We would like to first thank the anonymous reviewer for taking the time to review our manuscript in detail, and provide feedback that will help us to improve the manuscript. Thank you.

Our comments are organized below to answer the original reviewer comments (shown in *italics*), with our responses in **bold**.

As part of the comments from this reviewer as well as reviewer 2, we will be making the following major changes in the manuscript:

1. introducing an additional 12 independent catchments for validation, for a total of 24;
2. supplementing the validation analysis with the Wilcoxon rank sum test (effectively implemented as a Mann-Whitney U test) to test the hypothesis that the distribution of results within each independent catchment are identical, performed for both calibration and validation results

More minor adjustments to the manuscript based on reviewer comments will include:

1. adjusting the language to use the mean KGE across catchments to inform model configuration development, but is not used as far as formal testing of independent catchments
2. removal of the identifiability section (Section 3.5)
3. inclusion of a discussion on general differences between catchments, i.e. that some have higher and more consistent performance while others are more variable
4. inclusion of basic hydroclimatological metrics (annual average precip, PET, runoff ratio, etc.) with the tables of catchments in the paper

We believe that these changes will strengthen the manuscript and particularly the conclusions around presentation of an improved blended model. The results for some of these updates are presented as part of the response below.

1 Reviewer #1

For the comments submitted by Reviewer #1, we respond to the three primary categories of comments provided, as numbered from 1-3 in the initial reviewer's comments. Additional comments provided in the form of an annotated PDF file have been responded to within the annotated PDF file, and changes have been reflected in the updated manuscript.

Overall, we thank the reviewer for their detailed analysis of our manuscript, and we believe addressing these comments will strengthen the manuscript and results of this study.

1.1 Selection of process equations to test are poorly justified

The small sample of catchments suggest that the process equations used in the modeling framework will be selected based on our understanding of which processes are important in these basins, and which equations are appropriate to model these processes. There is some mention in the manuscript that expert knowledge informed at least some of the changes made to the Raven framework used in this paper, but what this knowledge was and how it informed some of these decisions is unclear. More bluntly, the approach largely seems to come down to testing various options for no other reason than that they are available in the Raven framework and seeing what sticks.

If this approach is meant to bridge the gap between machine-learning and traditional models, it must be clarified why and how the modeling equations that are being tested were chosen to be tested in the first place.

Unclear if the reviewer is concerned about our sample size here so we will address that first. The sample of catchments (12 MOPEX catchments) of Duan et al. [2006] is relatively small compared

to something like the CAMELS dataset [Addor et al., 2017], though these basins were historically selected to capture varying hydrological dynamics and are thus a reasonable set of catchments for our experiment. Using hundreds of catchments for this type of study, which focuses on model development, would render the study computationally impractical due to the large computational requirements per catchment. The effect of using a relatively small number of catchments is also accounted for by the use of independent catchments for validating the improved model, and strengthened further by increasing (now doubling) the number of independent catchments used.

Regarding the selection of equations and the empirical approach to develop the blended model configuration, our options for experimentation were limited due to the infeasibility of testing every possible combination of blended and non-blended model options available in Raven (see Mai et al. [2020] for an estimate of the number of plausible model structures that can be produced in Raven, and the Raven User's Manual [Raven Development Team, 2023] for information). Thus, each model configuration was selected based on the performance of recent configurations, the weights associated with those options within the blended model, and expert knowledge in selecting equations that were different enough to be useful when grouped together. In any case, the use of an empirical approach to guide the process is a defensible one, and the performance results for model configurations bear that out.

1.2 Metrics for success are poorly justified

The manuscript uses multiple metrics to assess the success of different model configuration, but it is unclear to me why the main one of these (average KGE scores across the samples) is appropriate. The number of catchments is quite small and performance changes for any of the model configurations seem a mixed bag at best: performance goes up in some catchment and goes down in others. Given that (1) there is considerable sampling uncertainty in these scores at the best of times, and (2) that it is known that KGE scores are difficult to compare between different flow regimes, it must be justified why looking at mean KGE changes is an appropriate thing to do.

This applies especially strongly to sections 3.4 and the conclusions, where the selection of model configuration 36 as the new blended model is declared a success based on very mixed, but on average slightly positive, changes in KGE scores.

This point was strongly considered in the updates provided, particularly to the validation of performance (Section 3.4) in our manuscript. We maintain the the use of mean KGE across catchments and trials is a reasonable metric to guide model development and generate new model configurations, though other information (such as process weights) was also used in developing model configurations (see Section 2.2).

With respect to the validation approach, we have added 12 additional independent catchments for validation, and opted to use the two-sided Wilcoxon rank sum test (implemented as a Mann-Whitney U test) to compare results within each of the 24 independent catchments, both for calibration and validation. These results are presented in the Table 1, and also displayed graphically in Figure 1.

Table 1: Results of the unpaired, two-sided Wilcoxon rank sum test in the independent validation catchments shown. A confidence level of 95% is used. The median KGE performance for both the original blended model configuration 0 and the selected blended model configuration 36 are shown, as well as the p-values for testing whether model config. 0 is different than the model config. 36.

Watershed Index	Model Config. 0 Median KGE	Model Config. 36 Median KGE	p-value
Calibration KGE comparisons			
13	0.900	0.919	3.94e-08
14	0.871	0.895	6.79e-08
15	0.912	0.908	1.37e-01
16	0.917	0.944	1.45e-11
17	0.894	0.946	1.02e-10
18	0.816	0.824	1.27e-01
19	0.902	0.911	1.48e-01
20	0.751	0.838	2.76e-10
21	0.885	0.878	1.19e-04
22	0.783	0.820	2.32e-08
23	0.894	0.884	2.00e-06
24	0.887	0.891	4.95e-01
25	0.797	0.906	1.45e-11
26	0.916	0.941	1.45e-11
27	0.796	0.864	1.02e-10
28	0.738	0.777	1.45e-11
29	0.861	0.892	1.76e-08
30	0.568	0.705	1.45e-11
31	0.850	0.890	1.33e-08
32	0.647	0.728	1.45e-11
33	0.901	0.903	6.40e-01
34	0.833	0.901	1.02e-10
35	0.660	0.841	1.45e-11
36	0.847	0.852	7.18e-02
Validation KGE comparisons			
13	0.871	0.894	7.47e-04
14	0.902	0.905	2.21e-01
15	0.834	0.837	4.45e-01
16	0.836	0.898	2.83e-09
17	0.765	0.818	5.41e-09
18	0.592	0.639	1.12e-02
19	0.750	0.771	2.63e-02
20	0.494	0.405	4.91e-02
21	0.701	0.710	7.58e-01
22	0.739	0.781	4.60e-02
23	0.749	0.738	2.65e-01
24	0.810	0.848	2.39e-03
25	0.787	0.870	1.74e-10
26	0.845	0.877	3.36e-05
27	0.761	0.724	2.11e-02
28	0.720	0.672	1.45e-11
29	0.744	0.772	3.36e-05
30	0.711	0.756	1.59e-03
31	0.743	0.857	5.41e-09
32	0.621	0.638	3.03e-08
33	0.826	0.824	8.83e-01
34	0.686	0.761	2.12e-07
35	0.515	0.704	1.45e-11
36	0.601	0.658	1.33e-08

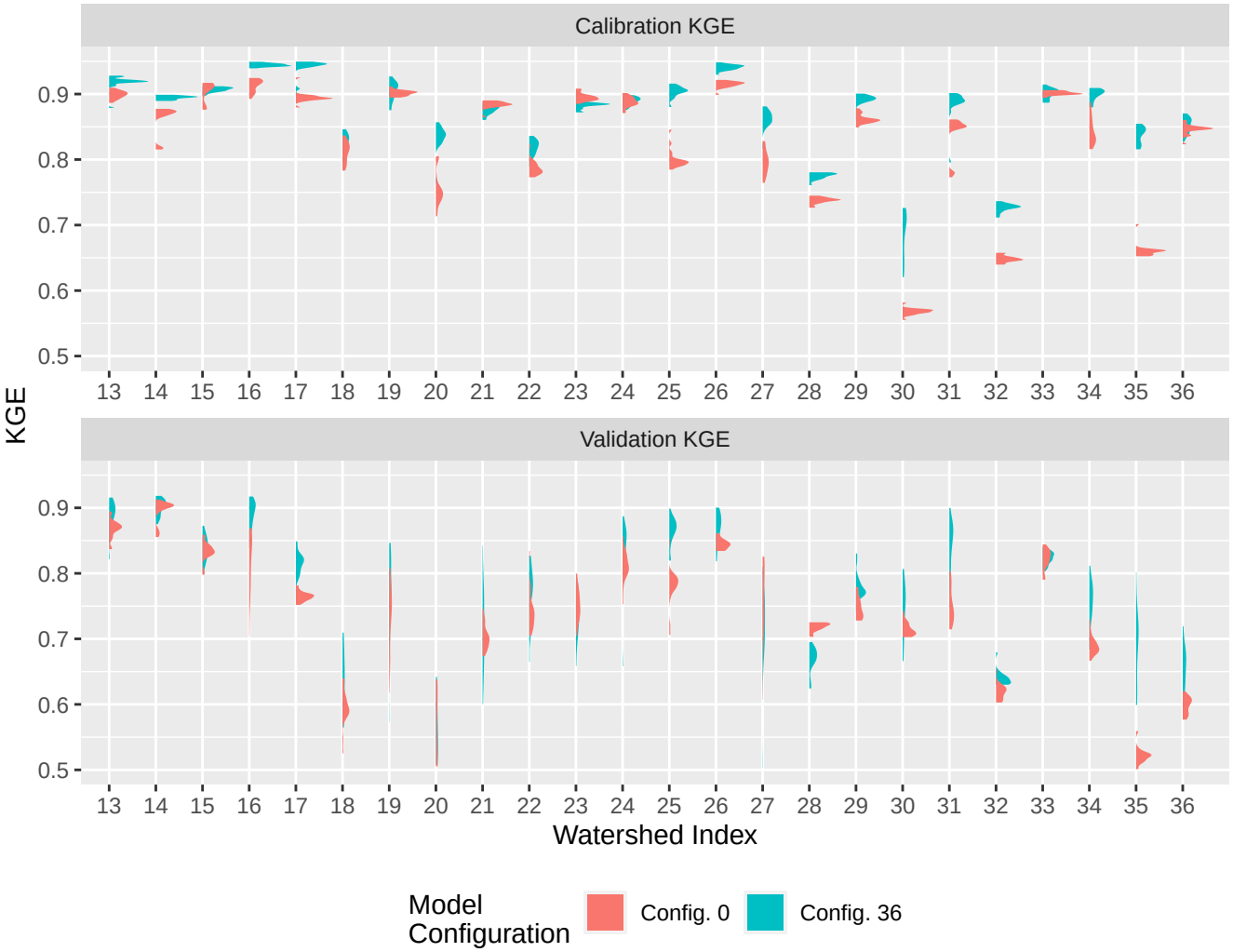

Figure 1: Comparison of KGE performance distributions for the original blended model configuration (configuration 0) and the selected blended model configuration 36. Distributions are shown as density functions for both calibration and validation.

The table shows that at the 5% significance level, there are 16/24 independent catchments where the calibration performance is significantly improved, and also 16/24 where the validation performance is significantly improved. There are 2 catchments where calibration performance is significantly greater in model config. 0 than 36, and 3 where the same is true for validation performance. Overall, the selected model configuration 36 performs better (significant p-value with the median greater for model configuration 36) or statistically similar (insignificant p-value, i.e. greater than 0.05) to the original configuration in 22/24 (92%) of the independent catchments for calibration performance, and 21/24 (88%) for validation performance. With these results, we can avoid the use of averaging KGE across catchments to present our main conclusion, which is that the selected model configuration 36 is an improved version. We trust that this will satisfy the core of this concern.

1.3 Results are strongly conditional on practical limitations of the calibration algorithm

The stated goal of this paper is to find a new blended model configuration to use as a default starting point for further work. Hence multiple possible new blended models are calibrated and tested, to select this new best one. The authors are open about the fact that the calibration algorithm sometimes struggles to find the optimum solution during the combined calibration of weights and parameters (line 281-284). However, I think simply stating this is insufficient in this case. According to Table A1, the model selected as the new blended model configuration to use (#36) is a subset of another model configuration (#24). In other words, model configuration #24 has all the capabilities of model configuration #36 and then some, yet it is not the one selected. The reason for this is that model config #24 performs worse than config #36 during calibration (Figure 3) and on the chosen evaluation metrics (Figure 5). Unless I misunderstand something, logically there can be no other reason for the lacking performance of #24 than that the calibration algorithm failed to find the proper weights and parameters.

I do not think that for a study such as this such a weakness in the calibration part of the work can be brushed aside. The small number of catchments adds to this problem, because the calibration outcomes seem somewhat chance-based, and it is impossible to know if these findings would hold across much larger samples.

As a technical correction to the comment, model configuration 36 is not a subset of model configuration 24 due to non-structural changes made (not reflected in Table A1), though we can still discuss this comment on the main point about limitations of calibrations algorithms. In this section, we present only model configurations 17-24, such that all model configurations presented in this section are subsets of the conglomerate model (configuration 24).

The failure of optimization algorithms to find the precise global optimum is ubiquitous in hydrologic modelling, and thus the solution in research is to consider this limitation when conducting modelling experiments. Here, we use a budget of $10,000$ model evaluations. The purpose of including a conglomerate model was to show that we cannot simply include every option under the sun in the blended model due to this practical limitation of our calibration routines. While the conglomerate model would outperform the other models in model calibration that are subsets of it given infinite model evaluations (though perhaps not in validation), we instead wanted to focus on building a reasonably parsimonious model that can be used in research and practice with reasonable calibration exercises, while using a large enough budget to make it likely that the model runs did approach a global optimum.

To emphasize this argument, we conducted a mini-experiment to assess the results of the conglomerate model (model configuration 24) in the event that we used an order of magnitude more calibration budget. We ran the same experiment for the conglomerate model just in MOPEX catchment #5, and calibrated this model with a budget of $100,000$ runs (10x the budget of all other model runs) and compared the results for calibration and validation in this catchment. We selected model configurations 17-23 as they are subsets of the conglomerate model, and the conglomerate model with a budget of $10,000$ runs (model configuration 24) to compare results. The performance results are shown in Figure 2; the same Wilcoxon rank sum test to compare the performance of each model to the large budget conglomerate model (results table omitted for brevity).

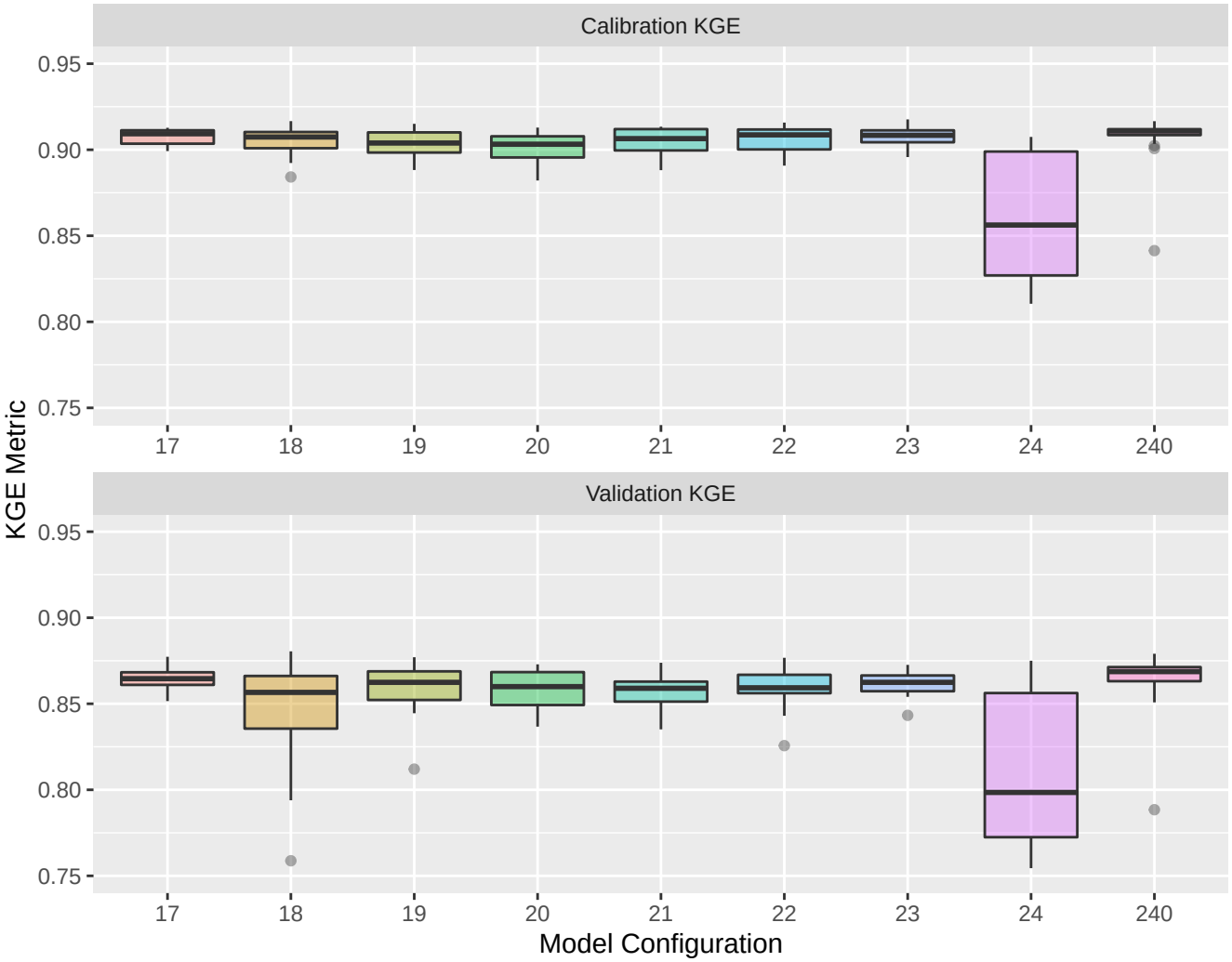

Figure 2: Comparison of KGE performance distributions for several models prior to the conglomerate model which are subsets of the conglomerate model (configurations 17-23), the conglomerate model (24), and the conglomerate model with a calibration budget of $100,000$ runs (labelled as model configuration 240).

The figure shows a substantial improvement in the conglomerate model with the calibration budget increased by an order of magnitude, indicating that convergence of the calibration algorithm is indeed a large factor of the conglomerate model's performance. However, even with this vast increase in budget for just the conglomerate model, it still fails to significantly outperform 4 out of 7 of the other model configurations in calibration, and 2 out of 7 of the other model configurations in validation, not including the conglomerate model with the original budget, based on the Wilcoxon rank sum test performed. It is also worth noting that the large budget conglomerate model has a number of poor outliers in calibration and validation. This suggests that even with a very large budget, the complex model is not immune to these shortcomings of the calibration algorithm, i.e. getting stuck in local optima.

This returns us to the question of, why are we building these models, and under what conditions will they be used? Our goal in this study is to develop a high-performing version of the blended model that researchers and practitioners will be able to use in their studies or applied work. The comparison of models under equal calibration budgets provides a way to test whether the added complexity in the models justifies the improved performance, and ensures that the model developed will still be useful to others with limited calibration budgets, as is always the case in research but especially in practice. The mini-experiment results show that, at a minimum, at least some models that are a subset of the conglomerate model are still statistically indistinguishable in performance from the large budget conglomerate model, suggesting that there is value in considering these budget constraints and carefully curating our models, rather than taking the most complex model possible and increasing the calibration budget. It is worth noting that the increase in calibration budget from $10,000$ runs to $100,000$ requires approximately 15x (not 10x) the allocated runtime in our server to run successfully, due to the uncertainty in running computational jobs with longer runtimes.

Overall, in the context of developing models for practical use and high performance, we trust that this discussion and exercise has convinced reviewers and other readers of the need for accepting the limitations of calibration algorithms and considering performance within a given budget as a real criteria in this experiment.

1.4 Summary

None of the items above are necessarily bad on their own (I understand that sometimes one just needs to get started somewhere), but combined they cast serious doubts on the validity and usefulness of this work. Put bluntly, it is unclear to me why the chosen methodology is appropriate for what is being investigated and what scientific advance is being made.

I believe these concerns may possibly be addressed by substantially increasing the number of catchments in the analysis, some re-thinking of how to quantify success, and inclusion of a more conceptual discussion about the purpose of blended models and this investigation in particular, but I think the amount of changes needed go beyond a simple revision.

We thank you again for the comments on our manuscript, as they were constructive and have helped us greatly to strengthen it, particularly in the validation of results and conclusions sections. We trust that the updated analysis, as well the mini-experiment presented here with respect to calibration budgets, has adequately addressed your concerns regarding our manuscript.

References

Nans Addor, Andrew J Newman, Naoki Mizukami, and Martyn P Clark. The camels data set: catchment attributes and meteorology for large-sample studies. Hydrology and earth system sciences, 21(10):5293–5313, 2017. ISSN 1607-7938.

Q. Duan, J. Schaake, V. Andréassian, S. Franks, G. Goteti, H. V. Gupta, Y. M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O. N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian,

and T. Wagener. Model parameter estimation experiment (mopex): An overview of science strategy and major results from the second and third workshops. Journal of Hydrology, 320(1-2):3–17, 2006. doi: 10.1016/j.jhydrol.2005.07.031.

J. Mai, J. R. Craig, and B. A. Tolson. Simultaneously determining global sensitivities of model parameters and model structure. Hydrology and Earth System Sciences, 24(12):5835–5858, 2020. doi: 10.5194/hess-24-5835-2020. URL https://hess.copernicus.org/articles/24/5835/2020/.

Raven Development Team. Raven: User's and Developer's Manual v3.7, 2023. URL http://raven.uwaterloo.ca/.