Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-635-AC2, 2019


Hydrology and Earth System Sciences
Discussions

# *Interactive comment on* "Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across a large-sample of catchments in Great Britain" *by* Rosanna A. Lane et al.

**Rosanna A. Lane et al.**

r.a.lane@bristol.ac.uk

Received and published: 26 April 2019

We thank reviewer 2 for taking the time to review our manuscript, and provide useful feedback. Our responses to each review comment are given in bold below.



Summary

This study compares four structures from the framework FUSE in 1100 UK catchments. This is, in itself, a significant achievement. The authors highlight which structures perform best in different regions (Results Section) and then discuss more generally why models fail and which improvements would be necessary to improve performance (Discussions Section). I think that, ultimately, the goal of such model intercomparison is to provide guidance on i) model selection (i.e., can specific models/modules be recommended based on basin attributes?) and ii) model development (i.e., are there specific process parameterisations that are currently missing, but are needed to improve the simulations?). In my view, the latter point is addressed quite well (although I suggest restructuring the text to makes these results stand out more, and to go beyond FUSE structures by discussing modelling decisions more generally) but the former point could be addressed in a more systematic and comprehensive way. Overall, I consider that, after revisions, this paper has the potential to become a timely and welcome addition to the literature.
**Response: We thank reviewer 2 for these helpful comments, and for taking the time to review our manuscript.**

Main comments

Model intercomparison vs. benchmarking: Since the authors use the term "benchmarking" in the title and throughout the manuscript, I encourage them to clarify in the introduction what differentiates model benchmarking from model intercomparison. As the authors compare FUSE structures with each other, isn't their study rather a model intercomparison? Do the authors mean that their runs can be used as benchmark by future studies, as suggested on P12L12? Please clarify.

**Response: We agree that this could be made clearer, and will include clarification of this in the introduction. We used the term 'benchmark' to highlight that these results can be used as an indicator of the ability of lumped models, which future studies may use when evaluating the performance of other models (that are perhaps more complex or include additional processes). For example, our results would inform a modeller that gaining an NSE of 0.7 in SE England is a good achievement, whereas gaining the same score in west Wales is not an achievement as most models can easily gain higher NSE scores for these catchments. The use of simple models as benchmarks has been advocated in previous studies, for example Seibert et al., (2018).**
**Seibert, J., Vis, M. J., Lewis, E., & Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. Hydrological Processes.**

Model evaluation using NSE: Since the authors aim to better understand "where and why these simple models may fail" the choice of NSE is somewhat suprising, since NSE is a measure of overall performance, which provides limited insights into the reasons for high or low performance. Although an evaluation based on hydrological signatures would have enabled a more process-based diagnostic of model failures, I am not requiring this, since it would imply significant additional analyses. However, if the authors stick to NSE (or use KGE), I suggest that they use benchmarks (as suggested by Seibert et al., 2018) to account for the fact that high NSE/KGE values can be relatively easy to reach depending on the catchment and the season. I believe this would enable a more fair and enlightening assessment of the hydrological models across the catchments.
**Response: We originally selected NSE as it is a widely used and easy to interpret measure of performance. However, we agree that in order to better understand model failures we will need to consider additional measures of performance. Therefore, we plan to also present correlation, variance and mean bias, as called for by the first reviewer, to support the seasonal analysis of model performance**



**that we have already carried out. As our focus is on reasons for model failures, we feel that these additional decomposed metrics will be more informative than the use of benchmarks.**

Relevance for the broad hydrological modelling community: A challenge here is to provide guidance for model selection, which is also relevant for modellers not using FUSE. Overall, the most interesting question is not really which FUSE model performs best, but why. I encourage the authors to discuss and highlight specific model elements that contribute to poor/good simulations, rather than focussing FUSE models themselves (e.g., TOPMODEL or PRMS). For instance, the fact that ARNO-VIC performs particularly well in high-BFI catchments is only an intermediary result, which is mostly relevant to FUSE users. The reasons why this is the case (e.g., last paragraph of Section 5.2), on the other hand, are relevant to a much wider group. I suggest a stronger emphasis on modelling decisions, as opposed to FUSE models, in particular in the most critical parts of the manuscripts (abstract and conclusions).
**Response: We agree that highlighting specific model elements that contribute to poor/good simulations would be of great use to the broad hydrological modelling community. To address this, we will explain in more detail how the modelling decisions differ between the four structures in the methods section. We will also clarify these different modelling decisions when discussing model performance in the abstract and conclusion However, as we have looked at the four parent models within the FUSE framework, and not all possible combinations of modelling decisions which would be very challenging over such a large sample of catchments, there is a limit to how far we can analyse which particular model elements contribute to good/poor results. Many elements differ between the model structures, and it is not always possible to decipher which individual element is responsible.**

Which process parameterisation are missing to capture the range of hydrological behaviours across the UK? The authors identify catchments in which the four model structures perform poorly, and reflect on characteristics of these catchments to which the poor performance can be attributed (e.g., chalk, snow, high human impacts). I suggest that the authors dedicate a subsection in the Discussion Section to these findings, which are relevant for both model development and selection. Can they formulate hypotheses on why annual maximum flows are underestimated, which could be tested by future studies?

**Response: Findings on which process parameterisations are missing to capture the range of hydrological behaviours are discussed in section 5.1. However, we agree that these should be made easier to find, and therefore we will split discussion section 5.1 into "Benchmarking performance of multiple lumped hydrological models across GB" and "Identifying missing process parameterisations for modelling in Great Britain". As suggested by the first reviewer, the choice of NSE could result in underestimation of flood peaks, and we will therefore comment in the discussion on how our choice of metrics could be a factor leading to the underestimation of flood values.**

How critical is the selection of model structure? There are cases of great equifinality (i.e., high NSE for all structures, mostly for humid catchments). As mentioned above, a high NSE is not a guarantee that the model structure is adapted, but as long as this is recognised (and this could be clearer throughout the manuscript), I think it is fine for this study. But in other (more interesting) catchments, some model structures clearly outperform other structures, and there, model choice is critical. I think this should be stressed more prominently, since these are cases in which the inadequacy of the model structure cannot be overcome by parameter tuning. Given the general tendency of using the same model structure across very diverse environments (as discussed e.g. by Addor and Melsen, 2019), I think this is an important result, which could be underscored more. A related question is: which catchment characteristics explain



these large NSE differences between model structures?

**Response: We explored the importance of model structure selection in figures 3, 5, 8 and 9, with figure 5 looking at catchment characteristics which were related to differences between the model structures. However, we agree that the question of how critical the selection of model structure is for different catchments is not well addressed in the manuscript. Therefore, we will add a short paragraph discussing this into section "5.2 Insights from applying an ensemble of model structures across a large sample of catchments" and make this key finding clearer in the abstract and conclusions of the paper.**

This leads me to a set of comments related to the use of catchment attributes to explain model performance. Just like hydrological behaviour, model performance is not determined by a single catchment characteristic, but rather, by the interaction of multiple catchment characteristics. So, firstly, would it be possible to consider a wider range of catchment attributes? So far, the authors employ the BFI, annual rainfall, the wetness index and the runoff coefficient, but many more attributes could be used to describe each catchment (e.g., Beck et al., 2015). I encourage the authors to add other attributes, which they might have computed for other studies or retrieved from the UK hydrometric register, which they mention in Table 1, in order to describe the landscape in a more complete fashion (indicators of human interventions would also be useful, see below).

And secondly, I think it would be beneficial to better account for the interactions between these attributes. The authors combine several attributes in Figure 7 to explain model performance, which I find particularly interesting. Maybe that the analyses they will perform when revising this study will lead to more figures of this type, and enable a more systematic analysis of the interactions between these predictors (perhaps using regression trees, see Poncelet et al., 2017). This is critical to go from describing where models fail and to explaining why they fail.

**Response: We selected the attributes of BFI, annual rainfall, wetness index**

**and runoff coefficient as they were observed to have the largest impact on model performance. However, it is possible to include additional attributes, to further explain model performance. We do not want to add more figures into the manuscript as we feel that this may detract from the main messages of the paper. However, we will produce results for a wider variety of attributes and include them as supplementary information. These will be in the form of scatter plots, showing interactions between different characteristics as suggested, in the same style as figure 7.**

Anthropogenic activities are repeatedly mentioned to explain poor model performance (e.g., P12L29, P14L16, P15L3). This is indeed plausible, but if qualitative or maybe quantitative indicators of the extent of human interventions could be included, so that their impacts on streamflow and model performance could be demonstrated or maybe even quantified, it would strengthen the study.
**Response: We agree that this is required to strengthen comments made regarding reasons for model failures. We have information on factors affecting runoff for all catchments in the hydrometric register. However, this only gives an indicator of which factors may affect runoff, and not to what extent, and therefore we decided not to include it in the original manuscript. In response to reviewer comments, we will investigate either a) providing maps of this information to assist in the interpretation of the results or b) Incorporate a summary of this information into figure 7 through changing the marker shape.**

Minor comments

I find the introduction too long. It attempts to cover too much material, and hence ends up being too general and its different parts are not very well connected. I suggest that



the authors focus on what is really necessary to introduce their study, transfer parts of the text to the rest of the paper (e.g. the methods), and delete the rest.
**Response: We agree that the introduction could be more concise and will shorten it in the revised manuscript of the paper.**

Outlook: it might good to mention that, although this study focusses on four FUSE models, it is possible build additional FUSE model to transition progressively from one model to the next, and establish which modelling decisions contribute most to the differences in the simulations.
**Response: this will be added.**

Data availability: “This study provides a useful benchmark of the performance and associated uncertainties of four commonly used lumped model structures across GB, for future model developments and model types to be compared against”. I agree. But then, I think that instead of saying that “All model outputs from this study are available upon request from the lead author”, the authors should make the runs available online, and provide the doi, before the paper is published. This is expected by AGU journals, and I think it is good practice in order to avoid data loss.
**Response: We completely agree with this, and are currently obtaining a DOI for the data. As also recommended by reviewer 3, we will add a table summarising the results for each catchment to be made available as supplementary information.**

Other suggested changes

Title: the field is “large-sample hydrology”, but here it should be “large sample”
**Response: Thank you, this has been changed.**

P1L15: add “and support model selection”
**Response: This has been added.**

P2L13: such as **Response: Thank you for spotting this, it has been changed.**

P2L29: impacted by what?
**Response: We have clarified the sentence by adding “impacted by future changes to the hydrological regime.”**

P4L12-17: this belongs to Data and Methods
**Response: These sentences have been moved to the data and methods section.**

P4L20: I suggest removing “(i.e. the number of storage components)” as it an arbitrary measure of complexity.
**Response: This has been removed.**

P5L22: discharge
**Response: We have clarified that it is a discharge data set.**

P6L5: please define “sufficient”
**Response: This was explained in the following sentence. We have re-arranged these sentences to make this clearer, now saying “Of these, 1128 had sufficient information (more than 2 years of discharge data available within the time period) available to include in this analysis.”**

P7L2: I suggest mentioning here that none of these four models includes a snow



Routine
**Response: We have added this, “They all close the water balance and include the same processes, for example none of the models have a snow routine.”**

P7L4: please define “dynamically different” and what makes them “equally plausible”
**Response: By “dynamically different” we meant that the models all represent the landscape in a different way, and have quite different and distinct structures as shown in figure 2. By “equally plausible” we are referring to the fact that we have no reason to expect one structure to behave better than the others, as all model structures are equally complete in terms of processes and all based on widely applied model structures. We will clarify this in the text.**

P8L13: please be more explicit about how this 13
**Response: We will include a better description of how this +/- 13% was calculated (see response to reviewer 1).**

P9L21: saying “snowmelt module” implies that accumulation is simulated but melt is not, use “snow module” instead.
**Response: We agree, and have changed “snowmelt module” to “snow module.”**