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 #3

Received and published: 18 March 2019

Summary:

This paper provides a detailed investigation into the performance of four lumped conceptual models over large number of catchments in the UK. It demonstrates some very interesting findings, such as the fact that all four models have very similar performance on a catchment-by-catchment basis, and that only one of the models is deemed suitable for catchments with very high BFI. This paper is generally well written, set out and easy to follow, and the graphics provided assist the reader well in the interpretation of the results, I particularly like Figures 5 and 7. The discussion section should be syn-

thesised as it feels repetitive of the results section. Overall, I feel that the motivations of the research, and the implications of the results are not very well reasoned. The authors need to think a bit more carefully about how others may make use of these results, and in particular, should publish the model performance scores as supplementary information (see my comments below).

Major Comments:

1. You've "benchmarked" performance, but you haven't provided these benchmarks. If I were to now go and simulate a UK catchment, I still cannot easily compare my results with yours to see if I have a better model. For you to have achieved your aims, I would expect a supplementary table of the best scores the models achieved in each catchment, and the parameter values that produced them.

2. Section 3.2 – why NSE?

3. Section 3.2 – "results are stored for a number of additional metrics not reported here". Stored where? Why would I care about this if you haven't made them available to me? I suggest you summarise these additional metrics in supplementary information. This may also address the issue of only reporting on NSE here.

4. Your statement in the abstract L23 that NSE scores of 0.72-0.78 were achieved for all catchments is misleading. How useful a measure is the "median maximum NSE for all the catchments"? It's pretty cryptic. There are catchments in E Scotland, and Anglian region that are showing pink/red for all 4 models, so NSE must be <0.5. Having got to page 9 I now see what you meant, but it isn't clearly stated. The sentences on P12 L16-17 are a better summary of the performances across catchments. Same issue on P16 L 32.

5. Catchment characteristics and climate – do all FUSE models maintain the water balance? Can you comment on the existence of models that don't (e.g. GR4J), and how those may overcome such problems? What are the implications of maintaining

vs not maintaining water balance in conceptual lumped models? Are the four models you've chosen actually quite similar to each other? I think you need to make more of this somehow.

6. P8 L23 – only a 1 year warm up period? This is not sufficient for many GW dominated catchments in the SE.

7. P6 L6 – 2 years of data was your criteria for catchment selection, this doesn't seem sufficient to me. . .

8. Reading through your discussion seems very repetitive of the results chapter. Can these be better synthesised, to reduce the discussion section?

9. P2 L32 "a national scale model" – you're talking about applying a catchment model nationally. Can this be classified a national scale model?

10. P3 L16 - "Secondly, evaluating more complex hydrological models relative to benchmark performance of simple models ensures that the relative difficulty of simulating different catchments is implicitly considered (Seibert et al., 2018)." I don't think I understand what you're saying here.

11. P10 L22-24 – "For very low values of the ARNO-VIC 'b' exponent (AXV_BEXP) as seen for high BFI vales in Fig. 6 for behavioural model distributions means that only at very high, near full upper storage levels is any larger extent of saturated areas predicted" – I don't follow this sentence either.

12. P8 L3 - Can you explain conditional probabilities in more detail?

13. P11 L 23 – "the top row of plots" – there is only one plot in Fig 8!

14. P12 L 7-8 "However, variations between years are less apparent when looking at 25th and 75th percentiles in Fig. 8." We can't distinguish variation between years from Fig 8?

15. Please provide more sensible y axis labels for fig 8 and 9, e.g. "AMAX discharge

score", and "AMAX percentage overlap" respectively. Multiply Fig 9 y axis by 100 to make it an actual percentage value, as you have referred to it as such in the text.

16. Your discussion is longer than the rest of the paper put together!

17. P13 L 3 – you've made no reference to anthropogenic influences in Scotland. This statements seems a bit throwaway.

18. P13 L9 – it is not just the Thames basin that is affected by abstractions! A lot of Anglian region is VERY heavily influenced.

19. P13 L12 "we found that the ensemble of model structures produced better results overall than any single model" – can you validate that statement from your figures?

20. P13 L15 – "The ensemble of model structures was able to take advantage of this" - this seems to be a contradictory argument to the previous statement that the models all have similar performance to each other on a catchment by catchment basis. I think you need to tease these two arguments out better somehow. E.g. in some situations the choice of a different model can yield better results (e.g. high baseflow), but in other situations, none of the models can do well (e.g. abstractions). What are the implications of this?

21. P17 L 11-14 "We also evaluated model predictive capability for high flows, as good model performance in replicating the hydrograph, assessed using Nash-Sutcliffe efficiency, does not necessarily mean models are performing well for other hydrological signatures. We found that the FUSE models tended to underestimate peak flows, and there were variations in model ability between years with models performing particularly poorly for extremely wet years." – so what? What are the potential implications?

Typos and grammar:

1. P2 L27 – CAMELS and MOPEX datasets (what are they datasets of?)

2. P5 L14 - "these" should be "those"

3. P5 L22 - remove "Environment Agency", a catchment is a catchment, the EA don't own the catchments, even if they do own the gauges!

4. Amend "Rainfall is highest in the West and North of GB and lowest in the East and South varying from a minimum of 500mm to a maximum of 4496mm per year (see Fig. 1)" to "On average, rainfall is highest in the north and west of GB, and lowest in the south and east, with GB totals varying from a minimum of 500mm to a maximum of 4496mm per year (see Fig. 1).

5. P6 L1 - remove the "of" after "South-East"

6. P6 L12 – they are the "UK Met Office" not the "UK Meteorological Office".

7. P6 L14 and L20 – replace "laid" with "lay"

8. L6 L21 and elsewhere – "data" is plural, and should be followed by "were" instead of "was"

9. P8 L15 – "observational uncertainty certainty bounds" huh?? Can you not just remove the word certainty here?

10. P9 L15 – you haven't introduced the abbreviation "SAC"

11. P10 L5 – I'd call that northeast Scotland, not central Scotland

12. P11 L29 – "behavioural model" should be "behavioural models"

13. P13 L7 – do you mean model "structures"?

14. P16 L17 – "we also shown how"

15. P16 L19 – refer to Fig 6

16. P16/17 – "The performance of the four models was similar, and all models showed similar spatial patterns of performance, and there was no single model that outperformed the others across all catchment characteristics and for both daily flows and peak flows." – and, and, and.

17. P17 L8 – "we found models performed poorly for catchments for catchments with unaccounted losses"

HESS REVIEW CHECKLIST

1. Does the paper address relevant scientific questions within the scope of HESS? Yes

2. Does the paper present novel concepts, ideas, tools, or data? Yes

3. Are substantial conclusions reached? Nearly, the wider implications, and utility of the research need to be better considered

4. Are the scientific methods and assumptions valid and clearly outlined? Yes

5. Are the results sufficient to support the interpretations and conclusions? Yes

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? Yes

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution? Yes

8. Does the title clearly reflect the contents of the paper? Yes

9. Does the abstract provide a concise and complete summary? Yes

10. Is the overall presentation well-structured and clear? Yes

11. Is the language fluent and precise? Yes

12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? Yes

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? Yes, the discussion should be reduced

14. Are the number and quality of references appropriate? Yes

Interactive comment

Printer-friendly version

Discussion paper

15. Is the amount and quality of supplementary material appropriate? No