We thank the reviewer for the valuable comments which will help to improve the paper. The responses (R) to the different comments (C) of the reviewer are provided below.

1 General comments

C 1. The manuscript 'Incorporation of rating curve uncertainty in dynamic identifiability analysis and model structure evaluation' presents an application of GLUE and DYNIA for model structure evaluation. Although the presented research covers the title, the use of 'incorporating' might be misleading, since the rating curve uncertainty is only used to determine 'limits of acceptability' for the former methodologies. It is only in the 'discussion and conclusions' section that the effect of the use of the rating curve uncertainty on the model structure evaluation is highlighted.

> **R 1.** *The paper will be restructured to overcome the current chaotic behaviour of the text. The general structure of the paper is split in Introduction, Materials & methods, Results, Discussion and Conclusions. Material and methods first introduces the data (study catchment) together with the rating curve uncertainty derivation (a in Figure 1). Bringing the latter to the materials and methods section clarifies for the reader that these data uncertainty boundaries are used as a starting point of the further analysis. This point was also raised by reviewer 1 and is discussed in detail in Reply to comment 1 of the first reviewer. Furthermore, the title will be adapted to 'Dynamic identifiability analysis based model structure evaluation considering rating curve uncertainty' to bring more focus towards the DYNIA and model structure evaluation, and down tune somewhat the focus on the rating curve uncertainty derivation.*

C 2. The topic addressed in this manuscript is quite broad, since a number of different techniques are combined and compared. The focus on dynamic identifiability is certainly attractive for model development and the comparison with the more traditional approach (of an evaluation of the model structure and parameters over the whole calibration period) shows that this methodology can be a great contribution. The use of the rating curve uncertainty additionally overcomes a traditional pitfall of determining the uncertainty on the observations.

> **R 2.** *We appreciate this comment and the value the reviewer sees in this work. The main added value of the paper is the combined approach of the applied methods. The information coming from the different methods enables an enhanced model evaluation. Nevertheless, we admit the risk of losing the overview of how these methods are interconnected and propose to add Figure 1 to the introduction to enable the reader to get a clear overview of how the applied methods should connect.*

C 3. In general the manuscript is well written (despite the relatively long list of technical corrections/suggestions) and clear. Nevertheless, the structure is sometimes a little bit awkward, with literature review, methodologies and results mixed in several sections, however still clear enough.

R 3. *This will be handled by a restructuring of the paper outlined in Reply to comment 1.*

2 Specific comments

C 4. p11446, L13: What is the corresponding area of the catchment?

R 4. *362km^2, this was added*

C 5. p11447, L7-16: You give the impression that you start describing your work here with respect to the uncertainty envelope, but after reading this whole section 4, it seems that this paragraph just represents the general methodology. Adapt your formulations in a sense that this becomes more clear from the manuscript.

R 5. *For the analysis, the method for deriving the rating curve uncertainty was performed with a similar approach as was done for the general methodology, following the method of Pappenberger et al. (2006). This apparently leads to some confusion. In the revised version, the differences will be emphasized and the uncertainty derivation of the rating curve will be more condensed to avoid misinterpretation.*

C 6. p11448, L2: What do you actually mean with this !uncertain membership region'? Is this a sort of uncertainty band on the calibration measurements wherein the rating curve(s) should fit? Is it possible (and clear) to plot these as e.g. whiskers on Figure 3? The latter would make it already more clear for the reader (although I do understand that you need some differentiation from the terminology uncertainty bands' on the time series, which are derived from the uncertainty envelope).

R 6. *The uncertain membership region is indeed the selected uncertainty band on the calibration measurements. Due to the underlying selection and corresponding to other literature, this term was selected. Furthermore, it clarifies the difference with the later defined uncertainty boundaries on the flow which could confuse the reader.*

C 7. p11450, L17: I would suggest to use the original reference for the Sobol quasi-random sampling: Sobol, I.M., 1967. On the distribution of points in a cube and the approximate evaluation of integrals. USSR Computational Mathematics and Mathematical Physics, 7: 86112. Sobol, I.M., 1976. Uniformly distributed sequences with an additional uniform property. USSR Computational Mathematics and Mathematical Physics, 16: 236-242.

R 7. *The references will be interchanged.*

C 8. p11450, L17: Which distribution did you apply for the sampling of the parameters?

> **R 8.** *Sampled from uniform distributions, this will be added.*

C 9. p11450, L27: Is the range of -2/2 for the limits of acceptance not too wide? You allow the predicted values to become double of the 95th percentile and half of the 5th percentile. That will result in about a 99% confidence interval for the observations (under the assumption of a standard normal distribution), which you even allow to cross 10% of the time steps. Provide additional information on the choice of this interval and the acceptability in your study.

> **R 9.** *This is a trade-off between acceptance and model evaluation as stated in the paper itself. This was phrased in the discussion as follows:*
>
> > *'A restriction in the application of the limits of acceptance approach is the need for relaxation of the initial limits of acceptance to avoid rejection of all model simulations 25 (both in terms of parameterization and structure), as was also needed in Blazkova and Beven (2009) and Liu et al. (2009). However, since the focus is on model improvement, the approach is based on rejection rather than optimization to identify and focus on particular parts of the time series that are not well simulated (Beven, 2008).'*
>
> *As an initial selection step to perform the DYNIA approach, these conditions enables to derive sufficent information about parameter identification, giving the opportunity to punish overall poor models, but still avoids the risk of rejecting an adequate model realization. It is outside the scope of the paper to fully layout the effect of the decisions made to extend the acceptance conditions and a comparable balance was used as in the mentioned literature references. In the revised document, these arguments will be added.*

C 10. p11451, L19: Figures 6 and 7 are not really clear, not even in colour. Additionally, the presentation of the results in this paragraph should stick to the results itself. You already start discussing the results and do not highlight the specific observations from the graph (in particular for Figure 7).

> **R 10.** *The authors are convinced that the current layout supports the message sufficiently, but will test if an empirical cumulative distribution instead of the current histograms supports the message better, since it only needs one line in the visualisation instead of the different histogram bars (Figure 2).*

C 11. p11452, L2: Did you also perform this analysis for the validation period? Adding this to the results could improve the manuscript and would substantiate the findings of the study.

> **R 11.** *Since this was a preliminary result before evaluating the DYNIA approach itself, this was only done on the calibration period. Validation is only performed on the last part with*

visualisation of the prediction uncertainty to validate the conclusions made about the model structures based on the DYNIA approach. We are convinced that adding these results also on the validation period would overload the paper. Moreover, the current selected calibration years already embrace different climatic conditions.

C 12. p11452, L4: It is not really clear to me why you have such an extended literature review in this section. In particular, a number of DYNIA applications are discussed in detail without really adding value to this part of your manuscript. It is not really clear to me, where you want to bring us by presenting this. It is not necessary to delete this whole section, but a reduction of both the number and the level of detail of cited articles might be appropriate.

R 12. *The author wanted to give an overall overview of previous DYNIA applications, mainly on those working with similar model structures. It enables the reader to compare the current results with previous applications in literature and get a general overview of the possibilities of the method and variations done. As such, readers get an overall impression based on this paper. Therefore, this will be moved to the introduction part to support the scientific background.*

C 13. p11453, L5-8: If you keep this sentence in the manuscript, you should elaborate on this, because it is not clear what you mean by this statement.

R 13. *The authors agree on the potential misinterpretation of this statement. The original reference itself is rather diffuse about it. Basically it expresses the idea that a conceptual model is only valid for a limited space and time. Major changes of the underlying system (e.g. landscape changes) are mostly not covered by the conceptual approach. In the revised version, this point will be made more clearly.*

C 14. p11454, L8-9: Explain this statement more, because in this way, it is not clear (e.g. do you mean 'time' or 'parameter space' regions?)

R 14. *parameter regions, this will be adapted*

C 15. p11455, L4-6: Something is missing in this sentence. What 'minimum level'? Can you explain or prove this statement?

R 15. *'of performance', this will be added. By doing so, the DYNIA approach gets a two-step application with a prior selection of a subset of model simulations. On one hand, this decision comes from our practical experience, making it more robust to the time-window that needs to be selected for the different parameters when applying DYNIA. On the other hand it also makes the analysis more relevant in the combined approach, since it highlights* *the compensation of parameter values conditioned by overall performance. The latter is in the end what is used to derive the prediction limits.*

C 16. p11455, L19: Although you have a discussion section, you start to discuss the result already in this section. Try to avoid this. (Similar remark for p11456, L1-5, p11458, L16-18 and p11459, L1-5)

R 16. *These remarks will be taken under consideration and the interpretation of the physical reasoning behind parameter changes will be converted to the discussion part.*

C 17. p11456, L5: Is the compensation of the Lmax parameter by the CKBF parameter only caused by a deficient model structure? Or can this be also related to overparameterization? How can you distinguish both?

R 17. *Overparameterization and model structural deficiencies are directly related. Overparameterization leads to deficiencies in the model structure and compensation of parameters. On the other hand, overparameterization is not the only reason of model structure deficits, sometimes the model is just not capturing the real dynamics. The main idea is that the parameters are identifiable for the processes that they tend to describe. Overparameterisation (and thus model structure failure) emerge through the occurrence of parameter compensation.*

C 18. p11457, L9: Is this an application of GLUE or DYNIA? I guess GLUE, because I dont see how you could get parameter sets for the validation period with the DYNIA approach. Nevertheless, you should make this more clear.

R 18. *The prediction uncertainty is indeed a result of the GLUE application on the final selected set of parameters. In the revised paper, the difference between parameter sets for DYNIA and GLUE approach will be better expressed.*

C 19. p11457, L16: I dont think you are underestimating the flow peaks in March! Opposite, you seem to overestimate them!

R 19. *Indeed. Correction will be made in the revised document.*

C 20. p11458, L7: Similar as for the remark on p11450, L27: it is questionable if the range of -2.5/2.5 for the limits of acceptance is not too wide?

R 20. *The iterative process of deriving these limits was not added to the paper and is drawn based on expert decision and previous literature (Liu et al. (2009) applied relaxation until 4 times the initial estimated uncertainty). Furthermore, the relative over and underestimation of the flow by the model in terms of flow magnitude remains the same (uncertainty on high flows remain relatively larger than those on low flows). Further research is definitely needed to improve insight.*

3 Technical corrections

R 21. *We would like to thank the reviewer for his detailed technical correcting, the quality of the manuscript will benefit a lot from the effort made. We will consider and include the different remarks in the revised paper.*

C 21. General: Look out with the mixing of the past and the present tense. This should be consistent, which is not always the case.
...

R 22. *This will be proof read in the revised paper*

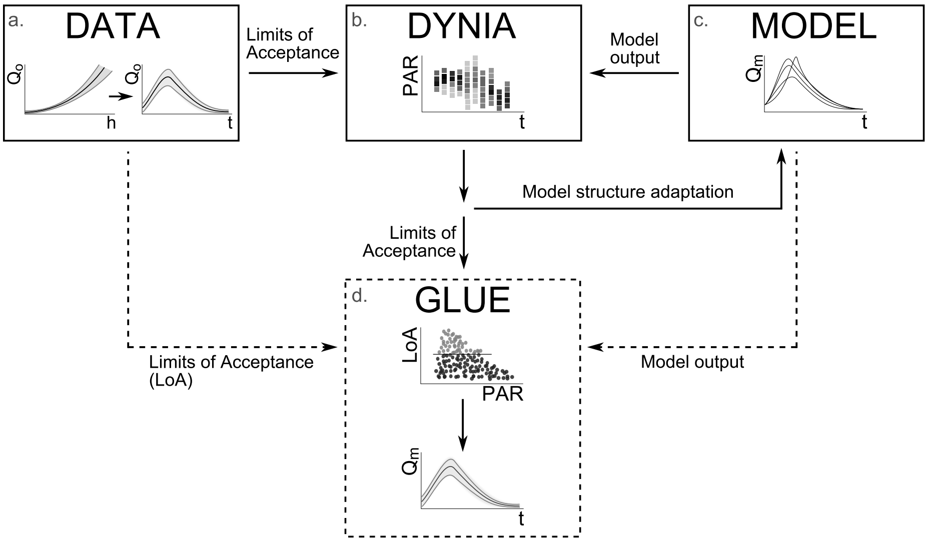
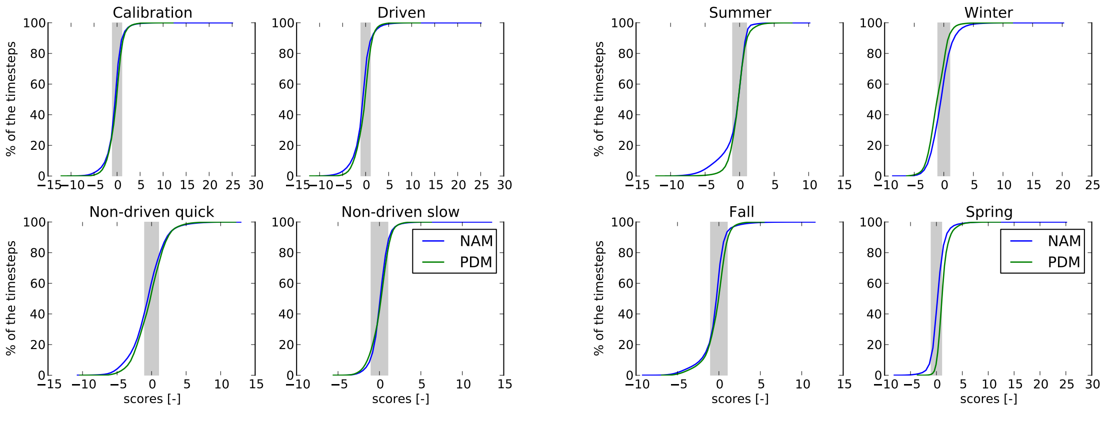


Figure 1: Schematic overview of the paper. The uncertainty in the rating curve is used to derive initial limits of acceptance (a), the DYNIA approach is used to evaluate the model structure and parameter identification (b) based on a Monte Carlo set of model runs (c). The prediction uncertainty is then determined based on the GLUE approach using the combined limits of acceptance information from the rating curve uncertainty and DYNIA approach (d).



(a) Driven and non-driven periods

(b) Seasons

Figure 2: Revised version of Figures 6 and 7 of the initial paper, showing the histograms as empirical cumulative distribution functions.