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.

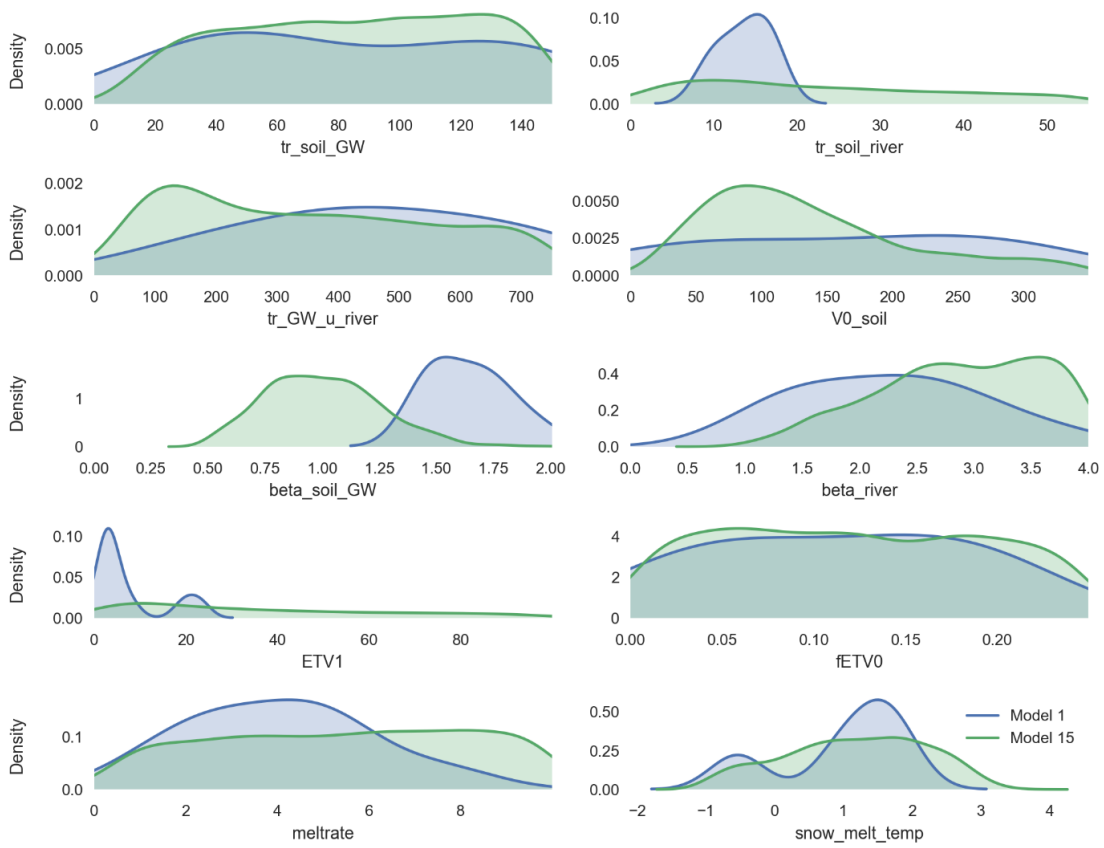


Figure 1: Distribution of all parameters shared by Model 1 (blue) and Model 15 (green), fitted with kernel density.

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.

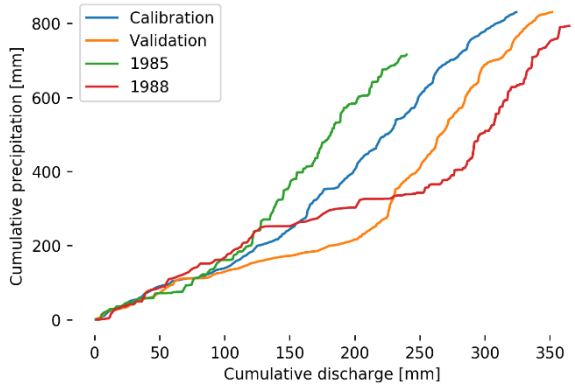


Figure 2: Cumulative discharge plotted against cumulative precipitation for the calibration and validation period and two years with extreme behavior. For the calibration and validation period, the cumulative discharge and precipitation are the average of the corresponding years.

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.

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).

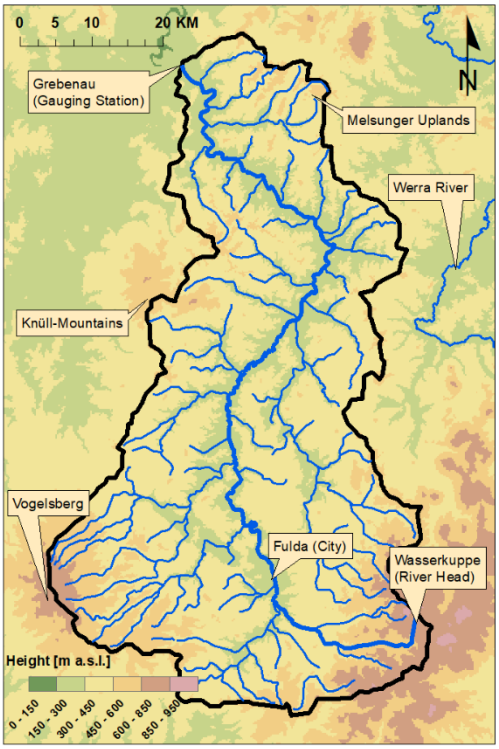


Figure 1: Relief map of the Fulda catchment for the gauging station Grebenau (black border).

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.

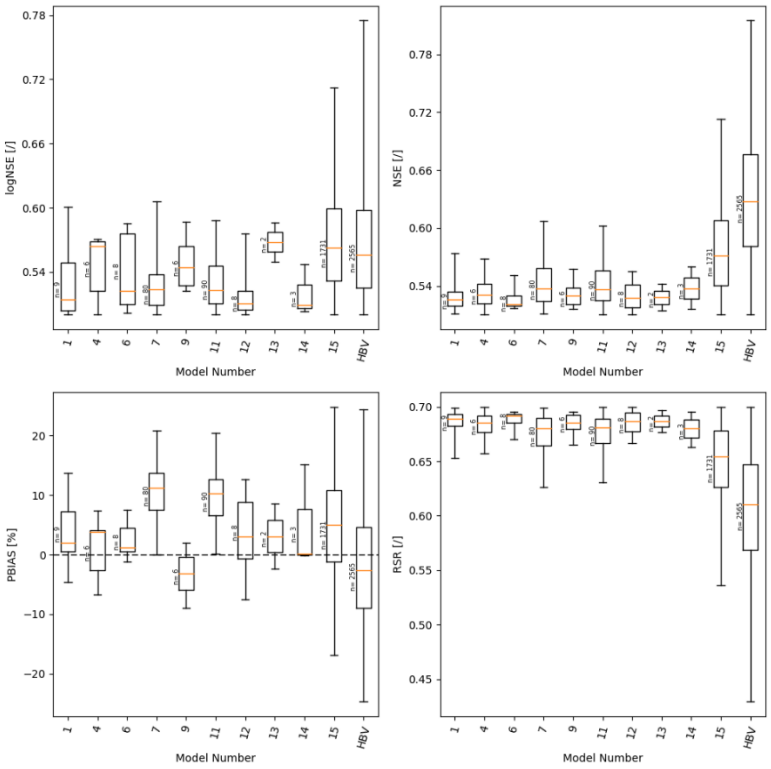


Figure 2: Boxplots of the objective functions for all models with behavioural runs in the calibration period. The yellow bar marks the median. Number of behavioural runs noted on boxplot.

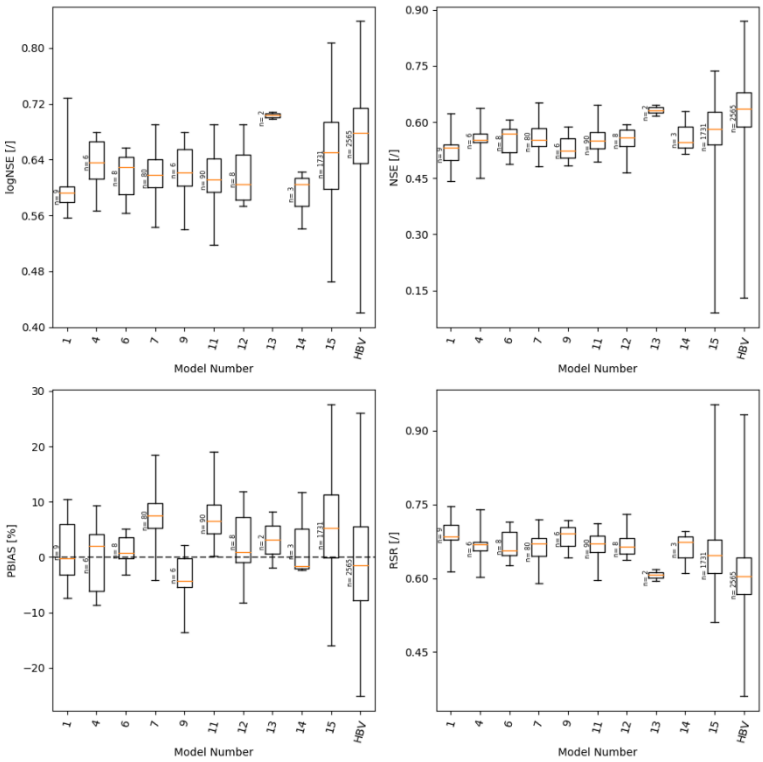


Figure 3: Boxplots of the objective functions for all models with behavioural runs in the validation period. The yellow bar marks the median. Number of behavioural runs noted on boxplot.

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.

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.

Incremental model breakdown to assess the multi-hypotheses problem

Florian U. Jehn[1], Lutz Breuer[1,2], Tobias Houska[1], Konrad Bestian[1], Philipp Kraft[1]

[1]Institute for Landscape Ecology and Resources Management (ILR), Research Centre for BioSystems, Land Use and Nutrition (iFZ), Justus Liebig University Giessen, Heinrich-Buff-Ring 26, 35390 Giessen, Germany
[2]Centre for International Development and Environmental Research (ZEU), Justus Liebig University Giessen, Senckenbergstrasse 3, 35392 Giessen, Germany

Correspondence to: Florian U. Jehn (florian.u.jehn@umwelt.uni-giessen.de)

Abstract. The ambiguous representation of hydrological processes have led to the formulation of the multiple hypotheses approach in hydrological modelling, which requires new ways of model construction. However, most recent studies focus only on the comparison of predefined model structures or building a model step-by-step. This study tackles the problem the other way around: We start with one complex model structure, which includes all processes deemed to be important for the catchment. Next, we create 13 additional simplified models, where some of the processes from the starting structure are disabled. The performance of those models is evaluated using three objective functions (logarithmic Nash-Sutcliffe, percentage bias and the ratio between root mean square error to the standard deviation of the measured data). Through this incremental breakdown, we identify the most important processes and detect the restraining ones. This procedure allows constructing a more streamlined, subsequent 15th model with improved model performance, less uncertainty and higher model efficiency. We benchmark the original Model 1 ~~with~~ and the final Model 15 with HBV-Light and find that the incremental model breakdown leads to a structure with good model performance, fewer but more relevant processes and less model parameters.

1 Introduction

In the world of hydrological modelling, scientists construct models and apply them for a specific research question. Sometimes, these models are modified or extended afterwards, but the core components stay the same. This approach has existed from the earliest days of simple equations until the models of connected, conceptual elements used today (Todini, 2007).

During the development of hydrological models, the issues of parameter and input data uncertainty were often in the center of the scientific debate and numerous methods for assessing this uncertainty have been proposed. Structural uncertainty has been investigated in the past decade (Breuer et al., 2009; Son and Sivapalan, 2007) and gained more momentum in the last few years (e.g Clark et al., 2015; Fenicia et al., 2011; Hublart et al., 2015). It was noted that problems often arose from the focus on trying to build one model that was meant to work equally well for all catchments (Fenicia et al., 2011).

In order to better scrutinize problems associated with the model structure, the theory of the multiple hypotheses was introduced, first by Beven (2001, 2002), and more recently picked up by Clark et al. (2011). This theory enables a more structured approach

to model building, as it identifies a given model not as a single hypothesis, but as an assemblage of coupled hypotheses. Hence, Clark et al. (2011) proposed that a model should be constructed in a way that allows the testing of every single hypothesis of every process separately. In addition, the interactions of single elements within such a model should also be considered to better understand why a certain model works or fails (Clark et al., 2016).

When the idea of multiple hypotheses emerged, there was no easy way to construct models with interchangeable components (Buytaert et al., 2008) except for some comparison inside the TOPMODEL model family (Beven and Kirkby, 1979). We now have model frameworks at hand that facilitate such a design, e.g. SUPERFLEX (Fenicia et al., 2011), Structure for Unifying Multiple Modelling Alternatives (SUMMA) (Clark et al., 2015b, 2015a), or the Catchment Modelling Framework (CMF) (Kraft et al., 2011). SUPERFLEX targets the construction of lumped conceptual models (van Esse et al., 2013; Gharari et al., 2014). SUMMA and CMF support the generation of multi scale approaches from plot over hillslope to basins and from lumped to fully distributed models. SUMMA focusses on the comparison of process-based models with predefined parameters sets and is up to now mainly tested for surface-atmosphere interactions (Clark et al., 2015b, 2015a). CMF is a programming library to build hydrological models from building blocks with both, process-based and conceptual models. It can be used for subsurface and surface water fluxes, surface-atmosphere exchange and solute transport. So far, it has been applied in studies to better understand hydrological processes (Holländer et al., 2009; Maier et al., 2017; Orlowski et al., 2016; Windhorst et al., 2014), to simulate solute transport (Djabelkhir et al., 2017; Kraft et al., 2010) and to capture hydrological lateral and vertical transport processes in coupled complex ecosystem models (Haas et al., 2013; Houska et al., 2014, 2017; Kellner et al., 2017). All toolboxes enable a stepwise modification of the model structure. Additionally, they allow an easier ~~comparability~~ comparison of different models, as they are all constructed from the same parts and a more straightforwardly handled through interfaces (Buytaert et al., 2008). Recently, some studies tried to tackle the multi-hypotheses problem within a model framework (e.g. van Esse et al., 2013; Fenicia et al., 2008; Gharari et al., 2014; Hublart et al., 2015; Kavetski and Fenicia, 2011). Most of these studies built their models incremental from bottom up to find out, if small modifications allow a better simulation (Bai et al., 2009; Westerberg and Birkel, 2015). Others compared predefined model structures (van Esse et al., 2013; Kavetski and Fenicia, 2011). In all cases, researchers stopped improving the models once a sufficient performance was reached. ~~However, there is a chance that they have missed an even better model performance by including further modifications.~~ Clark et al (2015ab) propose with the SUMMA concept another approach to test multiple hypotheses. Here, the number and type of subprocess stay static, yet the mathematical formulation of the process description are scrutinized by exchange~~.~~

Despite having the potential to create a wide range of models with such toolboxes, only a minor quantity in the vast space of possible model structures is currently explored. However, this thorough exploration is needed to find appropriate model structures for any catchment, as it seems that current hydrological knowledge does not allow to construct a model that works equally well for all environmental conditions, especially when using lumped models (Beven, 2000, 2007, 2016; Buytaert et al., 2008; Fenicia et al., 2014).

To better use the existing understanding of a given catchment and to test more complex models, this study turns the incremental approach of adding more process-understanding to a model upside down. First, we develop a conceptual model from current hydrological understanding that contains all ~~structures~~ subprocesses that might be important for the functioning of a catchment. Then, parts of this model are disabled through incremental model breakdown, and the reduced model structures are tested for their simulation performance. A subprocess is marked as necessary, when models lacking it are rejected. On this base, a subsequent model is constructed which uses ~~the insights gained from~~only meaningful subprocesses ~~those previous models with disabled processes~~. Incremental model breakdown is therefore a rejectionist approach, built on the learning from failure and not an optimization process. Beven (2006) ~~,~~ assumed that a rejectionist approach is generally better suited to gain insight about process hypotheses. ~~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?~~ To allow comparability of the incremental breakdown method with common modelling approaches, the subsequent model is finally benchmarked with HBV-Light.

The objective of this study is to demonstrate that incremental model breakdown allows a detailed examination of model structures, an easier identification of the most important hydrological processes, and thus the construction of an improved model. 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. Ultimately, this approach also enables a better hydrological understanding of the catchment, as different structures, flaws and errors of a first modelling approach become obvious, even if a theoretical optimal model structure is still unknown..

2 Material and Methods

2.1 Study area

The study area is an upper section (AEO 2~~.~~,977 km^2, gauging station Grebenau) of the Fulda catchment (Figure 1), a catchment with Mid-European temperate climatic conditions. Relevant processes and catchment characteristics to be considered included the contribution of snowfall to precipitation, a mix of land uses with open and closed vegetation cover, and urban regions that impact hydrology through non-gradient driven fluxes (e.g. water abstraction for drinking water supply, sewage treatment works, reservoirs or sealed areas).

Precipitation input is influenced by the surrounding low mountain ridges of the Vogelsberg, the Wasserkuppe, the Knüll-Mountains and the Melsunger Uplands, leading to a significant contribution of snowfall in winter. The elevation ranges from about 150 m a.s.l. at Grebenau to 950 m a.s.l. at the Wasserkuppe. Wittmann (2002) used tritium as a tracer and found that the Fulda catchment has two distinct groundwater reservoirs: A large one reacting slowly and a smaller one with faster reaction. Land use is dominated by agriculture (37 %) and forests (41 %).

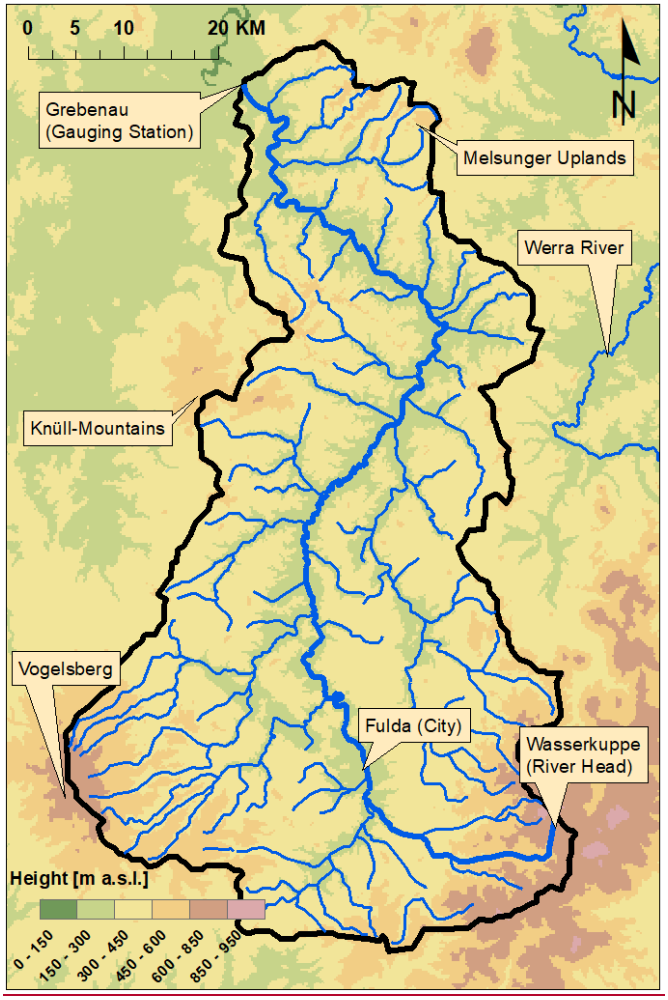


Figure 1: Relief map of the partial catchment of the Fulda River (black border) with side streams, ridges and parts of the Werra River.

2.2 Model input and validation data

Discharge data for the gauging station Grebenau, temperature and precipitation data were obtained from the Hessisches Landesamt für Naturschutz, Umwelt und Geologie (HLNUG). The point measurements for precipitation and temperature of the 108 measurement stations were extrapolated over the whole catchment, using kriging with altitude as an external drift (Hudson and Wackernagel, 1994). Finally, the extrapolated values were averaged over the whole catchment to get a single, lumped value per day.

The Fulda catchment has a humid, temperate climate. ~~. The climatic conditions during the calibration (1980-1985) and validation period (1986-1988) were rather similar~~ with an annual precipitation of 838 mm. The annual runoff coefficient ranges

between 0.3 to 0.6 (average 0.39), which is in the range for comparable catchments (e.g. Rawlins et al., 2006). The discharge and groundwater of the catchment are influenced by drinking water abstraction for 80,000 inhabitants (Rhönenergie Fulda GmbH, 2017).

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 2). The calibration period is wetter and contains six of the seven large rainfall events (> 30 mm d^{-1}) are located here. In addition, in both periods there is one year which represents more extreme weather conditions. 1985 for the calibration period with very little discharge in comparison with the precipitation and 1988 with very much discharge in comparison to the precipitation. Still the precipitation stays in the long term range for this catchment for all years (Fink and Koch, 2010).

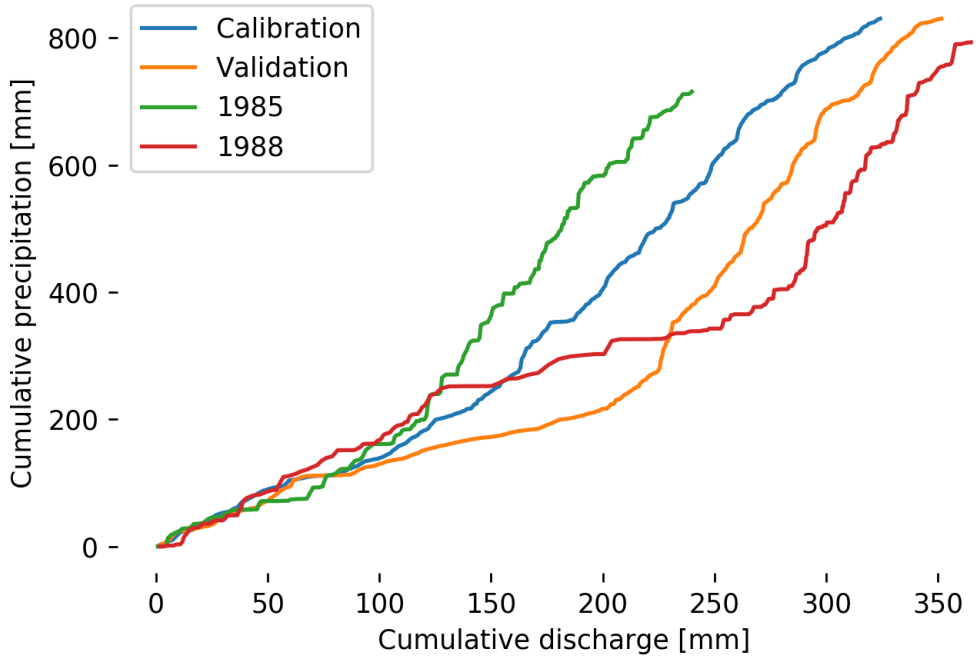


Figure 2: Cumulative discharge plotted against cumulative precipitation for the calibration and validation time period and two years which deviate most from the other years. For the calibration and validation period, the cumulative discharge and precipitation are the average of the corresponding years.

2.3 Model development using Catchment Modelling Framework (CMF)

For the construction of all models and all numerical calculations (except HBV-Light), we used CMF. CMF is a modular framework for hydrological modelling developed by Kraft et al. (2011) (see also (CMF, 2017)). For solving the differential

equations of models constructed with CMF, several numerical solvers are embedded in the toolbox. To avoid numerical problems (Clark and Kavetski, 2010; Kavetski et al., 2011; Kavetski and Clark, 2011) we selected the CVode Integrator (Hindmarsh et al., 2005) for all models. The CMF version used for this study was 0.1380.

In a first model set up (Model 1, Figure 3~~1~~) all processes are reliant on different flow connections. The incoming precipitation is saved in a snow storage in case the air temperature is below freezing point and rereleased to the surface storage after snowmelt. All other precipitation is split between the canopy or reaches the surface directly, depending on canopy closure. From the surface, the water is either directly routed to the river or enters three serial soil/groundwater layers, which in turn route water to the river as well. In addition, a fixed amount of water is abstracted from the lower groundwater to simulate drinking water extraction, which in turn is routed to the river. The river then routes all water to the outlet. Thus, it contains the implementations of processes for evapotranspiration, a canopy, snow, surfaces, a river, upper- and lower groundwater body (Figure 3~~1~~).

Following the findings of Singh (2002) all connections in the model with a flow curve (Figure ~~1~~3) are described as kinematic waves (Equation 1) (Singh, 2002), except for the infiltration and the drinking water abstraction.

$$Q = \frac{1}{t_r} Q_0 \left(\frac{V - V_{residual}}{V_0}\right)^{\beta} \quad (1)$$

where Q is the ratio of transferred water, $V_{residual}$ [m³] is the volume of water remaining in the storage, V_0 [m³ d^{-1}] is the reference volume to scale the exponent, V is the current volume of water in the storage [m³], and β is a parameter to shape the response curve [-]. The parameter ~~t_r~~ Q_0 ~~controls the reaction time of the storage, in case of β = 1, it is the mean residence time in days~~ is the flux in [m³ d^{-1}], when $\frac{V - V_{residual}}{V_0} = 1$.

Water that reaches the surface, i.e. throughfall or snowmelt, is routed into the upper soil as infiltration, with the following limits applied:

- Infiltration excess expressed by a maximum surface permeability K_{sat}
- Saturation excess expressed by a limiting factor calculated from the water content of the first subsurface water storage using a sigmoidal function

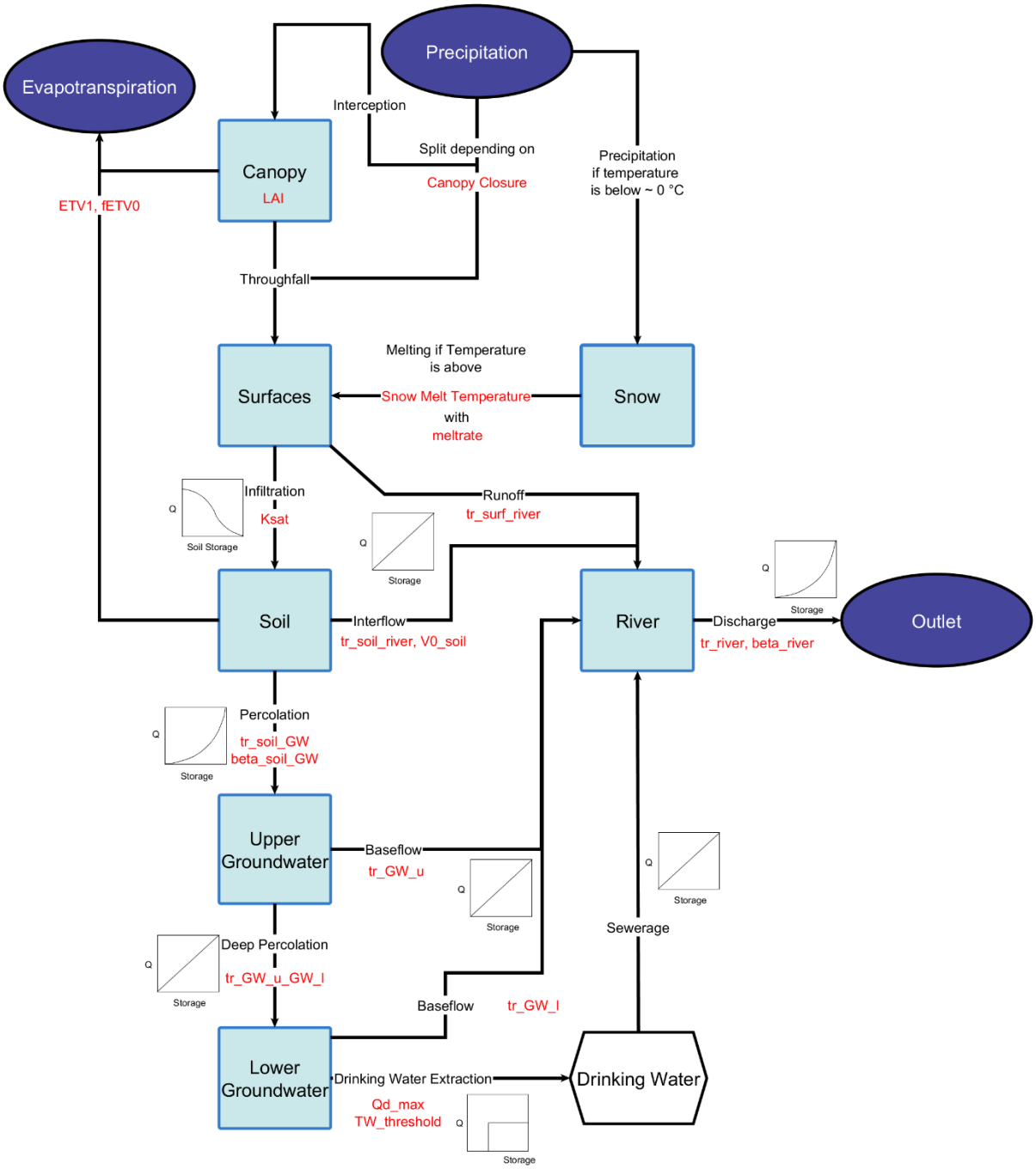


Figure 3: Structure of the Model 1 with water storages (light blue), boundaries (dark blue), temporary storages (white) calibrated parameters (red), fluxes (black arrows) and flow curves (for all applicable fluxes). **Water reaching the outlet is shifted one day into the future.**

-

Drinking water abstraction is implemented as a fixed amount of water. As the influence of the drinking water abstraction is not known, the amount of water abstracted is calibrated. It is transferred to the drinking water storage as long as the amount of water in the groundwater storage is above a threshold. As this threshold is not exactly known, we included it as a parameter for calibration. From the drinking water storage, all water abstracted for a given day is routed to the river.

Snowmelt uses a simple degree-day method (see API in CMF (2017)). The snowmelt temperature parameter was calibrated. Interception from the canopy is realized as Rutter interception (Rutter and Morton, 1977). Potential evapotranspiration was calculated with the modified Hargreaves equation by Samani (2000).

The devised model was tested by using fluxogram graphs. Fluxograms allow creating animated graphs that resemble model structures with all their storages and fluxes and they were used to analyse the implemented model processes (for more information about the fluxogram graph see Jehn (2018)). The size of the storages and fluxes change for each time step according to the amount of water stored/moved. For the fluxogram animations we used the model with the highest logarithmic Nash-Sutcliffe Efficiency (logNSE). ~~A detailed explanation on the construction and source code of fluxograms (and CMF models in general) can be found in the tutorial section of CMF (2017), including a number of examples.~~

2.4 Incremental model breakdown

To test the influence of different structure elements in Model 1, we used the concept of a one-at-a-time sensitivity analysis, i.e. disabling one process after the other, to track changes (Figure 4) . This resulted in 13 additional models with varying disabled processes (Table 1). In a second step, we disabled up to four processes, to scrutinize the interplay of processes, as proposed by Clark et al. (2011). 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.

-For each simplified model, the model performance was evaluated. If the model performance was getting worse, the deleted process was valued essential for the model and vice versa. If the performance did not change, the process was rated unimportant. This allowed us to separate influential processes from unnecessary ones and thereby assess if the chosen complexity was justifiable for the catchment (~~Table~~ Figure 5, Figure 6~~2~~). 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. 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 ~~marks~~ that subprocess ~~as~~ is irrelevant for this application. The ~~knowledge gained was then used to construct a~~ final Model 15 is constructed from Model 1 by removing the irrelevant subprocesses ~~containing only processes we found relevant to mimic the real-world system.~~

Figure 4: Flow chart for the method of incremental model breakdown

Table 1: Structural elements of the models and amount of parameters. GW = Groundwater, DW = Drinking Water. ET = Evapotranspiration. Light gray indicates active components. Dark grey indicates disabled components.

	Rain distribution					Groundwater (GW)				Number
	Canopy	Surfaces	Snow	Soil	River	upper GW	lower GW	DW	ET	Parameters
Model 1 (start)	x	x	x	x	x	x	x	x	x	19
Model 2 (no GW)	x	x	x	x	x				x	13
Model 3 (no ET)	x	x	x	x	x	x	x	x		18
Model 4 (no river)	x	x	x	x		x	x	x	x	17
Model 5 (no rain distribution)				x	x	x	x	x	x	13
Model 6 (no surfaces)	x		x	x	x	x	x	x	x	17
Model 7 (no canopy)		x	x	x	x	x	x	x	x	17
Model 8 (no snow)	x	x		x	x	x	x	x	x	17
Model 9 (no DW)	x	x	x	x	x	x	x		x	17
Model 10 (no GW/river)	x	x	x	x					x	10
Model 11 (no canopy/surfaces)			x	x	x	x	x	x	x	15
Model 12 (no river/surfaces)	x		x	x		x	x	x	x	15
Model 13 (no lower GW)	x	x	x	x	x	x		x	x	17
Model 14 (no lower GW/DW)	x	x	x	x	x	x			x	15
Model 15 (final)	x		x	x		x		x	x	10

-	Behavioural	Number	Calibration						-	Validation				
Model	runs	parameters	logNS [/]		PBIAS [%]		RSR [/]		logNS [/]		PBIAS [%]		RSR [/]	
1	9	19	0.53	± 0.04	3.97	± 5.55	0.68	± 0.01	0.61	± 0.05	1.13	± 5.95	0.69	± 0.03
2	-	13	-		-		-		-		-		-	
3	-	18	-		-		-		-		-		-	
4	6	17	0.55	± 0.03	1.28	± 5.60	0.68	± 0.02	0.63	± 0.04	0.13	± 7.27	0.67	± 0.04
5	-	13	-		-		-		-		-		-	
6	8	17	0.54	± 0.04	2.46	± 3.31	0.69	± 0.01	0.62	± 0.04	1.29	± 2.76	0.67	± 0.03
7	80	17	0.53	± 0.02	10.75	± 4.30	0.68	± 0.02	0.62	± 0.03	7.3	± 4.31	0.66	± 0.02
8	-	17	-		-		-		-		-		-	
9	6	17	0.55	± 0.03	3.29	± 4.12	0.68	± 0.01	0.62	± 0.05	4.13	± 5.61	0.68	± 0.03
10	-	10	-		-		-		-		-		-	
11	90	15	0.53	± 0.02	10.16	± 3.97	0.68	± 0.02	0.62	± 0.04	6.91	± 3.61	0.67	± 0.02
12	8	15	0.52	± 0.02	3.71	± 6.89	0.68	± 0.01	0.62	± 0.05	2.36	± 6.46	0.67	± 0.03
13	2	17	0.57	± 0.03	3.10	± 7.64	0.69	± 0.01	0.70	± 0.01	3.17	± 7.22	0.61	± 0.02
14	3	15	0.52	± 0.02	5.08	± 8.78	0.68	± 0.02	0.59	± 0.04	2.59	± 7.96	0.66	± 0.04
15	1731	10	0.57	± 0.05	4.86	± 8.03	0.66	± 0.04	0.65	± 0.06	5.56	± 7.94	0.65	± 0.05

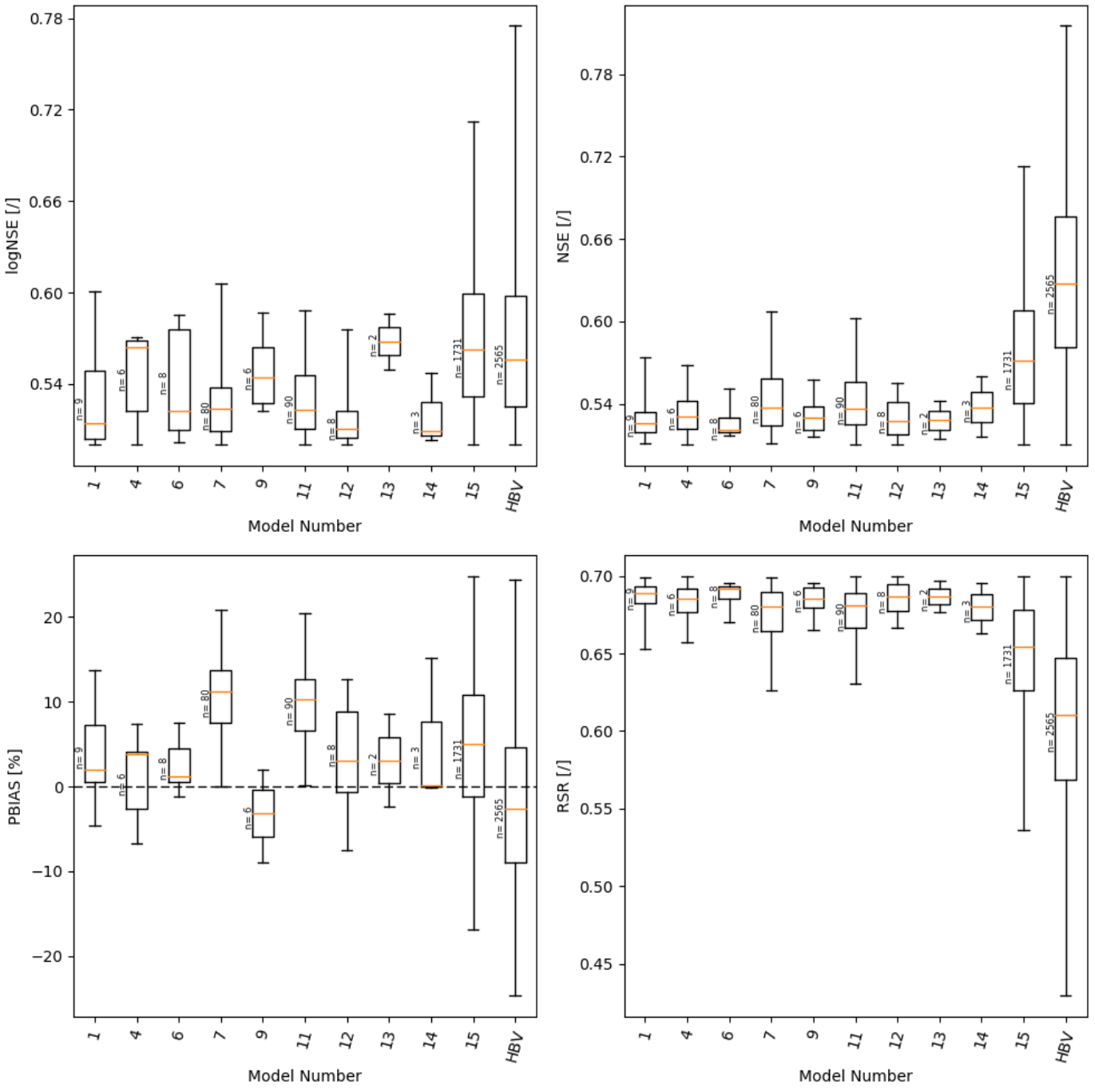


Figure 5: Boxplots of the objective functions for all models with behavioural runs in the calibration period. The yellow bar marks the median. Number of behavioural runs noted on boxplot.

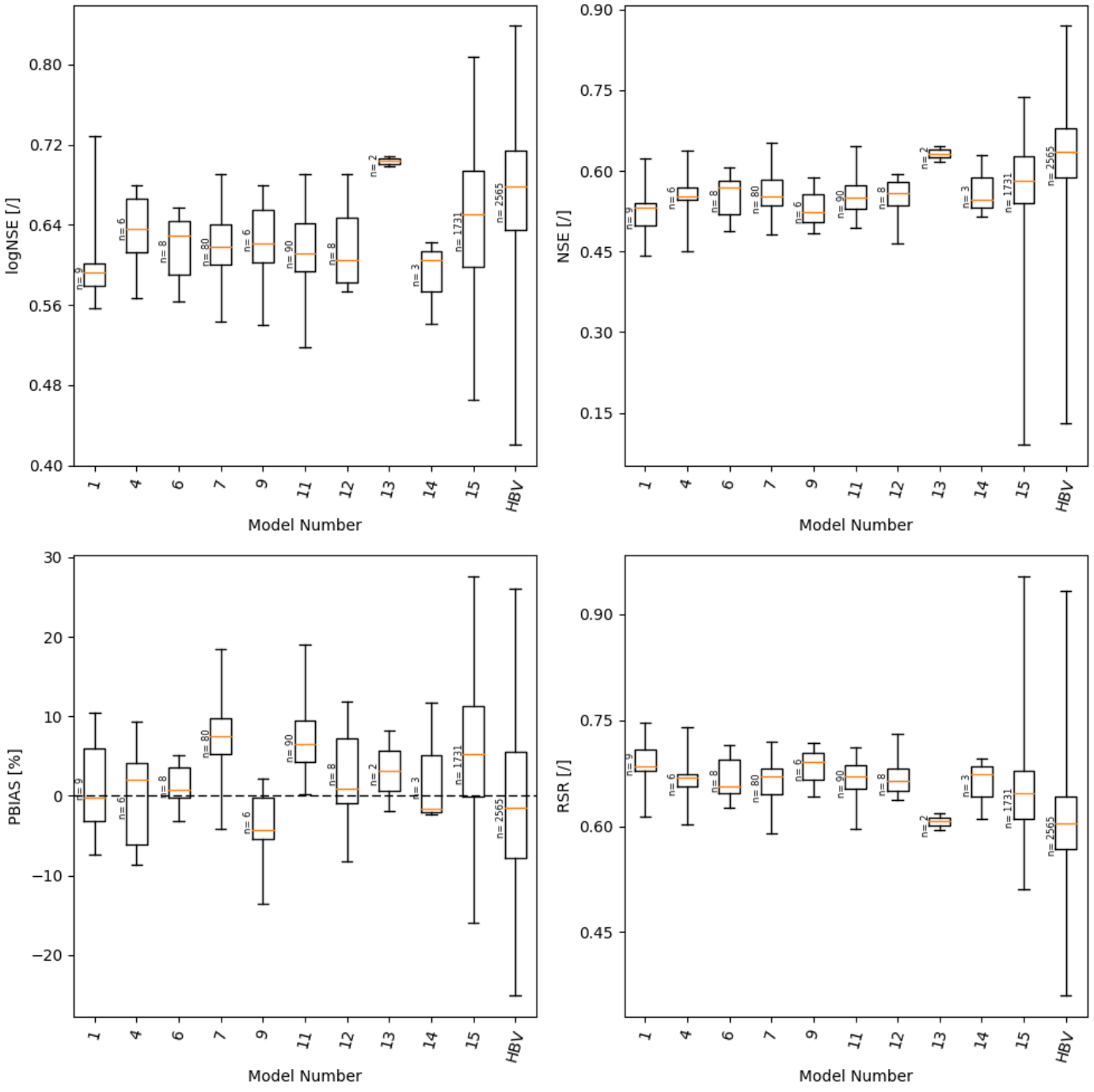


Figure 6: Boxplots of the objective functions for all models with behavioural runs in the validation period. The yellow bar marks the median. Number of behavioural runs noted on boxplot.

2.5 Calibration and validation

A model run was separated into warm-up period of one year (1979), a calibration period of six years (1980-1985) and a validation period of three years (1986-1988). First CMF model runs showed that many simulated discharge peaks occurred one day ahead compared to observed data. This is caused by rainfall occurring in the later time of a day, that leads to a reaction in the hydrograph of the following day as water needs time to reach the gauging station (Ficchì et al., 2016). The model, however, reacts directly to this as its input data is resolved in a 24 h time step. Therefore, we shifted the simulated time series one day into the future as proposed by Bosch et al. (2004). This led to better calibration results. This was not needed for HBV-Light, in which the MAXBAS parameter accounts for shifting peak discharge.

We used the Generalized Likelihood Uncertainty Estimation (GLUE) methodology (Beven and Binley, 1992) to find behavioral parameters sets for the calibratin period. 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. As single-objective calibration lowers the identifiability of model parameters and structural elements (Efstratiadis and Koutsoyiannis, 2010) and often hide shortcomings of models (Ritter and Muñoz-Carpena, 2013), we pursued a multi-objective calibration procedure. Following the concept of Moriasi et al. (2007), a model run was deemed behavioural, if the ~~logarithmic~~ Nash-Sutcliffe-Efficiency (~~log~~NSE) was >0.5 (optimal value: 1; range: 1 to - ∞), the percentage bias (PBIAS) was below/above ±25% (optimal value: 0; range: 0 to ± ∞) and the ratio between root mean square error to the standard deviation of the measured data (RSR) was <0.7 (optimal value: 0; range: 0 to ∞). As an additional constrained we also included the logarithmic Nash-Sutcliffe-Efficieny (logNSE) (optimal value: 1; range: 1 to - ∞) to allow a better evaluation of low flows. The ~~log~~NSE focuses on ~~low~~ peak flows~~, the RSR depicts peak flows~~ and the PBIAS considers the overall model deviation from observed data. It should be noted though, that this study does not aim on finding the optimal parameter sets for a single model, but to use the knowledge gained from calibration and validation to identify the most important processes in the model structure and use this to improve the model structure and reduce the number of parameters used. 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.

The sampling of the parameter space for calibration was done by Latin Hypercube Sampling (McKay et al., 1979) implemented via SPOTPY (Houska et al., 2015). ~~The CMF~~All models were run 300,000 times each, using a High Performance Computing Cluster. See the tutorial section of CMF (2017) for more detailed information on the coupling of CMF with SPOTPY for model calibration. Implemented parameter boundaries for Model 1 are given in Table ~~3~~2 and remained fixed for all further developed model structures to ensure comparability.

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 the findings of 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.

2.6 Benchmarking model

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.

Table 2: Lower and upper parameters bounds of all models and their indented meaning. GW = Groundwater

Name	Unit	Intende~~n~~t meaning	Min	Max
tr_soil_GW	day	Residence time from soil to upper GW	0.5	150
tr_soil_river	day	Residence time from soil to river	0.5	55
tr_surf_river	day	Residence time from surfaces to river	0	30
tr_GW_l	day	Residence time from upper GW to river/outlet	1	1000
tr_GW_u	day	Residence time from upper GW to river/outlet	1	750
tr_GW_u_GW_l	day	Residence time from upper to lower GW	10	750
tr_river	day	Residence time from river to outlet	0	3.5
V0_soil	mm	Field capacity of the soil	15	350
beta_soil_GW	/	Exponent ~~the~~ which changes the shape of the flow curve	0.5	3.2
beta_river	/	Exponent ~~the~~ which changes the shape of the flow curve	0.3	4
ETV1	mm	Volume under which the evapotranspiration is lowered	0	100
fETV0	%	Factor by what the evapotranspiration is lowered	0	0.25
meltrate	mm $°C^{-1}$ day^{-1}	Meltrate of the snow	0.15	10
snow_melt_temp	°C	Temperature of snow melt	-1	4.2
Qd_max	mm day^{-1}	Maximal drinking water extraction	0	3
TW_threshold	mm	Amount of water that cannot be extracted	0	100
LAI	/	Leaf area index	1	12
CanopyClosure	%	Canopy closure	0	0.5
Ksat	m day^{-1}	Saturated conductivity of the soil	0	1

3 Results

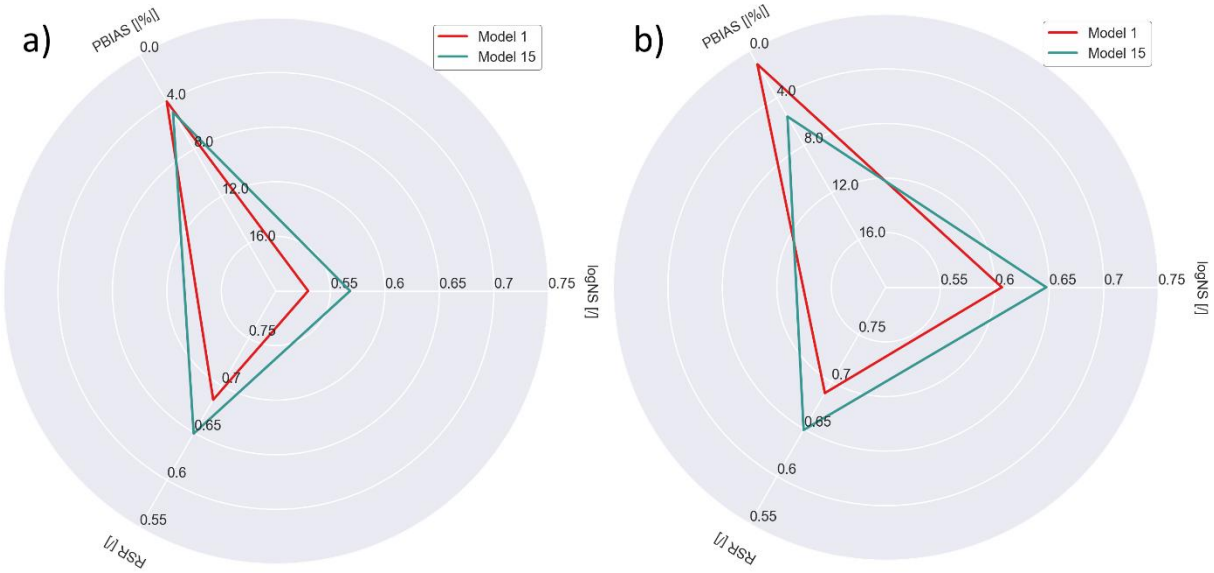
3.1 Behavioural runs of Model 1 to 14

Model 1 was able to achieve nine behavioural runs. The model has a better performance in the validation period (Figure 5~~2~~, Figure 3~~6, Table 2~~) This is true for all other models as well.. The simulated discharge is rather erratic (Figure 3~~7~~), i.e. it reacts directly on small changes in precipitation. Those quick reactions are timed correctly. However, they overestimate the discharge from small precipitation events, while underestimating large ones. These differences are larger in summer than in winter. This behaviour leads to underestimated high flows and many overestimated small peaks, while the overall simulated amounts are unbiased. Investigation of storages and fluxes (fluxogram-graph: https://youtu.be/cP0PfDpfW88) show that most of the water is stored in the upper groundwater storage, while the lower groundwater storage is removed directly by the drinking water production, as soon as it is above the threshold. Only very small amounts of water are stored in the surface storage, the canopy

storage and the soil storage. From the soil storage and the canopy storage large amounts of water evaporate, often exceeding the flow to the outlet. The river storage is mostly recharged from the groundwater and the drinking water storage. The soil storage contributes significantly to the river storage only at large precipitation events or during snowmelt. Overall, Model 1 slightly overestimates ~~base~~ low flows and the evapotranspiration, while largely underestimating the peaks.

Most of the deleted model processes from the most complex Model 1 led to more behavioural runs (~~Table 2~~Figure 5, Figure 6). Model 1, 4 (no river storage), 6 (no surface storage), 9 (no drinking water simulation), 12 (no river and surface storages), 13 (no lower groundwater storage) and 14 (no groundwater storages and drinking water simulation) have between two to nine behavioural runs. Model 7 (no canopy) and 11 (no canopy and surface storage) are able to produce 80 and 90 behavioural runs respectively (~~Table 2~~Figure 5, Figure 6). The remaining models 2 (no groundwater storages), 3 (no evapotranspiration), 5 (no rain distribution), 8 (no snow) and 10 (no groundwater and river storages) were not able to produce behavioural runs.

~~The~~ Most, but not all simplified models tend to show better performances for their ~~mean~~ median values of the logNSE, NSE and the RSR (at least in the validation period) than Model 1, while Model 1 has a PBIAS better than ~~most of~~ the other models. Especially ~~model~~ Model 13 (no lower groundwater storage) median values for the ~~logNSE~~ objective functions outperform Model 1. ~~Also all simplified models make use of less parameters than the first model.~~



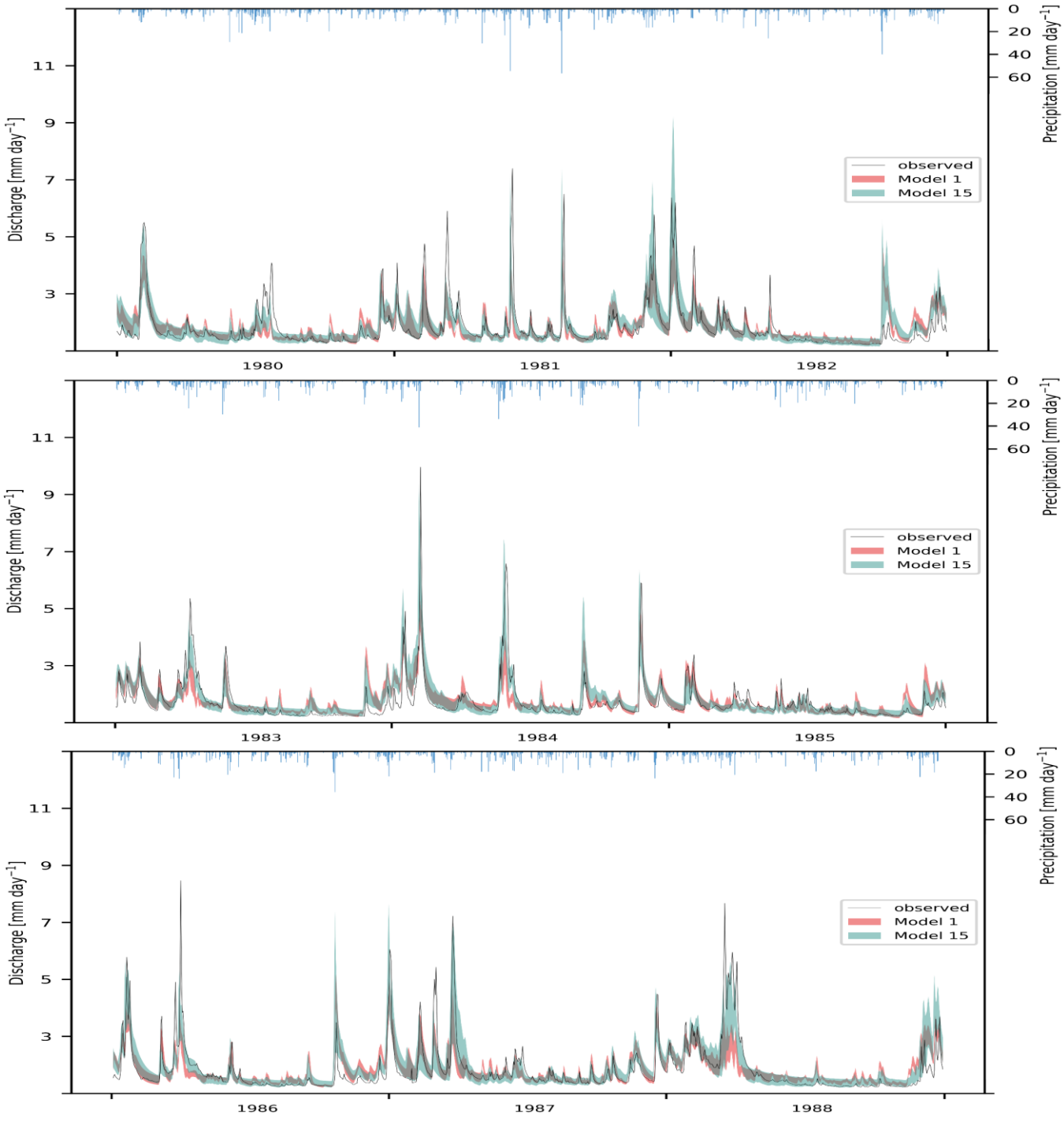


Figure 7: Hydrograph of the Fulda River for 1980-1988. Coloured areas depict the ~~uncertainty~~ range of the behavioural model runs (5th to 95th percentile). Calibration time period is from 1980 till 1985 (first two subplots). Validation time period is from 1986 to 1988 (third subplot). Observed discharge is depicted as black line. Precipitation is drawn with an inverted y-axis.

3.2 Construction and behavioural runs of Model 15

We can report that representation of model processes for the upper groundwater body, evapotranspiration and snow have a positive impact on model performance in the Fulda catchment, given the increased ~~number of behavioural runs and mean~~ median values of the objective functions (~~Table 2~~Figure 5, Figure 6), as when those processes are excluded the models struggle to produce behavioural runs. The exclusion of the canopy and drinking water have a more or less neutral impact on the median performance of the behavioural runs (Figure 5, Figure 6~~Table 2~~). Whereas the chosen implementations of the river, the surfaces and the lower groundwater affect the model quality negatively (Figure 5, Figure 6~~Table 2~~). The structure of Model 15 was created after all the other models had been evaluated. For this, we used the process knowledge gained from the reduced models (see discussion) and constructed Model 15 with only those processes, which had proven to have positive impact on the quality of the results. Therefore, Model 15 consists only of those processes most important for the given Fulda catchment (Figure ~~4~~9). In comparison with the model structure of Model 1, the processes surface water storage, lower groundwater storage, drinking water extraction, river storage and the simulation of the canopy were disabled.

Profiting from the insights of the models with disabled processes, Model 15 performs better than Model 1. The RSR, NSE and the logNSE depict better values, both in the validation and calibration period, while the PBIAS is slightly worse for both cases (Figure 5, Figure 6~~Table 2, Figure 2~~). Especially the maximal values are for the logNSE, NSE and RSR are much better than Model 1. ~~Further, it has more behavioural runs than the best of the reduced models (Table 2).~~ As for all other models, the performance increases from the calibration to the validation period for Model 15. The simulated hydrograph is a lot less erratic than the one from Model 1 (Figure ~~3~~7). In addition, the peaks fit better than in Model 1. However, summer peaks are less likely to be predicted than those during the rest of the year (Figure ~~3~~7). The overestimation of ~~baseflow~~ low flow in Model 1 is ~~only~~ apparent on fewer days. In Addition, to this increase of the performance in comparison with Model 1, Model 15 uses nine parameters less (Table ~~2~~1). The remaining ten parameters in Model 15 behave different from the same ones in Model 1 (Figure 8). 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.

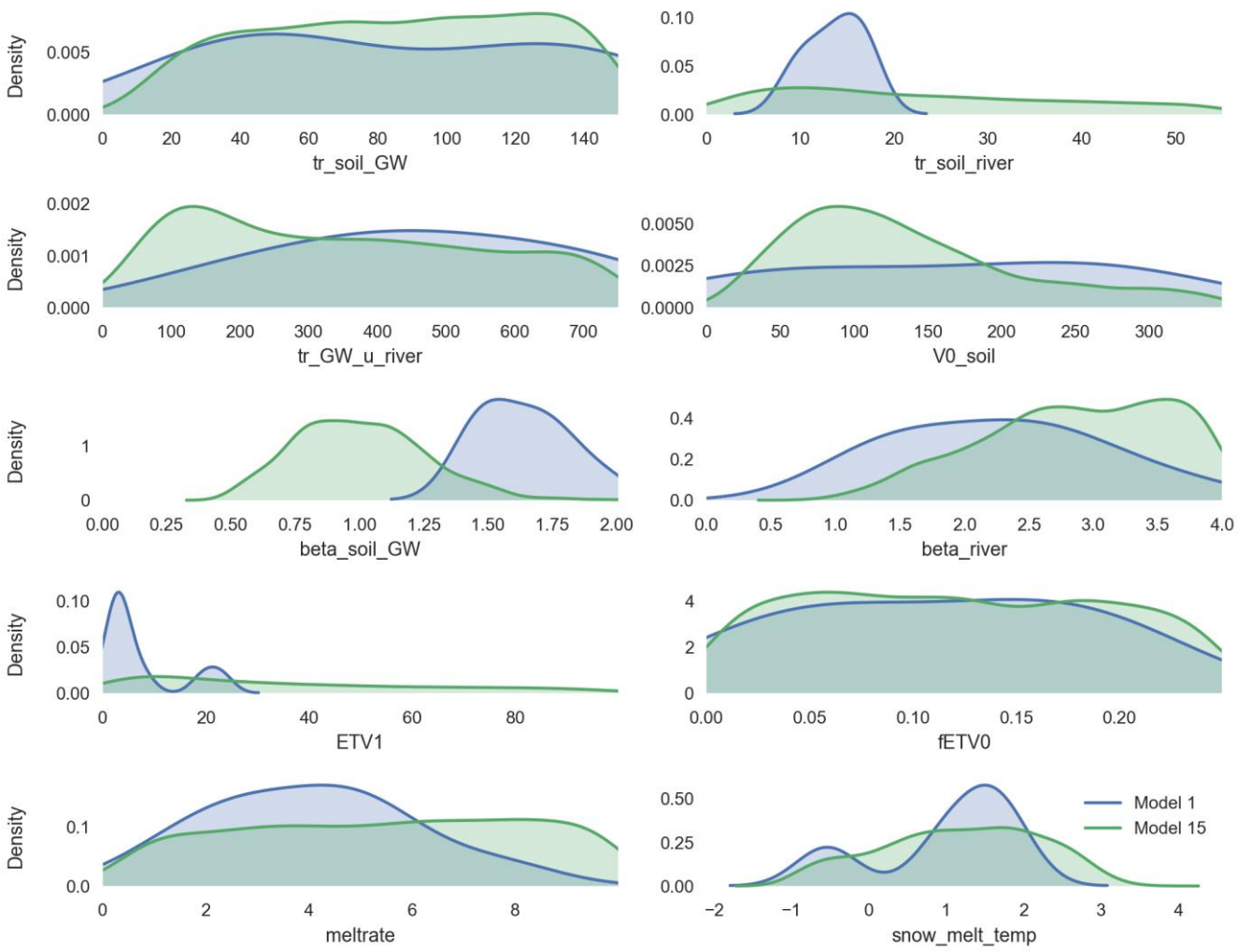


Figure 8: Distribution of all parameters shared by Model 1 (blue) and Model 15 (green), fitted with kernel density.

3.3 Behavioural runs of HBV-Light

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 (Figure 5, Figure 6). 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.

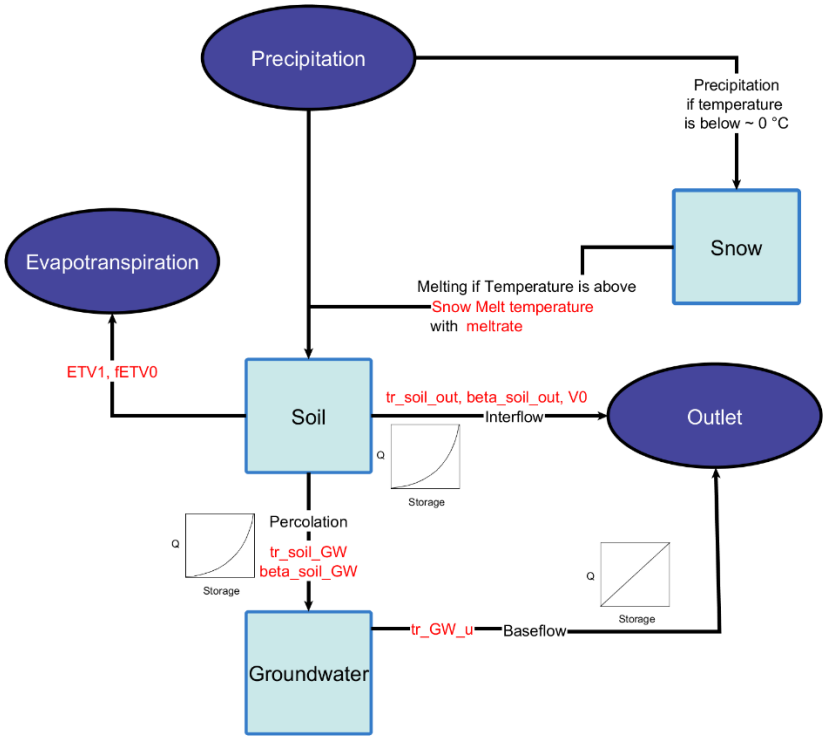


Figure 9: Structure of the final Model 15 with water storages (light blue), boundaries (dark blue), temporary storages (white) calibrated parameters (red), fluxes (black arrows) and flow curves (for all applicable fluxes. Water reaching the outlet is shifted one day into the future.

4 Discussion

4.1 Overview

The results show that Model 1 fell short on simulating the catchment correctly. Mainly caused by a slow reaction to precipitation events, which reduced discharge peak prediction and caused the model to focus on evapotranspiration to handle the excess water. Still, it was a good basis to determine relevant processes by incremental model breakdown. Insights from this led to an improvement in the performance of Model 15, while at the same time allowed a reduction of the number of parameters (from n=19 to n=10)(Table ~~2~~1). Nevertheless, even this improvement did not allow Model 15 to outperform HBV-Light. Still, ~~T~~this suggests that the method of incremental model breakdown is a good way to improve model performance and reduce equifinality ~~in the posterior parameter distribution~~through parameter reduction, which improves the identifiability of the model structure (Ambroise, 2004). It enables insight to which processes are important for discharge simulations in a given catchment. It also allows revealing errors, using them "as a means of discovery" of false model assumptions (Elliott, 2004). It

should be noted that this method does not necessarily lead to an improved model performance, but it allows creating a model, which relies on fewer processes, parameters and assumptions, thus being an application of "Occam's razor" (Clark et al., 2011).

4.2 Inspection of internal processes

Even though Model 1 did give sufficient but not excellent results, it was a good foundation for the construction of Model 15. Due to the implementation of many processes in Model 1, all those processes could be examined on their effect on the simulation. Upper groundwater, evapotranspiration and the simulation of the snow storage and snowmelt were identified as the most important processes, as for example models 2, 5 or 10 could not achieve behavioural runs without those processes (Figure 5, Figure 6~~Table 2~~). Model processes of drinking water and the canopy showed only minor impact on model discharge simulation performance (Figure 5, Figure 6~~Table 2~~). Improved values for the objective function were found for models 4, 6 and 13 with no river, no surfaces and no lower groundwater, as those processes likely hindered the models from being better. Excluding these processes make the models react slower and with this, more accurate to precipitation inputs. The drinking water storage's minor influence might simply be due to the rather low population of 159 persons per km² in the region, and neither water withdrawal for irrigation nor water-consuming industries are relevant players in the region's water cycle. The canopy, however, is commonly regarded as an important factor, as interception can cause 25 % and more of the rainfall not to reach the ground (Link et al., 2004). However, Model 15 was able to get better values for the objective functions than Model 1 even though the canopy was disabled (Figure 5, Figure 6~~Table 2, Figure 2~~). Fenicia et al. (2008) showed that canopies have a large effect in dry regions, which is underpinned by models developed for humid regions neglecting the canopy and still performing well, e.g. HBV-Light (Seibert and Vis, 2012). Also, the current implementation of the canopy in CMF assumes a fixed canopy storage for the whole year. A more realistic approach should be implemented for future applications, as was for example realized in plot scale CMF application coupled to plant growth models for winter wheat (Houska et al., 2014) and perennial grassland (Kellner et al., 2017).

The river storage is most likely too small to be an important reservoir in comparison to the catchment. The surface storages probably do not contribute to the runoff itself, because the catchment is mostly vegetated, which impedes overland flow. Lower groundwater was included because of the ability of the sand- and limestone in the catchment to store large quantities of water and because of the tritium based tracer experiments of Wittmann (2002). He found two distinct groundwater aquifers in the catchment. Their study comes to the results that the lower one of the aquifers must be very large. However, our posterior parameter boundaries indicate a very slow response.

Model 15 falls short in predicting peak flow in summer (Figure ~~3~~7). Due to that, the model has too much water and needs to compensate for this by overestimating baseflow and evapotranspiration (Figure 7~~3~~). The problem of not predicting the peak flow in spring completely right is probably caused by the lumped and simple implementation of snowmelt. Most of the snow in the Fulda catchment is stored in a small area along the ridges, while the lumped model does not make such a spatial distinction~~.~~. A further, possibly influential, discrepancy between our lumped modelling assumptions and reality is that the

snowmelt occurs evenly distributed over the whole catchment, so that the complete snowmelt in the model takes only a few days, often even in only one day. It might also be linked to the evapotranspiration. . In times of low evapotranspiration, the water is forced to leave the model as discharge. Therefore, large precipitation events are directly transferred to large peaks. During times of high evapotranspiration, much water can be released into the atmosphere, and as the water in the soil storage of Model 15 flows proportionately more if the storage is already high, this allows the water to stay longer in the soil, which in turn allows more evapotranspiration.

The fluxograms showed that Model 1 did not use the drinking water storage and used the canopy storage only rarely. These observations underline the demand of Clark et al. (2011) that the internal procedures of a hydrological model should be inspected as well, to better understand its functioning. The fluxograms helped to detect that canopy and drinking water are not used by the model.

When examining the ~~mean~~ median model performance of all reduced models and the resulting Model 15, one can see that the ~~mean~~ median values of the objective functions of Model 15 are similar to those of Model 13 (Figure 5, Figure 6~~Table 2~~). Model 15 is considered to be the better representation of the catchment than Model 13, as it has a more streamlined structure and seven parameters less. In addition, the good values for Model 13 are mainly caused by the low number of behavioural runs (n = 2), allowing one very good run to distort the ~~mean~~results. Also, Model 15 reaches higher maximal values for the objective functions.

~~Model 15 falls short in predicting peak flow in summer (Figure 3). Consequently, the model has too much water resulting in overestimated baseflow and evapotranspiration. The problem of not predicting the peak flow in spring completely right is probably caused by the lumped and simple implementation of snowmelt, which causes the snow to melt quicker than in reality. It might also be linked to the evapotranspiration. . In times of low evapotranspiration, the water is forced to leave the model as discharge. Therefore, large precipitation events are directly transferred to large peaks. During times of high evapotranspiration, much water can be released into the atmosphere, and as the water in the soil storage of Model 15 flows proportionately more if the storage is already high, this allows the water to stay longer in the soil, which in turn allows more evapotranspiration.~~

This improved performance of Model 15 in comparison with the Model 1 is overshadowed though, by the higher equifinality of some parameters in Model 15 (Figure 8). In Model 1 for example the parameter EVT1 has two very distinct peaks, while Model 15 distribution for this parameter is spread out widely. 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.

4.3 Comparison with HBV-Light

All three models show a distinct behavior (Figure 5, Figure 6), 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.

4.~~3~~4 Does model incremental breakdown allow the construction of improved models?

The improved performance of Model 15 shows that a priori model selection is not useful, as the models with different process implementations deliver very different results. This is in line with the findings of Ley et al. (2016), who used predefined model structures on a large amount of different catchments and found that no model was able to simulate all catchments well. Similar results were also found by Kavetski and Fenicia (2011) and Fenicia et al. (2014), who showed that lumped models need to be tailored for single catchments as they are often over-simplified.

Lumped models have the advantage of an easy set-up and low data requirements, but this comes at the cost of not being able to address the spatial heterogeneity of the catchments (Ley et al., 2016) and that the parameters and structures have no direct equivalence in the real world (Bergström and Graham, 1998). Therefore, a lumped model structure might simply have been too simple for the upper section of the Fulda, calling for a semi-distributed or even distributed model set up. This is also hinted by a study by Fink and Koch (2010), who where able to model the Fulda Catchment quite well with a modified semi distributed version of SWAT. Overall, we think that the proposed method of incremental model breakdown led to an improvement in model performance. In addition, the model complexity and amount of parameters and with this equifinality were reduced. 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. Both topics are often stated as the main goals of model development e.g. Efstratiadis and Koutsoyiannis (2010) and Gupta and Nearing (2014).

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 during 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. 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 (2006) ~~Beven (2005)~~ to gain a robust method.

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.

5 Conclusion

This study shows that the process-based incremental breakdown of a hydrological model using fluxograms and a multi-objective calibration allows the identification of important hydrological processes in a model and the reconstruction of the starting model structure to a ~~less uncertain and~~ more efficient version. We conclude that the method provided offers a useful approach in the identification of relevant hydrological processes. Model frameworks such as CMF facilitate the development of such an approach.

The incremental model breakdown can be used best in two cases: (1) finding out why an existing good model does produce good results in the sense of a diagnostic tool to assess model structures; or, as in this study, (2) determining which processes are most relevant, to allow the streamlining of a model.

One goal of this study was to find another strategic way to test the multiple implementations of catchment functioning. We were able to distinguish between unnecessary and relevant model processes. Further, it became clearer what causes those problems, by examining the model piece by piece as proposed by Clark et al. (2016). Therefore, this method can be seen as a useful third way, in addition to step-wise model building (Bai et al., 2009; Westerberg and Birkel, 2015) and the comparison of predefined structures (van Esse et al., 2013; Kavetski and Fenicia, 2011), to explore the realm of multiple hypotheses. We propose future research should consider an automatic assemblage of model structures to test not only a manually manageable number of models but rather scan a larger variety of feasible combinations, which in turn would allow a completely exhaustive exploration of the space of possible model structure. ~~.~~

Data availability. Datasets are available by contacting the Hessian Agency for Nature Conservation, Environment and Geology (HLNUG) (https://www.hlnug.de/service/english.html).

Competing interests. The authors declare that they have no conflict of interests.

Acknowledgements. We thank the "Hessisches Landesamt für Naturschutz, Umwelt und Geologie" for providing the meteorological and discharge data and Lieke Mielsen, ~~.~~ François Anctil and one anonymous referee for their valuable comments which allowed to improve this paper.

References

Ambroise, B.: Variable 'active' versus 'contributing' areas or periods: a necessary distinction, Hydrol. Process., 18(6), 1149–1155, doi:10.1002/hyp.5536, 2004.

Bai, Y., Wagener, T. and Reed, P.: A top-down framework for watershed model evaluation and selection under uncertainty, Environ. Model. Softw., 24(8), 901–916, doi:10.1016/j.envsoft.2008.12.012, 2009.

Bergström, S. and Graham, L. P.: On the scale problem in hydrological modelling, J. Hydrol., 211(1–4), 253–265, doi:10.1016/S0022-1694(98)00248-0, 1998.

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

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6(3), 279–298, doi:10.1002/hyp.3360060305, 1992.

Beven, K. J.: Uniqueness of place and process representations in hydrological modeling, Hydrol. Earth Syst. Sci., 4(2), 203–213, 2000.

Beven, K. J.: On hypothesis testing in hydrology, Hydrol. Process., 15(9), 1655–1657, doi:10.1002/hyp.436, 2001.

Beven, K. J.: Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system, Hydrol. Process., 16(2), 189–206, doi:10.1002/hyp.343, 2002.

Beven, K. J.: Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process, Hydrol. Earth Syst. Sci., 11(1), 460–467, doi:10.5194/hess-11-460-2007, 2007.

Beven, K. J.: Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrol. Sci. J., 61(9), 1652–1665, doi:10.1080/02626667.2015.1031761, 2016.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant, Hydrol. Sci. Bull., 24(1), 43–69, doi:10.1080/02626667909491834, 1979.

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.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M. and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Adv. Water Resour., 32(2), 129–146, doi:10.1016/j.advwatres.2008.10.003, 2009.

Buytaert, W., Reusser, D., Krause, S. and Renaud, J.-P.: Why can't we do better than Topmodel?, Hydrol. Process., 22(20), 4175–4179, doi:10.1002/hyp.7125, 2008.

Clark, M. P. and Kavetski, D.: Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes: Numerical daemons of hydrological modeling, 1, Water Resour. Res., 46(10), doi:10.1029/2009WR008894, 2010.

Clark, M. P., Kavetski, D. and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling: Hypothesis testing in hydrology, Water Resour. Res., 47(9), doi:10.1029/2010WR009827, 2011.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J. and Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept: A unified approach for process-based hydrologic modeling, Water Resour. Res., 51(4), 2498–2514, doi:10.1002/2015WR017198, 2015a.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Gochis, D. J., Rasmussen, R. M., Tarboton, D. G., Mahat, V., Flerchinger, G. N. and Marks, D. G.: A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies: A unified approach for process-based hydrologic modeling, Water Resour. Res., 51(4), 2515–2542, doi:10.1002/2015WR017200, 2015b.

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., Freer, J. E., Arnold, J. R., Moore, R. D., Istanbulluoglu, E. and Ceola, S.: Improving the theoretical underpinnings of process-based hydrologic models: Narrowing the gap between hydrologic theory and models, Water Resour. Res., 52(3), 2350–2365, doi:10.1002/2015WR017910, 2016.

CMF: Catchment Modelling Framework Website, <http://fb09-pasig.umwelt.uni-giessen.de/cmf>, accessed February 2017, 2017.

Djabelkhir, K., Lauvernet, C., Kraft, P. and Carluer, N.: Development of a dual permeability model within a hydrological catchment modeling framework: 1D application, Sci. Total Environ., 575, 1429–1437, doi:10.1016/j.scitotenv.2016.10.012, 2017.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrol. Sci. J., 55(1), 58–78, doi:10.1080/02626660903526292, 2010.

Elliott, K.: Error as Means to Discovery, Philos. Sci., 71(2), 174–197, doi:10.1086/383010, 2004.

van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, F., Kavetski, D. and Lobligeois, F.: The influence of conceptual model structure on model performance: a comparative study for 237 French catchments, Hydrol. Earth Syst. Sci., 17(10), 4227–4239, doi:10.5194/hess-17-4227-2013, 2013.

Fenicia, F., Savenije, H. H. G., Matgen, P. and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44(1), doi:10.1029/2006WR005563, 2008.

Fenicia, F., Kavetski, D. and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development: Flexible framework for hydrological modeling, 1, Water Resour. Res., 47(11), doi:10.1029/2010WR010174, 2011.

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.

Ficchì, A., Perrin, C. and Andréassian, V.: Impact of temporal resolution of inputs on hydrological model performance: An analysis based on 2400 flood events, J. Hydrol., 538, 454–470, doi:10.1016/j.jhydrol.2016.04.016, 2016.

Fink, G. S. M. and Koch, M.: Climate change effects on the water balance in the Fulda catchment, Germany, during the 21st centruy, conference paper at Symposium on sustainable water ressource management and climate change adaption, Nakon Pathom., 2010.

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H. and Savenije, H. H. G.: Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, Hydrol. Earth Syst. Sci., 18(12), 4839–4859, doi:10.5194/hess-18-4839-2014, 2014.

Gupta, H. V. and Nearing, G. S.: Debates-the future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science, Water Resour. Res., 50(6), 5351–5359, doi:10.1002/2013WR015096, 2014.

Haas, E., Klatt, S., Fröhlich, A., Kraft, P., Werner, C., Kiese, R., Grote, R., Breuer, L. and Butterbach-Bahl, K.: LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale, Landsc. Ecol., 28(4), 615–636, doi:10.1007/s10980-012-9772-x, 2013.

Hindmarsh, A. C., Brown, P., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E. and Woodward, C. S.: SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers, ACM Trans. Math. Softw. TOMS, 31(3), 363–396, 2005.

Holländer, H. M., Blume, T., Bormann, H., Buytaert, W., Chirico, G. B., Exbrayat, J.-F., Gustafsson, D., Hölzel, H., Kraft, P., Stamm, C., Stoll, S., Blöschl, G. and Flühler, H.: Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data, Hydrol. Earth Syst. Sci., 13(11), 2069–2094, doi:10.5194/hess-13-2069-2009, 2009.

Houska, T., Multsch, S., Kraft, P., Frede, H.-G. and Breuer, L.: Monte Carlo-based calibration and uncertainty analysis of a coupled plant growth and hydrological model, Biogeosciences, 11(7), 2069–2082, doi:10.5194/bg-11-2069-2014, 2014.

Houska, T., Kraft, P., Chamorro-Chavez, A. and Breuer, L.: SPOTting Model Parameters Using a Ready-Made Python Package, edited by D. Hui, PLOS ONE, 10(12), e0145180, doi:10.1371/journal.pone.0145180, 2015.

Houska, T., Kraft, P., Liebermann, R., Klatt, S., Kraus, D., Haas, E., Santabarbara, I., Kiese, R., Butterbach-Bahl, K., Müller, C. and Breuer, L.: Rejecting hydro-biogeochemical model structures by multi-criteria evaluation, Environ. Model. Softw., 93, 1–12, doi:10.1016/j.envsoft.2017.03.005, 2017.

Hublart, P., Ruelland, D., Dezetter, A. and Jourde, H.: Reducing structural uncertainty in conceptual hydrological modelling in the semi-arid Andes, Hydrol. Earth Syst. Sci., 19(5), 2295–2314, doi:10.5194/hess-19-2295-2015, 2015.

Hudson, G. and Wackernagel, H.: Mapping temperature using kriging with external drift: Theory and an example from scotland, Int. J. Climatol., 14(1), 77–91, doi:10.1002/joc.3370140107, 1994.

Jehn, F.: Zutn/Fluxogram: First Public Version Of The Fluxogram, , doi:10.5281/zenodo.1137703, 2018.

Kavetski, D. and Clark, M. P.: Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing, Hydrol. Process., 25(4), 661–670, doi:10.1002/hyp.7899, 2011.

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.

Kavetski, D., Fenicia, F. and Clark, M. P.: Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, Water Resour. Res., 47(5), doi:10.1029/2010WR009525, 2011.

Kellner, J., Multsch, S., Houska, T., Kraft, P., Müller, C. and Breuer, L.: A coupled hydrological-plant growth model for simulating the effect of elevated CO 2 on a temperate grassland, Agric. For. Meteorol., 246, 42–50, doi:10.1016/j.agrformet.2017.05.017, 2017.

Kraft, P., Multsch, S., Vaché, K. B., Frede, H.-G. and Breuer, L.: Using Python as a coupling platform for integrated catchment models, Adv. Geosci., 27, 51–56, doi:10.5194/adgeo-27-51-2010, 2010.

Kraft, P., Vaché, K. B., Frede, H.-G. and Breuer, L.: CMF: A Hydrological Programming Language Extension For Integrated Catchment Models, Environ. Model. Softw., 26(6), 828–830, doi:10.1016/j.envsoft.2010.12.009, 2011.

Ley, R., Hellebrand, H., Casper, M. and Fenicia, F.: Is Catchment Classification Possible by Means of Multiple Model Structures? A Case Study Based on 99 Catchments in Germany, Hydrology, 3(2), 22, doi:10.3390/hydrology3020022, 2016.

Link, T. E., Unsworth, M. and Marks, D.: The dynamics of rainfall interception by a seasonal temperate rainforest, Agric. For. Meteorol., 124(3–4), 171–191, doi:10.1016/j.agrformet.2004.01.010, 2004.

Maier, N., Breuer, L. and Kraft, P.: Prediction and uncertainty analysis of a parsimonious floodplain surface water-groundwater interaction model, Water Resour. Res., 10.1002/2017WR020749, doi:10.1002/2017WR020749, 2017.

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.

Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D. and Veith, T. L.: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, Trans. ASABE, 50(3), 885–900, doi:10.13031/2013.23153, 2007.

Orlowski, N., Kraft, P., Pferdmenges, J. and Breuer, L.: Exploring water cycle dynamics by sampling multiple stable water isotope pools in a developed landscape in Germany, Hydrol. Earth Syst. Sci., 20(9), 3873–3894, doi:10.5194/hess-20-3873-2016, 2016.

Rawlins, M. A., Willmott, C. J., Shiklomanov, A., Linder, E., Frolking, S., Lammers, R. B. and Vörösmarty, C. J.: Evaluation of trends in derived snowfall and rainfall across Eurasia and linkages with discharge to the Arctic Ocean, Geophys. Res. Lett., 33(7), doi:10.1029/2005GL025231, 2006.

Rhönenergie Fulda GmbH: Trinkwassergewinnung im Fulda Einzugsgebiet <https://re-fd.de/trinkwasser/der-weg-des-trinkwassers> accessed January 2017, [online] Available from: https://re-fd.de/trinkwasser/der-weg-des-trinkwassers, 2017.

Ritter, A. and Muñoz-Carpena, R.: Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments, J. Hydrol., 480, 33–45, doi:10.1016/j.jhydrol.2012.12.004, 2013.

Rutter, A. J. and Morton, A. J.: A Predictive Model of Rainfall Interception in Forests. III. Sensitivity of The Model to Stand Parameters and Meteorological Variables, J. Appl. Ecol., 14(2), 567, doi:10.2307/2402568, 1977.

Samani, Z.: Estimating solar radiation and evapotranspiration using minimum climatological data, J. Irrig. Drain. Eng., 126(4), 2000.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16(9), 3315–3325, doi:10.5194/hess-16-3315-2012, 2012.

Singh, V. P.: Is hydrology kinematic?, Hydrol. Process., 16(3), 667–716, doi:10.1002/hyp.306, 2002.

Son, K. and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data: Improving model structure through auxiliary data, Water Resour. Res., 43(1), doi:10.1029/2006WR005032, 2007.

Todini, E.: Hydrological catchment modelling: past, present and future, Hydrol. Earth Syst. Sci., 11(1), 468–482, doi:10.5194/hess-11-468-2007, 2007.

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.

Westerberg, I. K. and Birkel, C.: Observational uncertainties in hypothesis testing: investigating the hydrological functioning of a tropical catchment: Observational Uncertainties in Hypothesis Testing, Hydrol. Process., 29(23), 4863–4879, doi:10.1002/hyp.10533, 2015.

Windhorst, D., Kraft, P., Timbe, E., Frede, H.-G. and Breuer, L.: Stable water isotope tracing through hydrological models for disentangling runoff generation processes at the hillslope scale, Hydrol Earth Syst Sci, 18(10), 4113–4127, doi:10.5194/hess-18-4113-2014, 2014.

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.