Reply to Wouter Knoben’s comments

We thank Wouter Knoben for his comments on our manuscript, which will help to improve its overall quality. Below we give our replies (in blue) to his comments (in black) on how we modified the paper to account for his suggestions and recommendations.

Thank you for your considering the review comments carefully. I have read your response and am mostly satisfied with this.

Thank you for this feedback.

I do have some further broad comments listed below, and some detailed ones as annotations in the uploaded pdf.

1. I believe the manuscript would benefit from some further polishing. This partly involves further clarifications (see annotations; mostly suggestions about transferring specific replies made in the response to reviewers into the main manuscript) and partly a thorough check of spelling and grammar.

We clarified any necessary points based on your comments.

2. I believe the manuscript needs a brief discussion section that outlines the possible impacts of some of the main limitations of the study. While limiting the scope of the work is justifiable, I believe that explicitly listing how this may impact the findings of the work is needed.

We agree that some limitations may be highlighted. Therefore, based on your comments, we have distilled here and there points of perspective or discussion about this work, either in the methodology section or by adding an appendix on the uncertainty linked to the KGE score.

3. The mandatory Data availability section needs to be added. I would also strongly suggest adding a Code availability section, to enhance reproducibility of the work.

We added a data availability and a code availability sections to this end.

More generally, I think the manuscript can be clarified here and there, and possibly streamlined a bit (though I am unsure what to scrap - it just seems to have a large number of figures).

We clarified our statements and improved the figures.

L16: with limited **anthropogenic** influence

L20: a simple averag**ing** method

L21: The most efficient lumped model → **The lumped model with the highest efficiency score**

L50: data are used to **derive** model structures

L56: consists **of**

Terminology and syntax were checked and corrected.

L113: Surely the uncertainties must be quantified somehow to decide if the multi-model approach beats the single model. Should this sentence therefore read "we decided here not to focus on quantifying these uncertainties _individually_ [but do look at the aggregated impact of all uncertainties through comparing the accuracy of the model ensemble vs the single model]"?

We propose here to rephrase this sentence: “However, we decided here not to focus on quantifying these uncertainties individually (as it could be done with a probabilistic ensemble) but on looking at the aggregated impact of all uncertainties by comparing the deterministic averaging combination of several models with single models.”.

L153: Just noting that also here the choice of "efficiency" seems a little out of place to me. "Performance" sounds better to me

We agree. We changed this in the manuscript.

L155: Single-sentence paragraphs look weird. Perhaps consider merging this with the previous paragraph.

We agree. We merged these paragraphs.

L185-196: I believe this section needs a bit more attention, in particular to address my longer question about how the semi-distributed routing is implemented. Copying my earlier comment, with some additional clarifications:

Judging from Figure 2, there are 2 distinct cases:

- Case 1: a catchment has exactly 1 upstream sub-catchment somewhere along the main river stem [e.g. the bottom right plot in Fig. 2]. Question: is flow from the upper catchment routed through the lower catchment somehow, or are these flows assumed to magically appear at the outlet of the lower catchment and there combined with simulations from the lower catchment, or does upstream flow somehow feed into the hydrological model that simulates the downstream catchment as an extra (lateral) input?

- Case 2: a catchment has multiple upstream sub-catchments on different branches of the river [e.g. the top middle plot]. In these cases a large part of the [grey] catchment (the [most downstrteam] one) will have travel times for its runoff comparable to travel times in the [dark] gray subcatchments. How is routing handled in these cases?

These implementation details matter because they correspond to the level of realism the models attempt to provide, as well as being important for reproducibility of this work.

We clarified this point.
Both cases are handled in the same way. Here, the semi-distributed approach consists in performing lumped modelling on each sub-catchment by linking them through a hydraulic routing scheme. Thus, the two cases to be distinguished are:

1) the sub-catchment has no upstream inflows (i.e. is an upstream sub-catchment, Figure 2, dark grey)
2) the sub-catchment has upstream inflows (i.e. is a downstream sub-catchment, Figure 2, light grey)

In the first case, we only apply a lumped rainfall-runoff model, and in the second case the rainfall-runoff model is applied after integrating the upstream inflows using a runoff-runoff model (hydraulic routing scheme).

It is therefore important to differentiate between the routing part of hydrological models (enabling to distribute in time the quantity of water contributing to the streamflow in the sub-catchment studied: intra-sub-catchment propagation) and the hydraulic routing scheme (enabling to propagate the streamflow simulated at one outlet to downstream catchment: inter-sub-basin propagation).

L196: The authors mention in their reply that they also tested more complex routing schemes than the chosen one. Perhaps these tests can go into Supporting Information to provide more support for the chosen approach. A sentence like "This approach is fairly simple but offers comparative performance to that of more complex routing models that account for x,y,z (results not shown for brevity)"

We added a sentence to that effect at the end of the section.

L207: adapted **to** the hourly time step

L209-210: Since the various catchments used for this study **do not experience much snowfall**, no snow module was implemented.

L220: First, a **coarse** screening

Terminology and syntax were checked and corrected.

L222: Some information of stopping criteria of the calibration algorithm is still missing. This is needed to allow reproducibility of the work, and provides the reader with a (marginaly) better understanding of how close to optimal they can expect the single calibrated parameter set to be.

The algorithm used for model calibration is available in the airGR package which is freely available (see Calibration_Michel function for more details[1]). We clarified this in the manuscript. To answer briefly, the stopping criterion depends on the number of iterations (up to 100 times the number of parameters) and parameter variation (if the parameter variation in the transformed domain becomes less than 0.01).

L223: The choice to stick with a single parameter set is understandable but this is a limitation that needs to be explicitly stated and discussed. The authors' reply states:

"We did not focus explicitly here on parameter uncertainty, i.e. we did not use multiple parameter sets based on Monte Carlo simulations for example. We agree that this would be interesting to consider, but we thought this would make the article too complex. We will better explain our choice in the article."

I have not been able to find this explanation either here or in the discussion, and believe this needs to be added.

We apologize for this oversight. We corrected it.

[1] https://archive.softwareheritage.org/swh:1:cnt:15a89bf43d074a8b95fbf909b7f92efbe02795fa;origin=https://cran.r-project.org/package%253DairGR;visit=swh:1:snp:b14ede66c1ce5d036e4068297411cc78f06c6771;anchor=swh:1:rel:83fc526950e22151a11cd502d747dc455984a513;path=/airGR/R/Calibration_Michel.R

L266: This practice consists **of** separating

Syntax was checked.

L287-288: The reader is left hanging here. Adding a single sentence to explain why this approach was not developed further would help.

We agree. We changed our approach by choosing the one-size-fits-all model as a benchmark, which will be easier to understand.

L310: I would suggest to include the figure that shows the full uncertainty distribution shown in the response to reviewers document in this paper as well (possibly in the SI).

We added an Appendix C to deal with the uncertainty in KGE scores. We included this figure in this section.

L326: The benchmark **has** a median

Syntax was checked and corrected.

L326: This seems small, especially considering the sampling uncertainty has a similar order of magnitude (line 310). Does this suggest that going through the extra effort of calibrating multiple models for different basins may not lead to substantial changes in model accuracy?

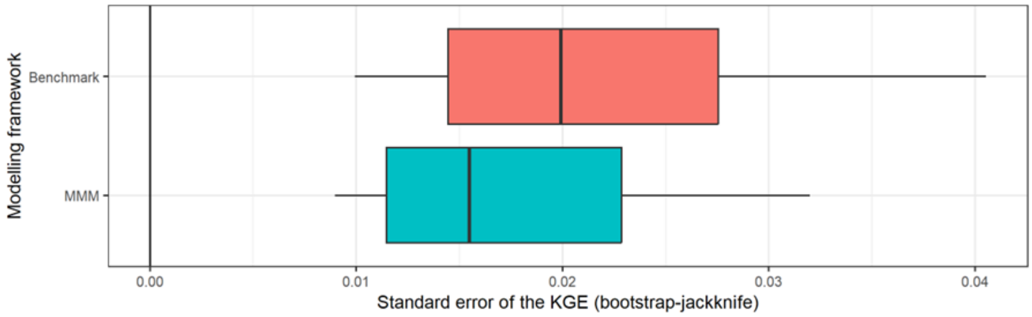
Here we compare the best combination of models with the best single-model approach (which could already be considered as a multi-model approach, since all models have been tested and a different model can be selected for each catchment). When compared with the one-size-fits-all model approach, the gains are greater, exceeding the sampling uncertainty that has been established. Therefore, it is always worth putting extra effort into calibrating multiple models for different catchments.

Figure 5: I believe it would be more helpful to include the full boxplot of the benchmark model here. This provides relevant context needed to compare the change in model performance to the estimated KGE sampling uncertainty. This would also apply (and fit better) with Figures 7-9.

We agree. We changed it.

L374: It's my understanding this sampling uncertainty value of 0.02 is specifically the sampling uncertainty associated with the benchmark (one model everywhere) approach. To full make the claim shown here the sampling uncertainty of this LMM approach needs to be calculated too, so that a comparison can be made between benchmark+uncertainty on the one hand, and LMM+uncertainty on the other.

Indeed, the sampling uncertainty of 0.02 is linked to the one-size-fits-all model (which will now be considered as our benchmark). The following results show the comparison of sampling uncertainty between our benchmark (one-size-fits-all model) and the mixed multi-model approach. The mixed multi-model approach therefore also reduces uncertainty about the KGE value.



We included this result in the Appendix C.

Figure 12: Also here adding this boxplot would be helpful

We agree. We changed it.

L461: Single-sentence paragraphs can be merged with the preceding one.

We agree, we merged these paragraphs.

L556: Is there a word missing after "streamflow"?

We corrected it as: “possible streamflow **simulation** on a catchment”.

L578: I'm missing a brief discussion of how any study limitations may impact these findings. Two particular ones that come to mind are:
1. Using only a single parameter set per model per catchment

We discussed this issue as in section 2.3 by mentioning Monte-Carlo simulations, which could explicitly account for parametric uncertainty, as a perspective.

2. Any possible inherent bias in the selection of the models used. These models have shown reasonable behaviour in the French context before, but perhaps there are other options out there that could provide extra variability in the ensemble.

This work is based on a large sample of models to reduce the bias. However, we agree that some limitations may be linked to the selection of the models used. Therefore, we added some sentences in the conclusion section to clarify that the inclusion of machine learning or physically-based models may add value to the multi-model approach.

L600: It would be appropriate to cite the Kratzert et al (2018) paper in HESS here, as this is the first application (I know of, at least) of LSTMs in hydrology.

We agree.

L652: Noting that Copernicus requires a "Data availability" statement and strongly recommends a "Code availability" statement (see: https://publications.copernicus.org/services/data_policy.html).

This section from the authors' reply should go in the second:
“Summary sheets containing the structural scheme of the different models, the pseudo-code and the table of free parameters are however available on request.”

We added a data availability and a code availability section.