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
Italicized orange font:	Updated manuscript wording, underline for changes to original

---

We greatly appreciate the time and detail that the three reviewers put into their evaluation of our manuscript. We have addressed all of the comments on a point-by-point basis, which we detail below. In particular, we have made substantial revisions throughout the manuscript to address several major concerns expressed by the reviewers, including:

- A more nuanced presentation of the specific experiments conducted in this work, emphasizing that we conduct sensitivity analyses rather than formal projections under climate change;
- Improved context for our results and experiments, in particular with respect to the machine learning literature on distribution shifts and causality;
- More detail on how our results highlight the strength and weaknesses of DL models to simulate hydrologic responses under changing snow accumulation and melt processes with warming; and
- A more thorough treatment of parametric uncertainty in the process models, and how it impacts the interpretation of DL model responses to warming.

We think these revisions, along with several others, have served to significantly improve the manuscript.

Reviewer #1

This study examines the behavior of different models under a hypothetical scenario where 4°C are added to the daily minimum and maximum temperatures. In doing so, the contribution finds that models with more explicit representations of hydrological processes are likely to exhibit more realistic behaviors under this shift.

I find this kind of study very important and very timely (as other discussion papers show; see e.g.: Reichert et al. 2023). On top of that, the execution is done well: The work is, by and large, well motivated; the idea is good; and, all tables and images are clear; almost everything is documented. I therefore think that the study should definitely be published on HESS. In terms of critique I have one point about the literature that I think is crucial, and some small questions/comments. The latter are, however, not so important.

We thank the reviewer for their overall positive, constructive, and speedy review. We greatly appreciate the feedback and believe it has served to significantly improve the manuscript.

Major Comment

The references are quite thorough with regard to the recent use of deep learning in hydrology. I complement the authors for that. They do, however, ignore large amounts of work from the outside the field. Normally this would not be a concern --- since one feeds into the other --- but here it does skew the motivation somewhat. As of now the introduction/motivation of the work reads as if current researcher are not aware that one can increase the temperature by some degrees and then test what the model would do under such circumstances. This is however not the case. For example, the group I am involved with, did not conduct such counterfactual experiments because we knew that deep learning models are out of the box not be able cope with arbitrary shifts in the covariance structures of the inputs. Statistical learning hinges on the idea that the future looks similar to the past --- and in a counterfactual setting this property is not given by design.

I strongly believe that the paper should give a better overview of the current machine learning literature and use that to discuss the merits and limits of the study design. This would give readers a much richer picture of what the proposed evaluation can probe.

Specifically, I am thinking that the paper should reference current work on (a) causality and (b) distribution shifts; and then use it feed into the discussion of the limitations of the current study. The reason why I think of (a) and (b) is that both research branches are fundamental to understand the study design: (a) Causality is important because the examination is a true counterfactual in that the adopted input has not --- and will never be --- observed in reality (remember, the daily values of the min and max temperatures change by adding exactly 4°C to all basins, while inputs like the radiation, wind, precipitation, and vapor pressure remain entirely the same). (a) The research on distribution shifts is important because adding 4°C to each day is a prime example of a covariate shift. Detecting, handling, "robustifiyng" and/or adapting to distribution shifts is an active area of research and should be seen as an open problem. Roughly speaking, results from (a) and (b) provide a counter point to the current motivation of the research in that they suggest that dara-driven models should per-se not be able to withstand a counterfactual examination. I think this would help readers to understand that the "physical plausible" response of the catchment model is measured with a "physically implausible" counterfactual signal (which is not observed in any catchment no matter what and will force the models into a sort of "extrapolation regime"). I believe that only then readers will understand that this is a very special form of test --- and that is very impressive that it is possible to design data-driven models that already show promising result in this setting, while having just a few more inductive biases than the current LSTM based rainfall-runoff models. In this regard, I do not want to force the authors to cite any particular work, but beg them to align their work with these branches of research (even if it means that they need to relativize their a-priori expectations)

We are grateful to the reviewer for making this suggestion. We agree that the literature on causality and distribution shifts in machine learning is extremely relevant to our study design, and in particular to its interpretation, limitations, as well as a fertile ground for future work. The

last two concluding paragraphs in our original manuscript were an attempt to address some of these issues, although admittedly this was not done to the extent necessary or with reference to the large body of work on these topics in the broader ML literature that the reviewer notes here. Consistent with one of the last reviewer suggestions below, we have taken this opportunity to significantly revise (i.e., large rewrite) our Discussion and Conclusion section, removing some of the older content and replacing it with a more robust treatment of the issues raised here. In the process of this revision, we tried to integrate our discussion around physics-informed machine learning (PIML) into the broader discussion on distribution shifts and causality, as we view PIML as one set of approaches (among several) that falls under the broader umbrella of causal deep learning methods.

We also note that in response to this comment and another by Reviewer #2, we have significantly revised the text throughout our manuscript to better convey what our experiment actually tests: the sensitivity of these models to imposed shifts in temperature and associated changes in potential evapotranspiration, rather than internally consistent climate changes across all meteorological variables. We believe these revisions also support the general points being made in this reviewer's comment.

Minor Comments

L. 85-86. Please add a reference to this sentence (or an explanation why no reference is given). You make the claim that "many argue" without even giving a single example.

We have added a recent paper (Nearing et al., 2022) that makes this argument based on past literature (or an adjusted version of this argument, see response to comment below), and also changed "many" to "some" to avoid overstating this claim.

L. 85-86. I think the meaning of "state-of-the-science" should be outlined. As far as I am aware it is not common terminology in hydrology (I, for one, had to look it up and am still not sure what is meant with it in this context).

We have changed the wording here to be more explicit, removing 'state-of-the-science' and instead replacing it with 'most accurate and extrapolatable'

L.100-101. I disagree with the claim about the corollary. Maybe it is an implication? I am not sure however: (a) Given the noise in the data, even without new climate conditions the predictions might be physically implausible. (b) Just because a ML model is "physically plausible" in in a out os sample setting does not mean that it remains so under a shift setting. What do you think about writing something like "From these results one might think that ..." or "If we spin these results further one could think that...".

We agree a wording change is warranted here. From the suggestions provided, we have altered this sentence to read:

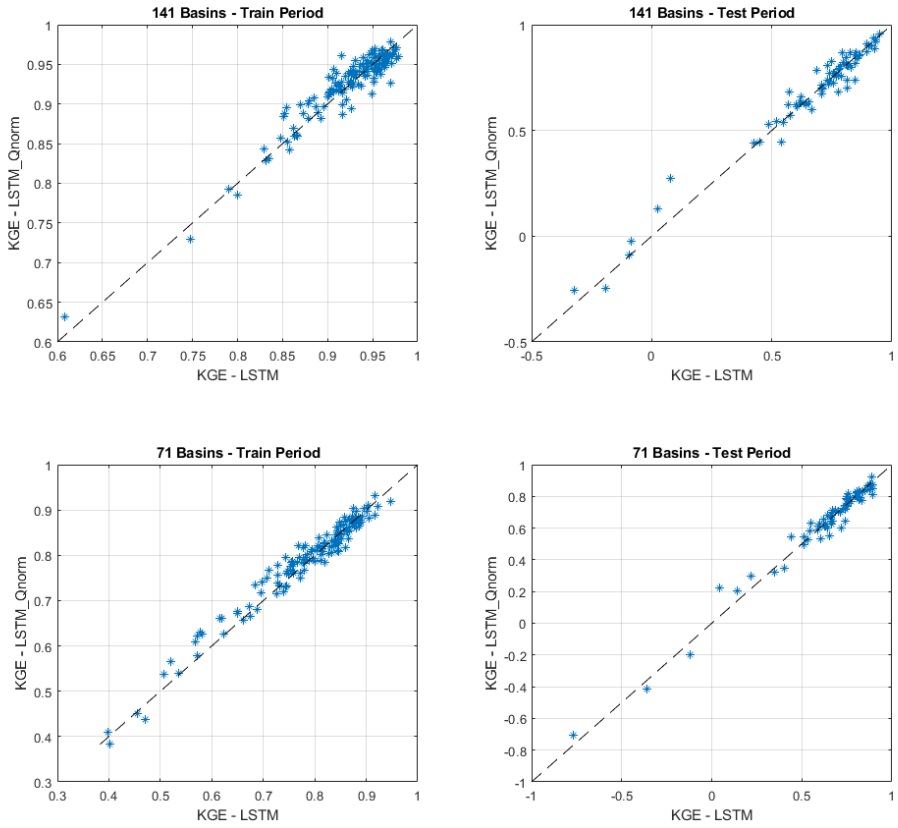
> *A potential implication of this finding might be that these models can produce physically plausible streamflow predictions under new climate conditions.*

L.108ff Is it correct that, from a hydrological perspective, this assume that there are no Glaciers/permanent snow in the basins (which, I think is, e.g., not true for CAMELS US as used in Liu et al. 2022)? The mechanism would be that as long as there is more melting happening, we should see higher water levels with higher temperatures.

That is correct, and our study in Wi and Steinschneider (2022) does point out (and highlight in the results) that this assumption of streamflow loss under warming would not apply to watersheds that drain regions with glaciers or over-year snow cover. We have adjusted this line to highlight this exception.

L. 334. Would you be so kind and make a comparison of the (normal) LSTM performance with normalized streamflow and without it? I once made a similar test, where I trained an LSTM on CMALES US without setting the standard deviations to one. And, got very bad results... However, your performance seems to be comparable to the ones reported in Mai et al. (2022). To me it is not really clear how you did that (especially since you used a relatively small learning rate and since the linear layer requires much bigger parameters in your setting). Maybe it is because the magnitude and behavior of the GRIP-GL rivers are much less diverse than the ones in CAMELS US?

Sure thing. We went ahead and refit the LSTM using normalized streamflow, employing the exact same training process as for the LSTM described in the article. We've pasted below a comparison of performance (KGE) between the original LSTM (without normalization) and the LSTM with normalization, showing results separately for training and testing sites and training and testing periods. We do not see any meaningful difference between the two. We agree that perhaps this result is driven by the fact that the diversity across sites in our domain is less than across the entire continuous US. Related to this, the flow in our region is much less skewed compared to some other arid and semi-arid regions, which might be contributing to this result.



L. 165ff. I know this is a choice of style and I will not mention this for the other occurrences, but: I would appreciate if you could already sketch the outcome of the experiments here (and in the other instances where you hypothesize about properties that one actually already knows at the time of writing).

We have adjusted the wording throughout the Introduction to state the outcomes we found, rather than posit them as hypothesized outcomes.

L. 176. Maybe adjust sentence a bit. I pretty sure that Frame et al. 2022 did not made an argument that physical constraints are not needed in for generating plausible projections under climate change. And, this sentence could easily be misread in that way.

To avoid any confusion or misinterpretation, we have simply removed the reference to Frame et al. 2022 (and arguments that physics-informed constraints are unneeded) altogether. We think the sentence stands fine on its own without this added clause.

L.268ff & L.350-351. It is probably an oversight on my side, but cannot find the code for this analysis in the zenodo repository.

This was an oversight on our part. The code for the National LSTM has now been added to the Zenodo repository.

L.344ff. Can you add a description or table with the grid you searched the hyper-parameters for to the supplementary?

Yes, we have added a table in the Supporting Information that shows the grid search used for the hyper-parameters, and now reference this SI table in the main manuscript.

L.377. I would recommend to explicitly write about σ and $\hat{\sigma}$ here so that readers know what you are referring to.

We now include the equation for $\hat{\sigma}(\cdot)$ in relation to $\sigma(\cdot)$, and provide a brief explanation of its output.

Table 2. I think the MC-LSTM KGE for "Testing Sites: Testing Period" should also be marked in bold since it is also 0.72 (the decimals that follow and are not shown should not be considered for a tie breaker here).

We agree, we have bolded the 0.72 being referred to in this table. We also slightly revised the manuscript text to be consistent with this change.

> *In particular, the LSTM outperforms the MC-LSTM and MC-LSTM-PET for NSE and FLV (as well as KGE in the training period), the MC-LSTM-PET outperforms the LSTM and MC-LSTM for PBIAS, and either the MC-LSTM or MC-LSTM-PET are the best performers for FHV.*

L.497ff Please describe the actual changes that you made to the static attributes either here or in the supplementary. I can see the changes in the data, but that requires readers to reconstruct what you did.

We have added a section to our Supporting Information to more clearly describe the changes made to the static attributes, and refer to this section here in the main article.

L.497ff I am probably missing something here, but to me its is not obvious why you changed the snow fraction of the precipitation with temperatures below 0°C? If the model gets an input with -3°C it should not matter to this whether this value was the true input or the counterfactually modified one; no?

The static input frac_snow is defined as the fraction of precipitation falling on days with mean daily temperatures below 0°C, i.e., the total amount of precipitation falling on days with T < 0C divided by the total amount of precipitation falling on all days. Under our warming scenario, the number of days with precipitation falling when temperatures are below 0°C declines, and thus, so does frac_snow. We now clarify this in our revised manuscript (see response to comment directly above).

L.656 consist -> consistent

This has been corrected.

L. 803ff. Is it really necessary to discuss short-wave radiation for so long here? You also did not consider that the thermic and dynamic behavior of the atmosphere and hence, the precipitation patterns would, for example, change over the whole region. I think you could abbreviate this paragraph considerably by just stating that the input modification is pragmatic and intuitiv, but does not reflect how the meteorological behavior would actually play out under climate change. This would then also my proposed literature references if you decide to include it.

We have taken the reviewer's suggestion, and have significantly shortened our focus on radiation here in favor of a broader treatment of the issues of distribution shifts and causality, as mentioned in the reviewer's main comment.

Upon reflection I would like to add that I think it would be highly beneficial if you could add some representative Hydrographs to an Appendix. This is, for one because I am interested to see some because of my personal experience with mass-conserving models; but secondly I also genuinely believe that it would help readers to put the performance and interventions into perspective.

When describing the results in Figure 7, we now reference individual hydrographs for specific sites (at both daily and monthly timescales), which are provided in the SI. We reference these SI figures while highlighting changes to key attributes of streamflow (FLV, FHV, COM) under warming, in an effort to better show what some of these differences in flow statistics mean in terms of daily flow time series.

References

- Reichert, P., Ma, K., Höge, M., Fenicia, F., Baity-Jesi, M., Feng, D., and Shen, C.: Metamorphic Testing of Machine Learning and Conceptual Hydrologic Models, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2023-168, in review, 2023.

- Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), Hydrol. Earth Syst. Sci., 26, 3537–3572, https://doi.org/10.5194/hess-26-3537-2022, 2022.

Reviewer #2

General comments:

The study provides a comparison of various deep learning models with process-based models across a large number of catchments. It provides insights into their strengths and weaknesses for prediction under "climate change" conditions (that is temperature mean shift). The general conclusion of the study is that careful consideration of their architecture and large sample learning is important to ensure physical plausibility of projections under different scenarios. I believe that the content and findings of the research would be valuable and may be of interest to the HESS readership.

We thank the reviewer for constructive and thorough review. The feedback has helped us to significantly improve the quality of our manuscript.

However, there are key concerns that I believe detract from the overall quality of the manuscript. The primary issues are the heavy emphasis on the role of PET while omitting snowmelt, and the potential overstatement in labeling the scenarios as "climate change".

1) The study appears to rely too heavily on PET as the primary determinant in understanding the impacts of climate change on streamflow. While PET is undoubtedly critical, it is only one of many factors influencing hydrologic responses, particularly snowmelt, that may be important to streamflow generation in the Great Lakes region (https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2016GL068070). For example, lines 513-521 suggest that the assumption is primarily concerned with the model's ability to discriminate between differences in water loss based on different PET projections under similar warming conditions. However, in regions where snowmelt may play a critical role in determining streamflow, temperature sensitivity could have dual implications - one for PET and another for snowmelt dynamics. Ignoring the latter could bias the results. In particular, I think it could explain the results in Figure 7 g and h that the authors didn't explicitly explain (lines 660-664): for process-based models that rely on physical processes, early snowmelt can significantly shift the seasonal pattern of streamflow as temperature increases. However, for machine-learning model, which mainly make predictions on the possible seasonal correlation, didn't present such a significant shift due to the seasonality of T and PET does not change. Therefore, if some models, especially process-based models, inherently account for snowmelt while others don't, then the comparison may not be apples to apples.

That is to say, the observed differences between process-based and machine learning models could be due in part to the fact that some models capture snowmelt dynamics while others don't. Therefore, extrapolating the study's findings to broader climate change impacts may be premature, especially if the full range of factors isn't accounted for, which is related to my second concern.

The reviewer is entirely correct that snowmelt dynamics in regions like the Great Lakes will change under warming. Our focus on PET and long-term average water loss was motivated by

the need for a strategy to *prove* whether or not the hydrologic responses to warming from the DL models are credible. In many cases, it is difficult to know *a priori* the true magnitude of hydrologic responses to different types of climate change. Therefore, unless the DL models estimate a hydrologic change that disagrees in sign with estimated (and well-accepted) changes from a process model, it is difficult to draw firm conclusions on the credibility of DL-based estimates of hydrologic change.

In our previous work (SW22), we focused on the sign of streamflow change to determine if the DL models were estimating reasonable hydrologic changes under strict warming. In this work, we tried to add nuance to this assessment by focusing on the magnitude of average water loss under warming, leveraging the distinct responses that should be expected under temperature-based and energy budget-based PET estimates to evaluate the reasonableness of estimated responses.

When designing our experiment, we carefully weighed which attributes of streamflow change to include, trying to balance two factors. On the one hand, we wanted to provide a broad set of statistics that captured different aspects of streamflow change, since we thought this would be of interest to a wider readership, especially since these types of assessments of DL models under warming are very rare in the literature. On the other hand, we wanted to limit the number of streamflow attributes to only those that were directly relevant to our core arguments, namely that DL models struggled to estimate long-term average water loss under warming due to their inability to separate historical temperature and PET correlations that will likely change in the future.

We ultimately selected four attributes to focus on (long-term average streamflow (Q.AVG), low flows (FLV), high flows (FHV), and seasonal streamflow timing (COM)). The comment by this reviewer is seeking additional analysis focused specifically on snow accumulation and melt dynamics. We note that Reviewer #3 requested that we limit our analysis only to metrics (Q.AVG and maybe FLV) directly related to the core focus of this paper (hydrologic response differences to different PET series). We understand this range of views, and in our response we have tried to balance how to address them. In general, we agree with this reviewer that a broad readership may want to see other hydrologic responses under warming beyond just those related to long-term mean daily flow or low flows, and so are inclined to include some additional analysis and discussion related to snow dynamics (as requested here). However, we balance this with the points of Reviewer 3 by limiting the degree to which these additions expand the scope of our study.

For changes in snow accumulation and melt dynamics, we had intended the center of mass (COM) statistic to act as a proxy for shifts in snow accumulation and melt, as this statistic reflects seasonal streamflow timing that is heavily influenced by snow processes in the Great Lakes (as the reviewer notes). Our results highlight a few important points in this regard:

- All models suggest a shift to more streamflow earlier in the year because more precipitation is falling as rain and not snow under warming and snowpack that does accumulate melts off earlier in the spring.

- The process models suggest a much more prominent shift in streamflow timing than the Great Lakes DL models (Figure 7g,h). However, there is not a clear way to know definitively which models are producing the correct response.
    - We note here that the DL models do not get time series information on seasonal timing / time of year as inputs, but only meteorological inputs, so they are not predicting these shifts based on "seasonal correlation", but rather as a direct response to changing meteorology.
- The National LSTM model, which is fit to a larger set of more diverse sites than the Great Lakes DL models, estimates shifts in streamflow timing that are more like the process models than the Great Lakes DL models (Figure 8g,h).

In our original manuscript, we never made an association between the changes in COM to shifting snow accumulation or melt dynamics. This was an oversight. In addition, the COM metric on its own is a coarse metric that does not fully show the hydrologic shifts specifically in the winter and spring that are heavily influenced by snow dynamics. Therefore, to address the reviewer's concern, we have made two significant revisions to the manuscript:

1) First, we have added language in the Methods, Results, and the Discussion and Conclusion section making a more direct connection between the COM statistic and model estimated changes to snow dynamics. We highlight in these additions that the DL models do not represent snow processes (or any hydrologic processes) explicitly, but rather learn these processes from the data. We note here (and also in our revisions) that others (e.g., Lees et al., 2022) have correlated internal states from LSTMs to independent measures of snow and find very strong relationships, suggesting that these models do learn the patterns of snow accumulation and melt directly from precipitation, temperature, and streamflow data. However, we note in our Discussion and Conclusion that it seems the National LSTM may do this better than the Great Lakes DL models, given the stronger consistency with the process models in terms of changes to the COM statistic under warming (although differences were still notable).
2) Second, when discussing the results in Figure 7 regarding the COM statistic, we also now refer to a new figure in the Supporting Information that shows the shift in the monthly hydrology between the baseline and 4C warming scenario across all models. This figure more clearly shows the shifts in flow during the winter and spring, and how this shift differs between the process and DL models, in order to more clearly demonstrate how these models are simulating hydrologic response under changing snow conditions with warming.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

2) While the study examines the sensitivity of hydrological models to temperature changes, it may be misleading to equate this solely with climate change. Climate change is multifaceted and includes more than just temperature changes. Although the authors attempted to indicate this as a

limitation in constructing climate change scenarios, the use of the term "climate change" in the title, abstract, and elsewhere could inadvertently downplay the myriad ways in which climate change affects hydrologic systems. For example, factors such as land surface changes due to elevated CO2 have been shown to play a more dominant role in changing runoff (https://www.nature.com/articles/s41558-023-01659-8). Therefore, I think it may be more accurate to frame the study as a sensitivity analysis of hydrologic models to temperature and related PET shifts, rather than an examination of so-called climate change scenarios.

The reviewer makes a valid point, and to address this, we have significantly revised our manuscript to better convey what our experiment actually tests: the sensitivity of these models to shifts in temperature and associated changes in potential evapotranspiration. There are several places throughout the manuscript where we have made such changes, including:

- The Title, which now reads: *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*
- In the Abstract, Methods, Results, and Discussion and Conclusions, we have made numerous changes to emphasize the fact that this study is conducting a sensitivity analysis rather than a formal set of projections under internally consistent climate change scenarios. Where appropriate, we removed reference to the phrases “climate change”, “projections”, and “future predictions” to avoid convoluting our analysis with actual future projections under climate change, when we are really referring to a sensitivity analysis of model responses to warmed historical temperatures and associated PET changes.
- In addition, we have added significant discussion and additional literature review to our Discussion and Conclusion section to address the concerns expressed here (and by Reviewer 1). Specifically, we now discuss the machine learning literature on causality and distribution shifts, specifically in the context of our experimental design and its limitations to address the full spectrum of climate change, and the potential for that design to influence the results of this work.

I would therefore suggest a major revision to explicitly state the assumptions made about snowmelt in each model, and to include snowmelt dynamics in the discussion of runoff differences. In addition, if the label "climate change" is to be retained, the study should consider a broader range of factors that might be influenced by climate change, not just uniformly increasing temperature.

Please see our two responses to the major comments above.

Specific comments follow,

Abstract: I can't find a word limit for the abstract for HESS, but as the submission guidelines say, "An abstract should be short, clear, concise...". The abstract in its current form is too long.

We recognize that the abstract for this paper is longer than most conventional papers. The journal does provide guidelines that “abstracts should be short, clear, concise…”, as the reviewer notes. However, they also promote more comprehensive abstracts so that readers can get a fully sense

of the article's content from the abstract, and then don't impose page limits on the full manuscript, which together is their strategy to promote both conciseness and completeness (see here: https://www.hydrology-and-earth-system-sciences.net/about/faqs.html):

"*What major advantages does HESS offer to the readers and scientific community?*

- *promotion of scientific conciseness and completeness at the same time by including comprehensive abstracts rather than imposing strict page limits.*"

One can see examples of HESS articles with much longer abstracts than is conventionally accepted by other journals (see examples below):

- Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), Hydrol. Earth Syst. Sci., 26, 3537–3572, https://doi.org/10.5194/hess-26-3537-2022, 2022.
- Lehmann, F., Vishwakarma, B. D., and Bamber, J.: How well are we able to close the water budget at the global scale?, Hydrol. Earth Syst. Sci., 26, 35–54, https://doi.org/10.5194/hess-26-35-2022, 2022.
- Miglietta, M. M. and Davolio, S.: Dynamical forcings in heavy precipitation events over Italy: lessons from the HyMeX SOP1 campaign, Hydrol. Earth Syst. Sci., 26, 627–646, https://doi.org/10.5194/hess-26-627-2022, 2022.
- Bentivoglio, R., Isufi, E., Jonkman, S. N., and Taormina, R.: Deep learning methods for flood mapping: a review of existing applications and future research directions, Hydrol. Earth Syst. Sci., 26, 4345–4378, https://doi.org/10.5194/hess-26-4345-2022, 2022.
- Jean, V., Boucher, M.-A., Frini, A., and Roussel, D.: Uncertainty in three dimensions: the challenges of communicating probabilistic flood forecast maps, Hydrol. Earth Syst. Sci., 27, 3351–3373, https://doi.org/10.5194/hess-27-3351-2023, 2023.

With that said, we acknowledge that our abstract could be made more concise, and therefore we have significantly reduced the length of our abstract by over 25% to try and accommodate this goal.

L105-124, I understand that this study may build on the previous WS22 study. However, the depth of detail should be balanced, as many of the methods, challenges, and conclusions of the earlier work are now repeated in the new paper. The introductory section should focus primarily on setting up the current study.

We have significantly reduced the text summarizing the results of WS22 by ~30%, with the goal of only highlighting the most salient features of that article needed to set up the present work.

L141, please specify what you mean by "this work". If it is what is shown in the current manuscript, it is strange to discuss the results in the introduction. If it is still from WS22, the opinions seem reiterated again.

We were not referring the current manuscript or to WS22. Rather, we were referring to the previously mentioned in-line citations (Lofgren et al., 2011; Shaw and Riha, 2011; Lofgren and Rouhana, 2016; Milly and Dunne, 2017; Lemaitre-Basset et al. 2022) that investigated the effects of different PET estimation methods on the magnitude of PET and runoff change in a warming climate. To better clarify this, we have replaced “this work” with “these studies”, so the reference we are making is clearer.

L155, the assertion that temperature-based PET methods "significantly overestimate future projections of PET" compared to energy budget-based methods is a strong one. It might be beneficial to provide more evidence from literature.

We have added several of the in-line citations we previously cited in this paragraph to the end of the line in question, as all those articles show convincingly that temperature-based PET methods overestimate PET and water loss under warming.

L170-172, as noted in my previous comments, the hypotheses may suggest that PET is the sole or overwhelming cause of declining streamflow.

While we agree that other factors (like snow accumulation and melt dynamics) will significantly impact streamflow behavior under warming, we do still assert that changes in PET (which could include changes in plant transpiration under changes in CO_2, as mentioned by the reviewer earlier) are the dominant cause of *long term average* streamflow decline under warming, absent any precipitation changes. We emphasize ‘long term average’ to distinguish our focus from declines in streamflow in certain seasons, which could be balanced by increases in streamflow in other seasons. This is exactly what one might expect due to changing snow accumulation and melt dynamics, where warmer temperatures lead to more water running off into the stream during the winter and early spring months, which is then balanced by less runoff from snowmelt later in the spring and early summer. In this situation, precipitation entering the watershed is being redistributed in terms of the timing of streamflow, but the long-term average streamflow is not significantly impacted. The major way that climate can drive a long-term average decline in streamflow across multiple years is by either changing the long-term total amount of water entering the system (i.e., precipitation) or by changing the long-term total amount of water that leaves the system before it can reach the stream (i.e., evapotranspiration).

In our experiments where only temperatures are warmed and PET adjusted (and precipitation is left unchanged), these climate shifts will mainly influence the ET sink, and are less likely to have direct impacts on the other possible long-term sinks of water in the watershed (e.g., inter-basin groundwater flux). In the context of the paragraph in question, we are directly discussing the hypothesized effects of our sensitivity analysis around warming temperatures and PET shifts with historical precipitation, which justifies our focus on PET as the driver of flow declines. However, we recognize that perhaps our emphasize on long-term average streamflow declines (rather than seasonal declines or changes) was not clear. Therefore, we have adjusted the text here as well as throughout the manuscript to emphasize our focus on long-term average streamflow declines for this specific experiment. In addition, we direct the reviewer to our response to the first major comment, where we detail our other revisions to more thoroughly

address the concern that we are not sufficiently treating other forms of climate change (i.e., impacts of warming on snow accumulation and melts).

L180-182, it would be beneficial to show how the correlation between temperature and PET shifts with different estimation methods.

In the figure below, we show the distribution of Pearson correlation values between daily average temperature (the mean of daily tmax and tmin) and PET calculated using either the Hamon method or the Priestley Taylor method, shown across all basins in the Great Lakes region. The average correlation between temperature and Hamon PET and Priestley Taylor PET is 0.94 and 0.83, respectively. We think the most natural place to present this information is when discussing the calculation of Hamon and Priestley Taylor PET in Section 3.3, after the other meteorological data and sites in the study domain have been introduced. We have now included the mean correlation values in this section in the revised manuscript.

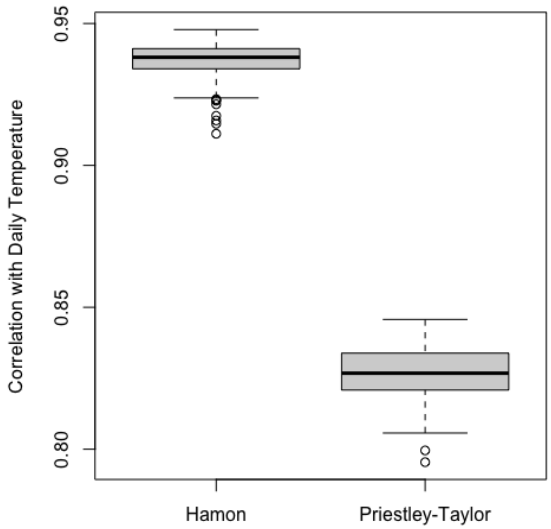


Fig. 2, I am confused by the presentation of the timing, as it seems to suggest that the cumulative flux of warming stops increasing after a certain day of the water year.

We think the reviewer is referring to the fact that the red line under warming reaches 100% of total flow and then stops at an 'early' day of the water year, as depicted on this plot. We agree that this is misleading, and so have changed Figure 2 to have the red cdf reach 100% at the day of water year as the blue curve.

L408-409, is there any post-examination of the trash cell (if it is constrained by PET or some value) to support this assumption?

We do evaluate the trash cell for both the MC-LSTM and the MC-LSTM-PET models in Figures 4 and 5. Importantly, in Figure 5, we see that the trash cell in the MC-LSTM model tends to accumulate significantly more water loss on certain days as compared to the MC-LSTM-PET model, showing how the PET constraint in the MC-LSTM-PET model limits water loss from the system on certain days as compared to a model without this constraint.

However, the reviewer is asking more specifically about whether we can evaluate the outcome of the trash cell in these two models to support the assumption in the MC-LSTM-PET model that the maximum allowable water lost from the system should not be allowed to exceed PET (i.e., water can't be lost via other terminal sinks, like aquifer recharge and inter-basin groundwater flux or human withdrawals and inter-basin transfers). This assumption is difficult to verify directly, because we have no direct observations of water lost to these other terminal sinks. However, we do note that the MC-LSTM-PET model has the lowest long-term bias of all DL models (see new Table 3), suggesting that this constraint helps rather than hurts the accounting of water entering and leaving the system on a long-term basis. This is evidence (albeit indirect evidence) in support of the assumption behind the constraint imposed in the MC-LSTM-PET, and we now point this out in our revision.

For a broader discussion of this point, please refer to our response to the third general comment from Reviewer #3.

L412, I think it depends on how many catchments of your study are natural without human intervention.

We have revised this text to provide a more nuanced explanation of when these assumptions likely hold:

> *Here, the ReLU activation ensures that any water in the trash cell (h_D) which exceeds PET at time t is added to the streamflow prediction $y[t]$, but the streamflow prediction is the same as the original MC-LSTM (Eq. 5) if water in the trash cell is less than PET. This approach assumes that the maximum allowable water lost from the system cannot exceed PET, and therefore ignores other potential terminal sinks (e.g., inter-basin lateral groundwater flows; human diversions and inter-basin transfers). This assumption is more strongly supported in moderately-sized (> 200 km^2), low-gradient, non-arid watersheds where inter-basin groundwater flows are less impactful (Fan 2019; Gordon et al., 2022), such as the Great Lakes basins examined in this work. However, we discuss the potential to relax the assumptions of the MC-LSTM-PET model in Section 5.*

In addition, we have significantly revised out Discussion and Conclusion section to revisit these assumptions, and discuss ways they might be relaxed or improved in future work.

L494-495, but this adjustment is not reflected in Figure 2. Why does the correlation bias have to take effect here, while the possible correlation between temperature and other input variables is omitted in the previous experiments?

In SW22, we carefully checked the correlation between temperature and all other input variables to the National LSTM (this is described in detail in that paper). The correlations were weak for almost all combinations of dynamic inputs except for minimum temperature and vapor pressure, because vapor pressure was a calculated input that depended on minimum temperatures. For this reason, we adjust vapor pressure with minimum temperature, but leave all other inputs unchanged.

We have adjusted Figure 2 to reflect the fact that vapor pressure also changes.

L499, please justify the selection of static input features that need to be changed, is this done by examining the dependence between mean temperature and other static input features?

The static input features that need to be changed are those that depend on temperature or PET in their calculation. In our warming scenarios we are changing temperature and PET, and so one could argue that the static features that were calculated from temperature and PET time series should also be adjusted accordingly. We have slightly reworded this sentence to make this clearer. In addition, we have added a new section in the Supporting Material to more thoroughly describe the adjustments to these static features.

> *We also consider changes to ~~some of~~ the static input features that depend on temperature and PET in their calculation (e.g., pet_mean, aridity, t_mean, frac_snow; see Table 1 for feature descriptions and Supporting Information S1 for details on adjustments to these features), and then run all models using two settings: 1) with changes only to the dynamic features, and 2) with changes to both dynamic features and to static features that depend on those dynamic features.*

L510, does the "year" here indicate calendar year or water year, it seems water year based on Figure 2.

Yes, water year. We have revised the text to state that explicitly.

L617-618, references are needed here to support the argument.

We have added the relevant citation (Allen et al., 1998) for this point.

L651-653, the explanation is not convincing for me, since the process-based model shows a more obvious change compared to the deep learning-based model, while no explanation of the difference between the two types of models was provided here.

It is true that the three process models do show significantly larger changes for low flows. However, they disagree on the sign of change, and so we don't think one can refer to the estimates of change in low flows from the process models as "more obvious". SAC-SMA shows declines in low flows under Hamon PET and a mix of increases and decreases in low flows under Priestley-Taylor PET. Conversely, HBV and HYMOD generally show increases in low flows across both PET methods. Therefore, one is left with a very unclear signal of change for this streamflow metric from the process models. The DL models show a much tighter range of change that sits within the range estimated by one of the process models (SAC-SMA), which so happens to be one of the more accurate models in terms of low flow estimation (as shown in Table 2). However, there is not clear evidence which of the models (process-based or DL) are producing a more realistic low flow response.

Our intent in this paragraph is simply to report the results as they are in Figure 7, rather than explain why the models differ in their responses. A thorough investigation and explanation of why we are seeing the differences in FLV changes between the process model responses, or between the process model and DL responses, is somewhat out of scope of our analysis. However, we do now provide individual hydrographs for specific sites in the SI, and highlight changes to key components of the hydrograph (including low flows) under warming, in an effort to better show what some of these differences in flow statistics mean in terms of daily flow time series. We reference these SI figures in this text to provide context for the differences in low flows we are seeing between the process-based and DL models.

L663-664, as I mentioned in the general comments, this may be due to the role of snow dynamics being treated differently between the process-based and deep learning models. Unfortunately, analysis on snowmelt was not included in the study.

Please see our response to the reviewer's major comment above regarding how we now better address hydrologic responses to changing snow accumulation and melt in our analysis of model sensitivity to warming.

L669-671, this finding is interesting, since if we really want to use the DL model to do the climate change impact analysis, considering the future surface climate and surface context variables changes are necessary, but the DL models do not seem to learn physically plausible relationships here when doing cross-region learning. It is also nice to see that a physically informed strategy can help mitigate the problem.

Thank you, yes, we have been getting requests from some practitioners asking us to consider using these DL models for climate change analyses. We are hoping these results help communicate that these models likely need some type of physics-informed adjustments before they can be used for this purpose.

L672-673, perhaps the additional implementation of test sites in the test period can be informed earlier around L609.

We have taken the reviewer's recommendation and moved reference to our assessment of the test sites in the test period earlier in the Results (i.e., to the line suggested here).

L803-805, it is also important to note that the constructed climate change scenarios break the Clausius-Clapeyron scaling, so I would suggest not calling them "climate change".

Agreed. Consistent with the reviewer's second major comment, we have avoided using the term "climate change scenario" throughout the manuscript when referring to the scenarios used in our sensitivity analysis. In this instance, we changed our wording to reference the "warming scenarios" used in this work. More broadly, we now discuss more thoroughly the climate change signals not addressed in this work in the last paragraph of the Conclusion, including Clausius-Clapeyron scaling and its potential impact on extreme precipitation under warming.

Reviewer #3

Review of the manuscript "On the need for physical constraints in deep learning rainfall-runoff projections under climate change"

This manuscript compares the simulations of unconstrained and physically-constrained deep learning models with the simulations of the three conceptual hydrological models in the Great Lakes Region under climate change scenarios with the goal to investigate versatility of the former in changing climatic conditions.

I find the premises of the experiment is very interesting and have a potential to provide useful insights on the fitness of the state-of-the-art models under changing climatic conditions. However, I find the experimental set up somewhat inconsistent: 1) national LSTM model does not use the same input (i.e., not the same PET data) as all the variants of Great Lakes Models, making it hard to disentangle the true reason behind its different behavior; 2) The implementation of PET-constrain in the LSTM model is rather crude and is based on the assumption that evapotranspiration and streamflow is the only way how the water can leave the system, which might not hold universally. Moreover, there are some subjective and ambiguous choices (e.g., choice of the performance metrics; choice of PET methods; choice of the conceptual hydrological models as baseline) in the experiment set up that needs to be clarified. Finally, although I find it interesting to compare behavior of the deep learning and conceptual hydrological models for future simulations, I find the results of the study rather inconclusive, because the differences between the simulations of the three conceptual hydrological models (e.g., Figure 8) seems to be very large (sometimes these differences even larger than the difference with the deep learning models), making one question the reliability of these conceptual models as the baseline. Please, find my detailed comments below.

We thank the reviewer for their thorough and constructive comments. We discuss and address each of the issues identified above in detail in our responses to the comments below.

General comments

Inconsistent set up of the national LSTM: The national LSTM model was driven by temperature, radiation and vapor pressure (Line 270-271), but not by either temperature-based or energy-based PET that were used for Great Lakes LSTMs. When comparing its simulations with the simulations of other LSTM-variants it is not possible to disentangle the origin of the observed differences, making one of the conclusions of the manuscript, that more diverse set of catchments might to some extent support learning physically-based processes, rather questionable, because the differences might be just as well be due to the difference in the forcing data.

We understand the reviewer's concern, but we'll note that both temperature-based (Hamon) and energy budget-based (Priestley-Taylor) PET methods used in this work are based entirely on input variables available to the National LSTM (temperature, radiation). Therefore, any PET would be entirely collinear with the inputs already available to that model, and so would not actually provide the model with additional information over what it already has.

Furthermore, some of the inputs that were developed specifically for the Great Lakes Intercomparison Project, which spans gauged sites across both the United States and Canada, are not available for the CAMELS dataset, and vice versa. Therefore, it is difficult to develop completely comparable models across these two datasets without first removing inputs from both datasets to ensure a completely overlapping set of inputs. However, by taking this step, the models produced in this work would not be directly comparable to either the models developed for the Great Lakes Intercomparison Project (an auxiliary goal of this work, as stated in the manuscript), or comparable to some of the accepted benchmark LSTM models commonly referenced in the literature (like the one from Kratzert et al. 2021 used in this study). We see this as a potential drawback of this approach.

Similarly, because the Great Lakes Intercomparison Project spans locations in two countries, the datasets used to define certain inputs are also very different than what is available for CAMELS (e.g., the meteorological data for the Great Lakes is based on the Regional Deterministic Reanalysis System v2, while for CAMELS three other meteorological products are available). This is a large source of uncertainty and potential discrepancy between any models developed on the Great Lakes basins and CAMELS basins, and would be very difficult to resolve without significant time and resources to develop a consistent dataset across both sets of basins. And again, by changing the data sources for the meteorological inputs in either the Great Lakes basins or the CAMELS basins, our results would not be directly comparable to benchmarking models for either region.

For all the reasons above, we argue that even if the LSTM was retrained (and hyperparameters re-estimated via cross validation, at significant cost) on the CAMELS basins with PET included, there will still be large discrepancies with the models developed for the Great Lakes basins, precluding an apples-to-apples comparison. And we contend that given the collinearity between PET and the inputs already available to the National LSTM, the likelihood of substantial differences in the National LSTM output is small. Therefore, in response to this concern, we have instead opted to include a paragraph in our Discussion and Conclusion that highlights the potential discrepancies between the National LSTM and the Great Lakes models as a limitation of this work, and also includes some of the nuances mentioned above in that discussion.

> *To properly interpret the results of this work, there are several limitations of this study that require discussion. First there were differences in the inputs and data sources between the National LSTM and all other Great Lakes models, including the source of meteorological data and the lack of PET as an input into the National LSTM. While this latter discrepancy might be less impactful (i.e., the National LSTM was provided meteorological inputs that together completely determine Hamon and Priestley-Taylor PET), the difference in meteorological data across the two sets of models is a substantial source of uncertainty and could lead to non-trivial differences in hydrologic response estimation, complicating a direct comparison of the National LSTM to the other models. Future work for the Great Lakes Intercomparison Project should consider developing consistent datasets with other (and larger) benchmark datasets like CAMELS to address this issue.*

Choice of PET methods: I very much like the idea of comparing temperature-based vs energy-based method for PET estimation. However, there is no rationale provided on the choice of the particular method (Hamon and Priestley-Taylor, Line 453). In my experience the difference between different temperature-based approaches can be huge. Likely, the same is true for the energy-based methods. I think using several methods for each type of the PET estimation would strengthen the argument of the manuscript.

We agree that different temperature-based PET methods can differ significantly in their estimates of PET. This is shown nicely in Shaw and Riha (2011). However, we don't believe that using multiple PET methods for all of our models is necessary to achieve the objectives of this work. Specifically, we sought to demonstrate that a standard LSTM, if provided two very different inputs on PET, would not respond differently under warming because the model learned a relationship between temperature and water loss rather than PET and water loss. This point can be made with only two PET methods that differ in their estimates of PET under warming, as was done in this work.

Furthermore, we note that the two PET methods selected (Hamon and Priestley-Taylor) lie towards the lower and upper bounds of temperature sensitivity across multiple PET methods (again, see Shaw and Riha, 2011; their Figure 1, copied below for convenience). Since our work shows that the LSTM model under either Hamon or Priestley-Taylor PET shows water losses consistent with process-model estimated under Hamon PET, and Hamon PET is one of the most sensitive PET methods to temperature while Priestley-Taylor is one of the least sensitive, our results as presented can be interpreted as approximately bounding the differences we might expect to see between the LSTM and process models in regards to long-term average declines in flow.

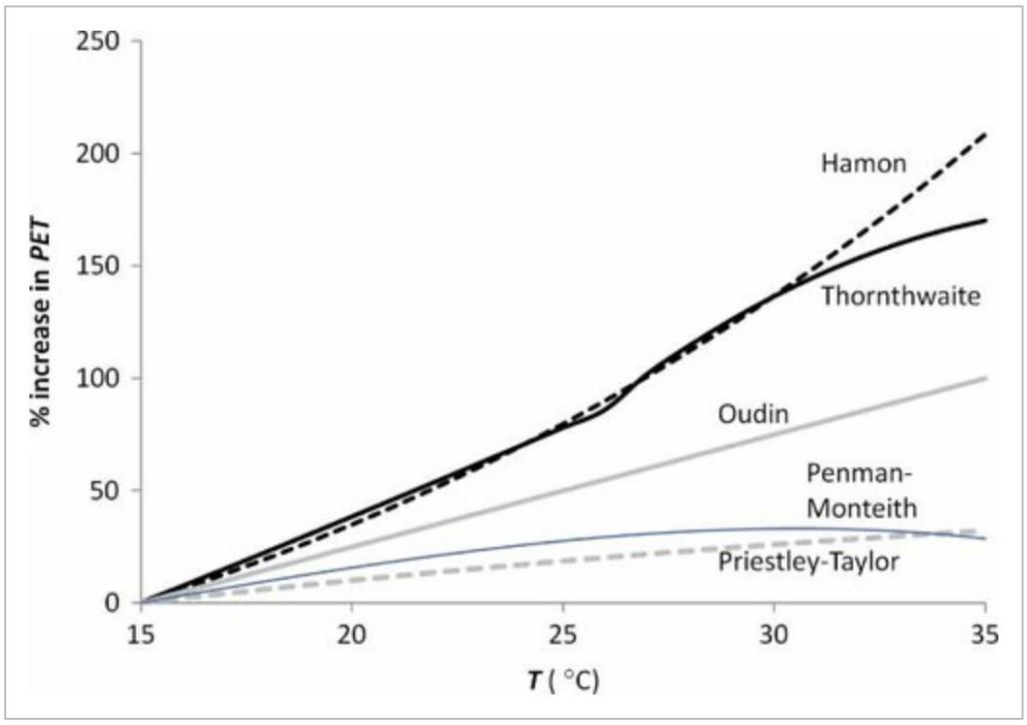


Figure 1 from Shaw and Riha, 2011, showing the temperature sensitivity of different PET methods.

Therefore, in light of the added cost of running similar experiments with multiple PET methods, we believe the value of these extra experiments would be marginal and not worth the cost of purchasing the required allocation on HPC resources. With that said, we have revised the manuscript to provide a clearer rationale for our choice of the Hamon and Priestley-Taylor PET methods examined in this work, including some of the points raised above. These changes can be found in the text introducing the two PET methods:

> *All Great Lakes models in this study are trained twice with different PET estimates as input, including the Hamon method (a temperature-based approach; Hamon, 1963) and the Priestley-Taylor method (an energy budget-based approach; Priestley and Taylor, 1972). We select the Hamon method because of its stronger dependence on temperature compared to other temperature-based approaches that also depend on radiation (e.g., Hargreaves and Samani, 1985; Oudin et al., 2005). We select the Priestley-Taylor method based on its widespread use in the literature (Wu et al., 2021; Su and Singh, 2023) and its approximation of the more physically-based Penman-Monteith approach (Allen et al. 1998). Together, these two approaches lie towards the lower and upper bounds of temperature sensitivity across multiple PET approaches (see Shaw and Riha, 2011).*

Hargreaves, G.H., and Samani, Z.A. (1985). Reference crop evapotranspiration from temperature. Applied Engineering in Agriculture 1: 96–99.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andreassian, V., Anctil, F., and Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modeling. Journal of Hydrology 303: 290–306.

Shaw, S.B. and Riha, S.J. (2011), Assessing temperature-based PET equations under a changing climate in temperate, deciduous forests. Hydrol. Process., 25: 1466-1478. https://doi.org/10.1002/hyp.7913

Su, Q., & Singh, V. P. (2023). Calibration-free Priestley-Taylor method for reference evapotranspiration estimation. Water Resources Research, 59, e2022WR033198. https://doi.org/10.1029/2022WR033198

Wu, H., Zhu, W., and Huang, B. (2021), Seasonal variation of evapotranspiration, Priestley-Taylor coefficient and crop coefficient in diverse landscapes, Geography and Sustainability, 2(3), 224-233, https://doi.org/10.1016/j.geosus.2021.09.002

PET-constrain for LSTM instead of hybrid models: The motivation for using PET constrains for LSTM by means of a trash cell that assumes that there are exclusively evaporative water losses (which might not always be the case, Jaschecko et al., 2021 https://doi.org/10.1038/s41586-021-03311-x) is not clear to me. Using hybrid models that seem to be exactly the tool that combines the strength of deep learning models and concepts of hydrological processes and therefore, are potentially more fit for producing reliable future simulations under changing conditions, seems like a more straightforward choice to me. One sentence in the discussion merely mentioning the existence of such models is definitely not enough in my opinion, because their existence and comparable performance with deep learning models question the need to develop any constrained variants of LSTM as done in this study.

The reviewer is absolutely correct that the MC-LSTM-PET model, as forwarded here, potentially over-constrains the model by limiting all water losses to be below PET. This structural constraint will lead to inductive biases in the MC-LSTM-PET that negatively impact the generalizability of the model to locations with recharge to deep aquifers and inter-basin groundwater flux, or in basins with human water withdrawals and inter-basin transfers.

However, we would argue that a similar set of issues is also entirely possible with hybrid differentiable models like those presented in Feng et al. 2022. In fact, those models inherent more structural assumptions from the process-model backbone on which are they built. This is both a benefit but also a potential drawback. The benefit is if the prior information embedded in that process model backbone helps the model generalize better to new locations, particularly in regions with little data (as shown in the PUR experiments in Feng et al., 2022), or under new climates. The drawback is if those underlying assumptions lead to structural biases between the model and reality. An example of this latter situation was recently presented in Feng et al., 2023, including in regions around the Great Lakes. As written in the abstract of Feng et al., 2023:

> ‘Nevertheless, relative to LSTM, the differentiable model was hampered by structural deficiencies for cold or polar regions, and highly arid regions, and basins with significant human impacts.’

That is, a hybrid framework only adds value if the prior information imparted by the process model backbone is a useful representation of reality in specific basins. There may be several instances where this is not the case, particularly if the structure of the process model is limited by having to be differentiable. In such cases, DL models with fewer assumptions (like the MC-LSTM-PET model) could prove beneficial.

Further, we note that one could view the MC-LSTM-PET model as a very simple version of a hybrid differentiable model. As stated in Feng et al. 2022, hybrid differentiable models can be viewed as either a constrained DL model, with constraints adopted from existing process models, or a more flexible process model, with flexibility added through the use of DL for certain modules. The MC-LSTM-PET model can be viewed as a specific instance of that first view, where two constraints (mass balance and a limit on water loss via PET) are adopted from a process model perspective into the DL architecture.

For the reasons above, its not clear whether the MC-LSTM-PET model or a variant of it (e.g., by allowing water loss to be limited by the sum of PET and estimates of deep aquifer recharge and inter-basin groundwater flux or human water withdrawals and imports/exports) will be “better” or “worse” than some of the recent differentiable hybrid models proposed in the literature, in terms of regional generalization or for hydrologic projections under climate change. We argue that such a comparison would be extremely useful but is beyond the scope of this article, because the contribution of this work is not limited to the introduction of the MC-LSTM-PET. Rather, a major contribution of this work is that we demonstrate, for the first time, how vanilla LSTMs may struggle to reproduce realistic predictions of hydrologic response under warming due to spatially pervasive correlations between temperature, radiation, and PET that could change under climate change. The MC-LSTM-PET model is then forwarded as one way to show that adding physical constraints into the DL model can help address this problem, but we do not argue it is the only way or the best way to do so.

Therefore, we do not feel its appropriate to try and introduce an entirely different modeling framework (hybrid differentiable models) into the current study. However, we agree with the reviewer that many of the issues brought up above deserve more treatment than they were given in the original draft of this work. Therefore, we have significantly revised (i.e., largely re-written) our Discussion and Conclusion section to do just this, and in the process, we have tried to embed many of these issues in the context of broader challenges when using DL models for causal prediction. Specifically, we have revised the Discussion and Conclusion section to:

- More broadly introduce the concepts of out-of-distribution generalization, covariate shifts, and how these issues in DL impact our experimental design.
- Introduce causal deep learning as an emerging family of techniques to help address this challenge.
- Forward physics-informed machine learning (PIML), like the MC-LSTM-PET and hybrid differentiable models, as a subset of approaches that falls under this family of techniques.

- Highlight some of the challenges with both the MC-LSTM-PET model (related to the reviewer's comment above), as well as some of the challenges with hybrid differentiable models (as mentioned in our response here).
- And then highlight other methods that fall under the umbrella of causal deep learning as alternative approaches that also might be worth consideration given some of the challenges with PIML for hydrologic modeling.

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. Water Resources Research, 58, e2022WR032404. https://doi.org/10.1029/2022WR032404

Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., Liu, J., Pan, M., Lawson, K., and Shen, C. (2023). Deep Dive into Global Hydrologic Simulations: Harnessing the Power of Deep Learning and Physics-informed Differentiable Models (δHBV-globe1.0-hydroDL), Geosci. Model Dev. Discuss. [preprint], https://doi.org/10.5194/gmd-2023-190, in review.

Choice of conceptual hydrological models, their parametric uncertainties and discrepancy in their simulations: The choice of the three conceptual hydrological models used as a benchmark is not clear. What is the rationale for selecting exactly these three models? What are the major structural differences between them? Are there any studies indicating the fitness of these models for future simulations/ET simulations? These questions have to be addressed to justify the choice of the baseline. Moreover, although the authors have accounted for uncertainty of training of deep learning models by running a 10 members ensemble, the parametric uncertainty of conceptual models (that can be very substantial) is completely ignored by using only one best simulation for each model, instead of using e.g., X% of best performing models (or so called behavioral parameter sets, Beven and Freer, 2001 https://doi.org/10.1016/S0022-1694(01)00421-8). Accounting for parametric uncertainty of the conceptual models might shed a light on large discrepancies between the simulations of conceptual models (e.g., Figure 8) that sometimes is even larger than the differences with the deep learning models.

We respond separately to the two main threads in this comment, which include: 1) why did we select the three chosen conceptual models as benchmarks; and 2) questions around the treatment of parametric uncertainty in the process models.

In terms of choice of process-based models, there were a few reasons why we selected the ones we did. First, in the Great Lakes Intercomparison Project (Mai et al., 2022), HYMOD was one of if not the best performing process model of all those tested, justifying its use in our study as a benchmark model. It performed second best across all process models in terms of KGE for streamflow, and it performed best among process models for actual evapotranspiration (AET). We selected the SAC-SMA model because of its widespread use in the United States by NOAA – it is the core hydrologic model used by River Forecast Centers in the Hydrologic Ensemble Forecasting System (HEFS; Demargne et al., 2014). We also found in Wi and Steinschneider (2022) that AET from SAC-SMA matched the seasonal pattern of MODIS-derived AET well across California. Similar to SAC-SMA, HBV is also an extremely popular model (Seibert and

Bergström, 2022), is used for operational forecasting in multiple countries (Olsson and Lindstrom, 2008; Krøgli et al., 2018), and performs very well (and at times the best) in hydrologic model intercomparison projects (Breuer et al., 2009; Plesca et al., 2012; Beck et al., 2016, 2017). We now provide this justification in the revised manuscript.

As to the issue of parametric uncertainty in the process models, the reviewer brings up an excellent point. We have revised our work to address this line of inquiry in three ways. First, we have revised our introduction to better address the issue of uncertainty in process-based models, and its implications for uncertainty in projections of future hydrology under climate change (see our response to a comment further below for more detail).

Second, we have altered the way that we train the process models. Whereas before we only conducted a single training trial, now we perform ten separate training trials under different random starts of the generic algorithm, similar to our approach for the DL models. We now present all results in the main manuscript as the ensemble mean of those 10 training trials for the process models.

Finally, and most directly in response to the reviewer's comment, we have used the ten training trails for each process model to explore how parametric uncertainty influences our interpretation of changes in hydrologic metrics, both between the process models and between the DL and process models. These results, shown in the Supporting Material (Figure S7) and referenced in the main manuscript, reveal three important insights:

- For long-term mean average flow (Q.AVG), high flows (FHV), and seasonal streamflow timing (COM), the uncertainty in the change of these metrics under warming for either Priestley-Taylor or Hamon PET linked to process-model parametric uncertainty explains much of, but not all of, the differences across process models.
- For changes in low flows (FLV) under warming, parametric uncertainty is quite large, but still does not explain the differences we see between SAC-SMA and the other two process models (HYMOD and HBV) under warming.
- When we compare changes in all statistics estimated by the DL models versus the process models, none of our conclusions in the main article change in terms of how the DL models fundamentally differ from the process models in their predictions under warming for some statistics (Q.AVG, FHV, and COM). That is, the parametric uncertainty in the process models is not large enough to explain the differences we see from the DL model predictions under warming.

We have revised both the Methods and the Results section of our manuscript to reflect these changes, and to highlight some of the key insights listed above.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A. (2016), Global-scale regionalization of hydrologic model parameters, Water Resour. Res., 52, 3599–3622, doi:10.1002/2015WR018247.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models (2017), Hydrol. Earth Syst. Sci., 21, 2881–2903, https://doi.org/10.5194/hess-21-2881-2017.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Adv. Water Resour., 32, 129–146, https://doi.org/10.1016/j.advwatres.2008.10.003, 2009.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., et al. (2014). The science of NOAA's operational hydrologic ensemble forecast service. Bulletin of the American Meteorological Society, 95(1), 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

Krøgli, I. K., Devoli, G., Colleuille, H., Boje, S., Sund, M., and Engen, I. K.: The Norwegian forecasting and warning service for rainfall- and snowmelt-induced landslides, Nat. Hazards Earth Syst. Sci., 18, 1427–1450, https://doi.org/10.5194/nhess-18-1427-2018, 2018.

Olsson, J., and Lindstrom, G. (2008), Evaluation and calibration of operational hydrological ensemble forecasts in Sweden Journal of Hydrology, 350 (1–2), 14-24.

Plesca, I., Timbe,E., Exbrayat, J.F., Windhorst, D., Kraft, P., Crespo, P., Vachéa, K.B., Frede, H.G., and Breuer,L. (2012). Model intercomparison to explore catchment functioning: Results from a remote montane tropical rainforest, Ecol. Model., 239, 3–13.

Seibert, J. and Bergström, S. (2022). A retrospective on hydrological catchment modelling based on half a century with the HBV model, Hydrol. Earth Syst. Sci., 26, 1371–1388, https://doi.org/10.5194/hess-26-1371-2022.

Choice of model performance metrics: The choice of the performance metrics is also not very clear to me. I can imagine that the inadequate partitioning of evaporative fluxes might especially affect the mean and the low flows, but what is the rationale for examining high flows and the seasonality of the flows? This needs to be clarified. There is also a discrepancy in how low and high flows are defined (98th percentile and the 30th percentile) that also needs clarification.

Our interpretation of the reviewer's comment is that they are questioning the inclusion of certain metrics of hydrologic response under warming (high flows, seasonal streamflow timing), since there is an expectation that those responses will not be impacted as strongly by evapotranspiration (which is a major focus of this article). We understand this concern, and when designing the study, we also debated the inclusion of other response metrics that may not be as impacted by ET. However, we ultimately decided that a broad readership may want to see other hydrologic responses under warming beyond just those related to long-term mean daily flow or low flows, especially because these types of comparative studies of DL and process models under warming conditions are relatively absent in the literature.

As proof of this, we point the reviewer to the requests of Reviewer #2, who was dissatisfied with the scope of changes we considered in our work and requested that we consider additional metrics or patterns of streamflow change beyond those we initially considered. Specifically, Reviewer #2 asked that we provide a more thorough treatment of snow accumulation and melt processes and their responses to warming across the suite of models in our work. This requested analysis is related to both metrics of seasonal streamflow timing (COM) and high flows (FHV).

To provide a more thorough analysis that would be of interest to a broad readership interested in using DL rainfall-runoff models under climate change, we have opted to be more inclusive in the metrics included in our analysis rather than more restrictive. Accordingly, we have chosen to include additional analysis related to the timing of flows during the snow accumulation and melt season, in response to Reviewer #2's concern (please see our response to their major comments for more details). However, we have taken this reviewer's comment into consideration when making these adjustments, seeking to control the degree to which the scope of our analysis grows beyond its core focus on streamflow losses and changes in temperature and PET. In addition, in direct response to the concerns expressed in this comment, we have added text justifying the inclusion of metrics beyond those that might be strongly impacted by PET changes.

Finally, with respect to the definition of high flows (top 2%) and low flows (bottom 30%), these are common metrics defined here (Yilmaz et al. 2008) and used in many intercomparison studies (Frame et al., 2022; Gauch et al., 2021a; Klotz et al., 2022; Kratzert et al., 2021). We now make a note of this in our revised manuscript.

Frame, J.M., Kratzert, F., Gupta, H.V., Ullrich, P., & Nearing, G.S. (2022). On Strictly enforced mass conservation constraints for modeling the Rainfall-Runoff process. Hydrological Processes, 37, e14847, https://doi.org/10.1002/hyp.14847.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021a). Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network. Hydrology and Earth System Sciences, 25, 2045-2062. https://doi.org/10.5194/hess-25-2045-2021

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022.

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging in multiple meteorological data sets with deep learning for rainfall-runoff modeling. Hydrology and Earth System Sciences, 25(5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021.

Yilmaz, K. K., Gupta, H. V., and Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, 1–18.

Detailed comments

Line 28 and elsewhere I would not really call the hydrological models used in this study as process-based. These are conceptual hydrological models that require extensive calibration and that can be very physically-unrealistic as well.

We would argue that there is a distinction between 'process-based' and 'physical/mechanistic' models, and that it is reasonable to refer to lumped, conceptual hydrologic models like HBV, SAC-SMA, and HYMOD as process-based but not physical/mechanistic. A 'process-based' model is one that is attempting to represent process. Lumped, conceptual rainfall-runoff models are attempting to represent hydrologic processes (e.g., infiltration, interflow, percolation, baseflow) explicitly through their structural equations, albeit at a coarse watershed scale. What distinguishes physical or mechanistic hydrologic models is that they attempt to explicitly solve for processes like mass, momentum, and energy conservation, often at high spatial resolution, and through physics-derived equations for these processes. We note that conceptual hydrologic models can be developed in a distributed fashion (i.e., applied to many small HRUs), with water then routed across HRUs into the stream (see Wi and Steinschneider, 2022). Therefore, the real distinguishing characteristic of a physical or mechanistic model is the attempt to represent and solve conservation equations for mass, energy, and momentum derived more directly from physics, rather than conceptual equations that abstract key hydrologic processes into a set of structural equations.

With that distinction made, we note that physical/mechanistic hydrologic models often struggle to represent hydrologic processes well at the watershed scale. This is shown many times in the literature (e.g., see Towler et al., 2023; Yan et al., 2023), and summarized nicely in Clark et al. (2017):

> "The trend towards "hyper" resolution land models (Wood et al.,2011) ,e.g., 1 km or 100 m over large geographical domains, emphasizes the need for general parameterizations of hydrological processes on this scale. However, this is still an unsolved problem: we do not have firm evidence that the structure and parameter values of our element-scale equations correspond to hydrologic reality at those scales. One of the most important causes of this difficulty is the spatial heterogeneity in the initial and boundary conditions, and in the material properties of the medium."

That is, physical/mechanistic hydrologic models are actually not all that "physical/mechanistic", because the processes coded into those models often cannot correctly represent sub-grid spatial heterogeneity in the landscape (e.g., preferential flow paths; see Blöschl, 2022 for a good summary) and how that heterogeneity at a sub-grid scale influences system-wide hydrologic response at a watershed scale. As evidence of this in our case study domain, we note that in Mai et al. 2022 (the Great Lakes runoff intercomparison analysis on which our paper builds), the conceptual hydrologic models often outperform the physical/mechanistic models.

Given our position above, we do think its appropriate to refer to the SAC-SMA, HYMOD, and HBV models used in this work as process models, although we agree that it is important to emphasize that they are conceptual models. Therefore, we have revised the text in several places to make this distinction clear.

Blöschl, G.: Flood generation: process patterns from the raindrop to the ocean, Hydrol. Earth Syst. Sci., 26, 2469–2480, https://doi.org/10.5194/hess-26-2469-2022, 2022.

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, Hydrol. Earth Syst. Sci., 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, 2017.

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., and Zhang, Y.: Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States, Hydrol. Earth Syst. Sci., 27, 1809–1825, https://doi.org/10.5194/hess-27-1809-2023, 2023.

Yan, H., Sun, N., Eldardiry, H., Thurber, T. B., Reed, P. M., Malek, K., et al. (2023). Large ensemble diagnostic evaluation of hydrologic parameter uncertainty in the Community Land Model Version 5 (CLM5). Journal of Advances in Modeling Earth Systems, 15, e2022MS003312. https://doi.org/10.1029/2022MS003312

Wi, S., & Steinschneider, S. (2022). Assessing the physical realism of deep learning hydrologic model projections under climate change. Water Resources Research, 58, e2022WR032123. https://doi.org/10.1029/2022WR032123

Line 32 and elsewhere: The term water loss is rather unclear. Please clarify and use a consistent term for that throughout the manuscript.

In most instances, when we are referring to water loss we mean evaporative water loss, i.e., water that is removed from the watershed via evapotranspiration. However, DL models do not explicitly represent or distinguish water loss from evaporative water loss, and so water loss more generally means water that enters the watershed via precipitation but never contributes to streamflow, i.e., water that gets ‘lost’ to some type of terminal sink. This sink could be evapotranspiration, but it also could water lost to deep groundwater that never interacts with surface water system, or water that is abstracted from the ground and exported out of the watershed via human activities.

To better clarify this term, we have made two revisions to the manuscript. First, in the abstract, we explicitly state ‘evaporative water loss’ when we mean water lost from the watershed due to evapotranspiration, and use ‘water loss’ more generally to mean any water that enters the watershed but does not contribute to streamflow. Then, early in the Introduction, we now define water loss explicitly, and also introduce the term evaporative water loss in the context of more water lost due to higher evapotranspiration:

Based on past literature, WS22 posited that in non-glaciated regions, physically plausible hydrologic responses should show an increase in water loss, defined as water that enters the watershed via precipitation but never contributes to streamflow because it is 'lost' to a terminal sink. Specifically, WS22 assumed that evaporative water loss should increase and annual average streamflow should decline compared to a baseline simulation due to increases in potential evapotranspiration (PET) with warming (and no changes in precipitation).

We note that we were asked by Reviewer 2 to substantially shorten our abstract, and so to respect that request, we do not provide this definition for water loss in the abstract.

Line 42 and elsewhere: Is actual evapotranspiration meant here? Please clarify

Throughout the manuscript, we now consistently use the term actual evapotranspiration when we refer to the quantity of water that is removed from a surface due to the processes of evaporation and transpiration.

Line 45-47: At this point the application of national of LSTM in addition to the regional LSTM sounds rather inconsistent. Please clarify here the objective for that.

We have revised the abstract to clarify the objective of including the National LSTM in our study design.

We also explore similar responses using a National LSTM fit to 531 watersheds across the United States to assess how the inclusion of a larger and more diverse set of basins influences signals of hydrologic response under warming.

Line 48 and elsewhere: If the statistical test was not performed, omit term "significantly". Use term "considerably" instead.

We have changed our use of the word "significantly" here and throughout the manuscript to avoid implying the use of statistical testing.

Line 52 and elsewhere: Average is a very ambiguous term. Is it mean or median? Is it daily or annual flows? Please use a clearer term throughout the manuscript.

Here and elsewhere, we are referring to long-term mean daily flows. We have revised the text in the abstract to make this clear, as well as throughout the manuscript, so that our presentation is specific and consistent.

Line 58: Smaller than what? It would be more helpful to include more quantitative results in the Abstract (e.g., Lines 63-64).

We are a little confused by the first part of this comment, as the sentence in question states that the changes in high flows and streamflow timing from the DL models are smaller than the changes predicted by the process models. However, we have revised the abstract to quantify the

degree of differences we see in the results, particularly with respect to changes in the long-term average of daily flows, which is a core focus of this paper. We do not include numerical values for some of the other statistics because they span different units (days for COM and % for FHV and FLV), which we think would be cumbersome to explain in the abstract (and we were asked by Reviewer 2 to shorten the abstract, which we have tried to do).

Introduction: I feel that introduction is very one-sided focusing on the purely deep learning models and not paying enough attention to the problems that conceptual hydrological models have when simulating future (e.g., Merz et al., 2011 https://doi.org/10.1029/2010WR009505; Wallner and Haberlandt, 2015 https://doi.org/10.1002/hyp.10430). It also completely omits the field of hybrid models (Jiang et al., 2020 https://doi.org/10.1029/2020GL088229; Höge et al., 2022 https://doi.org/10.5194/hess-26-5085-2022) that in my opinion might be more fit for future predictions than the deep learning models or even their constrained variants. Moreover, the Introduction is very much focused on the previous study by the authors, but fails to clearly distinguish the difference between them.

We agree that it would be helpful to address broader uncertainties in hydrologic modeling under climate change, and their impact on our ability to develop good baselines we can use to evaluate the fidelity of DL-based predictions under climate change. We have made revisions in the Introduction to address this concern (see excerpt below), and also now address this in our results as well (detailed more thoroughly in our response to a major comment above).

> *It is challenging to assess the physical plausibility of DL-based hydrologic projections under substantially different climate conditions, because there are no future observations against which to compare. This challenge is exacerbated by significant uncertainty in process model projections under alternative climates, which makes establishing reliable benchmarks difficult. Future process model-based projections can vary widely due to both parametric and structural uncertainty (Bastola et al., 2011; Clark et al., 2016; Melsen et al., 2018), and even for models that exhibit similar performance under historical conditions (Krysanova et al., 2018). Assumptions around stationary model parameters are not always valid (Merz et al., 2011; Wallner and Haberlandt, 2015), and added complexity for improved process representation is not always well supported by data (Clark et al., 2017; Towler et al., 2023; Yan et al., 2023). Together, these challenges highlight the difficulty in establishing good benchmarks of hydrologic response under alternative climates against which to compare and evaluate DL-based hydrologic projections under climate change.*

In terms of introducing hybrid models, we have revised the introduction to briefly introduce such models. However, as stated and justified in our comment above, these hybrid models are not the focus of this work. Therefore, we relegate a longer (and significantly revised) discussion of these methods to the Discussion and Conclusion (again, please see our response to a comment above).

In addition, we have significantly reduced the focus on our previous article (SW22) in the Introduction (by ~30%), with the goal of only highlighting the most salient features of that article needed to set up the present work. We will note that in the Introduction, we do clearly distinguish the difference between the focus of SW22 and of this study (in the paragraph starting

"For all models considered in WS22"), and this distinction is further emphasized in the first paragraph of the Methods section.

Clark, M.P., Wilby, R.L., Gutmann, E.D. et al. Characterizing Uncertainty of the Hydrologic Impacts of Climate Change. Curr Clim Change Rep 2, 55–64 (2016). https://doi.org/10.1007/s40641-016-0034-x

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J. J. F., Clark, M. P., Uijlenhoet, R., and Teuling, A. J. (2018). Mapping (dis)agreement in hydrologic projections, Hydrol. Earth Syst. Sci., 22, 1775–1791, https://doi.org/10.5194/hess-22-1775-2018.

Bastola S., Murphy C., Sweeney J. (2011). The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments. Adv Water Resour., 34, 562–76.

Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F. and Kundzewicz Z.W. (2018) How the performance of hydrological models relates to credibility of projections under climate change, Hydrological Sciences Journal, 63:5, 696-720, DOI: 10.1080/02626667.2018.1446214

Wallner, M., and Haberlandt, U. (2015), Non-stationary hydrological model parameters: a framework based on SOM-B. Hydrol. Process., 29, 3145–3161. doi: 10.1002/hyp.10430.

Merz, R., Parajka, J., and Blöschl, G. (2011), Time stability of catchment model parameters: Implications for climate impact analyses, Water Resour. Res., 47, W02531, doi:10.1029/2010WR009505.

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, Hydrol. Earth Syst. Sci., 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, 2017.

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., and Zhang, Y.: Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States, Hydrol. Earth Syst. Sci., 27, 1809–1825, https://doi.org/10.5194/hess-27-1809-2023, 2023.

Yan, H., Sun, N., Eldardiry, H., Thurber, T. B., Reed, P. M., Malek, K., et al. (2023). Large ensemble diagnostic evaluation of hydrologic parameter uncertainty in the Community Land Model Version 5 (CLM5). Journal of Advances in Modeling Earth Systems, 15, e2022MS003312. https://doi.org/10.1029/2022MS003312

Line 121: Does ET mean actual evapotranspiration here? This is not clear and I think this acronym is not used later anymore. Please revise.

In response to the reviewer's previous comment, we have significantly reduced the Introduction's focus on the details of our past study, and in the process, no longer refer to ET here (which did previously mean actual evapotranspiration). However, as stated above, we have gone through the manuscript and now consistently refer to actual evapotranspiration where appropriate.

Line 147-150: The energy-based methods (although indisputably more realistic) are also based on empirical relationships, are they not?

That is true, at least for some methods. For instance, the Penman Monteith equation was derived directly from an energy balance equation, albeit with some simplifications, whereas the Priestley-Taylor method replaces the aerodynamic terms in Penman Monteith with an empirically derived constant α. We have revised a line directly below the one referenced in this comment to integrate this point into our arguments:

> *Energy budget-based methods, while imperfect and at times empirical (Greve et al. 2019; Liu et al., 2022), account for some or all of these factors in ways that are generally consistent with their causal impact on PET, while temperature-based methods estimate PET using strictly empirical relationships based largely or entirely on temperature.*

Line 172: Evaporative water loss instead? This term is unclear.

Correct, we have made this change.

Line 204: I do not think that the reference to CAMELS-GB is appropriate here. It was not created with the sole purpose to benchmark deep learning models, nor does it actually benchmark them. Please revise.

We have removed this line and the associated reference to CAMELS-GB.

Line 243: Acronym AET was never later in the manuscript. Consider omitting it.

The acronym AET was used 23 separate times in the original article (26 in the revised article). Based on this persistent usage, we have decided to retain use of this acronym.

Line 243-245: I think it is worth to note here that GLEAM can be also associated with considerable uncertainties. Therefore, validation using this product might be questionable as well.

We agree that its worthwhile to point out that the actual evapotranspiration from GLEAM will have considerable uncertainties (as will all watershed-scale AET products), although its use of remotely sensed data makes it a useful benchmark to compare AET from our rainfall-runoff models. We have edited the text to make this point.

While AET from GLEAM is still uncertain, it provides a useful, independent, remote-sensing based benchmark against which to compare rainfall-runoff model estimates of AET.

Line 253: It is not clear what is meant by hydrological losses here and if this term is different from the term "water losses" used earlier. Please clarify.

We have changed this to "water losses", a term which is now defined earlier in the manuscript as per our response to a comment above.

Figure 2: A much more comprehensive caption describing every step and every acronym presented in the Figure is needed.

We have significantly expanded the caption for this figure, in order to describe each step in the experimental design. We have also defined acronyms in the caption (e.g., Q, AET, PET) that are not already defined in the figure itself (e.g., trash cell, TC). To avoid an excessively long caption, we do not spell out each DL and conceptual model acronym, but we define them clearly in the caption as deep learning models or conceptual, process-based models.

Line 337: For the purity of the test, I suggest that all models (conceptual and deep learning models) should be trained on the same objective function.

We have taken the reviewer's suggestion and retrained all the process models with the objective to minimize MSE, rather than maximize the KGE, in order to match the objective function used by the DL models. We then recreated all figures and results associated with this work. The results did not change much compared to the original process models used in our original submission, and so this change did not alter any of the conclusions drawn in the study.

Line 351: It would be helpful to mention around here how many of these catchments overlap with the Great Lake catchment sample. Even better would be to indicate them in Figure 1.

We have now added this information to this section.

There are 29 CAMELS watersheds located within the Great Lakes basin, and 17 of those 29 watersheds were also used in the training and testing sets for the Great Lakes LSTM.

We now also highlight the 17 CAMELS gauges that are also part of the Great Lakes gauging set in Figure 1.

Line 412: This statement requires a reference

We have revised this text to provide a more nuanced explanation of when these assumptions likely hold:

Here, the ReLU activation ensures that any water in the trash cell (h_D) which exceeds PET at time t is added to the streamflow prediction $y[t]$, but the streamflow prediction is

the same as the original MC-LSTM (Eq. 5) if water in the trash cell is less than PET. This approach assumes that the maximum allowable water lost from the system cannot exceed PET, and therefore ignores other potential terminal sinks (e.g., inter-basin lateral groundwater flows; human diversions and inter-basin transfers). This assumption is more strongly supported in moderately-sized (> 200 km^2), low-gradient, non-arid watersheds where inter-basin groundwater flows are less impactful (Fan 2019; Gordon et al., 2022), such as the Great Lakes basins examined in this work. However, we discuss the potential to relax the assumptions of the MC-LSTM-PET model in Section 5.

In addition, we have significantly revised out Discussion and Conclusion section to revisit these assumptions, and discuss ways they might be relaxed or improved in future work (please see our response to a major comment above for more detail).

Line 435: The rationale for using both KGE and NSE as performance metrics is unclear to me

Both KGE and NSE (as well as FLV and FHV) are commonly used in many inter-comparison studies (Frame et al., 2022; Gauch et al., 2021a; Klotz et al., 2022; Kratzert et al., 2021). We now make a note of this when introducing the metrics.

Line 441 GLEAM estimates are not observations and can be associated with large uncertainties too.

We have edited the text (here and elsewhere) to avoid referring to GLEAM AET as "observed" AET. In addition, we have noted the uncertainty in AET from GLEAM when we introduce this product, as stated in a response to a previous comment.

Line 449: It is not clear how fraction of snowfall was adjusted. Please clarify. Moreover, please use full terms and not the acronyms that were not previously introduced.

We have added information regarding how all of these static features were adjusted in a new section in the Supporting Information, which we now reference in this line. The acronyms were previously introduced in Table 1, which we now state explicitly in this line to help the reader find the full definition for each feature.

Line 497-501: A table with the overview of all scenarios and setups would be helpful.

We have added a new Table 2 that describes each of the scenarios developed in this work.

Line 507-511: It is not clear what is meant by "average" here. Please clarify. Consider avoiding using so many acronyms, the manuscript is oversaturated with them, making it hard to understand.

We now more clearly define what is meant by average here, i.e., the long-term mean of daily streamflow across the entire series.

In terms of acronyms, we recognize the benefits of reducing their usage, although this must be balanced against commonly understood acronyms and redundancy throughout the manuscript due to repeated long phrases. To address this concern, we tallied the number of acronyms used in the paper and the number of times they were used:

- Deep learning – DL (84)
- Potential evapotranspiration - PET (174)
- Actual evapotranspiration - AET (26)
- Physics-informed machine learning - PIML (15)
- United States Geological Survey – USGS (1)
- Water Survey Canada – WSC (1)
- Regional Deterministic Reanalysis System v2 – RDRS-v2 (3)
- Canadian Precipitation Analysis – CaPA (1)
- digital elevation model – DEM (1)
- Global Soil Dataset for Earth System Models – GSDE (1)
- Global Land Evaporation Amsterdam Model - GLEAM (11)
- Great Lakes Runoff Intercomparison Project Phase 4 - GRIP-GL (3)
- Hydrologiska Byråns Vattenbalansavdelning – HBV (7)
- Sacramento Soil Moisture Accounting - SAC-SMA (7)
- Long short-term memory network – LSTM (76)
- Mass-conserving long short-term memory network - MC-LSTM (23)
- Mass-conserving long short-term memory network + PET limit - MC-LSTM-PET (30)
- Nash Sutcliffe Efficiency – NSE (8)
- Kling Gupta Efficiency – KGE (20)
- Percent Bias – PBIAS (9)
- Long-term mean daily flow – AVG.Q (13)
- Low flow bias – FLV (19)
- High-flow bias – FHV (16)
- Center of mass – COM (5)

Based on this analysis, and considering which acronyms are very commonly used in the literature, we removed the following seven acronyms from the manuscript entirely: USGS, WSC, RDRS-v2, CaPA, DEM, GSDE, GRIP-GL.

Table 2: I miss here the timing error introduced earlier.

The metrics used to evaluate model performance (shown in Figure 2) are described Section 3.2, and are slightly different than the metrics used to evaluate how models predict streamflow will change under warming (described in Section 3.3). The COM statistic is not one used to evaluate model performance, and so was not shown in Table 2.

Line 617-618: This statement requires a reference and it would be helpful if it will be presented in a more quantitative way.

We have added the relevant citation (Allen et al., 1998) for this point, and also now point out that these changes in Rn are generally less than 5% across all sites.

Figure 5 and 6: Provide the names and the locations of the selected watersheds. It would also be helpful to indicate them on Figure 1 to show their geographical location.

We now highlight these sites in Figure 1. In addition, we have added details for each of these locations in the Supporting Information (new Table S3).

Line 657: This is not really the timing of streamflow per se. It is rather a seasonality of the flow. Please clarify that.

We note that the center of mass statistic is often referred to in the literature as a streamflow timing statistic, but we agree this is specifically in terms of seasonal timing of flows. Therefore, in the manuscript we specify that the COM statistic is a measure of seasonal timing.

Line 681-690: This part is rather confusing and difficult to read. Please revise.

We have tried to make some educated guesses where the issues of confusion might be arising, and have made some edits to hopefully resolve these issues. These include splitting up sentences to make each shorter and easier to follow, and some additional clarifying text. We also note that we now introduce the 29 and 17 basins mentioned in this paragraph earlier when introducing the National LSTM, which also might help address the confusion here.

Line 696-698: This is rather vague. Please provide a more quantitative assessment. Moreover, nothing is mentioned about huge differences between the simulations of the conceptual models and how this affects the reliability of the baseline chosen in this experiment.

We have replaced the words “moderately larger” and “much larger” with the actual percent differences in the median decline in AVG.Q from the National LSTM and the median predictions of loss under the process models.

Please see our response to the major comment above in terms of the reliability of the baseline conceptual models.

Figure 8: Please explain all the acronyms in the caption.

We have provided descriptions for the acronyms for all streamflow statistics, and also deleted the acronym for “deep learning”. We also made the same changes to Figure 7.

Line 720-721 This part is rather confusing, please revise.

We have revised this sentence by splitting it into two separate sentences, and providing additional clarifying text to each, in order to help readers better understand what is being described.

Editorial comments

Line 27: state-of-the-art

This phrase was removed from the revised abstract.

Line 30: under exacerbating climate change

In other edits meant to shorten the abstract, we have removed any qualifier ahead of the phrase "climate change".

Line 32: overestimation

To shorten the abstract, we have changed this to: "temperature-based PET methods overestimate evaporative water loss"

Line 170: similarly large

We have made this change.

Line 334: by drainage area

This has been corrected.

Line 656: consistent

This has been corrected.

Line 657: changes in high flows

We have changed this to "more substantial declines in high flows", since we think specifying declines here is more specific and useful than "changes". We imagine that the reviewer might be taking issue with the words "larger" and "declines" being used together, hence our suggested edit.

Line 828: considerable errors?

In the process of significantly revising the Discussion and Conclusion section of the paper, this line was deleted from the manuscript.