Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-635-RC2, 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.

Anonymous Referee #2

Received and published: 5 March 2019

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.

Major 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.

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.

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).

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?

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?

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.

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.

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.

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.

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.

Other suggested changes

Title: the field is "large-sample hydrology", but here it should be "large sample"
P1L15: add "and support model selection"
P2L13: such as
P2L29: impacted by what?
P4L12-17: this belongs to Data and Methods
P4L20: I suggest removing "(i.e. the number of storage components)" as it an arbitrary measure of complexity.
P5L22: discharge
P6L5: please define "sufficient"
P7L2: I suggest mentioning here that none of these four models includes a snow routine
P7L4: please define "dynamically different" and what makes them "equally plausible"
P8L13: please be more explicit about how this 13P9L21: saying "snowmelt module" implies that accumulation is simulated but melt is not, use "snow module" instead

P10L29: SACRAMENTO

References

Addor, N. and Melsen, L. A.: Legacy, rather than adequacy, drives the selection of hydrological models, Water Resour. Res., 55, doi:10.1029/2018WR022958, 2019.

Beck, H. E., de Roo, A. and van Dijk, A. I. J. M.: Global maps of streamflow characteristics based on observations from several thousand catchments, J. Hydrometeorol., 16, 1478–1501, doi:10.1175/JHM-D-14-0155.1, 2015.

Poncelet, C., Merz, R., Merz, B., Parajka, J., Oudin, L., Andréassian, V. and Perrin, C.: Process-based interpretation of conceptual hydrological model performance using a multinational catchment set, Water Resour. Res., 53, 2742–2759, doi:10.1002/2016WR019991, 2017.

Seibert, J., Vis, M., Lewis, E. and van Meerveld, I.: Upper and lower benchmarks in hydrological modeling, Hydrol. Process., 32, 1120–1125, doi:10.1002/hyp.11476, 2018.

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