RC2:

I believe this is a significant paper, I am looking forward to its publishing, however, I do believe some improvements can and should be made first. The paper lacks a bit of specificity when it comes to implementation details of the metamorphic testing approach. In particular I find the choice of model types in the multi-model approach important, the merits of which are not particularly covered in the paper until the conclusion section. I would suggest this be done in a more sophisticated way around lines 88-89. Other considerations should also be specified and guidance provided as to what correct versus incorrect implementation entails, it is at the moment a bit too broad on this front.

> We agree to add a few sentences about the considered model structures and the underlying motivation. In particular that we can expect physically realistic responses from conceptual models as long as the modified inputs do not lead to changes in catchment properties, such as vegetation and soil structure, whereas the machine learning models bear the potential of covering such changes, but their accuracy/correctness beyond the calibration data is difficult to estimate.

Please see more details below:

Line 19: Odd source placement.

> The source is primarily about the hydrological data and catchment properties, not about all the research stimulated by that. For this reason, moving the citation to the end of the sentence would not be correct. We will add examples of the research stimulated by these data at the end of the sentence and hope that this will clarify both citation placements.

Line 24 and 30 please specify whether you are talking about deep learning, and if you do then clarify right away. Additionally, at its current location the Shen (2018) citation seems out of place.

> We intentionally wanted to be general here. We will add a sentence that these achievements were primarily based on deep learning methodologies and even more specifically on LSTM models. We will remove the Shen (2018) citation here.

Line 27: Good that you point this out, please also provide some examples.

> We will provide some references to the combination of both points (i) and (ii). Some of them are already cited in line 22 above.

Lines 33 to 36: Seems out of place, maybe move or integrate into discussion meaningfully.

> We will move this sentence further up after mentioning the success of the machine-learning (deep learning, LSTM) approaches (after the citation of Ma et al. 2021) and mention the availability of more data sets later.

Lines 82 and 85-88: I am glad you point these out!

> We agree that it is very important to be aware of the limitations of this approach also.

Line 88-89: I was hoping for more detail.

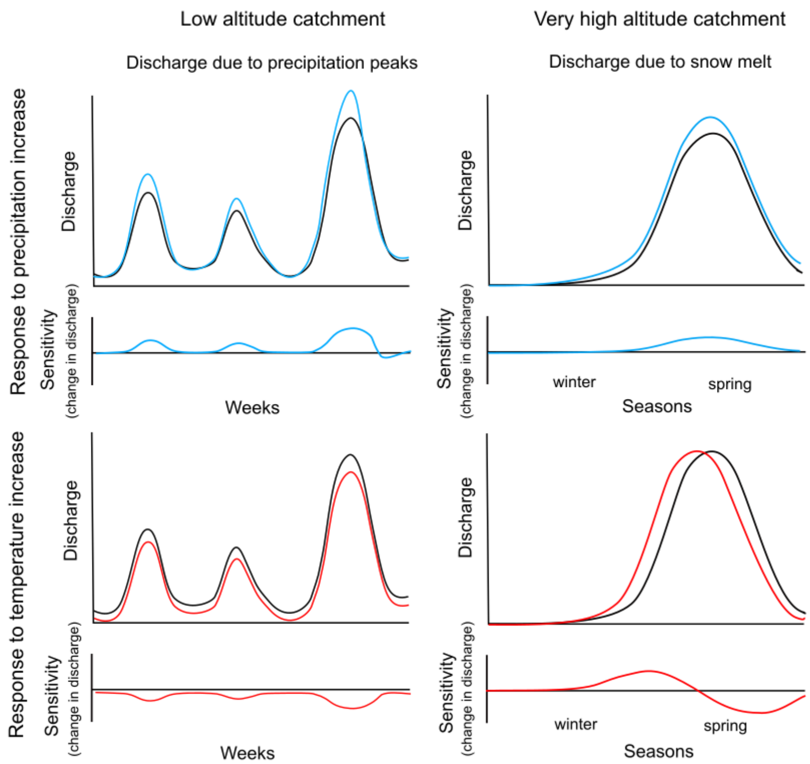
> Since the models are described in section 2.2 and extensively in the appendices A and B, and the calibration options were partly motivated by the results (and thus described in section 3.3, in particular in Table 1), we tried to avoid repetitions. In the new version, we will reference these sections and appendices so that the reader can easily find where to search for more details.

Line 137: Re-cite in the bracket

> We agree.

Line 150: Reduce the overall size of this paragraph and the one around line 130 by focusing only on the basins that will appear in the study. Perhaps a more concise mention near Line 170 and 175 of the expected effects would be best. A schematic can also be useful in showing the expected effects for the basins in your study.

We agree and suggest to add the following figure of the expected qualitative changes in the response at the catchment outlet and shorten the text accordingly.



The expected qualitative changes to the precipitation increase are indicated in blue, those to the temperature increase in red. The qualitative changes are expected to be less smooth than in these figures that just represent general trends due to shorter-term precipitation and temperature fluctuations.

Line 195: Please specify the choice of optimizer.

We will add a sentence to indicate that the chosen optimizer for our model is AdaDelta (lr=1.0, rho=0.9).

Line 220: Remain consistent throughout with your word-choice of training versus calibration. (However, you should keep the statement equating these terms for researchers from different backgrounds that may be used to one or the other.)

We agree. It should have been “training” here. We will explain that we use “training” for machine learning models and “calibration” for the conceptual models as it is typical in the corresponding literature, but that both are conceptually the same although often different algorithms are used.

Figure 2: Utilize symbols for black and white version distinguishability of basin classification.

We will adapt the symbols accordingly.

Figure 3/4/5: Particularly the top 2 panels are difficult to read in black and white, double check with the editor whether this would need to be adjusted.

We will adapt the color palette for color-blind readers.

Line 323-324: Perhaps use a different loss function with guaranteed convexity to avoid this problem. You can still use NSE for evaluation of course to stay consistent with the other models. (I would recommend you do this at least for the reanalysis)

> We agree that experimenting with different loss functions is important but this is not a major focus of this paper as we wanted to conform with common practice with LSTM modelling in hydrology, which so far was mainly based on NSE as a loss function.
>
> We also do not see that we could avoid the existence of multiple (many!) local minima of an LSTM model by changing the loss function as the existence of multiple minima originates from the complex model structure and not from the loss function applied to the residuals between model and data. Using different seeds at least partly demonstrates the variability that we can expect from ending up in different local minima.

Line 363: Instead of "was not investigated" saying "this was not the main aim" would be better.

> We agree.

Line 412-414: Make these claims and future research directions more specific. Why ML and which research will be necessary to achieve this?

> We will mention here, that adding catchment properties, such as vegetation and soil, to conceptual or physical hydrological models, could in principle improve the predictive properties of these models under strong input changes. However, the parameterization of the processes for vegetation and soil structures is very difficult and can lead to higher model structure uncertainty. For this reason, a ML approach that is trained by a large variety of catchments under different climatic conditions that lead to different catchment properties may be a promising approach. However, as seen in the current study, this potential is difficult to realize also and a hybrid approach could be promising.

Last paragraph: In the conclusion new ideas should not be first mentioned as it is done here. Citations should also not be in the conclusion section for this reason. Please mention these points earlier in the discussion and just summarize them in the conclusion.

> We will discuss these points already in the discussion and move the citations there.

Thank you for your submission! I am looking forward to reviewing the revised version of this manuscript!