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.

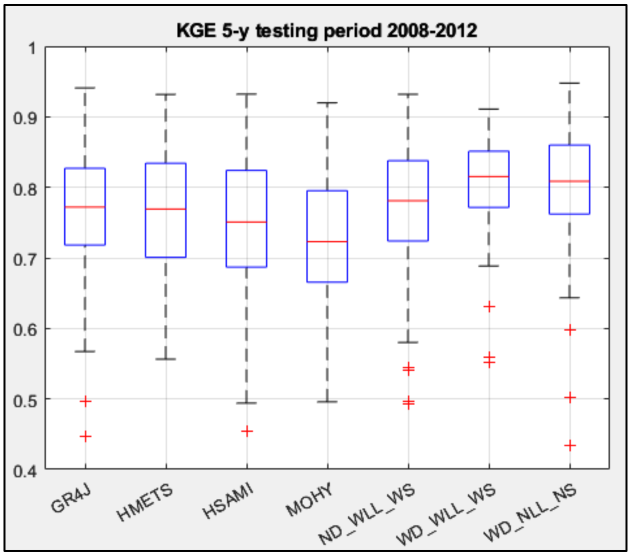

Figure RR1: This figure is the same as Figure 3a in the original paper (model performance over the testing period), with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate 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.

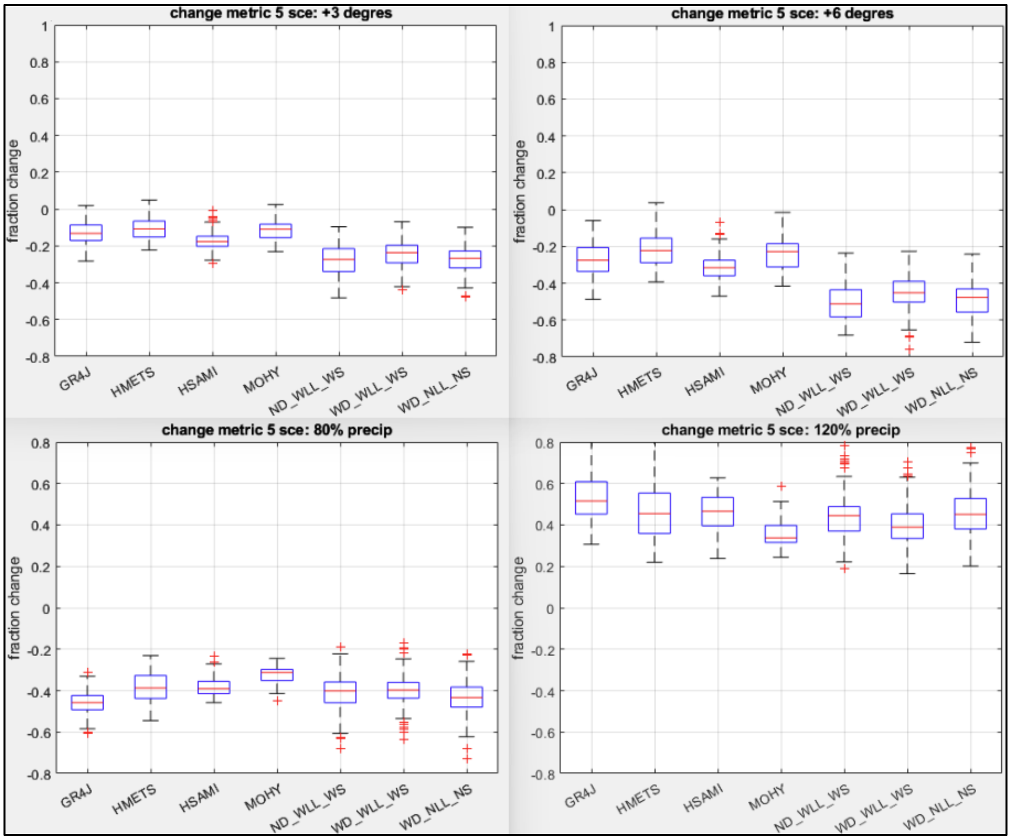

Figure RR2: This figure is the same as Figure 6 in the original paper (projected mean fall SON flows), with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors (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.

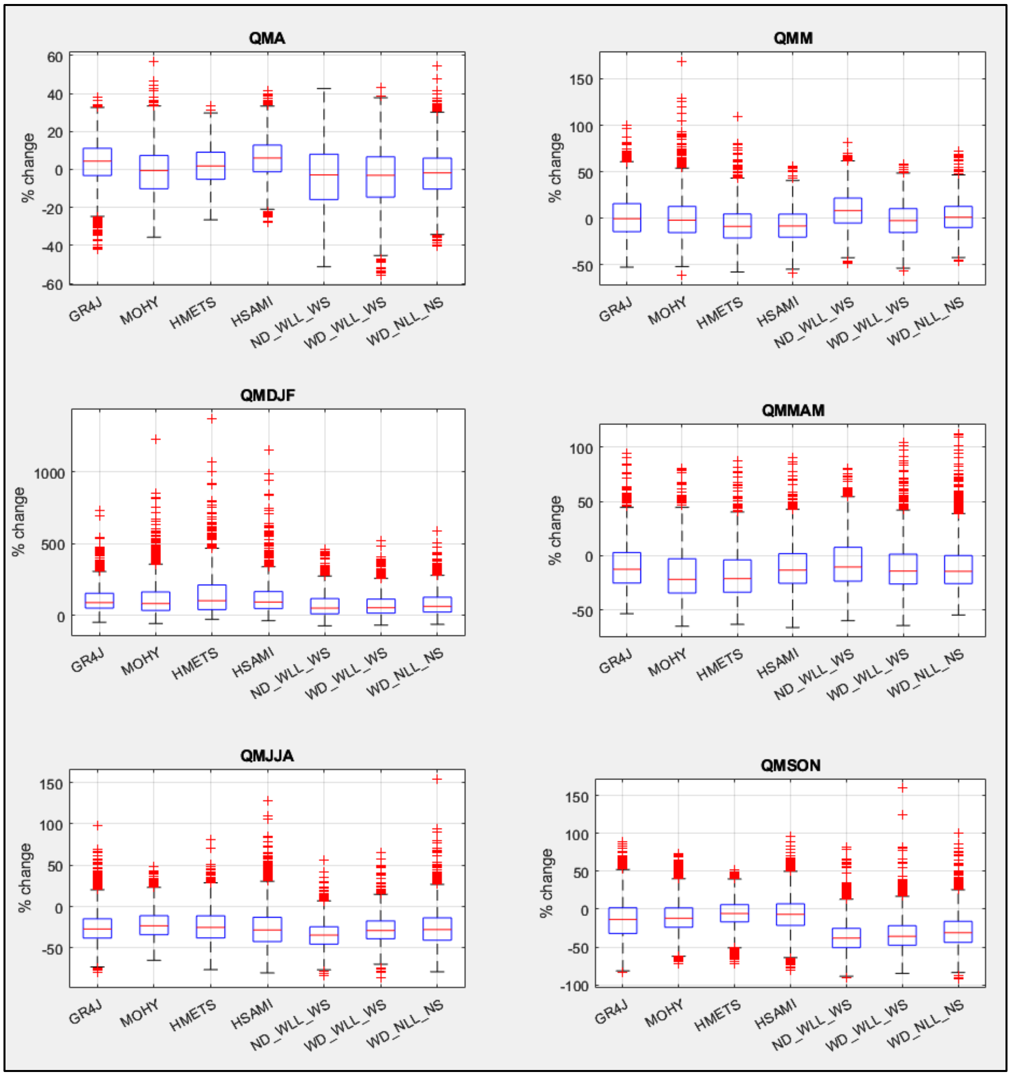

Figure RR3: This figure is the same as Figure 7 in the original paper with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors (NS)). This figure contains the aggregated 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 #1: https://doi.org/10.5194/egusphere-2024-2133-RC1

The study is sound and methodology is robust. The experimental design is clever, dataset curation (especially training/validation data setup) is suitable to answer the research questions and the rationality in choosing climate change scenarios and climate models is also reasonable. Language, presentation quality and significance are good, and results/conclusion are also sound and relevant. I recommend publication, however, after minor revisions.

Thank you very much for your positive comments and suggestions. Please refer to the point-by-point responses to your comments below.

1) you can merely infer from “between the lines” in the method section, that for the hydrological model runs evaluating climate change, the hydro-models are actually also trained on the hindcasts from the same GCMs that are used for the projections (as it should be done, so the model is trained from the same population that it is tested on). If it was done like this, then maybe just add a small paragraph or sentence to e.g. 2.4.2.3 to make this crystal clear. If I am mistaking and it is not done like this, we have a bigger problem and you need to redo the results and train on the hindcasts.

The training (or calibration) for the traditional hydrological models was performed on all available observations from the ERA5 reanalysis (i.e., precipitation and temperature between 1981 and 2007) using observed streamflow for the same period as the target variable. A small validation period between 2008 and 2012 was kept to have a fair comparison with the testing. Since hydrological models cannot be calibrated using climate model data (as there is no correlation between climate model data and observations on a day-to-day scale), the models are calibrated on the ERA5 reanalysis dataset (reanalysis of meteorological datasets, i.e. observations). Then, the climate model data is bias-corrected using the ERA5 data as reference, making it possible to run the model on the reference period GCM data. Finally, the bias-correction postprocessor is applied to the future climate model data and the resulting data is run in the hydrological model. This is a typical hydro climate processing chain, as used in the IPCC reports and most hydroclimatological sciences papers. Thus, it is not possible to calibrate directly on the climate model data directly, but the bias-correction step allows us to simulate the flow from the climate model in reference and future periods. This is also why we cannot evaluate day-to-day flows, but only long-term statistics, no matter the model (traditional or LSTM-based) used. This will be made clearer in the text.

2) The model seems overly complicated, given the fact that up to recently, the state of the art model only comprised one single LSTM layer (https://doi.org/10.5194/hess-25-2685-2021). It would be useful to benchmark (i.e. plot/compare) against the above-mentioned studies’ performance to see if it actually gives an advantage in performance. The structure obviously doesn’t hurt performance, as can be seen from the high scores shown in the manuscript, but does it help? As a comment without need for action: probably 90% of your network is inactive, but the remaining 10% is why it still works well.

It is true that a simpler model can give very good results as well, but trial and error have shown that adding complexity added more and more performance, although towards the end it was more marginal and thus, we stopped at this level. Also, the sheer amount of data points for training encourages the use of a larger model. We also implemented this architecture to act a bit like a

ResNet, using interconnected layers with residuals for a few layers, which helped robustness. The inconvenience is that it does add training time and demands more computing resources. To provide more concrete context, we did another run with a model similar to that of Kratzert et al. (2019), using a single LSTM layer with 256 cells, concatenated with statics and fed to a dense layer (rather than linear). We used the same 0.4 dropout and 0.001 learning rate. We also changed the 270-day look back window to a 365 day one to better account for the snow processes in Canada. The results show that the median testing KGE over all catchments was 0.71 (vs 0.77 for our model with no-donor catchments, and 0.82 with donor catchments). During the development of the models in the paper, we did this process with increasingly complex models and ended up with those used in the paper. We built upon this experience through various tests to provide this present model, even though it is indeed quite large and complex. But this is still a small model compared to models such as AlexNet and LLMs, and we believe that the addition of more data will always require more complex models to make use of it as best as possible. The figure below shows raw data for the various runs as in those in the "major modification" section at the top of this document, with the simple model runs in the last column for comparison.

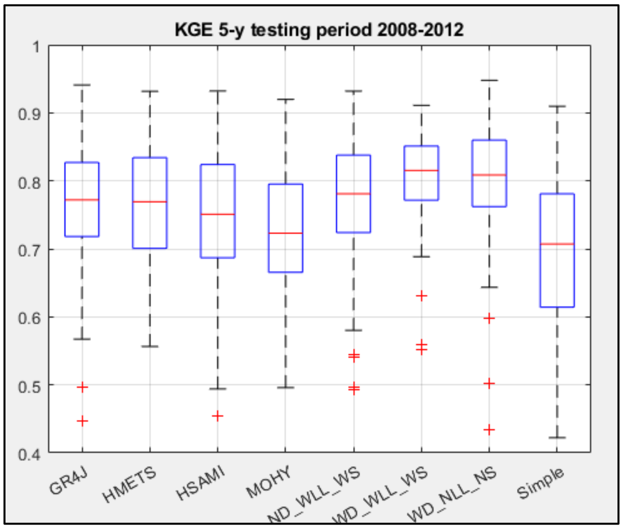

Figure RR4: This figure is the same as Figure 3a in the original paper, with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors (NS)), and the eighth boxplot adding the simple LSTM model described above. This model was only trained once and it is quite probable that performing multiple trainings would improve results to a certain extent, but not to the point of competing with the larger models.

3) Chapter 4.4.1.: results and literature stand next to each other disconnectedly; no true conclusion is drawn in the chapter. Insert it.

Thank you for this suggestion. Based on the recommendations from all reviewers, the manuscript will be restructured. This section will be revised, and a clear conclusion will be provided in the revised manuscript.

4) Results and discussion are formally separated from each other, but then there is another bunch of analysis and four (!) figures in the discussion section, which renders the separation of results and discussion irrational. Either clearly separate results from discussion, or – much preferred, because much easier to follow/ understand in general – make a single "results and discussion" section and discuss noteworthy point right next to the figures.

Agreed. In the revised manuscript, we will combine the results and discussion sections into a single section, ensuring clear discussions aligned with related literature.

5) the conclusion is a stump and contains mostly commonplace statements. Revise to revolve around actual key conclusions from your results.

We will rework the entire conclusion section to better reflect the key conclusions that are presented throughout the different results.

6) typos:

Line 576 "mpdels" à models

The typo "mpdels" will be corrected to "models".

Line 671 "streamflow electricity" à elasticity

The typo "electricity" will be corrected to "elasticity", and it will be revised throughout the manuscript. The auto-correct was quite original on this one.