Review for "Behind the scenes of streamflow model performance" by Bouaziz et al.

Model evaluation is an essential and important step in any hydrological modelling study and is typically based on streamflow data. The study of Bouaziz et al. takes a different perspective, in the sense that it assesses internal states and fluxes (streamflow, evapotranspiration, root zone soil water content, and total water storage) of twelve models that were shown to have a comparable (acceptable) streamflow model performance for the Meuse basin. The idea of multi-variable evaluation is not novel, however, it is rarely conducted in such a detailed and extensive way as in this study. I personally like the set-up of the multi variable analysis and I think that the final outcome of this study clearly contributes to improving our knowledge on model evaluation. The current state of the manuscript reflects the vast amount of results related to a multi-model/ multi-variable study and I encourage the authors to invest some more time into analysing and presenting results in a more concise way. I hope that the comments below will be helpful for the authors to improve their study.

Major comments:

1. The four variables streamflow, evapotranspiration, root zone soil moisture content, and total water storage are used in very different ways to evaluate model performance. For example, i) the temporal resolution varies from multiple years, to one year, to one season, or to a single month; ii) some variables are evaluated using magnitude related metrics others using dynamic related metrics; and iii) a different number of aspects of a particular variable is analysed. This diversity can be confusing and makes it difficult to understand the big picture as one is confronted with something new in every figure. I highly recommend to think about a way to analyse and present the data/results in a more consistent way.
2. The manuscript describes and presents the methods and results in quite some detail. This makes it difficult to separate the key messages from the more "nice to know" aspects. For example, i) the manuscript contains a total of 46 (sub)figures. Do you really think every single sub(figure) is needed to tell your story?; ii) there are multiple text sections that could easily be shortened by removing details that are not important for understanding the story (see detailed comments). I strongly encourage the authors to better filter information in order to improve the clarity of the manuscript.
3. The results are often only presented for one of the three catchments, which is related to the rather qualitative model evaluation. Showing the results for all three catchments would enhance the robustness and relevance of the results.
4. The study has two main objectives, whereby the second one is to evaluate models using "soft" measures. I didn't have the feeling that the "soft" measures were really an aim of this study – they were rather a tool (that was actually only used in the very last figure). If the use of "soft" measures was one of the objectives, then I would argue that you would have to do some serious testing of this measure that includes a comparison against other typically used evaluation metrics.
5. Model calibration was based on hourly streamflow data, whereas the evaluation was based on data with daily or monthly resolution. Don't you think that the discrepancy in temporal resolutions will lead to an even stronger "overfitting" to streamflow? In other words, would the performance for the internal states and fluxes be better if the model was calibrated against daily streamflow? Furthermore, model evaluation was conducted in a different period than model calibration, which introduces additional uncertainty. Do you think the results would be different if you did the evaluation in the calibration period?

Minor comments:

P2 L36-45: This paragraph lists a lot of studies on multi-variable calibration. However, as the list is not complete, I recommend to use "e.g." when listing references.

P3 L66-74: The use of "we" is confusing. Please make sure that it is clear what was part of the previous study and what is part of the current study. Furthermore, this is one of these paragraphs that could be shortened, also because the same information will be repeated in the methods section.

P4 L83-107: The reference for the land cover information is missing.

P5 L117: "…follow the same approach to extend the dataset…". Do your refer to temperature or ET?

P8 L234-252: i) This is one of the sections that could be shortened. E.g., the information about the proxy-basin test or about the evaluation from 2001-2003 and 2008-2010 is not relevant for this study; ii) How many parameter sets did you have per model?; iii) How similar are the calibration and validation period in terms of streamflow, evapotranspiration, root zone soil moisture content, and total water storage?

P9 L275-280: The description of the confusion matrix is rather confusing. I would directly jump to L280, where you actually say what you will evaluate (i.e, "…ratio of days when snow observed by MODIS is correctly identified by the model,…").

P10 L289: Is SR the range of relative root-zone soil moisture or the relative range of root-zone soil moisture? The equation shows a ratio and not a range.

P11 L318: You later mention that your goal was not the reject models or to find the best one. However, ranking models inevitably leads to a comparison. Therefore, wouldn't it be better to make a binary classification, i.e., accept or reject models for a particular variable?

P11 L 324: The results section is often mixed with discussion. Examples are L346-349, L356-359, L39-440, L 465-467, L481-483, L500-503.

P17 L535-536: Where do you show that the calibration strategy influences model performance?

P19 L575: Is it spatial resolution or spatial coverage?

Fig. 1: I think Fig. 1c is not needed as your description in the text is very clear. The variable EA in Fig. 1c is not defined.

Fig. 2: I would suggest to add labels to the individual tanks and fluxes, because colours are not intuitive. Why do wflow_hbv and FLEX-Topo have a different number of elevation bands?

Fig. 3: Why is the range in model performance much larger for NSE,logQ than for NSE,Q? What are the feasible parameter set? How many are they? How representative is Fig. 3d for all the other models (or what do we learn from it)?

Fig. 4: I would chose boxplots to show the interception values to be consistent with evapotranspiration. It is not clear what exactly the bars show (i.e, is it the 25-50% quantile?).

Fig. 9: Could you turn this figure into a heatmap? For which catchment is this figure?