Reviewer 2:

Reviewer 2 seems to be a quite senior scientist with in depth experience in catchment modelling. Hence, he/she addresses some very critical aspects of this study. In many cases, catchment models are not being used by individuals with such a background and that much experience. We tried to make the problems visible in the manuscript in order to bring the educational value of this exercise to the attention of the reader.

Comment 1:

The aim of this paper it seems is to compare underlying model assumptions in the prediction of ungauged basins, which is the first step of a 3 step exercise. However, surely there is an essential missing first step here, where a design rainfall and design catchment are used (which could of course be based on the dimensions and characteristics of Chicken Creek). Then there would be a benchmark set of results with which to compare the next step with, where the real catchment setup is attempted. This would be an easy and essential first step to make.

Answer:
We disagree with this argument. What we do is to test the quality of an a priori prediction in a given, physically real case. To do a model comparison with made-up data (mimicking a similar catchment) is in fact a brilliant idea which goes way beyond what can be done in this paper, which is already excessively long in the submitted form of manuscript. We keep the suggestion in mind for the up-coming project period. Such a model test on designed cases can also be done after the ongoing project phase

Comment 2:

There are effectively 2 things being tested in this paper at the same time: effects of (i) model structure (ii) modeller practice. This would be an obvious separation in terms of an introduction and discussion at the end.

Answer:
We integrated this point into the discussions.

Comment 3:

Where is the discussion? The discussion section is just a description of results. Please think about the implications of this work. Again there is a huge potential for a hard hitting paper here.

Answer:
See reviewer 1 Comment 1: We agree that is was absolutely needed to divide results and discussions. We hope this made it clearer and that the main issues are presented more explicit.

Comment 4:

Where is the sensitivity analysis!!? There are some suggestions of fiddling around with some model parameters, setup decisions to achieve more reasonable results, but surely an essential part of such an intercomparison exercise is (preferably a formal quantitative) sensitivity

analysis of model parameters, initial conditions, setup decisions etc. Poor modelling practice. Some of the models (especially the physically based ones) are overparameterised, so sensitivity analysis is essential. How can you draw conclusions about the effects of model structure/modeller decisions on discharge predictions without sensitivity analysis? How can we learn from this exercise otherwise?

Answer:
We argue the same way as done answering Comment 1 (Reviewer 2). It is obvious that more can be done and we intend to do more along these lines. But there is no way in packing more and more load to an already lengthy paper.

Comment 5:

Why was only one set of results per model shown in results? E.g. why weren't monte carlo simulations run from feasible parameter space. I have first hand experience of running physically based catchment models and it is just not that difficult with current computational resources. Again, this paper displays poor modelling practice. With such a multi simulation approach some estimate of the range of feasible results could have been presented and discussed (uncertainty). Then a really useful comparison of the different ranges provided by the different model structures could have been discussed. I think the conclusions may have been very different had this been done. I would suggest all future attempts at this PUB exercise at Chicken Creek are designed more carefully to take account of this.

Answer:
We argue the same way as done answering Comment 1 (Reviewer 2). It is obvious that more can be done and we intend to do more along these lines. But there is no way in packing more and more load to an already lengthy paper.

Comment 6:

Where is the measurement error in the catchment discussed (especially as this is a back-calculated value)! How can you compare absolute model results with absolute model predictions without taking into account the uncertainty?

Answer:
We discuss the measurement error of the lake outflow in section 3.3 and section 3.8. It is much smaller than that error which we introduce by back-calculating the outflow of the catchment from the lake outflow.

Comment 7:

Results are poorly presented. Please show me some graphical statistics – range of results for Q95, Q50, Q5, particular events etc.; deviation from benchmark. The results section is very dense to read and more graphical figures of any nature would be a benefit. By the very nature of this intercomparison experiment several figures are required to fully understand what's going on!

Answer:
We added more results to section 4.

Comment 8:

As already noted by Reviewer #1, English is clumsy in places. Please get manuscript checked by native speaker.

Answer:
Yes, this critique was justified. It is easy (easier) for a native speaker to write a manuscript in a "non-clumsy way". Instead of making depreciatory remarks about language, it would be sufficient just to require a re-editing by a native speaker. We suggest to the reviewer to be more considerate in phrasing this type of critique addressed to those who take the trouble of writing in a language which is understandable for those who most likely master not many foreign languages.

Comment 9:

Terminology. Throughout the manuscript poorly defined terminology is used. E.g. Discussion of model validation (p.3203, l.5): Mostly the term 'validation' is not used now and is replaced with 'model performance'. Explain why you discuss validation. E.g. you need to discuss how you are defining 'physcially based' and 'process based' models. What is the difference (sect 3.2)?

Answer:
We apologize for this and took care of the correct terminology.

Comment 10:

Section 3.2 seems to have been written by different people- each section needs to be in the same format and written in the same way (see point above about using coherent terminology). Although several tables have been provided later I think a simple figure comparing e.g. y-axis: conceptual-physically based and x-axis dimensionality, would place these models in some kind of context.

Answer:
Each section has been changed to the same format and terminology. We do not agree that the mentioned figure gives additional and more valuable information. Reviewer 1 also mentioned that the tables are useful. Reviewer 1 Comment 4: "The processes are sorted nicely into tables, …".

Comment 11:

Title is slightly misleading. This paper is about using the catchment in 'ungauged' mode and i think the title should focus on this, rather than using term 'sparse'?

Answer:
For the modeller it is both, it is ungauged and the data are sparse. In the planned second paper (step 2) the data will not be (very) sparse, but the catchment still ungauged. Hence we prefer to have this being prominently signalled in the title. The word ungauged shows up in the abstract and is therefore retrievable by a web search.

Comment 12:

The manuscript is incoherent in places. Not only is the discussion missing (and already noted above) but methodology is found in the results section and results are found in the discussion section. E.g. correction strategy for precipitation input is methodological (p. 3223). Overall the logic of the structure comes across as confusing. I was constantly turning pages to find the information that I needed to understand what was going on. I think a major restructure is required.

Answer:
Yes, we agree and as a consequence we rearranged sections 3.4, 4 and the discussions.

Comment 13:

It would be very useful to know the computational time needed for this exercise by the different models (preferably using comparable computational resources, or failing this just a description of what resources were used). This is a very important consideration for modellers attempting any exercise and especially for intercomparison exercises. In addition what time estimates are there for model setup and testing time?

Answer:
We added the computation time in section 3.9. All computer resources used by the modeller were standard personal computers. The computation time was always below 10 hours. The main testing time of the models was within one week. Hill-Vi needed to implement code so that the longer mainly due to the code development (section 3.9).

Comment 14:

Ksat parameter p. 3225. Good that the ranges of Ksat used are documented here, however as most modellers know very well Ksat is usually a very sensitive parameter and is certainly not interpretable as a 'physical' value. Ksat is an effective parameter which is known to e.g. vary with grid cell size. Thus the ranges discussed here are dependent upon the model used as well as other variables (such as grid size). Without a benchmark sensitivity analysis it is impossible to draw the conclusions that are attempted here.

Answer:
We fully agree with the reviewer. We tried to phrase it correctly to avoid the misunderstanding because we do not consider Ksat as a physical parameter. It is a parameter which is in almost all model applications related to a magic permeability of the soil and it is rarely ever based on a sufficient number of "measurements". In this study it is estimated the way it is estimated in the majority of the application cases. In addition, we stress that Ksat is not the **most sensitive** parameter in our study. The van Genuchten parameter**s** result in larger differences.

Comment 15:

Why was SWAT used if it isn't normally used for small catchments (p. 3227)

Answer:

This was the modeller's decision. Actually there are more modelers who intend to apply SWAT also in this "simple" case where the catchment property is not so complex and where they will be known after the exercise.

Comment 16:

Results. Needs to be separate sections on for example (i) what are the results – i.e. what are the range of overall discharge predictions (ii) how these compare to observations – thus model performance (iii) why this might be (links to representation of hydrological processes and model structure – which information should already have been given in setup section!)

Answer:
We tried to incorporate these recommendation by rearranging and partly rephrasing (shortening) the sections.

Comment 17:

The decision making of the modellers needs to be more explicitly looked at. In fact i think that this is such a valuable part that it has the potential to be a paper on its own: I would suggest using social science expertise to look at not only the obvious decision making but also some of the tacit assumptions of the practitioners in this exercise. I would encourage the authors to look into this. This would enable not only my queries about particular points in this paper (e.g. why did the hydrus-2d modeller use a daily input time series – i can't accept that computational time was enough of an issue to warrant this decision and if it was based on a hydrological decision then it needs explaining) to be answered, but also provide a broader look at the practice of hydrological modelling which is a very valuable exercise (similar exercises are currently underway for flood forecasting and climate modelling for example).

Answer:
We wholeheartedly agree with these statements and recommendations. We try to make these aspects visible but we do have a problem with including more aspects because of the manuscript size. In order to make clear what has been done and how we fill so many pages that some internal reviewers suggested deleting some of the presented information. We are convinced that deleting it would lead to an information deficit for the reader. In this sense, the above suggestions must be the topic of a follow up (as alluded by the reviewer). There will be a follow up in this sense and the remarks of this reviewer encourage us to do so.