Assessing the adequacy of traditional hydrological models for climate change impact studies: A case for long-short-term memory (LSTM) neural networks

We would like to thank the reviewers for their valuable and constructive feedback. We appreciate the time and effort that was put into the review. All major and minor concerns have been carefully addressed. Detailed responses to each of the reviewers' comments are presented below. For clarity, the reviewers' comments are presented in black font, with our responses in blue.

Sincerely,

Jean-Luc Martel, on behalf of all authors.

Major modification: important note on significant changes, not affecting the main conclusions of the paper.

The different reviewers highlighted a possible problematic aspect of the LSTM model implementation for this study, more specifically relating to the use of climate-based static descriptors in the model training and simulation. It was pointed out that using static climate-based descriptors in climate change was far from ideal and that they should be removed, or at least made to change along with the climate data. The same reasoning was proposed for the latitude and longitude static descriptors, where the model could learn some specificity for a given region rather than rely on the climate data to modulate flows. While we are expected to provide a detailed explanation on how we will revise the paper, this comment was too important, with the potential to influence the conclusion of our paper and needed to be investigated right away.

We therefore developed and trained a new model that does not use these data (climate-based statics and latitude/longitude descriptors). The only descriptors remaining are actual catchment descriptors such as drainage area, aspect, slope, elevation, soil porosity, land use and other such variables. The model was trained twice: for the regional LSTM model with only the initial 148 catchments (LSTM-R in the paper) and for the continental LSTM model using an extra set of 1000 catchments (LSTM-C in the paper). While we did not have the time to rerun all results, we tested the modified continental LSTM model (which incorporates additional donor catchments) on all 148 study catchments. This was conducted to assess performance over the reference period, using the four climate sensitivity experiments (+3 °C and +6 °C, -20 % and +20 % precipitation), as well as under the 22 climate models for the 2070-2099 period. The regional LSTM model (excluding additional donor catchments) was still running at the time of this writing. Additional optimization runs will also be completed before final publication. However, our results thus far clearly indicate that our conclusions remain unaffected by this methodological change. In the revised version, we will retain only the new runs (excluding latitude, longitude, and static climate variables), as we agree with the reviewers that this is a more reasonable approach for climate change impact studies. The various sections will be reworked accordingly.

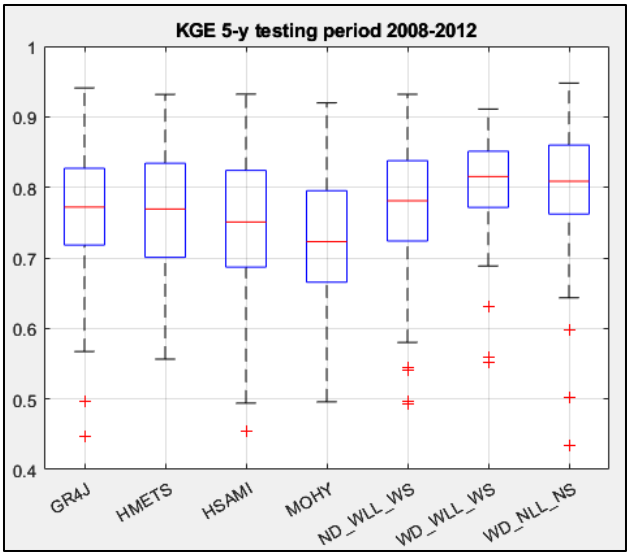

Figure RR1: This figure is the same as Figure 3a in the original paper (model performance over the testing period), with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors (NS)). As shown, there are some minor differences compared to the latitude-included boxplots, but these are minor and align with the expected variance from multiple runs. We will also retrain the model to further optimize it. We anticipated some performance decline over the reference period due to the omission of these static descriptors and were somewhat surprised that it was not the case. This indicates that the essential information is contained within the hydrometeorological historical time series (P, T, Q) and the true static catchment descriptors, and the LSTM model can effectively capture this information without the excluded descriptors. This in itself is a very interesting result.

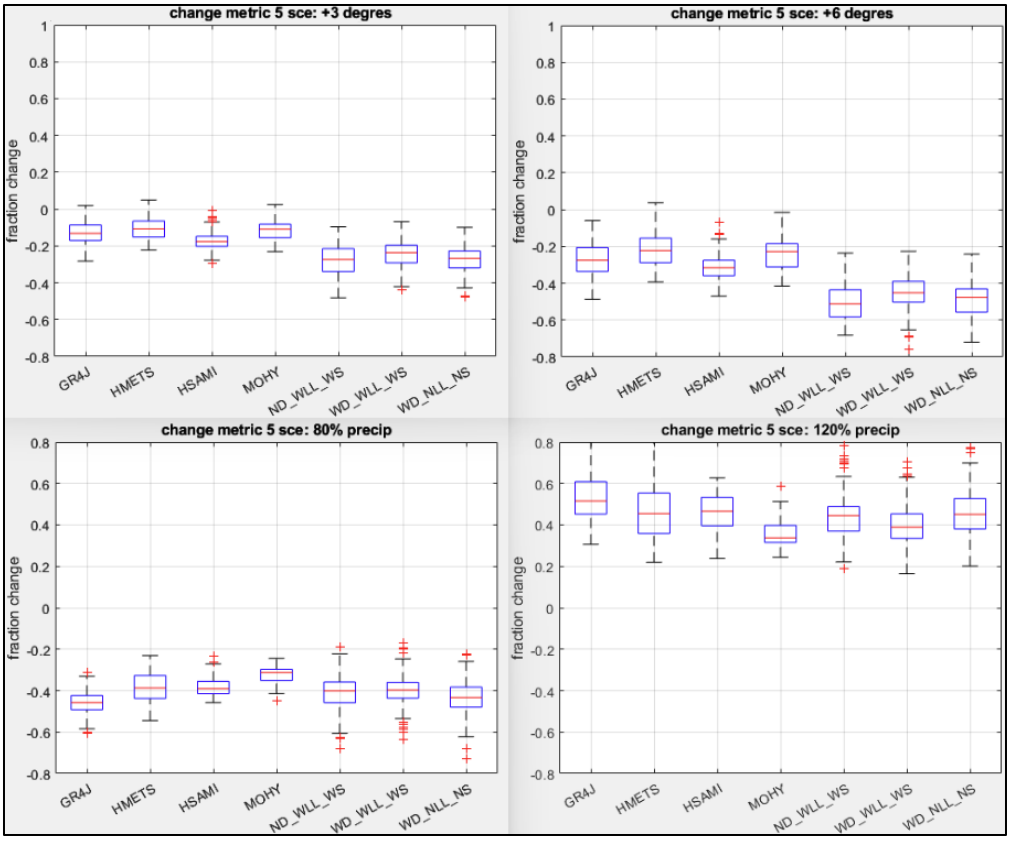

Figure RR2: This figure is the same as Figure 6 in the original paper (projected mean fall SON flows), with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors (NS)). This Figure shows that the modified LSTM keeps the same sensitivity to precipitation and temperature change. In particular, it preserves its added sensitivity to temperature compared to the process-based hydrological models.

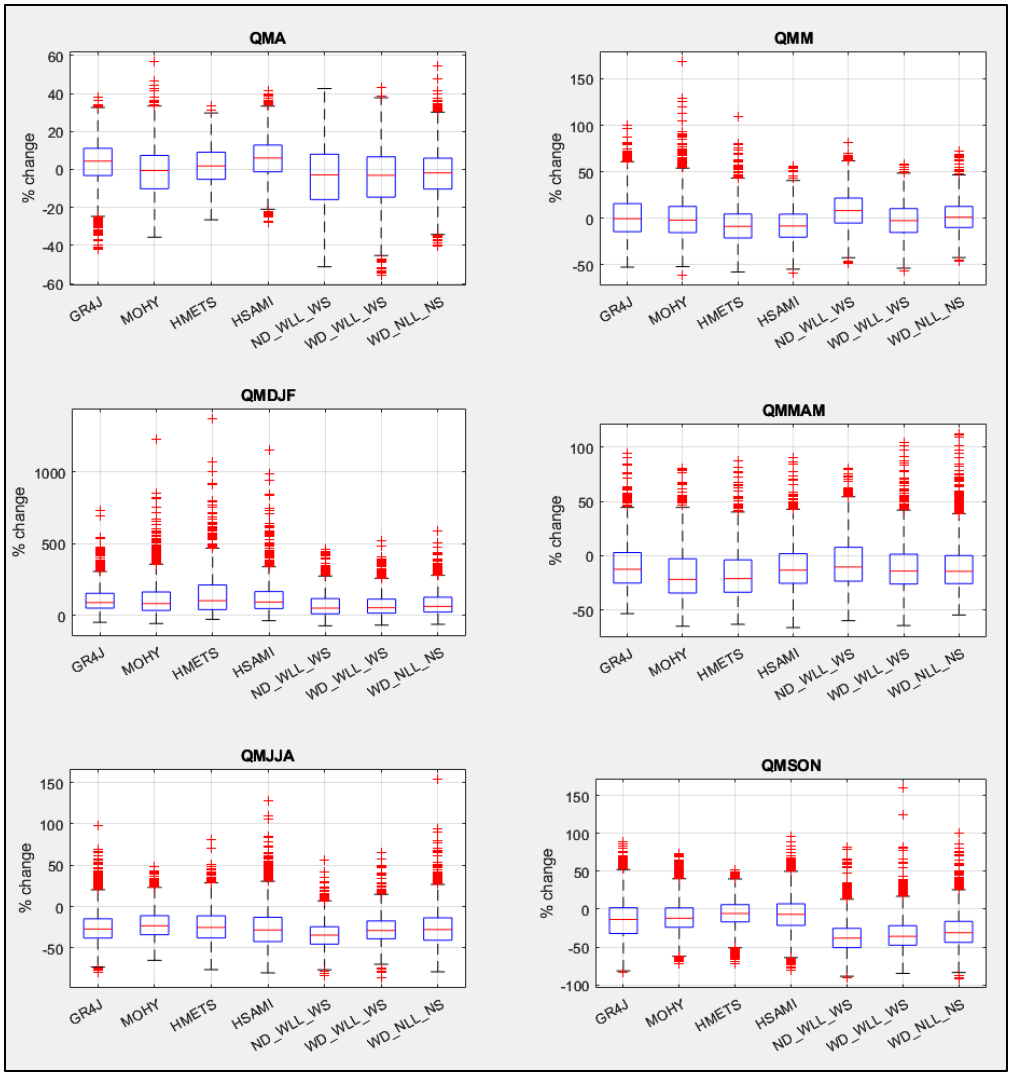

Figure RR3: This figure is the same as Figure 7 in the original paper with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors (NS)). This figure contains the aggregated results from the 22 climate models. It shows that the modified LSTM aligns itself with the original two, thus preserving the conclusions drawn in the paper.

After careful analysis, we came to the following conclusion: The addition of static variables seems to help the model to converge faster in terms of training, but the actual impacts on hydrology are caused by the meteorological forcings. The static catchment attributes (area, slope, land-use, porosity, etc.) actually condition these flows and impact their regime. The climate static variables, on the other hand, do not impact the flow modulation and only seem to help the model understand the meteorological properties better in order to converge faster. By removing them, we not only made the model more robust (fewer inputs that change over time) but also made it harder and longer to train. However, in the context of climate change studies, we believe this version of the model makes more sense, and as such, we will modify all figures and results to include this version only. The revised paper will present "clean" figures, as these were made quickly for the revision.

Reviewer #3: https://doi.org/10.5194/egusphere-2024-2133-RC3

The paper evaluates LSTM-based hydrological models and traditional hydrological models in Climate Change impact assessments. The models are assessed regarding their ability to simulate future streamflow and against their sensitivity to climatic changes. The authors conclude that LSTMs, given they are trained with sufficient data, are a viable and most likely a better alternative for climate change impact assessments.

Thank you very much for your comments and suggestions. We provide point-by-point replies to each of your comments below.

In my opinion the title does not fully express what you actually did in the study. You assessed the adequacy of both traditional models and LSTMs. Not sure, you might want to highlight that.

We agree that we assessed both lumped traditional hydrological models and LSTMs through the proposed methodology. However, our main goal was to highlight the potential effectiveness of LSTM neural networks compared to traditional hydrological models. Therefore, we consider that the current title reflects the intended take-home message for the readers.

The authors address an important topic and use a suitable dataset and methods for the evaluation. However, the main problem I have with the paper is the structure:

Nowhere in the paper do the authors refer back to the three objectives outlined at the end of the introduction. I would expect that you explain how you are going to achieve the objectives in the methods, show the respective results and discuss those, and ideally come back to the objectives in the conclusion. The paper seems unstructured in that regard and it is hard to understand which of the subchapters contributes to which objective, thus disconnecting the analysis from the objectives.

This is a valuable point also raised by Reviewers 1 and 2. Therefore, following the reviewers' comments, we will restructure the paper by combining results and discussion sections together, streamline the paper and move some figures in the supplemental materials.

Another structural problem is already evident in the abstract: Your last sentence of the abstract highlights the analysis of precipitation elasticity and catchment analogues. I like these two analyses, but I find it strange that these are shown (including figures) in the discussion only. I do not see a reason why these analyses should not be structured into methods, results, discussion.

This is a good point, and we will address it by combining the results and discussion sections, as mentioned in the previous reply.

Additional Comments:

l.130-134: It seems unclear at this point if the current study is also limited to >500km² and 30yr data. Suggest to add a brief explanation, particularly as you mention later (l.159) that you restrict to 20yr streamflow data.

This will be clarified by modifying this paragraph as follows: "*In the Arsenault et al. (2023) study, only those catchments with a drainage area exceeding 500 km^2 were included, thereby sidestepping potential issues related to scale and time lag in model regionalization efforts. Catchments also required at least 30 years of data over the 1979-2018 period to be selected in order to have sufficient data to train both the conceptual hydrological models and the deep learning implementations. These criteria were also used in this study, resulting in the selection of the same 148 catchments for the analysis.*". The 20-year limit was set only for the extra 1000 donor catchments to widen the set of available catchments for this analysis, but these were not as critical as the original 148 as they were not used for model testing. Therefore, this constraint was relaxed to 20 years for the extra set of catchments. This will be clarified in the text, especially at line 159.

l.136: I think the term 'scenario' is somewhat misleading in this context. I suggest to write something along the lines "An extra set of 1000 donor catchments was selected for an additional LSTM application."

Thank you for your suggestion. The term "*scenario*" will be replaced with "*LSTM configuration*" as suggested in the revised manuscript.

l.149: against the background of the previous sentence, it is unclear what you mean by 'common denominator for all models'

We will clarify the author's meaning by adding, "*(i.e., the one that corresponds to the intersection between both datasets)*" in the revised manuscript.

l.153-154: suggest to check Tarek et al. if that is generalizable for any catchment / region.

Thank you for your suggestion. The study presented by Tarek et al. was done for all North America, covering our entire study region. The study shows indeed a generally good performance over the whole region compared to observations, with some loss in performance in eastern U.S. compared to observations. Nonetheless, the performance remains satisfactory for the present study.

l.195: I assume the climatic descriptors are kept constant under climate change? Do you think considering a possible spatial shift of these climatic conditions under climate change could further improve the models? Perhaps a point worth discussing.

Please see the major modification presented at the beginning of this document for more information on how this was resolved and will be implemented in the revised version of the paper.

l.255: μ is missing in explanation

We will add in the revised manuscript that μ refers to the average of the simulation and observed streamflow, respectively.

l.289-290. I do not understand this part. Why is data combined from multiple catchments for the computation of the objective function? And why was the NSE used for the LSTM's objective function and not the KGE as for the classical hydrological models?

Since the model is a regional one (or continental for the one with the extra donors), it requires being fed data from various catchments. The order that these data are fed into the model for training is randomized to ensure stochasticity and to help convergence. Therefore, each batch needs to compute an objective function on a subset of the data sampled at random from hundreds of thousands of data points. To make this possible, the objective function needs to allow this. The objective function is a NSE-based metric that is slightly altered to ensure that the flow values are scaled appropriately, or else the larger catchments would weigh more than the smaller ones simply due to their size (and thus larger flows). Doing so with KGE is more complicated because the ratio of variability is highly impacted by the very small sample sizes of the batches. However, for evaluation, using KGE makes more sense as all flows are available, and the metric allows for more in-depth evaluation. For the traditional hydrological models, we would not use the same NSE-derived and scaled objective function as it is done one catchment at a time. Thus, we simply decided to use KGE directly. This means that the classical hydrological models are given a bit of an advantage over the LSTM models, but this is minor and, as the results show, the models are still comparable over the independent validation/testing period.

We propose to add a clarification to this effect in the paper.

l.305: Suggest to introduce LSTM-R and LSTM-C earlier in the manuscript and mention the respective 148 and 1000 catchments.

We will introduce LSTM-R and LSTM-C and the respective 148 and 1000 catchments configurations earlier in the manuscript as suggested.

l.322-327: I assume when implementing one test, all other variables were held constant (so no combination of TMP and PCP changes)? I suggest to mention that.

Yes, exactly. We will clearly state in the revised manuscript that no combination of temperature and precipitation changes was used in any of the tests. Changing both variables at the same time complicates the interpretation. The 22 GCMs provide projections where both variables change.

l.384-385: why no low flow metric, such as the annual minimum streamflow? Also, I suggest to add the time periods for which these metrics are calculated (I assume the hindcast and the future climate?)

In this work, we have computed and analyzed 51 different streamflow metrics. These metrics cover mean flow values, distribution quantiles, and low- and high-flow extremes. We chose to focus on mean annual and seasonal streamflows, as these are robustly simulated by process-based hydrological models and serve as key climate change indicators. Low flows present challenges for process-based models, with a high level of uncertainty in projecting low flows in a warmer climate, where hydrological model uncertainty is much larger than that of GCMs. We also included the mean of maximum annual streamflow, which, as a "moderate extreme," is well represented by hydrological models. More extreme low- and high-flow metrics would require additional validation, especially for LSTM models. We will provide a stronger justification for our choice of metrics in the revised manuscript.

l.395-397: You mention NSE and KGE at three locations now. I am confused what data and models are used for which metric. And if different metrics were used for different models, I see a significant bias here given that you compare other metrics (see 2 comments further below). I suggest to mention the KGE and NSE metrics only once in the manuscript to avoid confusion. Also, in l.403 you mention NRMSE which was not mentioned in the methods.

Considering that the objective function used to train the LSTM model incorporates all catchments at the same time, it is necessary to add a standard deviation weighting to avoid overfitting catchments with larger streamflow. No identical objective function can be fully used for both types of models. In all cases, the traditional hydrological models are favored by this methodological choice, as mentioned previously.

We also propose to add the NRMSE (as well as the NSE which was not clearly mentioned) in the method section to avoid any confusion.

l.410: You did not introduce the optimum values for each metric. I suggest to add this information to where the metrics are introduced first, or/and the optimum could be added as a line to the diagrams.

We will add the optimum values for each of the metrics in section 3.1 of the revised manuscript as suggested.

l.431: Could the difference you see in the variability ratio between conventional and LSTM models be due to the different objective functions you used (KGE vs NSE)? The tradeoff between the different performance criteria is interesting. For further discussion of the relationship between the performance criteria, you can look into Guse et al. 2020 (https://doi.org/10.1080/02626667.2020.1734204)

This is a good idea, and we will analyze the impact of the objective function on the variance ratio for the various models. We will then add these findings in the revised manuscript to better contextualize these results.

l.442-443: suggest to summarize the main results of the other metrics here. I would assume the QMM is of interest to many readers.

This is a good idea, and we will summarize the results for readers.

l.465: I do not understand why a more pronounced response to temperature changes implies providing more accurate projections. Without comparing this to observations, I think this is not defendable. Please explain.

We agree with your comment. This was definitely worded incorrectly. We will revise this paragraph accordingly.

l.500-501: I am not sure about this statement. The classical hydrological models project an increase in streamflow. That would mean precipitation is overruling temperature.

This comment was mostly based on the relative position of all 6 models. We will rephrase and expand to also discuss the increase/decrease of streamflow as shown on the Y-axis.

l.509: what do you mean by "near the surface"?

“Near the surface” refers to a height of 2 meters above the surface. We will clarify this as “near the surface (2 meters height)” in the revised manuscript.

l.532: Why did you evaluate this section only for QMA?

This paper is already too long (as mentioned previously by the reviewers). We have decided to select this variable to conduct this analysis. However, all data are available in the data repository and readers could evaluate the metrics of their choosing to determine if the model is appropriate for their needs. Internally, we analyzed 51 metrics (The 51 from Table 2 in Arsenault et al. 2020) and results are fairly consistent. We decided to only show QMA for space considerations, given that it is one of the main metrics used to evaluate impacts of climate change on hydrology. Had we gotten strange and unexplainable results for QMA, it would have been a clear sign that the LSTMs were perhaps not suited for climate change studies. Doing more metrics would be desirable but we must balance the amount of data and results presented with the overall message, we wish to expose.

Arsenault, R., Brissette, F., Chen, J., Guo, Q. and Dallaire, G., 2020. NAC2H: The North American climate change and hydroclimatology data set. Water Resources Research, 56(8), p.e2020WR027097.

l.561: I do not understand what you mean with "to prevent contamination"

This refers to “data leaking”, meaning that any data from the testing or validation sets cannot be used to help build the scalers or be used in the process of defining the model weights. Doing so would let the model include that data during training and thus make it more robust in testing, meaning that the model would be “cheating” by accessing “future” data it should know nothing about. This will be clarified in the paper, along with appropriate references.

l.576-577: Why do you write here that the LSTMs only use the climatic data? I suggest to add "..rain, and snow besides the catchment descriptors represent a fraction...".

Good point, we will add the descriptors to this list in the revised paper.

l.582-585: There are many hydrological models around that can make use of additional physical and temporal data. While I understand the advantage of complex LSTMs that can ingest all this data, I think you should not overstress this 'advantage' only because you chose traditional lumped hydrological models only, that cannot use additional data.

Actually, this section of the text points to a limitation of the LSTM models, in that while they can use a large variety of input data, that data also needs to be available for any simulation, forecast, or projection. This is a limiting factor that other models do not have (i.e. you could calibrate a hydrological model using remote sensing data to tweak parameters and then feed climate data to the model with no issue, but this is impossible with the LSTM-based models used here). Therefore, we propose to leave the text as is. As to the comment related to the ingestion of data, we will clarify elsewhere in the text (see one of the comments below) that other hydrological models can use more sources of hydrometeorological data than those used in this study.

l.625-629: Isn't it also reassuring that the different classical hydrological models performed similar? It could also mean that the projections can be considered robust.

This is a good point, and will be added to the final text.

l.634-635: Well, this is clearly beyond scope for your study, but you could at least discuss that there are studies that used historical streamflow change to evaluate models (see for instance Eyring et al. 2019, https://doi.org/10.1038/s41558-018-0355-yor; Kiesel et al 2020, https://doi.org/10.1007/s10584-020-02854-8 who both propose out-of-sample evaluations). And also, Krysanova et al (2018, already cited in your manuscript) provide a 5-step validation procedure that allows an assessment of how well models are suitable for climate change impact assessments. This might also be worth discussing.

This is a good idea in theory, but the problem is that we need to find a large set of catchments that have similar climate change signatures, and thus require long enough time periods. We can add a sentence in the paper to that effect.

l.646-650: I do not completely agree with this statement based on your study, considering that other traditional models exist that utilize additional data sources as well. I suggest to add to l.648: ... a theoretical advantage over the four traditional hydrological models used in this study.

We agree to a certain extent, but we still believe the theoretical advantage holds. The advantage comes not from the fact that more data variables can be ingested (which other models can do, such as humidity, radiation, wind and other such variables). The advantage comes from the fact that the LSTM model ingests data from over 1000 diverse catchments and can learn how different climates lead to different streamflow values. No other model can do this, unless we count large-scale regional models that are calibrated on a large and diverse region. But even then, the parameters are fixed in time and the changed meteorological data is still fed to a deterministic model that cannot modulate flows according to nonlinear processes between meteorology and hydrology. This is where the LSTM strives. We therefore propose to keep this in the text but to better contextualize it, as done in this response.

l.660: Why is this a disadvantage of conceptual hydrological models? You also fixed the LSTM parameters after training?

Yes, but it is adaptive and has many internal weights able to process the various seasonalities within. This would be more similar to a hydrological model whose parameters are calibrated to different values for every day, for example. This will be clarified in the text.

l.685-694: in l.690, could you mention percentages instead of "a few" or "most"? Also, you could simply calculate the elasticity ratios yourself based on your the historical data?

We can calculate the percentages indeed, and will do so in the revised version. However, calculating the elasticity is a difficult and time-consuming prospect given the detailed analysis required to determine the impact of a change of precipitation on the change in streamflow when many confounding variables are at play (antecedent soil moisture, snow, vegetation, etc.). We will therefore only rely on the existing literature on these values.

l.773-777: I suggest to also mention that you did not include more physically-based hydrological models that can utilize additional data which could react differently to the future climate forcings.

We will mention that more physically-based hydrological models were not included in this study due to the extensive work and computing power required to calibrate such models on 148 catchments. Nonetheless, we will highlight that different results can be obtained, as more physically-based hydrological models can utilize and process additional data such as elevation, land types, land uses, and more.

Minor comments:

l.31 consider "studies evaluate" to avoid repeating "assess"

We will replace "studies evaluate" to "assess" as suggested in the revised manuscript.

l.182 comprehnsive -> comprehensive

We will correct this typo in the revised manuscript.

l.280 ...performance gains in other... ?

We will correct this typo in the revised manuscript

l.370 you express pcp change in %, therefore *100 should be added to the calculation example.

We will add “ *100 ” as suggested.

l.511: wet vs dry models

We will correct this typo in the revised manuscript

l.576: mpdels -> models

We will correct this typo in the revised manuscript

l.671: electricity -> elasticity

We will correct this typo in the revised manuscript

l.720: unit should be m3 s-1 km-2 ?

This is correct; we will correct this typo in the revised manuscript.

l.730 and supplementary material Figures: suggest to add or mention the RMSE unit of normalized streamflow.

We will add this clarification in the revised manuscript.