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.

C 1. The paper is about incorporating of rating curve uncertainty in model evaluation using dynamic identifiability analysis. The question raised by the title is interesting, how incorporating rating curve may help modelers to avoid type I and II errors. However I have to express my concern about the content of paper as well as the structure.
The paper is not well structured in my point of view. The literature review of different method should not be spread over the entire paper; it should be presented in a way that gives readers the ability to understand the relevance of previous studies and this study.

> **R 1.** *We have the feeling that the organization of the paper prevented the reviewers from keeping the overview and grasp the messaged we wanted to convey. Therefore, the paper will be restructured to improve readability. To assist this, an extra figure will be added to the paper that gives a concise overview of the different elements in the paper and illustrating the underlying connections (Figure 1).*
> *The general structure of the paper will be split in Introduction, Materials & methods, Results, Discussion and Conclusions. All state of the art literature background will be compiled in the Introduction (see next comment).*
> *Material and methods first introduces the data (study catchment) together with the rating curve uncertainty derivation (Figure 1a). Moving 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. Secondly, the link between the rating curve uncertainty and the derived limits of acceptance for model evaluation is explained (arrows with LoA in Figure 1). Subsequently, the DYNIA method is explained and the differences with the approach of Wagener et al. (2003) are indicated, since in the presented paper an initial selection of the simulations is applied before adding the time window (Figure 1b). This is addressed as a two-step application as opposed to Wagener et al. (2003). Next, the determination of the prediction uncertainty with the GLUE approach based on the limits of acceptance is briefly introduced. Finally, the model structures are explained together with selected uniform parameter distributions, the sampling used and Monte Carlo runs (Figure 1d).*
> *The results part in the revised paper will be organised in two main sections: (1) the (two-step based) DYNIA approach (solid line arrows in Figure 1) and (2) the prediction uncertainty with GLUE (dotted line arrows in Figure 1). The first part explains the initial selection of model simulations (Figures 6 and 7 in the initial submission) to clarify the usefulness of this initial subset selection and then collects DYNIA results on this subset in more detail. The second part of the results evaluates the posterior uncertainty of the selected model structures for both the model calibration and validation period. Furthermore, the effect of the posterior parameter sets on seasonal selection is looked at in more detail.*
> *The discussion and conclusions will be reorganised to better state how the research questions have been organised (see further).*

C 2. The title and research question is interesting; however I did not find the methodology proper to answer the research question. Logically if the authors want to compare the effect of rating curve on model structure they should look how models perform for calibration on different discharges

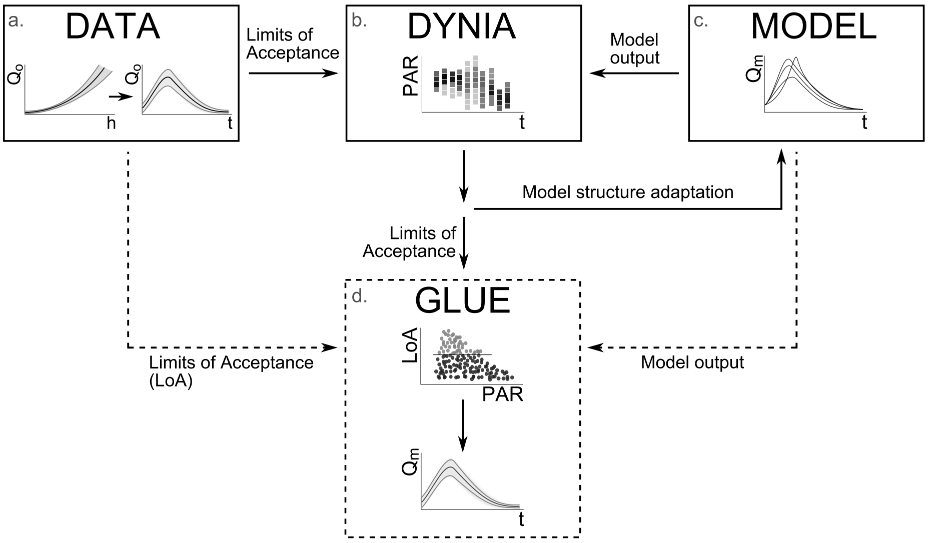

Figure 1: Schematic overview of the paper. The uncertainty in the rating curve is used to derive initial limits of acceptance (a), the DYNIA approach is used to evaluate the model structure and parameter identification (b) based on a Monte Carlo set of model runs (c). The prediction uncertainty is then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

obtained by rating curve and see how different the parameter ranges become. I completely missed the link between rating curve analysis and model evaluation.

> **R 2.** *The rating curve analysis is used as a starting point in the analysis. As explained in recent literature (Clark et al. 2011, Blazkova and beven 2009), the uncertainty of the data hampers the model evaluation. Furthermore, as stated by the reviewer, it gives rise to type I and II errors. The paper considers this rating curve uncertainty in the DYNIA approach and apply an adapted version of the DYNIA approach (to the knowledge of the authors, this combined approach has not been reported previously in literature).This will be better addressed in the paper.*

C 3. Flexible model structure is an interesting approach for hypothesis testing, it is a laboratory in which different model structures representing different hypothesis can be evaluated. However, it can be misleading if it is not handled carefully; higher uncertainty in parameter estimation does not necessary mean poorer model structure for chosen catchment. Moreover comparing different model structure within Flexible framework is subjected to careful scrutinization, and with model structures with completely different development background the comparison should be based on many other observed data which in this case study seems to be absent.

> **R 3.** *The authors admit that the approach of flexible model structures is not completely*

relevant in the presented case, since only two model structures are compared and the evaluation is done solely on discharge observations. As stated previously, other data is not always present. Hence, getting as much information for model evaluation from the information given by the flow observations is essential and the presented paper supports this.
The authors do not completely agree on the complete different background of the models. These lumped conceptual model structures are in essence always expressible in a simple set of Ordinary Differential Equations (ODE), as done for the implementation of the NAM and PDM versions used in the paper. They differ in their state variable declaration, mass balance of the used 'reservoirs' and the description of the internal fluxes, but could be easily included in a flexible approach as presented by Clark et al. (2008). However, since the main message of the paper is about model structure evaluation, it was decided to present the models in their original description to improve the understanding for the reader.
The authors agree on the risk of putting the wrong focus by bringing too much attention to the flexible approach. In the revised version, the authors will leave out the part on flexible model structures in the introduction and will rewrite the first part in function of general model structure evaluation. The extension to flexible model structures will be left for a separate paper.

C 4. The paper, simultaneously, tries to look at model consistency over different conditioned (wetting, drying and : : : periods). This opens another front of investigation, which is parameter stability over time, but the authors did not mentioned relevant studies which have been done in this regard.

R 4. *The DYNIA approach applied in the paper is in essence a method that evaluates the model structure based on the stability of optimal parameter values in time. This is not another front of investigation, but an essential part of this DYNIA approach. Relevant literature was present in the paper, but was partly included in the discussion section, which might be the reason the reviewer missed it (see pieces of text below).*

> *'Temporal analysis to evaluate the information content of the data and to extract signature information is a valuable procedure to identify potential model deficits. de Vos et al. (2010) use temporal clustering to identify periods of hydrological similarity. Reusser and Zehe (2011) propose an approach to relate types of model errors with parameter sensitivity and model component dominance to understand model structural deficits. Reichert and Mieleitner (2009) combine the estimation time dependent model parameters with the degree of bias reduction to identify model deficiency. The DYNamic Identifiability Analysis (DYNIA) developed by Wagener et al. (2003) builds on the GLUE by evaluating the parameter identifiability in a moving window.'*
>
> *. . .*
>
> *'Based on the variation in optimal parameter sets, both in seasonal variation and on storm level, and the relation between model states and optimal parameter combinations lead to the of introduction of time-variant and stochastic parameters (Beck and Young, 1976; Cullmann and Wriedt, 2008; Lin and Beck, 2007; Reichert and Mieleitner, 2009; Kuczera et al., 2006; Tomassini et al., 2009) and is associated with the Data-Based Mechanistic (DBM) approach using state-dependent parameters to identify non-linear systems (Young et al., 2001). The main argument for introducing stochastic parameter values is the inherent stochasticity of these conceptual models due to spatial and*

> *temporal averaging (Kuczera et al., 2006). Nevertheless, the use of time-variant parameters remains mainly useful in terms of model structure evaluation. The idea of allowing parameters to vary in time to gain information about potential model structural improvements goes back to Beck and Young (1976) and the potential of learning from the behavior of time-dependent parameters is higher than from corrections in model states (Reichert and Mieleitner, 2009). Cullmann and Wriedt (2008) argue to incorporate the state-dependent parameter changes in the formulation of conceptual model intended for continuous simulations, adapting to governing processes.'*

The authors believe that with the proposed revisions on the structure, the focus will be clearer for the reader in this regard. Parameter stability is used in the method and essential in the parameter identification.

C 5. The case study should be more transparent,

R 5. *The data-section will combine the information of the study catchment together with the rating curve uncertainty to bring together the information on the case study in one place in the paper. Since the model selection is focussing on the discharge observations, the study catchment is only shortly described. The selected years for calibration and validation are representative years with sufficient variability in between years to support the research question.*

C 6. the objectives of paper should be set up more clearly and the answer to the research question should be given more accurately backed with strong reasoning.

R 6. *The general idea of the paper is an [1] improved model structure evaluation, preventing the modeller from taking type I and type II errors and [2] attempting a more objective derivation of the prediction uncertainty. Different approaches for improved model evaluation and identification are possible: (1) Using different sources of information (other types of data); (2) using multiple (noncommensurable) objective function or limits of acceptability; (3) Taking into account the data uncertainty (input and output data) and (4) model structure evaluation by model parameter identification. The inter-relation between the different aspects decreases the transparency of the added-value. The paper proposes a combined approach (Figure 1). It should be noted that only flow data is available in this case as it is in many applications and as such the proposed method is not dependent from these extra data sources but is generic enough to be extended and include this.*
The paper supports the idea of using limits of acceptance (2), both by the rating curve application and the ability to propose evaluation functions that are able to discriminate the model structures on their performance. The latter information comes from the DYNIA approach, indicating where model structures have potential pitfalls. Instead of testing multiple objective functions hoping that differences will be seen, the DYNIA analysis indicates 'where' these difference can be found. Practically for the presented paper, the seasonal evaluation is essential to compare the performance of both models.
Uncertainty of the data (3) is applied in this paper by considering the rating curve uncer-

tainty. Rainfall uncertainty is not taken into account, but this is addressed in the discussion. Parameter identification (4) is directly evaluated by the DYNIA approach.