Reply to Anonymous Referee's comments (RC1)

We thank the anonymous referee (RC1) for their constructive comments and careful evaluation of our manuscript. Their feedback has been very helpful in improving the clarity and robustness of the paper. Below, we provide a detailed, point-by-point response to all comments. The referee's remarks are shown in black, and our responses and corresponding revisions are provided in blue.

This paper demonstrates and compares approaches to combine multiple hydrological models (i.e. multi-model mosaics vs multi-model combinations), answering the question of which multi-model approach performs best over a large sample of catchments in the US. First, I'd like to say that I really enjoyed reading this paper. It covers an important topic – how to improve streamflow simulations through multi-model approaches – in a novel way. I also appreciated the discussion of sampling uncertainty, which is often overlooked in modelling studies. The figures were excellent, well presented and very clear, and the paper was well-written. I would recommend that this manuscript is worthy of publication with minor corrections and clarification of the methods. Further comments and suggestions are outlined below.

We thank the referee for this positive evaluation of our work. We appreciate the constructive feedback and address all specific comments below.

Further justification is needed for the use of model structure with best median KGE as a benchmark. This is a difficult benchmark to beat – it already requires a multi-model approach running all 78 combinations of model structures and selecting the ones with the highest overall performance. The selected benchmark model is dependent on your catchment selection, and I wonder if this gives an unfair advantage to catchments which are the least similar to other CAMELS-US catchments (i.e. where the benchmark model structure is less suitable and therefore easier to beat). I am curious why you did not use the FUSE variants based on the four existing models (i.e. relating to VIC, PRMS, SAC and TOPMODEL) as benchmarks, as these may be a better representation of what we might expect from a single model approach which your multi-model approaches then build on.

We thank the referee for this thoughtful comment. We agree that the selected benchmark represents a strong baseline, as it is identified from a large ensemble of 78 model structures. Our intention was precisely to define a robust upper bound for a single-structure solution thereby providing a stringent reference against which the added value of more complex multi-model approaches can be assessed. In other words, we aimed to evaluate whether multi-model strategies offer clear improvements over the top-performing single model available within the ensemble, rather than over a legacy or arbitrarily chosen "one-size-fits-all" model. This also avoids implicitly promoting more complex multi-model approaches (with their associated computational cost and uncertainty) when a carefully selected single model may perform equally well in terms of performance metric values.
We acknowledge that the selected benchmark depends on the catchment sample considered. However, this dependence applies equally to all multi-model approaches tested here, as they are all calibrated and selected on the same set of catchments. Importantly, the benchmark does not systematically outperform other structures in every catchment and thus does not necessarily confer an advantage in regions where it is less suitable.
Regarding the use of the four parent FUSE configurations (VIC, PRMS, SAC, and TOPMODEL) as benchmarks, these models are embedded within the broader FUSE structural space but represent only a small subset of possible configurations. To improve transparency, we now explicitly highlight these four parent configurations in Figure 3, allowing readers to directly assess their relative performance within the full ensemble. Our objective was not to compare multi-model methods against legacy model formulations, but rather to assess whether more complex multi-model strategies provide added value

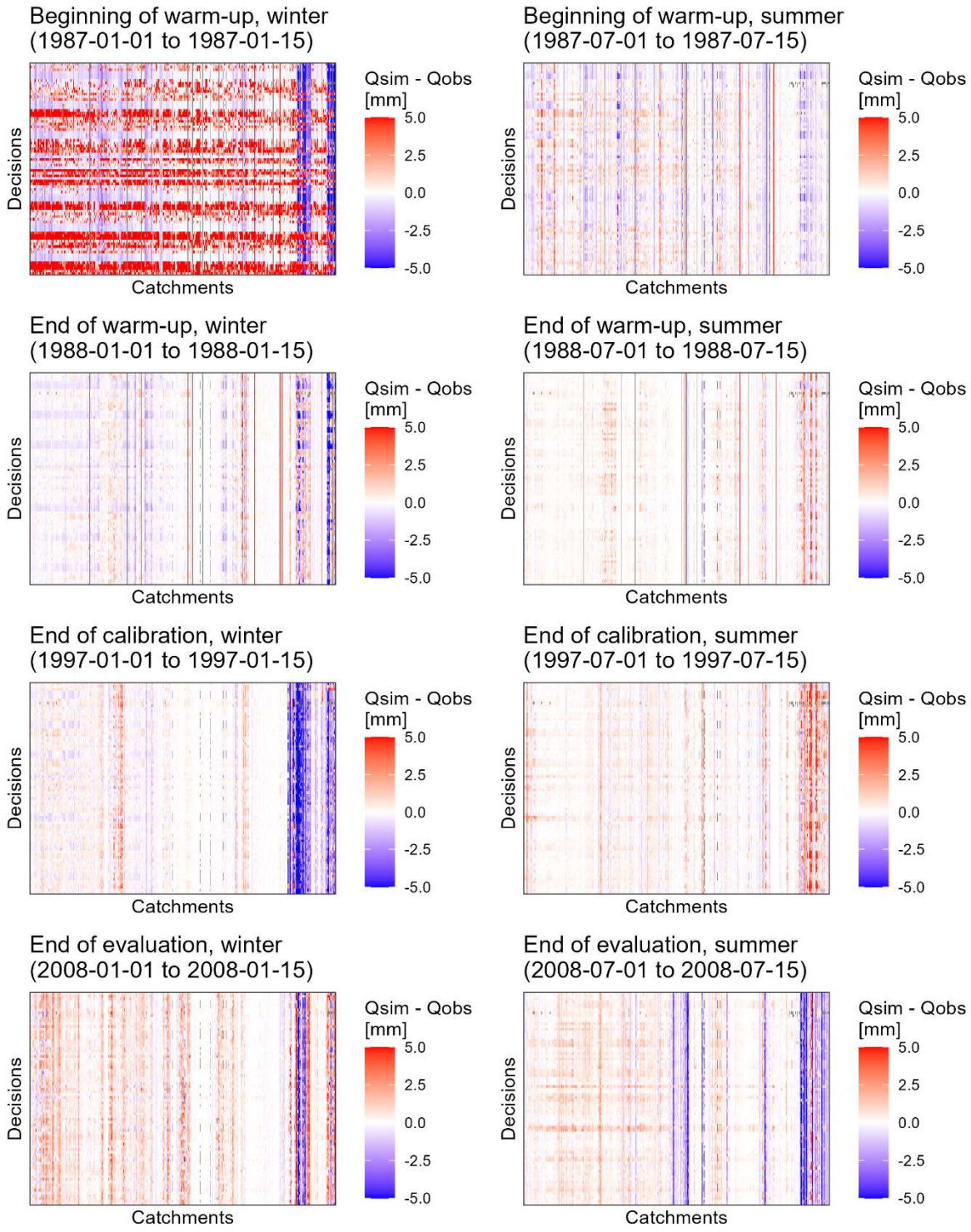
over the simplest one, using the best-performing single structure available within the considered ensemble. We have clarified this in the revised manuscript.

The FUSE model variants share many similarities: all lumped conceptual hydrological models run at a daily timestep with the same input data. It would be worth discussing their similarities as well as the key differences given in Table 2 – as a more diverse multi-model ensemble may have even greater benefits. This is briefly touched upon in the discussion, but I feel that it is also worth elaborating upon in the methods.

We thank the referee for this important remark. We agree that the 78 FUSE variants share a common modelling framework, as they are lumped conceptual models run at a daily timestep with identical forcing data. While this ensures methodological consistency and a controlled comparison of structural decisions, it also limits the diversity of the ensemble compared to combinations of fundamentally different modelling paradigms (e.g., with physically-based or machine learning models).
Within this common framework, structural diversity arises from differences in upper- and lower-layer architectures, baseflow parameterizations, surface runoff generation mechanisms, and percolation formulations (Table 2). These decisions lead to substantial differences in performance across catchments (Figure 3), yet they remain variations within a conceptual modelling family.
We agree that a more heterogeneous ensemble, including structurally distinct modelling paradigms, could potentially yield larger benefits from multi-model approaches. We have expanded the description of the FUSE framework in Section 2.2 to better clarify both the shared assumptions and the sources of structural diversity, and we now explicitly acknowledge this limitation earlier in the manuscript.

Section 2.3: “Each model is calibrated for each catchment over the period 1989-1998 with a preliminary warm-up period of two years.” Please could you further specify these dates – were they run over water years or calendar years, and was the warm-up period before the calibration i.e. 1987-1988 inclusive or the first two years of the calibration period 1989-1990? Is two years sufficient for a warm-up period for your catchments? We have found that some groundwater dominated catchments require longer warm-up times depending on the model initialisation, but I have no experience of modelling catchments in the USA. This choice of calibration period should be explained in the paper – 10 years is relatively short and may not capture particularly dry/wet years.

The calibration period covers calendar years 1989–1998, and the two-year warm-up period precedes calibration (calendar years 1987–1988). This has now been clarified in Section 2.3.
We agree that a two-year warm-up period may not be sufficient in all contexts, as the time required for model spin-up depends on model structure, climate, and catchment characteristics. To assess whether the chosen spin-up was adequate in our case, we analysed the temporal evolution of streamflow error (simulated vs observed) for different dates across the 78 structures and 559 catchments. This analysis indicates that model errors generally decrease rapidly and stabilise within the first few months of simulation for most structures and catchments. While we acknowledge that some groundwater-dominated catchments may require, in theory, longer spin-up periods depending on model formulation and initialisation, the results suggest that the two-year warm-up period is sufficient for the purposes of this large-sample comparative study. This behaviour is consistent with the lumped and conceptual nature of the FUSE configurations, which typically exhibit shorter memory than physically based models with explicit aquifer representations.



RC1 — Figure 1 : Mean streamflow error (Qsim − Qobs) computed over a 15-day window for selected dates, across the 78 FUSE structures (y-axis) and the 544 catchments (x-axis). Rows typically correspond to different years, while columns distinguish between winter and summer seasons. Decisions are ordered numerically, and catchments are arranged according to their identifiers.

Regarding the overall simulation period, we acknowledge that longer time windows may capture a broader range of hydro-climatic variability. Nevertheless, the selected period includes both relatively wet and dry years across CONUS for calibration and evaluation, providing a balanced compromise between representativeness and computational feasibility given the number of catchments and the size of the model ensemble. An additional constraint arises from data availability: the Daymet forcing

product provided within CAMELS spans 1980–2015, limiting the maximum possible simulation length to 35 years. In practice, however, near-complete streamflow observations are rarely available over such an extended period. The 22-year window used here (2 years warm-up, 10 years calibration, 10 years evaluation) was therefore selected to maximize the number of catchments with sufficiently complete streamflow time series, while maintaining consistency with previous large-sample CAMELS-based studies (e.g., Knoben et al., 2020). These aspects are now clarified in the revised manuscript.

Section 2.5 would benefit from a more thorough description of the multi-model approaches. In particular, I noted the following:

(1) Section 2.5.2.2 left me with questions such as what are the benefits of minimising the number of models, how exactly does the method reduce the number of models required, and how many model structures remained? On further reading I found that more details are given in appendix A – it would be helpful to refer to this in the main text.

We agree that the description of the performance-equivalence mosaic in Section 2.5.2.2 was too concise and required additional clarification.
The objective of minimizing the number of models is to identify the smallest subset of structures that can reproduce performance within the sampling-uncertainty bounds across catchments. This approach reduces the "noise" that may arise in a traditional performance-based mosaic, where a different best-performing model is selected independently for each catchment, which can give the impression that a large number of distinct structures are necessary, whereas many of them may be statistically equivalent. By identifying a minimal set of equivalent models, the method improves parsimony and interpretability at the domain scale.
We have expanded the description in Section 2.5.2.2 to clarify how the linear programming algorithm identifies the minimum set of equivalent models (i.e., by solving a set-cover type problem). We now explicitly report in the main text that only eight model structures are required to cover all catchments under the performance-equivalence mosaic. We also added an explicit reference to Appendix A, where further methodological details and supporting figures (Figures A3–A4) are provided.

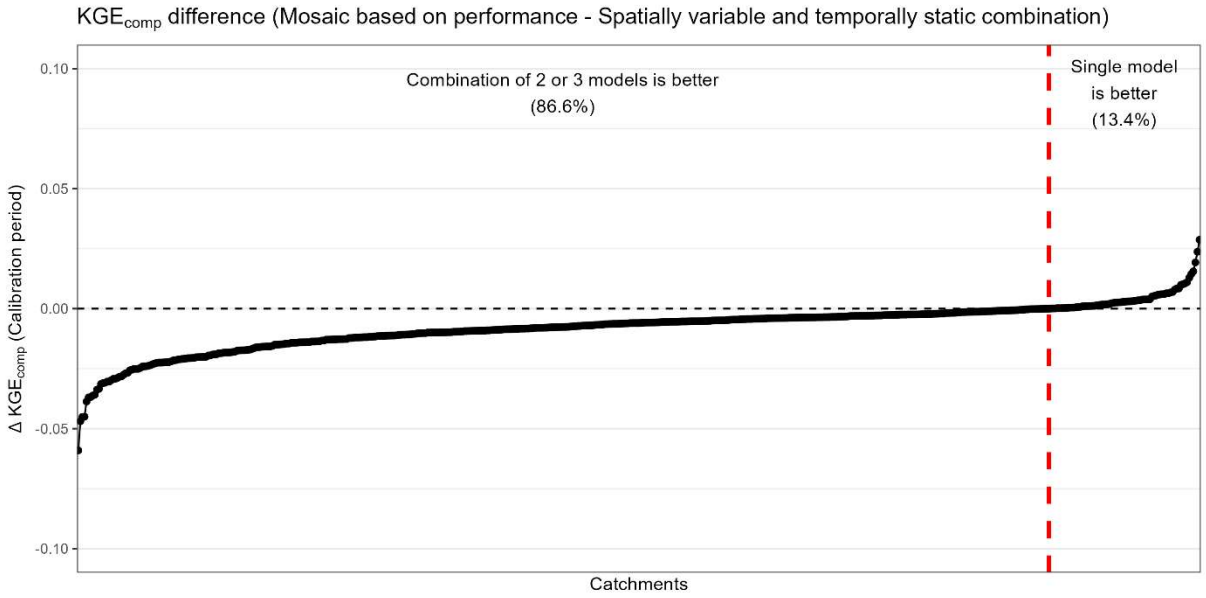
(2) Section 2.5.3.1. Could you clarify how the models were combined? "using a simple average of up to four models" – did you take an equally weighted mean of discharge values from all four models for each timestep?

In Section 2.5.3.1, the combination is computed as an equally weighted arithmetic mean of simulated streamflow from the selected structures. In other words, for a combination of n models, streamflow at each timestep is calculated as the simple average of the n individual simulations, with weights equal to 1/n and constant across both time and space for this approach. We have clarified this explicitly in the revised manuscript.

(3) Section 2.5.3.2. – the method selects "the combination of up to three models that yields the highest KGEcomp scores over the calibration period" – I'd be curious to know if there any cases where a single model is better than any combination of 2 or 3 models? And in this case would you use the single model as 'the best combination' or does this method require a minimum of 2 models? Again, this section could refer to appendix A.

In the present study, a single model (i.e., a combination of size one) was not considered within the combination framework and was therefore excluded from the search space. As a result, the top-performing combination always consists of two or three models. We have clarified this point in Section 2.5.3.2 and added an explicit reference to Appendix A.
We note that in a limited number of cases (73 catchments over 544), the best-performing single model achieves a higher KGE_{comp} value than any combination of two or three models over the calibration period. This indicates that while model combinations tend to improve performance overall, they do

not systematically outperform single structures in every catchment. Allowing combinations of size one could therefore lead to marginal improvements in a small subset of catchments. However, given the limited number of affected cases and the generally small differences in KGE_{comp}, the overall performance distributions and the main conclusions of the study would remain unchanged. This analysis was added in Appendix C.



RC1 — Figure 2 : Difference of performance KGE_{comp} between the top-performing single model and the top-performing combination (2 or 3 models) for each catchment over the calibration period. Each dot represents a catchment, and they are ordered by difference of performance. The dashed black line indicates equality; dots under show that the top-performing combination of 2 or 3 models is better than the top-performing single model and vice versa. The dashed red line highlights the tipping point.

Figure 9. This figure has a lot of information content with the locations of all gauges, but it is hard to see at a glance which methods are doing better and which are equivalent. I found myself trying to read and compare the numbers written above each map and struggled to see any patterns with so much information available. Could it be presented more clearly, e.g. as a table of pie charts/bar graphs rather than a table of maps?

We thank the referee for this helpful suggestion. We agree that the original presentation of Figure 9 made it difficult to directly compare the relative proportions of “better”, “equivalent”, and “worse” outcomes across methods.
To improve clarity, we have replaced the table of maps with bar graphs summarizing the percentage of catchments falling into each category for each pairwise comparison. This revised visualization allows a more immediate comparison of the relative performance of the different approaches.

Appendix line 646: “Interestingly, these are the two top-performing models in Figure A2, but model 72 is not selected in Figure A4, suggesting a large degree of similarity (i.e., equivalence) between both models.” Could this also be because model #72 is equivalent with other model structures (e.g. 96) – rather than necessarily being equivalent with #126? And does equivalence (i.e. similar KGE scores) necessarily mean similarity (i.e. similar hydrographs)?

We thank the referee for raising this important point. We agree that the absence of model 72 in Figure A4 does not necessarily imply direct equivalence with structure 126 alone. Model 72 may also be performance-equivalent to other structures (e.g., model 96 or others), and therefore may not be required in the minimized set identified by the linear programming procedure. We have revised the text to clarify this point.

We also agree that performance equivalence — defined here based on KGE_{comp} within sampling-uncertainty bounds — does not necessarily imply similarity in hydrograph shape or process representation. It indicates statistical indistinguishability with respect to the chosen metric, but does not guarantee identical dynamic behaviour. We have clarified this distinction in the revised Appendix.

Figure A10: how are the catchments ordered in this figure? Knowing if catchments are grouped by location, key characteristics, or performance would help with the interpretation of this plot.

In Figure A10, catchments are ordered according to their station identifiers (USGS IDs). This ordering does not reflect any deliberate grouping by hydrological characteristics or model performance, although USGS station numbers broadly follow a regional logic and therefore may roughly correspond to geographical organization.
We have clarified this in the revised figure caption to facilitate interpretation.