Response to Reviewers for: ‘On the need for physical constraints in deep leaning rainfall-runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration’

Sungwook Wi, Scott Steinschneider
Corresponding author: Sungwook Wi, sw2275@cornell.edu

Key

Black font:	Reviewer comments
Blue font**:**	Author responses
Orange font:	Excerpts from the manuscript, with key changes underlined and **bolded**.

We greatly appreciate all of the time and detail that the reviewers put into their evaluation of our manuscript, both in the first round of revisions and in this second round. We have carefully addressed the remaining comments, which we detail below. As stated in our first set of review responses, we think this process has really helped to improve the manuscript, and are grateful for all of the thoughtful feedback.

Notification to the authors:

Checking your paper, I noticed that your table 1 contains coloured cells. Please note that this will not be possible in the final revised version of the paper due to HTML conversion of the paper. When revising the final version, you can use footnotes or italic/bold font. For now, the process will continue, but please note that the final version cannot be published by using coloured tables.

We have removed the coloring in Table 1.

Please ensure that the colour schemes used in your maps and charts allow readers with colour vision deficiencies to correctly interpret your findings. Please check your figures using the Coblis – Color Blindness Simulator (https://www.color-blindness.com/coblis-color-blindness-simulator/) and revise the colour schemes accordingly. => Figs. 3, 4, 7 and 8.

We have made changes to color schemes of Figs. 1, 3, 4, 7, and 8 accordingly.

Reviewer #1

By and large, I have to say that authors did good job with addressing most of the reviewer comments. All my minor comments where well answered and the authors added a whole new discussion section to address my major comment. Alas, while the new discussion section is well written it is not what I asked for. What I wanted was that some basic ML knowledge is added to the introduction so that readers to not get the wrong impression that one SHOULD expect that the LSTM performs well under (counterfactual) distribution shifts. What I got was a discussion on potential limitations at the end of the manuscript. This choice shows me that the authors have a different (perhaps irreconcilable) view on this than I have. That said, at its core I still think that this is a good study. I especially like the construction of the MC-LSTM that considers the PET. Its simple and well thought trough.

Thus, writing a second review was difficult for me. I tried it several times and at the end I decided for what it is now. I recommend an accept with only technical corrections (see minor comments). As a general comment, I would just like to recommend to the authors to go over their references again and check for adequacy. I did know and/or check of the given references and found often inappropriate (see minor comments), but I am quite sure that there are more errors that I did not catch (maybe the other reviewers have?). Apart from that, I decided to provide a bunch of minor comments that are geared to making the manuscript more thorough. I let the authors/editor decide to which degree they should be adressed. I hope that this helps to improve the manuscript even further.

Good luck and all the best,
Daniel

We appreciate the thoughtful second read of our manuscript. We understand the difference of opinion in terms of where some of the new discussion should be placed in the manuscript. We carefully considered this when completing the revisions, and ultimately concluded that the introduction would become unwieldy and overly dense if we tried to insert this content into the Introduction. We were particularly concerned about how this would impact readability for interested early-career students. This is what motivated our choice of placement.

Regarding the references, we have gone through and considered the reviewer's suggestions. We do not completely agree with all instances where the reviewer deems certain references inappropriate, but we have made an effort to address those instances where we most see the reviewer's point.

Overall, we would again like to thank the reviewer for their very thorough and constructive review. As stated above, we sincerely feel this process has helped to significantly improve the work, and we are grateful.

Minor Comments

L. 141-L.144 I do not understand this sentence. How does the absence of tests for climate change conditions invalidate that physical constraints inhibit the ability of DL models to learn biases in the forcing data? Your test design does not probe for that and your results certainly do not indicate anything in this regard.

The absence of tests for climate changes do not invalidate the fact that physical constraints inhibit the ability of models to learn biases, and we did not intend to imply this from this sentence. We simply were noting that there are some downsides to physical constraints in DL models, but there might be some benefits in the climate change context. To avoid confusion, we have deleted the reference to learning biases, as its not critical for the point we are making. We have revised the text as follows:

> *For instance, Hoedt et al. (2021) introduced a mass conserving LSTM (MC-LSTM) that ensures cumulative streamflow predictions do not exceed precipitation inputs. Hybrid models present a related approach, where DL modules are embedded within process-based model structures (Jiang et al., 2020; Feng et al., 2022; Hoge et al., 2022; Feng et al., 2023a). In some cases, these architectural changes can degrade performance compared to a standard LSTM (Frame et al., 2021b; Frame et al. 2022; Feng et al.,*

> *2023b), but other times such changes can be beneficial (Feng et al., 2023a). **To date, the benefits of mass conserving architectures have not been tested when employed under previously unobserved climate change.***

L. 128. Karpatne et al. (2017) is an inadequate reference here since they do not use the term physics-informed machine learning.

We have added Karniadakis et al. 2021, which directly uses the term physics-informed machine learning.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L. et al. Physics-informed machine learning. Nat Rev Phys 3, 422–440 (2021). https://doi.org/10.1038/s42254-021-00314-5

L. 138. This is incorrect. Jiang et al. (2020) and Feng et al. (2022, 2023a) do not embed DL modules within a process model. As a matter of fact, for Jiang et al. (2020) it is literally the opposite: A conceptual model module is embedded in a DL framework.

Our interpretation of this comment is the use of the word 'embed'. To avoid confusion while retaining brevity in the text, we have changed this to 'combined'.

L. 187-189. I would actually argue that your results show the opposite, since model the National LSTM --- which is, amongst other things, also a little bit trained conditions where temperature and PET are less correlated --- already performs better. I would recommend to adapt the framing of Karpatne et al. (2017) here and say something like "..., which indicates that we either need to build or models on large data sets that comprise similar conditions to the ones under climate change or we need to guide the model selection using theory (see e.g., Karpatne et al., 2017)".

We agree with the proposed wording and have made this change.

L. 198 You say here that the primary goal is the experimental design, but that is not reflected in the abstract and in the introduction. Your previous work already introduced the experimental design. Why would you make it your primary goal again? Further, as I already mentioned in my first review your experimental design is NOT suitable to evaluate DL rainfall--runoff models for hydrological projections under climate change. In your revised discussion section you know say so yourself. So why are you keeping this part here, as it would indicate that you did not reach your primary goal.

We have decided to remove any statement about the overarching goal of this work, as the previous two paragraphs already state the purpose of our work (and also summarize the key results of our work). Therefore, we feel the line under discussion here is redundant from that perspective. In addition, by removing this line, the next sentence flows more directly from the end of the previous paragraph.

L. 200-201. Beven (2023) is not the correct reference.

We are not quite sure why the reviewer thinks Beven (2023) is not the correct reference, but we have decided to remove this reference anyway given that we removed the preceding line around designing benchmarking studies in response to the previous comment.

L. 281-284. Is this model driven with the same forcings as in Kratzert et al. 2021? If not, it should be mentioned right here that the change in forcings introduces a second covariate shift that the model is exposed too.

The forcings are in fact the same (besides the warmer temperatures, which we discuss below). Therefore, we make no changes here.

L. 303. "We develop ..." -> "We calibrate ..."

This change has been made.

L. 306 Delete conceptual here since these models are conceptual by nature and not by choice.

Another reviewer was very insistent that we highlight the conceptual nature of these models in multiple places (including this one), and so we have decided to retain our reference to 'conceptual' here to respect that reviewer's perspective and request.

L. 321. "... developed ..." -> "... calibrated ..."

This change has been made.

L.361ff. This paragraph basically suggest to readers that the LSTM from Kratzert et al. (2021) is not the same LSTM except for 1 different input (as is, for example, claimed in figure 2), but does ingest a substantial amount of different inputs. Please revisit every passage where this claim is made.

We are a little confused by this comment, as just a few lines down we state explicitly that the National LSTM was trained using a different set of data as compared to the Great Lakes models. However, we do see how one might interpret our caption in Figure 2 as overly simplifying the differences between the two sets of models, and so we have revised that caption to highlight the different inputs used.

L. 417f. The definition of \hat{sigma} is still wrong, since R is not a vector (see L. 412).

We have now clarified that \hat{sigma} is applied column-wise to the matrix R so that R is column-wise normalized.

L. 496ff. It is unclear what is meant here with stronger. Stronger than what?

Stronger than other temperature-based methods that also depend on radiation, and so produce PET that is less correlated to temperature than Hamon-based PET. We actually think this is pretty clear from the sentence as its worded, and so have decided not to make any changes here.

L. 542. The statement "... temperatures are warmed by 4°C ..." is wrong. You just add 4°C.

We have changed 'warmed' to 'increased'.

L.548ff. This is a bit of a repetition from what I wrote in my pervious review, but you should (again) mention here that this induces a covariate shift.

We have decided not to make this edit. Currently, we don't introduce the idea of covariate shifts until the Discussion. We think it would confuse the reader to introduce this concept here or above without having read the broader remarks around covariate shifts that is placed further below in the Discussion section.

L. 723-724. I would recommend to delete the last part of the sentence (everything from the but onward). You explained before that the FLV is very erratic for values near zero, and hence that a large change in value does not correspond necessarily to a hydrologically significant change. I think the last part does therefore not contribute anything to the argument, but might irritate readers or lead them to wrong conclusions.

We agree and have taken the recommendation to delete the latter part of this sentence.

L. 841-842. The results do not show what is claimed here. As you say yourself in the previous paragraph your results can, if at all, only show that LSTM are not able to predict physically plausible differences in streamflow under the assumptions that you nicely summary right in the sentence before this one.

We have changed this line to say that our results show that a standard LSTM did not predict physically plausible responses, rather than a standard LSTM is not able to predict physically plausible responses. We think this change gets at the important point that our results are not a final comment on the capabilities of LSTMs, for all the reasons discussed in the previous review.

L. 842-843. I would recommend delete this sentence since its obvious. Evaluating for streamflow prediction performance is not an indicator of how a good a model performs under distribution shift. If anything one would need to explain WHY the model performs well in this scenario if it did.

We tend to disagree here, mainly because we are not convinced this point is obvious to the broader readership of students (and professional hydrologists) that are still not well versed in DL hydrologic models. We know many who fit this description, and think this point is important to make for that audience.

L.844-847. This sentence is simply wrong. What your experiment can show is that, given an arbitrary (somewhat unrealistic) distribution shift, an LSTM based model is not able to adjust a way that we would expect by using a physically plausible rational. In no way or form do the experiments express anything about the general physical plausibility or implausibility of an LSTM per se. Also note that the claim is not in line with what you write a few paragraphs later (i.e., L. 892-906; where you literally write that your test is not well suited for testing the adequacy ML approaches).

We understand the concern and agree that we want to avoid making overly general claims given the issues around covariate shifts already discussed. However, we think this can be addressed simply by adjusting the last line of this paragraph, highlighting that this result applies to the LSTM predictions as produced in this work (and therefore, not to all LSTM predictions more generally).

L. 884. It is wrong to assume that a discrepancy in the inputs might be less impactful without testing. Just because something can be learned in theory does not mean that is has to be learned from the data.

We have revised this sentence to avoid speculating about the impact of the different data sources.

L. 895. Please use the appropriate references. Razavi, 2021 does uses the word sensitivity analysis only in passing when explaining that ANNs are black boxes and he not use the word metamorphic. How can this be a correct reference at this place? This is the second time that I arbitrarily looked at a reference of yours (the other was in the first manuscript) and again it is a reference that has no connection with what is written in the sentence. An example for a more appropriate list of reference for the topic would perhaps be Chen et al. (1998), since they introduced metamorphic testing; Murphy et al. (2008), since they seem to be the first to have used the concept in an ML context; and Yang and Chui. (2021) since they seem to be the first to have used it in the hydrological domain; and Reichert et al. (2023) because they basically forced you ;)

Here we respectfully disagree. Even if a reference doesn't use a specific term, the reference can still be appropriate if the content of their analysis follows the definition of that term. We believe that is the case for Razavi 2021. The reviewer made a similar comment regarding physics-informed machine learning above, and our response is similar to that concern. Furthermore, we do not think it necessary to add references that extend outside of hydrology, as a reader can pursue those themselves if they follow the Yang and Chui (2021) or Reichert et al., 2023.

L. 953. "Advancing causality" has no meaning in this context. I think what you want to say is something like "... other DL methods that make use of causal concepts ..."

We agree and have implemented the suggested wording change.

Reviewer #2

I would like to thank the authors for their thorough and detailed responses to my previous comments. I think this paper will make a good contribution to the literature on the applicability of machine learning-based hydrological models for climate impact studies. I have only one more minor editorial suggestion: throughout the manuscript, the terms "process model" and "process-based model" are used interchangeably. I recommend choosing one term and using it consistently for clarity ("process-based model" may be more common).

Thank you for the positive feedback and the very constructive review process. We have gone through the manuscript and now use the term 'process-based model' consistently throughout.

Reviewer #3

I have reviewed the revised version of the manuscript „On the need for physical constraints in deep learning rainfall-runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration" and I have to commend the authors for thoroughly addressing all my comments and suggestions. I am convinced now that this manuscript will be a valuable contribution of great interest for HESS readership. I have just a few minor comments that might help to clarify some ambiguous points in the text.

Kind regards,
Larisa Tarasova

We would like to thank you again for the constructive and helpful review process. Please see below for our revisions to address the remaining comments.

Minor comments

Line 26, 34, 594: I would suggest the author to use the term “conceptual, process-based models” directly in the first sentence of the abstract (similarly as it is done in Line 34). I think this is important because there are clearly two very different perceptions on what “process-based” models are. The set of models used in this study are not strictly process-based (although they do aim to resemble the actual physical processes), despite the explanation the authors provide in the rebuttal. I do acknowledge that there is a lot of literature that does use the term “process-based” for such models (as the authors demonstrate in their rebuttal), although in my opinion there is a confusion between model discretization (lumped vs fully-distributed) and model physicality (bucket concepts vs physical processes). The authors in their rebuttal rather refer to the former than to the latter. Therefore, I believe that the term “conceptual, process-based models” would be an acceptable compromise that will help to avoid any confusion among the readers.

We understand this viewpoint and agree that including “conceptual” in the first line of the abstract serves as a good compromise. Therefore, we have made this change.

Line 86: extrapolatable in space (since the authors are examining temporal extrapolation in this manuscript)

We have revised this sentence to read: “the most accurate and **spatially** extrapolatable rainfall-runoff models”

Line 268, 273, 574: evaporative water losses?

We have revised the text in these lines to explicitly reference evaporative water loss.

Line 308-314: Popularity is not really a comprehensive criterion for selecting something for a scientific experiment, as popularity is often the result of simplicity rather than scientific rigor. It would be much more useful to describe the differences in the structure of the benchmark models and whether or not they were reported to perform consistently in previous climate change studies. It would be important to cover a wide range of behaviors (as for example was excellently done with the choice of PET methods).

To address the concern here, we have made three changes to the paragraph in question, which we highlight below. First, we have added in a line and citation describing how SAC-SMA has been shown to outperform the National Water Model across the CAMELS dataset in out-of-sample performance, which we think is relevant in the context of how we discuss ‘scientific rigor’ of more complex, physics-based models. Second, we have deleted our reference to the popularity of HBV, which we agree may not be the best criterion. Finally, we now cite and describe the main result in Herman et al., 2013, which showed that HYMOD, SAC-SMA, and HBV can exhibit significant inter-model differences in behavior, dominant processes, and performance controls through time, even in situations where they share similar process formulations. This is perhaps the most relevant change that addresses the reviewer’s point, as they correctly argue that its important that the benchmark models cover a wide range of behaviors.

> *We develop three conceptual, process-based hydrologic models as benchmarks, including the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström and Forsman, 1973), HYMOD (Boyle, 2001), and the Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash, 1995) coupled with SNOW-17 (Anderson, 1976). These models are*

> *developed as lumped, conceptual models for each watershed, and were selected for several reasons. First, in the Great Lakes Intercomparison Project (Mai et al., 2022), HYMOD was one of the best performing process models for both streamflow and AET estimation. SAC-SMA is widely used in the United States, forming the core hydrologic model in NOAA's Hydrologic Ensemble Forecasting System (Demargne et al., 2014).* ***This model was also shown to outperform the National Water Model across hundreds of catchments across the United States (Nearing et al. 2021).*** *We also found in WS22 that AET from SAC-SMA matched the seasonal pattern of MODIS-derived AET well across California. HBV is also used for operational forecasting in multiple countries (Olsson and Lindstrom, 2008; Krøgli et al., 2018) and performs very well in hydrologic model intercomparison projects (Breuer et al., 2009; Plesca et al., 2012; Beck et al., 2016, 2017; Seibert and Bergström, 2022).* ***Importantly, the HYMOD, SAC-SMA, and HBV models can exhibit significant inter-model differences in behavior, dominant processes, and performance controls through time, even in situations where they share similar process formulations (Herman et al., 2013).***

Herman, J. D., P. M. Reed, and T. Wagener (2013), Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior, Water Resour. Res., 49, 1400–1414, doi:10.1002/wrcr.20124.

Line 723: Avoid using term "significantly" here.

We removed the latter part of the sentence including this term per another reviewer's recommendation.