Comments review Chlumsky et al. (2023) by Janneke Remmers

My points for clarification:

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.

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?

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.

4. Regarding their methods, I had some questions about the choices in this study. I have a background in motivations behind modelling decisions, so this focus is apparent in the questions I ask here. I leave it up to them, to what extent they want to incorporate or clarify the following aspects within their manuscript.

    a. 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?

    b. 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?

    c. 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?

    d. 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?

5. For the Results and Discussion:
    a. 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.

    b. 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).

6. For the Conclusions:
    a. 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?

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.

1. Figure 1:
    a. 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)*.
2. 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').

    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.

3. 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?

    b. Model configuration 16 seems to be an outlier, do they know what caused this?

4. 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?

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?

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?).

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.

Lastly, some textual remarks:

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.

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?

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

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.

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.

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

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.

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.