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.

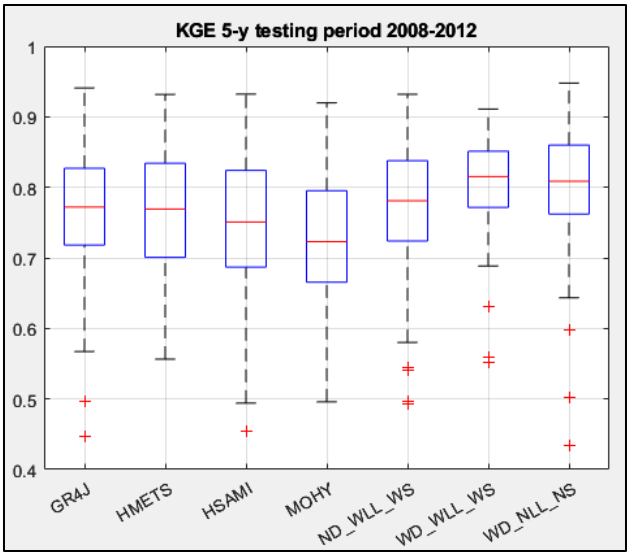

Figure RR1: This figure is the same as Figure 3a in the original paper (model performance over the testing period), with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate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

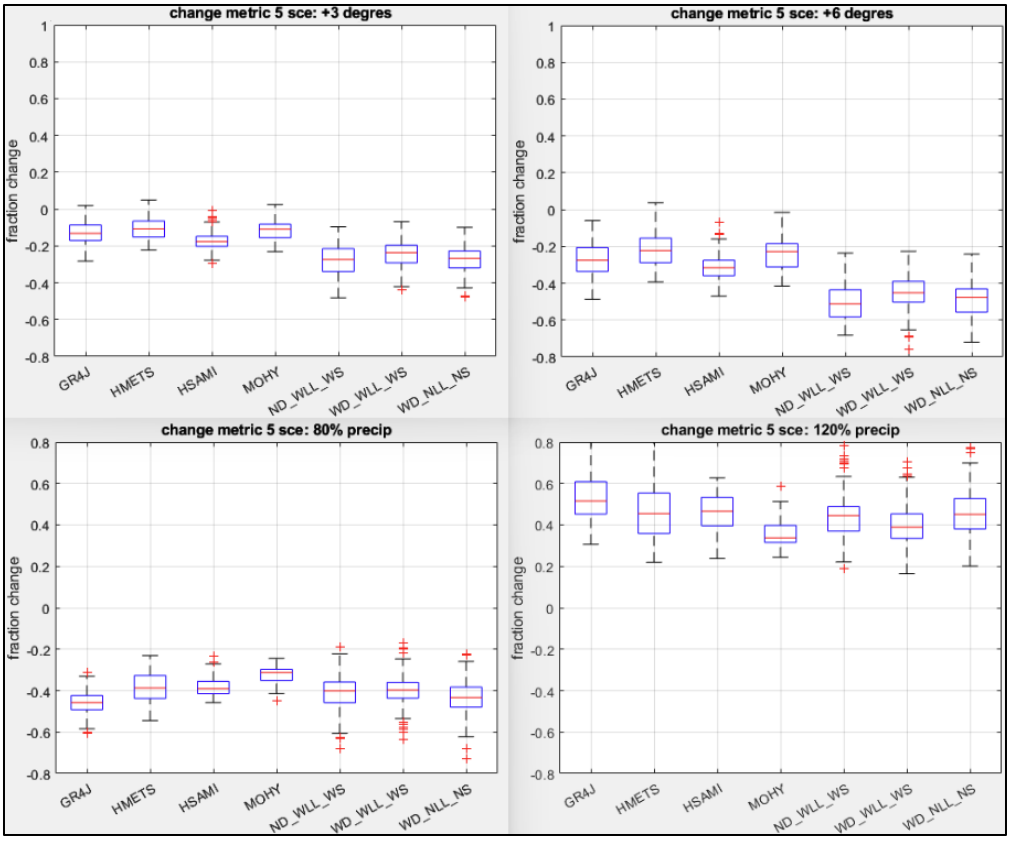

Figure RR2: This figure is the same as Figure 6 in the original paper (projected mean fall SON flows), with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors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

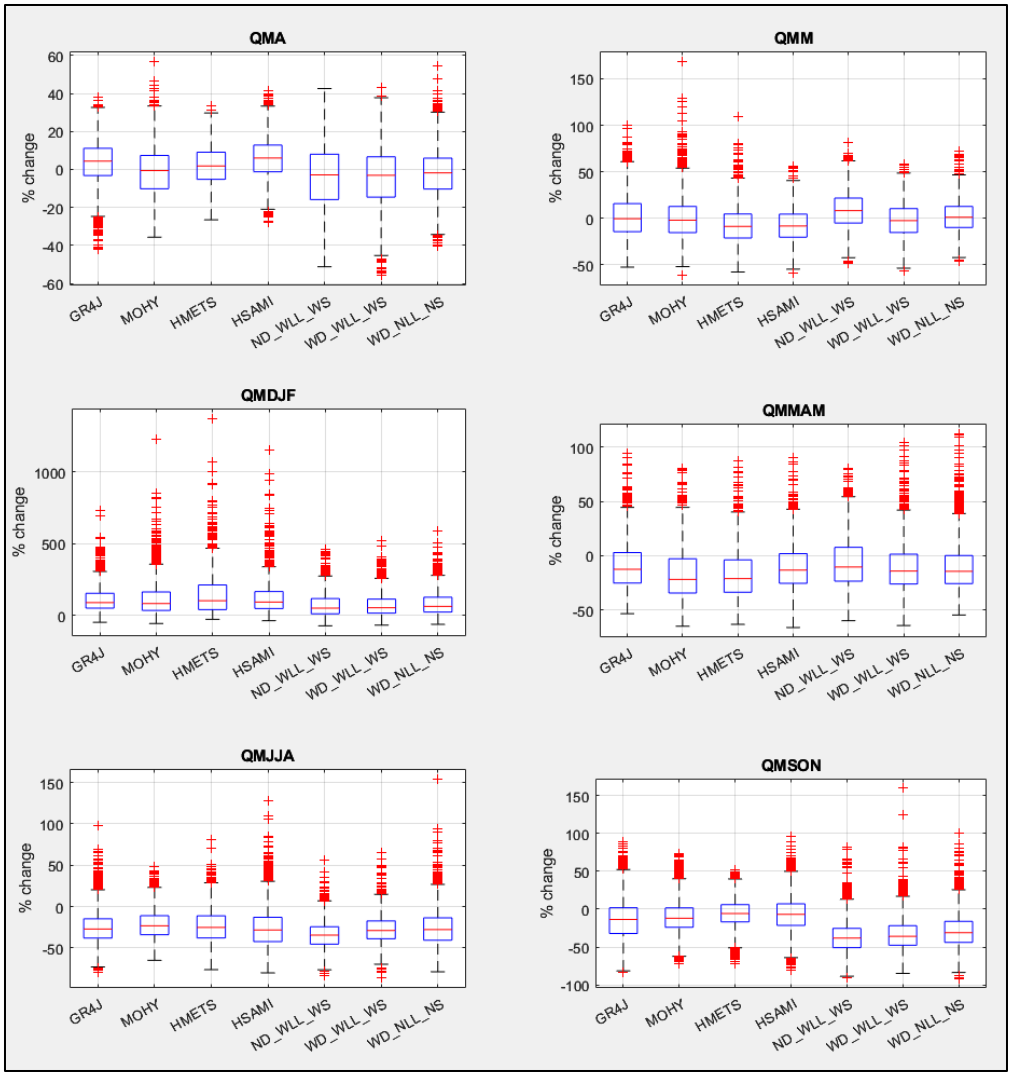

Figure RR3: This figure is the same as Figure 7 in the original paper with the exception of the seventh boxplot, which corresponds to the modified continental model (with donor catchments (WD), excluding latitude and longitude (NLL), and no static climate descriptors (NS)). This figure contains the aggregated 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 #2: https://doi.org/10.5194/egusphere-2024-2133-RC2

The study aimed to compare process-based models (PB) and machine learning models (ML) when used for scenario analysis. Specifically, they assessed three common PBs and two Long-Short Term Memory (LSTM) configurations. The researchers focused on examining the sensitivity of streamflow to different forcing perturbations. Their findings indicated that LSTM trained at a continental scale is a more reliable model due to its training on a wider range of variability. However, the study suggested that in most cases, the sensitivity between PB and ML is similar, with a few exceptions.

Thank you very much for your comments and suggestions to improve the manuscript. Below, we provide point-by-point responses to all concerns, comments and suggestions.

Major comments:

1. The use of latitude, longitude, and climatic static attributes in the LSTM models restricts the model to learn local information rather than capturing the spatial variability in the dataset. This could affect the model's ability to accurately analyze sensitivity under different climatic conditions. I suggest running the model without latitude and longitude and using the minimum number of climatic attributes while changing them according to the sensitivity being analyzed (ex. 20% increase in precipitation means a 20% of the mean annual precipitation).

This is a good observation. We were curious about this when reading your comments and have retrained the model without using the latitude and longitude in the variables. Please see the "major modifications" section at the top of this document for an in-depth response to this point and how the paper will be modified to reflect this.

2. There is an excessive number of figures and sections presented. Some results are repetitive, and certain sections may not add significant information due to high uncertainty (ex. 4.4.2). I recommend including only the figures and sections that directly support the main conclusion of the report while considering moving additional figures and analyses to the appendix or supplemental information.

Based on comments from Reviewers 1 and 3, we will rework and combine the results and discussion sections and move some of the figures into the supplemental information.

Minor comments:

Line 14. The comment about traditional hydrological models relying on historical climate data is misleading. ML models rely on historical data too.

Thank you for pointing this out. The sentence will be reworded in the revised manuscript as follows: "*Traditional hydrological models, which rely on simplified process parameterizations with a limited number of parameters, are scrutinized for their capability to accurately predict future hydrological streamflow in scenarios of significant warming.*"

Line 85. The definition of what is long and short is subjective. In fact, how long the memory in an LSTM model remains an open question.

This is a good point. The sentence will be reworded to "*Their unique architecture enables them to learn and remember over longer sequences of data compared to standard RNNs, making them highly effective for predictions of time series*" in the revised manuscript.

Line 109. The problem is not the new climatic scenario, the problem is how to define when LSTM is in extrapolation mode. Because we can have an extreme condition in one catchment but the same would be normal in another, so the model can infer the relationship. In that case, the model is still in interpolation mode.

Agreed. We will reformulate this in the revised paper. The advantage of using all of these extra donor catchments is to widen the training dataset characteristics such that the climate model-driven simulations are mostly in interpolation mode rather than extrapolation.

Line 118. You should have an introduction between sections and subsections. Probably a title as a dataset or data would match better with what you have.

We will add a short introduction between sections and subsections. However, we do not understand the second part of your comment. The section is titled "Study area and Data", as we describe the study area in 2.1 and the data in 2.2.

Line 132. More data is better, but how did you define this number?

It was not possible to take more than approximately 30 years of data without having to drop a significant number of watersheds. We will mention this in the revised manuscript as follows: "*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. The basin selection criterion was set to a minimum of 30 years of data to ensure not only a sufficient data length but also a robust sample of basins for performance assessment.*"

Line 161 – 166. You do not need that paragraph. You can use a reference to explain this in more detail.

We will make the paragraph significantly shorter and add the following reference: *Wickham, H., & Stryjewski, L. (2011). 40 years of boxplots. Am. Statistician, 2011.*

Figure 1. It needs a legend.

We will add a legend explaining the difference between study catchments and donor catchments (large circles with black outline vs small circles with gray outlines for the a and b panels, and green and orange boxes for the c and d panels respectively).

Line 193. This attribute shouldn't be used because you are anchoring the dynamic to a location which is exactly what you are trying to avoid. That can have serious effects on the sensitivity of your model. AND Line 195. How are you disentangling the correlation between catchment attributes and the meteorological forcing?

These two comments are related to the major comment related to the latitude and longitude statics as well as the climate-based static descriptors. Please see the major comment at the beginning of this document for a detailed response to this point.

Line 317. How did you apply that modification? only testing or in the entire period?

This was done over the entire period. This will be specified in the revised manuscript.

Line 404. This is not the reason for not presenting the validation period. All your results should be in a period that was never used during the training and validation.

This is correct. We proposed the following reformulation for this sentence: "Note that the validation period for the LSTM-based models is not shown as the validation data are contaminated by the training data, and thus, should not be investigated."

Figure 3. I recommend using CDF plots. This format has been widely used in streamflow models. I suggest adding a line where the best value is found.

We will test using CDFs, and if the graphs are still clear enough to interpret (i.e. not stacked too heavily one atop the other), we will consider using CDFs instead of boxplots.

Line 453 – 454. This could be a consequence of using latitude and longitude as input. The LSTM model is fixed to the location so it is less sensitive to precipitation because part of the precipitation correlation is shared with those attributes or with the climatic ones.

This is a good comment and will be resolved in the revised version. Please see the main major comment at the beginning of this response document for information on how this was handled.

Line 461 – 462. Something similar could be happening here with elevation. Temperature and elevation are very correlated. An interesting experiment would be increasing temperature and decreasing elevation by using an altitudinal gradient. Should the sensitivity increase or decrease?

See Supplemental material Figure S2 b) to see the distribution of elevation across all watersheds. All watersheds are within a 500-meter range more or less, thus the impact of elevation on temperature should be marginal, much less so than the impacts of climate change.

Figure 4. I would decrease the y-axis range. Probably [-60,20] for a and b. [-50,50] for c and d.

Agreed, the y-axis ranges in Figure 4 will be modified as suggested to improve the display of the results.

Figures 5 and 6. Move it to the appendix or supplement information.

We will modify the general structure of the manuscript based on all reviewers comments. Please refer to our reply to your main comment 2 for further details.

Line 479 – 480. Explain why it is clear to you.

This is based on Figure 2 which shows mean annual temperature and precipitation changes for each of the 22 GCM models. Median temperature increases range from +4 to +8.5 °C and from +4 % to +20 % for precipitation. Based on this, the +6 °C and +20 % sensitivity scenarios are a lot more realistic than the +3 °C and -20 % ones. We will rephrase and clarify this statement.

Figure 7. In many cases the differences look not significant, you should do hypothesis testing to check the level of significance. Moreover, you should mention something about the differences in the variability between some models. All the sub-figures must have the same y-axis range to do a fair comparison. I do not think you need all the sub-figures; you should show here only the most significant.

Good point for the figures. We will rearrange Figure 7 so that the range (maxY - minY) is the same for all 6 panels. We will also double-check the other similar Figures for the same issue. We performed statistical tests prior to submitting the preprint. More specifically, the nonparametric rank sum test for equality of median was used. The test showed that the LSTM results were in almost all cases statistically significantly different from the process-based models. We will expand on this in the revised version.

Line 537. No result can be considered a result, but in that case, you could move the entire section to supplemental information.

In this case, we prefer keeping this analysis here as it also supports the fact that the LSTM-based models and the classical hydrological models are providing similar responses to the same inputs. The fact that the results show no difference is actually very positive for the analysis. We therefore propose to leave this as-is, albeit in the new “results + discussion” section as discussed in another set of comments.

Line 558. But exactly for this reason you have the third period. Are you trying to say that this period is not representative enough? If this is the case, you should check that.

This sentence aims to convey the fact that in typical hydrological modelling, we have 2 phases: calibration and validation. But in Deep Learning, there are 3 phases required: Training (equivalent to calibration), validation, and testing (equivalent to validation for hydrological models). There is no equivalent of the “DL-Validation” in the hydrological modelling world. This data is used as a stopping criteria only, and is not used to define the scaler parameters nor is it used as training data to determine the model weights. It is used to stop the training process once the best validation is achieved, to prevent overfitting. Without this, models would converge to near perfection in training but would be completely useless in testing due to the massive overfitting. This will be clarified in the text.

Line 559 – 561. This is not part of the discussion; this is just part of the methodology.

We will remove this statement from the discussion and incorporate it into the LSTM model description in the revised manuscript.

Line 563 – 564. This is a strong statement without support. Remember that the number of parameters is not comparable between PB and ML models.

The vastly different number of parameters is what is referred to in the term "significantly greater number of degrees of freedom" in the original text. We proposed to rephrase the sentence as follows: "*From Figure S3, it is evident that LSTM training period performance surpasses that of conceptual models.*" and remove the notion of degrees of freedom.

Line 568 – 569. This sentence is exactly the opposite you said in line 558. You should put everything in one paragraph to tell a more consistent story.

Yes, the sentence is opposite to the previous one because the first one refers to the training period (i.e. calibration) where the performance can be essentially perfect with a large enough model. We state that the training performance should not be analyzed deeply because it is essentially unusable to verify model performance, contrarily to traditional hydrological models. In lines 568-569, we are talking about the independent testing performance, which can be used to assess model performance and where no signs of overfitting are seen. This is due to the precautions taken to prevent overfitting, using the validation period (the DL-validation, not the validation of the hydrological model) as a stopping criteria for training. We will rewrite this section to better clarify the intent.

Line 571 – 572. 1000 sounds like a large number of catchments so I disagree.

This sentence simply states that we saw an important improvement in model results when going from 148 to 1148 (+1000) catchments, and thus this confirms that adding a large number of catchments improves performance. We do not understand what this comment is referring to that could be disagreed with.

Line 575 – 576. Are you talking about distributed models? If this is the case, this would not be a fair comparison. If you want to add more degrees of freedom to PB, for example, you could combine the different sources of precipitation. Moreover, remember that the parameters in a PB encode the local descriptors (local characteristics) that the model needs.

It is true that we could use more sources of data to generate ensembles for example, which would allow averaging using a model averaging system that could indeed add more degrees of freedom. However here we meant that a single model simulation of a PB cannot make use of multiple precipitation datasets at once unless that model was specifically built to handle multiple precipitation inputs at once. The structure is fixed and cannot be increased. LSTM models afford this capability. And yes, the PB models encode the information directly into the model structure, but this also means that these are fixed in space and time, whereas an LSTM model can be applied

to an ungauged location by using the descriptors, and can add more if needed. Ex. some PBs do not use the soil porosity or hydraulic conductivity explicitly, whereas an LSTM model could add one with no issue as long as the data is available.

Line 580. I disagree. More variables increase performance but decrease interpretability. How could you do the same sensitivity analysis with 10 hydroclimatic variables highly correlated?

We agree with this, the interpretability would be much lower with many highly correlated variables. But this is not what we are advocating. We are expressing that adding mode data could (or most likely would, based on other papers) increase model performance, and could make the model more likely to respond accurately to climate change signals due to the extra information. A model using only precipitation would definitely see improvements in adding temperature. We are simply extending this reasoning to higher orders.

Line 590 – 591. You do not need to be sorry for finding that those models are the worst, this is just part of the results. Delete the sentence.

This sentence will be deleted.

Line 605 – 610. I disagree. The situation is exactly the opposite. You are not counting all the sensibilities; a temperature change can change the precipitation too, meaning that the final sensitivity of streamflow can be higher or lower. So, your analysis is a simplified sensitivity analysis which does not mean you are more accurate.

We certainly did not mean to imply that varying one-parameter at a time is better despite our poor choice of words. Of course, a temperature increase definitely affects precipitation in a myriad of different ways (mean, variability, extremes). A sensitivity analysis typically varies one parameter at a time, and we still believe that this is a useful approach (as stated in the paper) to better understand the potential difference in sensitivity between the classic hydrological models and the LSTM class of models.

We will rephrase the original sentence: "*By independently altering each variable—precipitation and temperature—we were able to more accurately evaluate the impact of each change…*"
with the following:

"By independently altering each variable—precipitation and temperature—we were able to quantify the impact of each change, ..."

Line 622. What family is that?

We will rephrase to: "*The four traditional hydrological models share similar structures, all being lumped conceptual models*".

Line 625. This is a contra argument about sensitivity coming from structure.

We are not entirely sure what this comment applies to. This sentence aims to show that the sensitivity was similar across the 6 models, including 4 conceptual ones with similar structures, and 2 LSTM-based models with extremely different structures. Thus, we do not think the sensitivity comes from the structure, but from the meteorological inputs.

Section 4.4. You focus just on which is the best. You must analyze the benefit of the ensemble of models (multi-representation approach). The concept of the best model does not exist.

While a multi-model approach can be used to provide a better evaluation of the overall uncertainty, our results support the conclusion that when using a single hydrological model, the LSTM-based model is likely to provide a better evaluation of the impact of climate change on the hydrological process. However, we agree that adding a section in the text highlighting the potential of a multi-model approach to evaluate the overall uncertainty is important. Also, we will propose to reformulate the section heading as follows: "Which type of hydrological models should we trust more for climate change impact studies?"

Line 653. That is not true. There is a lot of research on interpretability. We do not have yet the same level of interpretability as PB, but this does not mean we are not going to get it in the future.

This is a correct assessment. For some simple deep learning models there is starting to be some progress in interpretability, and the statement will be changed to clarify that this is not realistic at the moment but that with time this might be something that becomes possible.

Figure 9. I would prefer a table or a CDF figure showing the distribution. It is very hard to compare models within each group.

We will test using CDFs, and if the graphs are still clear enough to interpret (i.e. not stacked too heavily one atop the other), we will consider using CDFs instead of boxplots.

Line 701 – 703. I agree about the catchment attributes used however you considered only input similarity. This is not enough to define dynamic similarity, at least you should consider adding similarity in the streamflow signature.

We understand the reviewer comment. However, this is not possible for this assessment. The reason is that we are trying to see if the simulated flows under a future climate are well represented by the LSTM model. To do so, we need to find indicators that are independent of streamflow, as using a streamflow signature metric to find an analogue would ensure that the catchments found had similar hydrographs. This would defeat the purpose of ensuring that the analogues in terms of climate are indeed analogues in terms of hydrology.

Line 705. If this is the case, you should use a uniform weight distribution. Add information supporting your decision to use something different than uniform.

All of our catchments are located within Northeastern North America. A simple look at North American maps of mean annual precipitation and temperature clearly shows that mean annual precipitation changes little from North to South, whereas there is a huge mean annual temperature gradient therefore justifying our choice of weighing temperature more. As described in the paper, we tried different weighting schemes and picked the one which seems to result in the best analogues based on a limited number of catchments. Our evaluation was qualitative. Ultimately, the results were similar (with the continental LSTM model performing slightly better), even with a choice of analogues that did not seem optimal. We have to rerun the analogues with the newly developed LSTM models and will consider using equal weightings in the paper to make it simpler.

Figure 10. Given the huge difference between the analogues and the models, it is impossible to say that one model is better than the other. Moreover, if I suppose that the catchments presented are the best ones, it is impossible to infer more from this type of comparison.

Analogue analyses are, by definition, uncertain. We provide this analysis with the stated aim of finding adequate analogues (10) for each case to encompass the uncertainty of selecting the "best" analogue catchments. Looking at the hydrographs, it is clear that the models are all representing the hydrological cycle correctly, albeit with some key differences in select aspects of the hydrographs. We can definitely see that LSTM models perform better in some cases and classical models perform better in others, which is expected. However, it is clear that the LSTM based models are not consistently worse in any part of the hydrograph, but present a balanced comparison with the hydrological models. This is reassuring, in that the objective of the paper is to investigate if LSTM-based models can be used for climate change impact studies, and this analysis seems to confirm that they are at least not worse than classical hydrological models.

Line 759 – 764. Exactly for this reason I would drop the entire section. The results from this section are not different than before but with a huge uncertainty.

We respectfully disagree about dropping the entire section. While there are indeed some limitations to this analysis, it still provides valuable insight on the ability of the different types of hydrological models to simulate streamflow under a changing climate. It is one of the only insightful methods to assess the ability of models to represent climate change and to validate those results. Furthermore, these results are supported by the other results from this paper and thus contribute to the overall objectives of the paper. There is definitely some uncertainty, but this does not warrant, in our opinion, removing the analysis and its findings.