored. In practice, this has two limitations. 1) No performance information is provided for models 2, 3, 5, 8, and 10, for which the suppression of a structural component turned out detrimental. The issue is that we are provided no information on how much detrimental this operation is, which is quite important to the manuscript since model 15 is essentially built around them.

To our understanding, the parameter uncertainty of models 2, 3, 5, 8 and 10 is not worth to consider any further. None of the tested parameter sets has achieved the thresholds of the predefined objective functions. So why bother to evaluate these models? Therefore, we applied the SPOTPY software in such a configuration that unbehavioral model runs are not saved for further analyses. With the now included boxplots for all models with behavioural runs it is also more obvious that all models with behavioural runs were a good deal better than those without, were all model runs are below the lower whisker of the boxplots.

2) A small gain in performance may lead to a large increase in the number of behavioural runs. Information in Table 2 is not that informative because it reflects only the behavioural runs. For instance, we are told that model 13 should be dismissed even if its metrics are better than model 15, because of a much lower number of runs to compute metrics (line 4, page 16). It would be easier to address that by giving all the information (not just the mean and the standard deviation) for example in the form of a box plot. From an operational point of view, hydrologists are looking for the best possible model, and variant 13 may fit their needs better than variant 15.

This is a very valuable comment. We therefore deleted table 2 from the paper and added all behavioural models as boxplots (Figure 2, 3). It is now more obvious that Model 15 delivers runs with much higher values for the objective functions than Model 13. Descriptions referencing to table 2 are changed accordingly.

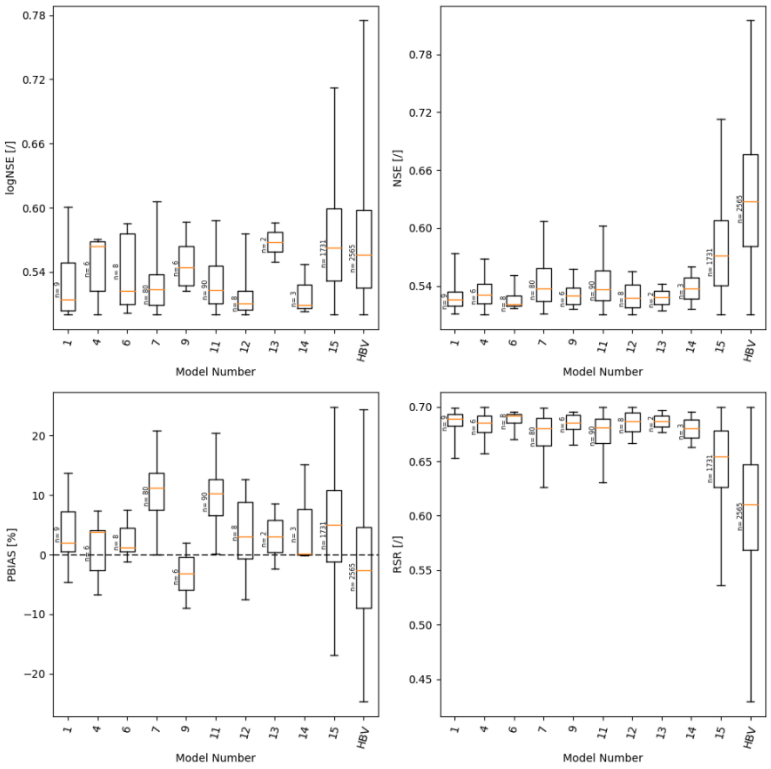

Figure 2: Boxplots of the objective functions for all models with behavioural runs in the calibration period. The yellow bar marks the median. Number of behavioural runs noted on boxplot.

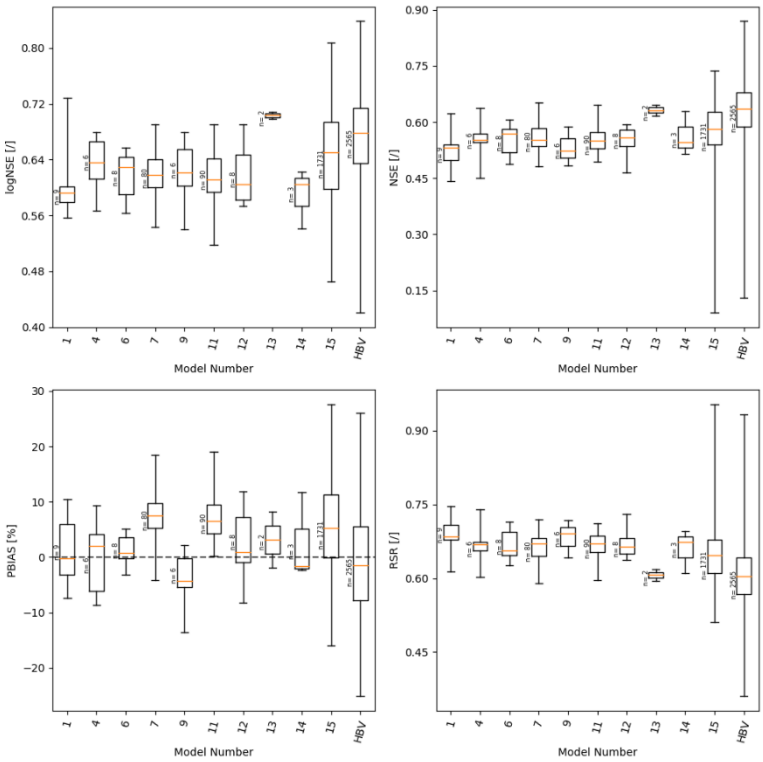

Figure 3: Boxplots of the objective functions for all models with behavioural runs in the validation period. The yellow bar marks the median. Number of behavioural runs noted on boxplot.

Minor comments:

Are the authors aware of any other hydrological studies on the same site that could offer some basis of comparison?

We found one additional conference paper by Fink and Koch (2010) which we added. The only other study that we are aware of by Wittmann et al. (2002) has already been cited.

Figure 2 is not much useful.

This is true. Therefore, we deleted the figure from the paper.

Figure 3 would be more intelligible if it would be split in two: a figure for model 1 and another one for model 15

We see the problem, the referee is mentioning. Having two time series in one plot is always a bit difficult to look at. Nevertheless, we think that it is necessary here to have both models in the same plot, as it allows a better comparison of the two.