We would like to first thank reviewer Janneke Remmers for taking the time to review our manuscript in detail, and provide the feedback to us to help clarify a number of points and improve our manuscript overall. Thank you.

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

Reviewer #2

0.1 General Comments

1. *1. Model space: in the introduction and methods, model space is mentioned to delineate this study and to show the added value of the choices they have made. However, afterwards, I missed the referral back to this concept. In section 3.5, this topic is touched upon with the identifiability in blended models. Still, I think this manuscript could benefit from a more explicit (short) discussion on model space: is it possible to sample the full model space? And how well does the blended model created in Raven reflect the full model space? Because the blended model is limited to the capabilities of Raven, therefore, it might not be possible to fully sample the model space.*

   A more complete discussion of model space, including its limitations, has been added to this section in the Introduction. We believe this adds a nice context for the blended approach in the theoretical landscape of hydrologic modelling. However, the Identifiability section is being removed and discussing this again in other sections does not seem like a good fit, so the discussion is kept to this Section 2.1 in this manuscript.

2. *2. In their testing, they found that some options for certain processes do not have to be included in the blended model. The addition of these options does not enhance the performance of the blended model. How would this reflect on models that do use these process options within their model structure? Can anything be said about this?*

   It is most likely that the process algorithm with low weights that was removed is not useful within that particular process group, perhaps because it's functionality was duplicated by another similar algorithm within the group (i.e. not much value added by a second similar algorithm to estimate a particular flux). This does not mean that the algorithm would not function well in another model configuration (especially a non-blended one), though it would be interesting to test that in a different study by swapping some of these algorithms and comparing results. A line has been added to address this question.

3. *3. In the blended model, a weighted average is used between different process options. I was wondering how this influences the overview of uncertainty and what this means for the processes itself. In nature processes are not averaged, how can the results of the blended model be interpreted? I could not find anything related to this in Mai et al. (2020) either, but maybe I missed this. Whether or not this will be included in the manuscript, I leave up to the authors, though it might improve the delineation of the usefulness of a blended model further.*

   I've added a couple of lines about this in Section 2.1. The interpretation of a blended process group is essentially the same as any process algorithm being used in a model, just a more complicated way of determining the flux for that process. For example, instead of using the Penman-Monteith equation to estimate PET, we can use a blended group that includes several PET equations, but the end result is still an estimate of PET, the same result as any other PET equation. The same applies for determining the flux between two storage units in the model (e.g. baseflow from soil to surface water), it is just a more complex equation to determine the flux but otherwise equivalent. In nature, none of the equations we use are 'true' but we use them anyway as (hopefully) reasonable representations of that process. The only difference here is that the blended approach has some flexibility to adjust the representation through weighting, but the end result is similar conceptually. With respect to uncertainty, the approach of blending opens the door to the application of sensitivity analysis and perhaps uncertainty analysis as well without the need for model ensembles, the former of which was demonstrated in [Mai et al., 2020].

4. *4a) Section 2.2, paragraph 2 (from line 128), the calibration and validation are described. I have two questions about this: i) The period chosen is respectively 1972-1983 and 1984-1989. Why did they chose these two exact periods? And not for example 10 years later for both calibration and validation. ii) For calibration, a 2 year warm up period is used. Why not the same for validation? Are the initial conditions copied from the last time step of the calibration?*

   In this case, the period was chosen to be consistent with the setup from Chlumsky et al. [2021], and the period of 12 years for calibration and 6 for validation is deemed long enough for a reasonable evaluation. A longer duration (e.g. 20 years calibration) would be feasible but we also wanted to keep the runtime of each calibration experiment shorter for computational purposes, given the number of model evaluations run for all configurations (¿30 configurations and 2.4 million evaluations per configuration).

   To the second question, yes the model is run continuously from calibration to validation (when the validation period is evaluated) so the model is already 'warm' without the need for an additional warm up period.

5. *4b) Section 2.2.1: in section 3.1 (line 268), 'expert consideration' is mentioned as input for the choices of the initial update. However, I did not get this from section 2.2.1, so I would recommend to add this already to the methods. On top of that, 'expert consideration' raises the following questions with me: whose expert considerations? How many options were added? Any processes still missing based on expert consideration?*

   The expert consideration is now mentioned in Section 2.2.1, including to mention the expert consideration of the co-authors. The added options can also be viewed in Table A1, now linked in that section as well. The remainder of the modifications were based on empirical evidence, thus the expert consideration was only specifically applied for that first modification of selected process equations, and is not a large focus on this study.

6. *4c) Section 2.4: in the final paragraph (line 231), the combination of the different metrics are explained. In this paragraph, I missed which metrics were combined (even though in the results this does become clear through figure 5). Also, would it matter if they combined the metrics in a different way? Or changed the order of how they combined them?*

   The four metrics are added to the first line of this last paragraph in Section 2.4 to clarify. The second question was interesting. We tested plotting the pareto plots in all possible combinations, and in any case the results do not change - model version 36 is always on the Pareto front in one plot, and the dominant point in a second plot. However, this was a good check on our method. A line has also been added in the results section to mention that the order of the metrics in plots does not change the final outcome.

7. *4d) General: in sections 2.2.1, 2.2.2 and 2.2.4, new options or different options are tested for the blended model strategy. I wondered why certain options were tested and not others or if this was all that could be tested. For example: i) Section 2.2.2: why were potential melt and potential evapotranspiration chosen to be blended? And not others? ii) Section 2.2.4: line 187 – 189, were all Raven options tested? And why were these kept non-blended?*

   The two main reasons for using potential melt and PET in blending (question i) were 1) they were considered the most influential in the hydrologic model, and 2) they both had many options to experiment with, where as many other forcings in Raven have only a few. A few lines have been add to capture this.

   On question ii), not all Raven options were tested for each process as this would result in more model configurations than could be reasonably tested, though for some groups like canopy drip, both options (i.e. all) algorithms in Raven were tested. These were kept as non-blended because of the limited variety and number of algorithms in Raven, as well as the consideration that blending would likely not add much value in that particular case (e.g., the difference between a blended depression seepage and a single depression seepage equation is likely not nearly as important as blended PET vs a single PET equation). Lines are added to Section 2.2.4 to capture some of this discussion

8. *5a) For the Results and Discussion: Section 3.1, line 277 – 280: here the mean calibration KGE is described to have been improved a certain amount. Yet, in figure 2 they showed the maximum calibration KGE and validation KGE. To me, this was confusing, because it is not consistent. Also, was the mean*

calibration KGE for configuration 36 the highest? Or did they mention this one, because it is the new model configuration? This confusion between mean calibration KGE in the written text and maximum calibration KGE in figure 2 applies for this whole paragraph

This confusion was noted by the other reviewer as well, and this section will be adjusted to discuss mean KGE only for consistency. In addition, the final results and discussion will lean away from using mean KGE to justify the model developments, and in the validation Section 3.4, we will instead employ a basin-by-basin comparison using statistical means to test the results within each basin.

9. *5b) For the Results and Discussion: Section 3.5, line 365 – 371: I think I understand what they mean after reading this part a couple of times (in some groups, all options were used in all trials, meaning that the weighs attributed to the options was more equal. But for other process groups, sometimes only 1 option was (mainly) used in certain trials, meaning the weights were not equally distributed.). Initially, I thought it would be difficult to define a preferred option with only 2 options available within 1 process group and wondered how this distribution could be wide (it only spans the 2 options).*

   This section on Identifiability is being removed from the manuscript, based on comments from another reviewer. However, this section is basically referring to the distributions of weights within the results of a given model configuration for a specific process group. In some groups, like baseflow with 2-3 options, the distributions all basically overlap and there is no clear preference (and much uncertainty), where as in other groups, there is much less overlap and narrower distributions, making the 'preferred' algorithm choice much more clear.

10. *6a) For the Conclusions: Line 395: "strategies to reduce the dimensionality of blended models." I did not understand what this is related to and where it came from. Do they mean the testing they conducted on which options were not used during multiple trials and multiple catchments?*

    This word has been replaced with complexity, which is now also defined in the manuscript more formally.

0.2 Specific Comments for Figures and Tables

1. *Concerning the figures, it is mainly the lay-out. For figure 2 and 3, I have some additional questions as well. First a general remark: I would recommend to change the axis of all figures a bit, so it aligns better. This would make the figures more visually appealing*

   The axes on these figures can certainly be updated to align better in the manuscript.

2. *Figure 1: There are several points for which the ID is not connected to the dot. This is sometimes confusing, especially for WV (4). I would recommend to make this consistent. The other points that are not connected are: ID (22), ME (13), and MD (2)*

   This is a function of the script we are using, but we can do our best to ensure that the points are connected for readability. Thank you for the suggestion.

3. *Table 1 a) Based on line 245/246, I expected all catchments to be included in this table, not just the 12 independent catchments. I would add all catchments to the table, as no information is provided in the whole manuscript (including appendices) about the MOPEX catchments. Another option would be to change these sentences (e.g. change 'selected' to 'independent' or 'validation')*

4. *Table 1 b) Because the catchments were chosen to represent diverse climate conditions, I would recommend to add some information about this or a reference in the table.*

   Something similar was suggested by reviewer 1, to include all catchments and include some basic characteristics of each catchment. This will be updated in the manuscript, including for 12 additional new validation/independent catchments that will be added.

5. *Figure 2 a) With model configuration 1, a drop in max calibration and mean validation KGE due to different data. What if someone else uses different data? Would this mean an initial drop in performance as well?*

In this case model configuration 0 uses the PET_OUDIN algorithm to estimate PET, which appears to do a better job estimating PET than the provided PET "data" used in model configurations 1-5. This indicates that the PET data may have been model-estimated or model-informed as well, and that the routine used here seems to perform better. It is difficult to extrapolate outside of this particular setup, though it would depend on the quality of the PET data available to the user. The model-estimated PET values for later model configurations does appear to work well under the conditions in this study.

6. *Figure 2 b) Model configuration 16 seems to be an outlier, do they know what caused this?*

   It is perhaps a particularly poor combination of POTMELT algorithms, though it would require a bit more investigation to be certain. Model configuration 16 is modified from model configuration 15 just based on the combination of PET and POTMELT algorithms, so it is likely the combination of one or both of those sets of algorithms.

7. *Figure 3 a) It would be great to improve the colourblind friendliness of this graph. The triangles are difficult to see against the inside of the boxplots. In line with this, I would also change "orange points" to "orange triangles" in the caption. This also applies to figure 6 (visibility of triangles and lines of the boxplots and the phrasing in the caption). Could, for consistency, the same colour be used for both validation triangles?*

   Figure 6 is being updated and will show the distributions for calibration and validation, thus will no longer have triangles. However, the colourblind friendliness of these figures, particularly with respect to the orange triangles, can be improved. Thank you.

8. *Figure 3 b) At MPX 8 (Idaho), I do not see any validation triangle for model configurations 0, 2, 3 and 24. Was validation not possible? Or did the point fall outside the graph?*

   In those cases, the point fell below the plotted threshold of 0.6 in validation. The axis limits are set to provide a reasonable view of the results without compressing the better performing ones to a point with no visibility, though a line has been added to the description for Figure 3 to mention this clarification.

9. *Figure 3 c) Why were these 8 model configuration chosen to be shown in this graph? For some of them, I can understand the reasoning (e.g. 24), but for others I do not (e.g. why both 7 and 15?).*

   They generally represent the stages between the model development steps, for example, model configuration 7 is the first model with blended PET and POTMELT, 15 is a mid point between that and the conglomerate model, model configuration 29 was the first configuration in reducing complexity, etc. However, other configurations could have been chosen for this plot as well to represent the same storyline (e.g., 14 instead of 15 would capture this as well).

10. *Figure 3 d) Is it known why in some catchments the validation is consistently lower (e.g. MPX 1 and 7)? And some more variable (e.g. MPX 8, 11 and 12)? This spiked my curiosity*

    This was mentioned by the other reviewer as well, and a discussion portion will be added to at least touch on this. Our thought is that some catchments are easier 'problems to solve' with respect to capturing the streamflow dynamics, and have perhaps more information content in their hydrographs than others, leading to the generally higher performance and more consistent performance in some catchments than others.

0.3 Specific Textual Remarks

1. *1. Line 71 – 78 and line 81 – 87: in both they give a more detailed description of what the methods section entails, but to me it felt repetitive, because it is right after each other. I would recommend to leave it out of the introduction or at least reduce it substantially.*

   The portion in the Introduction has been reduced to provide a broader overview of the manuscript, and the methods lines have been focused to describe the organization of the methods section.

2. *2. Line 144: "With the analysis of successive each blended model configuration", it seems as if the order of this sentence is not quite right. Maybe 'each' and 'successive' should be swapped?*

   Yes, these should be swapped and has been adjusted in the manuscript.

3. *3. Line 181: "whether indicate whether the maximization". I believe "whether indicate" should be deleted.*

   Agreed, this has been adjusted in the manuscript.

4. *4. Section 2.3: I understand the assumption explained in the first paragraph. However, I Line 205: I had to read this paragraph twice to understand it.*

   I have adjusted this paragraph slightly and removed the second half of the last sentence (originally line 205) to make this line clearer.

5. *5. Line 211 and 225: both have an explanation of the validation gap. Personally, I found the explanation at line 225 clearer. So, I would recommend using that explanation at line 211 and refer back to this in line 225. To me, it seemed quite repetitive, especially because the more elaborate explanation came later.*

   I hvae provided a very short explanation of the validation gap in the earlier section, and pointed to the more rigorous definition in the original lines 225 (Section 2.4) to reduce the repetition. Thank you for this suggestion.

6. *6. Line 253/254: "undertaken assess" should be "undertaken to assess" I believe*

   This line has been updated.

7. *7. Figure 5: in the caption, the concept calibration consistency is used, but this is not explicitly mentioned. I think it would be good to add this.*

   Thank you, this has been updated and is more consistent with the other descriptions.

8. *8. Line 319 – 326 (section 3.3): in discussing figure 5A, I would expect the usages of the terms validation gap and calibration consistency. At the moment, I thought something else was referred to in this paragraph.*

   These terms have been added to the paragraph for consistency.

References

Robert Chlumsky, Juliane Mai, James R. Craig, and Bryan A. Tolson. Simultaneous calibration of hydrologic model structure and parameters using a blended model. Water Resources Research, 57(5):e2020WR029229, 2021. doi: 10.1029/2020WR029229.

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/`.