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:

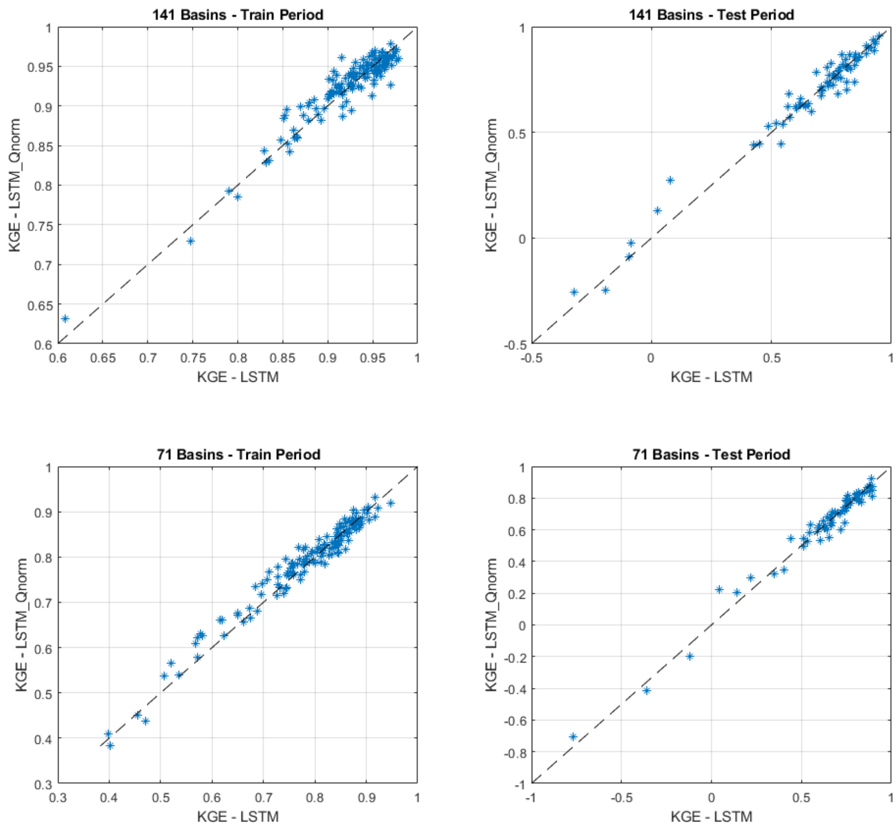
> *A <u>potential implication of this finding might be</u> that these models can produce physically plausible streamflow predictions under new climate conditions.*

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 <u>NSE and FLV (as well as KGE in the training period)</u>, 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.

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[1], **Scott Steinschneider**[1]

[1]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA

Abstract

Deep learning (DL) rainfall-runoff models ~~have recently emerged as state-of-the-science tools for hydrologic prediction that~~ outperform ~~conventional,~~ process-based models in a range of applications. However, it remains unclear whether ~~deep learning~~DL models can produce physically plausible projections of streamflow under ~~significant amounts of~~ climate change. We investigate this question ~~here, focusing specifically on~~through a sensitivity analysis of modeled responses to increases in temperature and potential evapotranspiration (PET), with other meteorological variables left unchanged. Previous research has shown that temperature-based PET methods ~~to estimate PET lead to~~ overestimate~~s~~ evaporative ~~of~~ water loss ~~in rainfall-runoff models~~ under warming~~,~~ ~~as~~ compared to energy budget-based PET methods. We therefore ~~Consequently, we assess the reliability of streamflow projections under warming by comparing projections with both temperature-based and energy budget-based PET estimates,~~ assume~~ing~~ that reliable streamflow ~~projections~~ responses to warming should exhibit less evaporative water loss when forced with smaller, ~~(~~energy budget-based ~~) projections of future~~ PET compared to temperature-based PET. We conduct this assessment using three conceptual, process-based rainfall-runoff models ~~rainfall-runoff models~~ and three ~~deep learning~~DL models, trained and tested across 212 watersheds in the Great Lakes basin. The ~~deep learning~~DL models include a ~~regional~~ Long Short-Term Memory network (LSTM), a mass-conserving LSTM (MC-LSTM) ~~that preserves the water balance~~, and a novel variant of the MC-LSTM that also respects the relationship between PET and evaporative water loss (MC-LSTM-PET). After validating models against ~~We first compare~~ historical streamflow and actual ~~watershed-scale~~ evapotranspiration, ~~predictions from all models under spatial and temporal validation, and also assess model skill in estimating watershed-scale evapotranspiration. We then~~we force all models with scenarios of warming, historical precipitation, and both temperature-based (Hamon) and energy budget-based (Priestley-Taylor) PET, and compare their ~~projections~~ responses ~~for changes~~ in ~~average~~ long-term mean daily flow, ~~as well as~~ low flows, high flows, and seasonal streamflow ~~seasonal~~ timing. ~~Finally, w~~We also explore similar ~~projections~~

responses using a National LSTM fit to ~~to a broader set of~~ 531 watersheds across the ~~contiguous~~ United States to assess how the inclusion of a larger and more diverse set of basins influences signals of hydrologic response under warming. The main results of this study are as follows:

1. The three Great Lakes ~~deep learning~~DL models ~~significantly~~ substantially outperform all process models in streamflow estimation~~ under spatiotemporal validation, with only small differences between the DL models~~. The MC-LSTM-PET also matches the best process models and outperforms the MC-LSTM in estimating actual evapotranspiration~~ under spatiotemporal validation~~.
2. All process models show a downward shift in long-term mean daily ~~average~~ flows under warming, but ~~this~~ median shifts ~~is~~ are ~~significantly~~ considerably larger under temperature-based PET (17% to 25%) ~~estimates~~ than energy budget-based PET (-6% to -9%). The MC-LSTM-PET model exhibits similar differences in water loss across the different PET forcings~~, consistent with the process models~~. ~~However~~Conversely, the LSTM exhibits unrealistically large water losses under warming ~~as compared to the process models~~ using Priestley-Taylor PET (20%), while the MC-LSTM is relatively insensitive to PET method.
3. ~~All deep learning~~DL models exhibit smaller changes in high flows and ~~streamflow~~ seasonal timing of flows as compared to the process models while ~~deep learning~~DL ~~projections~~ estimates of low flows are ~~all very consistent and~~ within the range ~~projected~~ estimated by the process models.
4. Like the Great Lakes LSTM, the National LSTM also shows unrealistically large water losses under warming (25%), but ~~. However, when compared to the Great Lakes deep learning models, projections from the National LSTM were~~it is more stable when many inputs ~~were~~ are changed under warming and better aligns~~ed~~ with process model ~~projections~~ responses for ~~streamflow~~ seasonal timing of flows. ~~This suggests that the addition of more, diverse watersheds in training does help improve climate change projections from deep learning models, but this strategy alone may not guarantee reliable projections under unprecedented climate change.~~

Ultimately, the results of this ~~work~~ sensitivity analysis suggest that physical considerations regarding model architecture and input variables ~~are~~ may be necessary to promote the physical realism of deep learning-based hydrologic projections under climate change.

Keywords

Deep learning, machine learning, Long Short-Term Memory network, LSTM, Great Lakes, climate change, rainfall-runoff

1. Introduction

Rainfall-runoff models are used throughout hydrology in a range of applications, including retrospective streamflow estimation (Hansen et al. 2019), streamflow forecasting (Demargne et al., 2014), and prediction in ungauged basins (Hrachowitz et al., 2013). Work over the last few years has demonstrated that deep learning (DL) rainfall-runoff models (e.g., Long Short-Term Memory networks (LSTMs); Hochreiter and Schmidhuber, 1997) outperform conventional process-based models in each of these applications, especially when those DL models are trained with large datasets collected across watersheds with diverse climates and landscapes (Kratzert et al., 2019a,b; Feng et al., 2020; Ma et al., 2021; Gauch et al., 2021a,b; Nearing et al., 2021). For example, in one extensive benchmarking study, Mai et al. (2022) found that a regionally trained LSTM outperformed 12 other lumped and distributed process-based models of varying complexity in rivers and streams throughout the Great Lakes basin. These and similar results have led ~~many~~ some to argue that DL models represent the most accurate and extrapolatable rainfall-runoff models available (Nearing et al., 2022)~~most accurate and extrapolatable rainfall-runoff models available (Nearing et al., 2022)~~.

However, there remains one use case of rainfall-runoff models where the superiority of DL is unclear: long-term projections of streamflow under climate change. Past studies using DL rainfall-runoff models for hydrologic projections under climate change are rare (Lee et al., 2020; Li et al., 2022), and few have evaluated their physical plausibility (Razavi, 2021; Reichert et al., 2023; Zhong et al., 2023). A reasonable concern is whether DL rainfall-runoff models can extrapolate hydrologic response under unprecedented climate conditions, given that they are entirely data driven and do not explicitly represent the physics of the system. It is not clear *a priori* whether this concern has merit, because DL models fit to a large and diverse set of basins have the benefit of learning hydrologic response across climate and landscape gradients. In so doing, the model can, for example, learn hydrologic responses to climate in warmer regions and then transfer this knowledge to projections of streamflow in cooler regions subject to climate change induced warming. In addition, past work has shown that LSTMs trained only to predict streamflow have memory cells that strongly correlate with independent measures of soil moisture and snowpack (Lees et al. 2022~~1~~), suggesting that DL hydrologic models can learn fundamental hydrologic processes. A ~~corollary~~ potential implication ~~to~~ of this finding ~~is~~ might be that these models can ~~may~~ produce physically plausible streamflow predictions under new climate conditions.

It is challenging to assess the physical plausibility of DL-based hydrologic projections under ~~significantly~~ 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.

Recently, Wi and Steinschneider (2022) (hereafter WS22) ~~addressed this challenge directly, forwarding~~forwarded an experimental design to evaluate the physical plausibility of DL hydrologic responses to new climates, in which DL hydrologic models ~~fit to 15 watersheds in California and 531 catchments across the United States~~ were forced with historical precipitation and temperature, but with temperatures adjusted by up to 4°C. Based on past literature ~~(Cayan et al., 2001; Stewart et al., 2005; Kapnick and Hall, 2010; Lehner et al., 2017; McCabe et al., 2017; Dierauer et al., 2018; Mote et al., 2018; Woodhouse & Pederson, 2018; Martin et al., 2020; Milly & Dunne, 2020; Rungee et al., 2021; Gordon et al., 2022; Liu et al., 2022)~~, WS22 posited that in non-glaciated regions, physically plausible hydrologic ~~projections~~ 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 ~~decline in total annual~~ average streamflow should decline compared to a baseline ~~historical~~ simulation~~,~~ due to increases in potential evapotranspiration (PET) with warming (and no changes in precipitation). Results showed that ~~the~~ an LSTM trained to the 15 watersheds in California often led to misleading increases in annual runoff under ~~significant~~ warming, while this phenomenon was less likely (though still present) in ~~the~~ a DL model trained to 531 catchments across the United States.

WS22 also conducted their experiment with physics-informed machine learning (PIML) models~~, in which data-driven techniques are imbued with process-knowledge constructs~~ (Karpatne et al., 2017), ~~. WS22 focused on two PIML strategies for the smaller case study in California,~~ using process model output ~~(e.g., soil moisture, evapotranspiration (ET))~~ directly as input to the LSTM (similar to Konapala et al., 2020; Lu et al., 2021; Frame et al., 2021a)~~, and also~~or as additional target variables in a multi-output architecture. The former approach had some success in removing instances of increasing runoff ratio with warming, ~~but~~ although this ~~depended heavily on the accuracy of~~was dependent on the ~~the~~ process ~~-~~model used~~ET~~.

Other PIML approaches that more directly adjust the architecture of DL rainfall-runoff models may be better suited for improving long-term streamflow projections under climate change without requiring an accurate process-based model. For instance, Hoedt et al. (2021) introduced a mass conserving LSTM (MC-LSTM) that ensures cumulative streamflow predictions do not exceed precipitation inputs. Hybrid models present a related approach, where DL modules are embedded within process models structures (Jiang et al., 2020; Feng et al., 2022; Hoge et al., 2022; Feng et al., 2023a). In some cases, ~~T~~these~~is~~ architectural changes cane ~~slightly~~ degrade performance compared to ~~underperformed~~ a standard LSTM ~~when predicting out-of-sample extreme events~~ (Frame et al., 2021b; Feng et al., 2023b), but other times such changes can be beneficial (Feng et al., 2023a). ~~and s~~Some have argued that these physical constraints may inhibit the ability of DL models to learn biases in forcing data (Frame et al. 2022)~~). Still~~, but the benefits of ~~this~~ such mass conserving architectures have not been tested when employed under previously unobserved climate change.

For all models considered in WS22, a major focus was evaluating the direction of annual total runoff change in the presence of warming and no change in precipitation. However, that study did not consider the magnitude of runoff change and how it relates to projected changes in PET. As we argue below, this comparison provides a unique way to assess the physical plausibility of future hydrologic projections. Several studies have investigated the effects of different PET estimation methods on the magnitude of PET and runoff change in a warming climate (Lofgren et al., 2011; Shaw and Riha, 2011; Lofgren and Rouhana, 2016; Milly and Dunne, 2017; Lemaitre-Basset et al. 2022). Broadly, ~~this~~ these ~~work~~ studies have~~s~~ shown that temperature-based PET estimation methods (e.g., Hamon, Thornthwaite) ~~significantly~~ substantially overestimate increases in PET under warming as compared to energy budget-based PET estimation methods (e.g., Penman-Monteith, Priestley-Taylor), and consequently lead to unrealistic declines in streamflow under climate change. This is because the actual drying power of the atmosphere is driven by the availability of energy at the surface from net radiation, the current moisture content of the air, temperature (and its effect on the water holding capacity of the air and vapor pressure deficit), and wind speeds. 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. The latter approach works sufficiently well for rainfall-runoff modeling under historical conditions because of the strong correlation between temperature, net radiation, and PET on seasonal timescales, even though this correlation weakens considerably at shorter timescales (Lofgren et al., 2011). Under climate change, consistent and prominent increases are projected for temperature, but projected changes are less prominent or more uncertain for other factors affecting PET (Lin et al., 2018; Pryor et al., 2020, Liu et al. 2020). Consequently, temperature-based PET methods ~~significantly~~ substantially overestimate future projections of PET compared to energy budget-based methods (Lofgren et al., 2011; Shaw and Riha, 2011; Lofgren and Rouhana, 2016; Milly and Dunne, 2017; Lemaitre-Basset et al. 2022).

As argued by Lofgren and Rouhana (2016), the bias in PET and runoff that results from different PET estimation methods under warming provides a unique opportunity to assess the physically plausibility of hydrologic projections under climate change. In this study, we adopt this strategy for DL rainfall-runoff models ~~and forward an experimental design~~through a sensitivity analysis in which both conceptual, process-based and DL hydrologic models are trained with either temperature-based or energy budget-based estimates of PET, along with other meteorological data (precipitation, temperature). These models are then forced with the historical precipitation and temperature series, but with the temperatures warmed by an additive factor and PET calculated from the warmed temperatures using both PET estimation methods. We ~~anticipate~~ show that the process models 1) ~~will~~ exhibit similar performance in historical training and testing periods when using either temperature-based or energy budget-based PET estimates; but 2) ~~will~~ exhibit ~~significantly~~ substantially larger long-term mean streamflow declines under warming when using future PET estimated with a temperature-based method. If the DL rainfall-runoff models follow the same pattern, this would suggest that these models are able to learn the role of PET on evaporative water loss. However, if DL-based models estimate similarly ~~and~~ large long-term mean streamflow declines regardless of the method used to estimate and project PET, this would suggest that the DL models did not learn a mapping between PET and evaporative water loss. Rather, the DL models learned the historical (but non-causal) correlation between temperature and evaporative water loss, and then incorrectly extrapolated that effect into the future with warmer temperatures. ~~Such~~ We show this latter ~~an~~ outcome to be the case, ~~would~~which indicates that some degree of PIML ~~is~~ may be necessary to guide a DL model towards physically plausible projections under climate change~~, in contrast to previous arguments against the need for such physical constraints (Frame et al. 2022)~~.

We conduct the experiment above in a case study on 212 watersheds across the Great Lakes basin, using both standard and PIML-based LSTMs. We ~~hypothesize~~ show that a standard LSTM ~~will~~ produces unrealistic hydrologic ~~projections~~ responses to warming because it relies on historical and geographically pervasive correlations between temperature and PET to ~~project~~ estimate streamflow losses~~ under warming~~. We also ~~hypothesize~~ show that PIML-based DL models ~~will be~~are better able to relate ~~future projections of~~changes in temperature and PET to streamflow change, especially those PIML approaches that directly map PET to evaporative water loss in their architecture.

The primary goal of this work is to forward an experimental design that can be used to evaluate the suitability of DL rainfall-runoff models for hydrologic projections under climate change, in line with a recent call to design benchmarking studies that assess whether models are fit for specific purposes (Beven, 2023). The Great Lakes provides an important case study for this work, given their importance to the culture, ecosystems, and economy of North America (Campbell et al., 2015; Steinman et al., 2017). Projections of future water supplies and water levels in the Great Lakes are highly uncertain (Gronewold and Rood, 2019), in part because of uncertainty in future runoff draining into the lakes from a large contributing area (Kayastha et al. 2022), much of which is ungauged (Fry et al., 2013). Improved rainfall-runoff models that can regionalize across the entire Great Lakes basin are necessary to help address this challenge, and so an auxiliary goal of this work is to contribute PIML rainfall-runoff models to the Great Lakes Runoff Intercomparison Project Phase 4~~ (GRIP-GL)~~ presented in Mai et al. (2022). This study currently provides one of the most robust benchmarks comparing DL rainfall-runoff models to a range of process-based models, and so we design our experiment to be consistent with the data and model development rules outlined in ~~the GRIP-GL~~that intercomparison project.

2. Data

This study focuses on 212 watersheds draining into the Great Lakes and Ottawa River, which are all located in the St. Lawrence River basin (Figure 1).~~ We note that this region is of similar spatial scale to other benchmarking datasets for DL rainfall-runoff models (e.g., CAMELS-GB; Coxon et al., 2020).~~ For direct comparability to previous results from the Great Lakes Runoff Intercomparison Project~~GRIP-GL~~, all data for these watersheds are taken directly from the work in Mai et al. (2022) and include daily streamflow time series, meteorological forcings, geophysical attributes for each watershed, and auxiliary hydrologic fluxes. Daily streamflow were gathered from the U.S. Geological Survey~~ (USGS)~~ and Water Survey Canada~~ (WSC)~~ between January 2000 and December 2017. All streamflow gauging stations have a drainage area greater than or equal to 200 km^2 and less than 5% missing data in the study period. The watersheds are evenly distributed across the five lake basins and the Ottawa River basin, and they represent a range of land use/land cover types and degrees of hydrologic alteration from human activity. In the experiments described further below, 141 of the watersheds are designated as training sites, and the remaining 71 watersheds are used for testing (see Figure 1). In addition, the period between January 2000 to December 2010 is reserved for model training (termed the training period), and the period between January 2011 – December 2017 is used for model testing (termed the testing period).

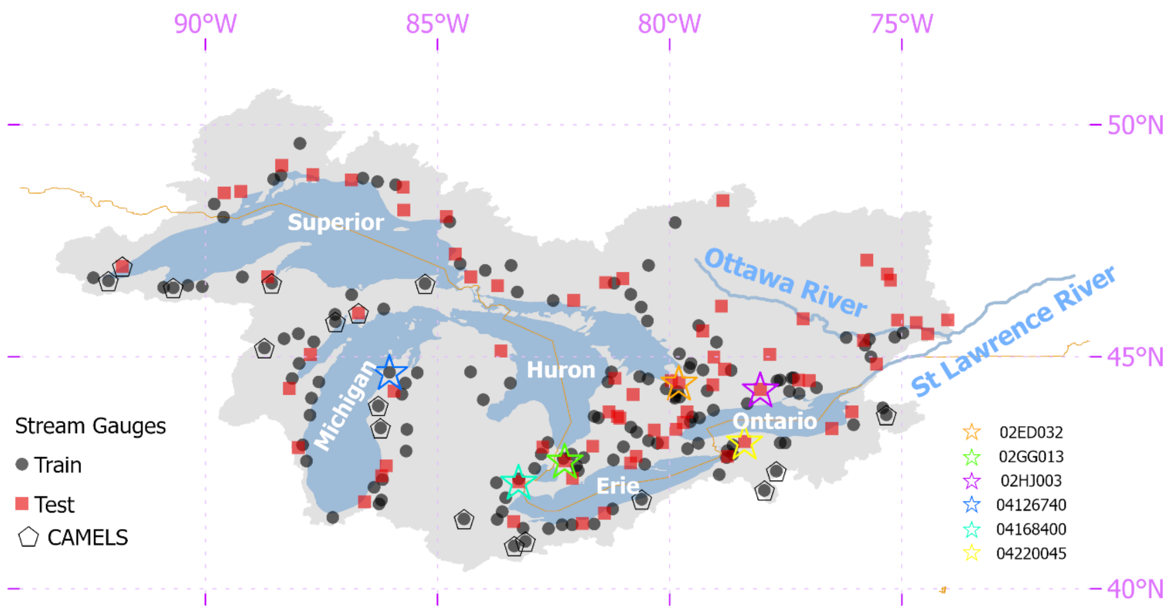


Figure 1. Great Lakes domain, with training and testing streamflow gauges used throughout this study. A subset of seventeen of these gauges that are also in the CAMELS database are highlighted, as are six sites used to present select results in Section 4.

Meteorological forcings are taken from the Regional Deterministic Reanalysis System v2 ~~(RDRS-v2)~~, which is an hourly, 10 km dataset available across North America (Gasset et al., 2021). Hourly precipitation, net incoming shortwave radiation (R_s), and temperature are aggregated into a basin-wide daily precipitation average, daily R_s average, and daily minimum and maximum temperature. We note that the precipitation data from the Regional Deterministic Reanalysis System v2 ~~RDRS-v2~~ is produced from the Canadian Precipitation Analysis ~~(CaPA)~~, which combines available surface observations of precipitation with a short-term reforecast provided by the 10 km Regional Deterministic Reforecast System. That is, the precipitation data is not model based, but rather is based on gauged data and spatially interpolated using information from modeled output.

Geophysical attributes for each watershed were collected from a variety of sources. Basin-average statistics of elevation and slope were derived from the HydroSHEDS dataset (Lehner et al., 2008), which provides a digital elevation model ~~(DEM)~~ with 3 arcsec resolution. Soil properties (e.g., soil texture, classes) were gathered from the Global Soil Dataset for Earth System Models (~~GSDE;~~ Shangguan et al., 2014), which is available at a 30 arcsec resolution. Land cover data at a 30 m resolution and based on Landsat imagery from 2010-2011 were derived from the North American Land Change Monitoring System (NALCMS, 2017). These geophysical datasets were used to derive basin-averaged attributes for each watershed, listed in Table 1.

Table 1. Watershed attributes used in the deep learning models developed in this work (adapted from Mai et al., 2022).

Attribute	Description
p_mean	Mean daily precipitation
pet_mean	Mean daily potential evapotranspiration
aridity	Ratio of mean PET to mean precipitation
t_mean	Mean of daily maximum and daily minimum temperature
frac_snow	Fraction of precipitation falling on days with mean daily temperatures below 0°C
high_prec_freq	Fraction of high-precipitation days (= 5 times mean daily precipitation)
high_prec_dur	Average duration of high-precipitation events (number of consecutive days with = 5 times mean daily precipitation)
low_prec_freq	Fraction of dry days (< 1 mm d-1 daily precipitation)
low_prec_dur	Average duration of dry periods (number of consecutive days with daily precipitation < 1 mm d-1)
mean_elev	Catchment mean elevation
std_elev	Standard deviation of catchment elevation

mean_slope	Catchment mean slope
std_slope	Standard deviation of catchment slope
area_km2	Catchment area
Temperate-or-sub-polar-needleleaf-forest	Fraction of land covered by "Temperate-or-sub-polar-needleleaf-forest"
Temperate-or-sub-polar-grassland	Fraction of land covered by "Temperate-or-sub-polar-grassland"
Temperate-or-sub-polar-shrubland	Fraction of land covered by "Temperate-or-sub-polar-shrubland"
Temperate-or-sub-polar-grassland	Fraction of land covered by "Temperate-or-sub-polar-grassland"
Mixed-Forest	Fraction of land covered by "Mixed-Forest"
Wetland	Fraction of land covered by "Wetland"
Cropland	Fraction of land covered by "Cropland"
Barren-Lands	Fraction of land covered by "Barren-Lands"
Urban-and-Built-up	Fraction of land covered by "Urban-and-Built-up"
Water	Fraction of land covered by "Water"
BD	Soil bulk density (g cm-3)
CLAY	Soil clay content (% of weight)
GRAV	Soil gravel content (% of volume)
OC	Soil organic carbon (% of weight)
SAND	Soil sand content (% of weight)
SILT	Soil silt content (% of weight)

Finally, we also collect daily actual evapotranspiration (AET) for each watershed in millimeters per day, which was originally taken from the Global Land Evaporation Amsterdam Model (GLEAM) v3.5b dataset (Martens et al., 2017). GLEAM couples remotely sensed observations of microwave Vegetation Optical Depth, a multi-layer soil moisture model driven by observed precipitation and assimilating satellite surface soil moisture observations, and Priestly-Taylor based estimates of PET to derive an estimate of AET for each day. The daily data were originally available over the entire study domain at a 0.25° resolution between 2003-2017 and were aggregated to basin-wide totals for each watershed. 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.

3. Methods

We design an experiment to test the two primary hypotheses of this study, namely that a standard LSTM will overestimate ~~hydrologic~~ water losses under warming because of an overreliance on historical correlations between temperature and PET, while this effect will be lower in PIML-based rainfall-runoff models designed to better account for water loss in the system. To conduct this experiment, we develop three different DL rainfall-runoff models to predict daily streamflow across the Great Lakes region, as well as three conceptual, process-based models as benchmarks, each of which is trained twice with either an energy budget-based or temperature-based estimate of PET. The DL models include a regional LSTM very similar to the model in Mai et al., (2022), an MC-LSTM that conserves mass, and a new variant of the MC-LSTM that also respects the relationship between PET and water loss (termed MC-LSTM-PET). After comparing historical model performance, we conduct a sensitivity analysis ~~force~~ on all models ~~with climate change scenarios~~in which ~~composed of~~ historical ~~precipitation and historical but warmed~~ temperatures are warmed by 4°C, ~~as well as~~ PET is updated based on those warmed temperatures, and all other meteorological variable time series are left unchanged from historical values. This is a similar approach to that taken in SW22, but in contrast to that study this work 1) focuses on the magnitude of streamflow response to warming under two different PET formulations; 2) considers a different set of physics-informed DL models in which the architecture (rather than the inputs or targets) of the model are changed to better preserve physical plausibility under ~~unprecedented~~ shifts in climate ~~change~~; and 3) evaluates an expanded set of hydrologic metrics to better understand both the plausibility and the variability of ~~climate change~~ responses across the different models. Finally, in a subset of the analysis, we also utilize a fourth DL model, the LSTM used in SW22 that was previously fit to 531 basins across the CONUS (Kratzert et al. 2021), which uses daily precipitation, maximum and minimum temperature, radiation, and vapor pressure as input but not PET. This model is used to evaluate whether a DL model fit to many more watersheds that span a more diverse gradient of climate conditions behaves differently under warming than an LSTM fit only to locations in the Great Lakes basin. Figure 2 presents an overview of our experimental design.

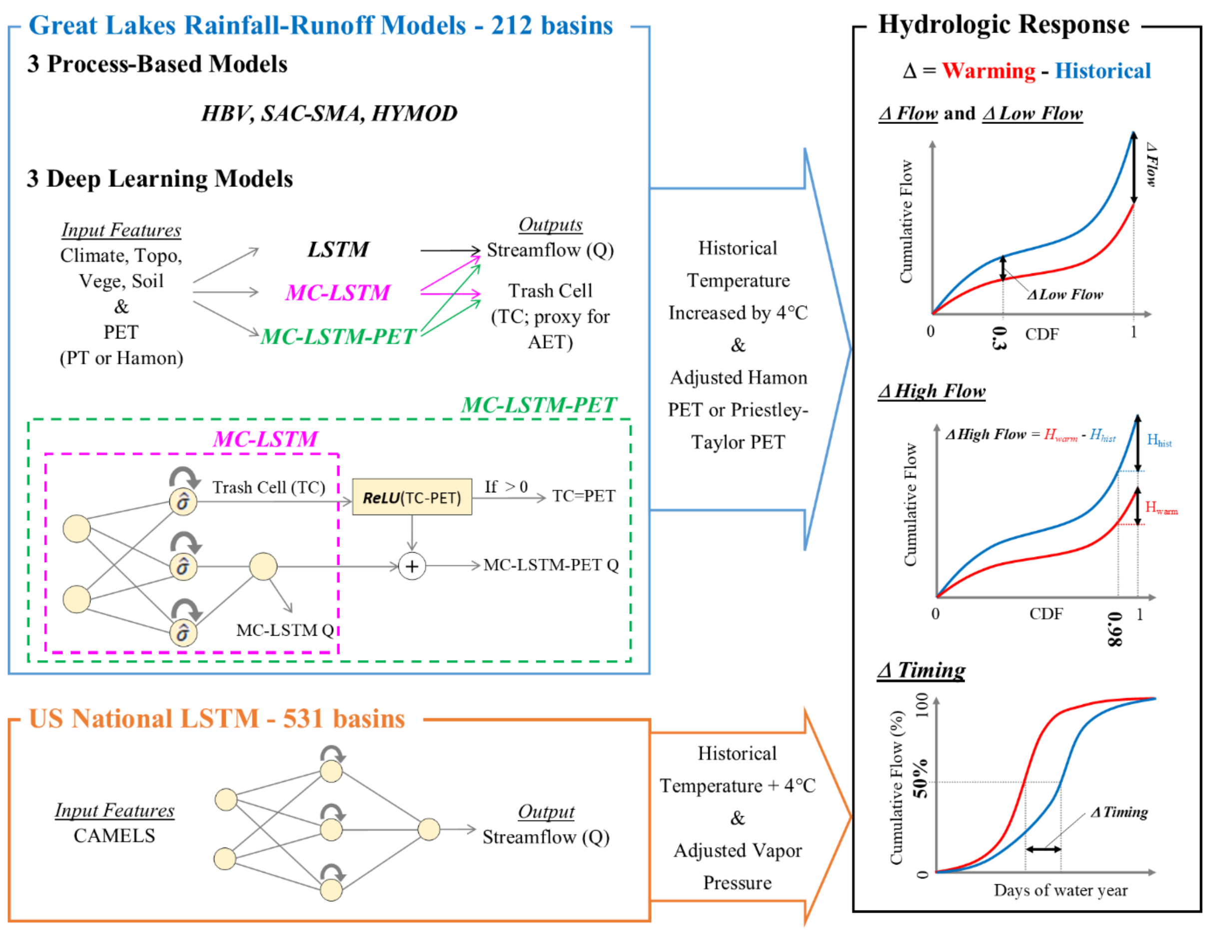


Figure 2. Overview of experiment design. Three deep learning rainfall-runoff models (LSTM, MC-LSTM, MC-LSTM-PET) and three conceptual, process-based models (HBV, SAC-SMA, HYMOD) are trained and tested across 212 watersheds throughout the Great Lakes basin. Models are validated by comparing predictions to streamflow (Q) and actual evapotranspiration (AET). All models are then forced with historical meteorology, but with historical temperatures warmed by 4°C and potential evapotranspiration (PET) updated based on those warmed temperatures using either the Hamon or Priestley-Taylor method. Hydrologic model responses across all models are then compared in terms of long-term mean daily flows, low flows, high flows, and streamflow seasonal timing statistics. The experiment is also repeated with an LSTM fit to 531 basins across the contiguous United States, except that model does not use PET as an input and vapor pressure is also adjusted along with temperature.

3.1. Models

3.1.1. Benchmark Conceptual Models

We develop three conceptual, process-based hydrologic models as benchmarks, including the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström and Forsman, 1973), HYMOD (Boyle, 2001), and the Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash, 1995) coupled with SNOW-17 (Anderson, 1976). These models are developed as lumped, conceptual models for each watershed, and were selected for several reasons. First, in the Great Lakes Intercomparison Project (Mai et al., 2022), HYMOD was one the best performing process models for both streamflow and AET estimation. SAC-SMA is widely used in the United States, forming the core hydrologic model in NOAA's Hydrologic Ensemble Forecasting System (Demargne et al., 2014). We also found in WS22 that AET from SAC-SMA matched the seasonal pattern of MODIS-derived AET well across California. 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 in hydrologic model intercomparison projects (Breuer et al., 2009; Plesca et al., 2012; Beck et al., 2016, 2017).

We calibrate the process-based models with the genetic algorithm from Wang et al. (1991) to ~~maximize~~ minimize the ~~Nash~~ mean-~~Sutcliffe~~ squared ~~Efficiency~~ error (~~NSE~~MSE), using a population size equal to 100 times the number of parameters, evolved over 100 generations, and with a spin-up period of 1 year. Each benchmark model is calibrated separately to each of the 141 training sites using the temporal train/test split described in Section 2, and training is repeated~~.~~ 10 separate times with different random initializations to account for uncertainty in the training process and to estimate parametric uncertainty. Benchmark models are developed for the 71 testing sites in two ways: 1) separate models are trained for the testing sites during the training period; and 2) each testing site is assigned a donor from among the 141 training sites, and the calibrated parameters from that donor site are transferred to the testing site. The first of these approaches enables a comparison between DL models fit only to the training sites to benchmark models developed for the testing sites, i.e., a spatial out-of-sample versus in-sample comparison. The second of these approaches enables a more direct spatial out-of-sample comparison between DL and benchmark models. We note that donor sites were used to assign model parameters to testing sites in the benchmarking study of Mai et al. (2022), and to retain direct comparability to the results of that work we use the same donor sites for each testing site. Donor sites were selected based on spatial proximity, while also prioritizing donor sites that were nested within the watershed of the testing site.

3.1.2. LSTM

We develop a single, regional LSTM for predicting daily streamflow across the Great Lakes region. In the LSTM, nodes within hidden layers feature gates and cell states that address the vanishing gradient problem of classic recurrent neural networks and help capture long-term dependencies between input and output time series. The model defines a *D*-dimensional vector of recurrent cell states $\boldsymbol{c}[t]$ that is updated over a sequence of *t*=1,…,*T* time steps based on a sequence of inputs $\boldsymbol{x} = \boldsymbol{x}[1], \ldots, \boldsymbol{x}[T]$, where each input $\boldsymbol{x}[t]$ is a *K*-dimensional vector of features. Information stored in the cell states is then used to update a *D*-dimensional vector of hidden states $\boldsymbol{h}[t]$, which form the output of the hidden layer in the model. The structure of the LSTM is given as follows:

$$\boldsymbol{i}[t] = \sigma(\boldsymbol{W}_i\boldsymbol{x}[t] + \boldsymbol{U}_i\boldsymbol{h}[t-1] + \boldsymbol{b}_i) \quad \text{(Eq. 1.1)}$$

$$\boldsymbol{f}[t] = \sigma\left(\boldsymbol{W}_f\boldsymbol{x}[t] + \boldsymbol{U}_f\boldsymbol{h}[t-1] + \boldsymbol{b}_f\right) \quad \text{(Eq. 1.2)}$$

$$\boldsymbol{g}[t] = tanh\left(\boldsymbol{W}_g\boldsymbol{x}[t] + \boldsymbol{U}_g\boldsymbol{h}[t-1] + \boldsymbol{b}_g\right) \quad \text{(Eq. 1.3)}$$

$$\boldsymbol{o}[t] = \sigma(\boldsymbol{W}_o\boldsymbol{x}[t] + \boldsymbol{U}_o\boldsymbol{h}[t-1] + \boldsymbol{b}_o) \quad \text{(Eq. 1.4)}$$

$$\boldsymbol{c}[t] = \boldsymbol{f}[t] \odot \boldsymbol{c}[t-1] + \boldsymbol{i}[t] \odot \boldsymbol{g}[t] \quad \text{(Eq. 1.5)}$$

$$\boldsymbol{h}[t] = \boldsymbol{o}[t] \odot tanh(\boldsymbol{c}[t]) \quad \text{(Eq. 1.6)}$$

$$\boldsymbol{y}[T] = ReLU\left(\boldsymbol{W}_y\boldsymbol{h}[T] + b_y\right) \quad \text{(Eq. 1.7)}$$

Here, the input gate ($\boldsymbol{i}[t]$) controls how candidate information ($\boldsymbol{g}[t]$) from inputs and previous hidden states flows to the current cell state ($\boldsymbol{c}[t]$); the forget gate ($\boldsymbol{f}[t]$) enables removal of information within the cell state over time; and the output gate ($\boldsymbol{o}[t]$) controls information flow from the current cell state to the hidden layer output. All bolded terms are vectors, and $\odot$ denotes element-wise multiplication. To produce streamflow predictions, $\boldsymbol{h}[T]$ at the last time step in the sequence is passed through a fully connected layer to a single-node output layer (i.e., a many-to-one formulation). We ensure nonnegative streamflow predictions using the rectified linear unit (ReLU) activation function for the output neuron, expressed as ReLU(x) = max(0,x). Importantly, there are no constraints requiring the mass of water entering as precipitation to be conserved within this architecture.

The LSTM takes K=39 input features: 9 dynamic and 30 static. The dynamic input features are basin-averaged climate, including daily precipitation, maximum temperature, minimum temperature, net incoming shortwave radiation, specific humidity, surface air pressure, zonal and meridional components of wind, and PET. The static features represent catchment attributes (see Table 1) and are repeated for all time steps in the input sequences $\boldsymbol{x}$. All input features are standardized before training (by subtracting the mean and dividing by the standard deviation for data across all training sites in the training period). Note that we do not standardize the observed streamflow, besides dividing ~~my~~ by drainage area to represent streamflow in units of millimeters.

We train the LSTM by minimizing the mean-squared error averaged over the 141 training watersheds during the training period:

$$MSE = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\left(\hat{Q}_{n,t} - Q_{n,t}\right)^2 \qquad (2)$$

where N is the number of training watersheds and T_n is the number samples in the n^{th} watershed. $\hat{Q}_{n,t}$ and $Q_{n,t}$ are, respectively, the streamflow prediction and observation for basin n and day t. To estimate $\hat{Q}_{n,t}$, we feed into the network an input sequence for the past T=365 days. The model was developed with 1 hidden layer composed of D=256 nodes, a mini-batch size of 256, a learning rate of 0.0005, and a drop-out rate of 0.4, and it was trained across 30 epochs. All hyperparameters (number of hidden layer nodes, mini-batch size, learning rate, dropout rate, and number of epochs) were selected in a 5-fold cross-validation on the training sites (see Table S2 for details on grid search). Network weights are tuned using the ADAM optimizer (Kingma & Ba, 2015). The model is trained 10 separate times with different random initializations to account for uncertainty in the training process.

For the evaluation of streamflow ~~projections~~ responses to ~~under climate change~~warming, we also use an LSTM taken from Kratzert et al. (2021) and employed in SW22, which was fit to 531 basins across the contiguous United States (hereafter called the National LSTM). This model was trained using a different set of data compared to our Great Lakes LSTM but also used a mix of dynamic and static features, all of which were drawn from the Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS) dataset (Newman et al., 2015). This model uses daily precipitation, maximum and minimum temperature, shortwave downward radiation, and vapor pressure as input but not PET. However, we note that temperature, radiation, and vapor pressure are the three major inputs (besides wind speeds) needed to calculate energy budget-based PET. 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 (see Figure 1).

3.1.3. MC-LSTM

Following Hoedt et al. (2021) and Frame et al. (2021b), we adapt the architecture of the LSTM into a mass conserving MC-LSTM that preserves the water balance within the model, i.e., the total quantity of precipitation entering the model is tracked and redistributed to streamflow and losses from the watershed. Using similar notation as for the LSTM above, the model structure is given as follows:

$$\hat{\boldsymbol{c}}[t-1] = \frac{\boldsymbol{c}[t-1]}{\|\boldsymbol{c}[t-1]\|_1} \quad \text{(Eq. 3.1)}$$

$$\boldsymbol{i}[t] = \hat{\sigma}(\boldsymbol{W}_i\boldsymbol{x}[t] + \boldsymbol{U}_i\hat{\boldsymbol{c}}[t-1] + \boldsymbol{V}_i\boldsymbol{a}[t] + \boldsymbol{b}_i) \quad \text{(Eq. 3.2)}$$

$$\boldsymbol{o}[t] = \sigma(\boldsymbol{W}_o\boldsymbol{x}[t] + \boldsymbol{U}_o\hat{\boldsymbol{c}}[t-1] + \boldsymbol{V}_o\boldsymbol{a}[t] + \boldsymbol{b}_o) \quad \text{(Eq. 3.3)}$$

$$\boldsymbol{R}[t] = \hat{\sigma}(\boldsymbol{W}_R\boldsymbol{x}[t] + \boldsymbol{U}_R\hat{\boldsymbol{c}}[t-1] + \boldsymbol{V}_R\boldsymbol{a}[t] + \boldsymbol{b}_R) \quad \text{(Eq. 3.4)}$$

$$\boldsymbol{m}[t] = \boldsymbol{R}[t]\boldsymbol{c}[t-1] + \boldsymbol{i}[t]\boldsymbol{x}[t] \quad \text{(Eq. 3.5)}$$

$$\boldsymbol{c}[t] = (1 - \boldsymbol{o}[t]) \odot \boldsymbol{m}[t] \quad \text{(Eq. 3.6)}$$

$$\boldsymbol{h}[t] = \boldsymbol{o}[t] \odot \boldsymbol{m}[t] \quad \text{(Eq. 3.7)}$$

Here, the inputs to the model are split between quantities **x**[t] to be conserved (i.e., precipitation), and non-conservative inputs **a**[t] (i.e., temperature, wind speeds, PET, catchment properties, etc.). Water in the system is stored in the *D*-dimensional vector **m**[t] and is updated at each time step based on water left over from the previous time step (**c**[t-1]) and water entering the system at the current time step (**x**[t]). The input gate **i**[t] and a redistribution matrix **R**[t] are designed to ensure water is conserved from $\boldsymbol{c}[t-1]$ and $\boldsymbol{x}[t]$ to **m**[t], by basing these quantities on a normalized sigmoid activation function ~~that sums to unity~~:

$$\hat{\sigma}(z_j) = \frac{\sigma(z_j)}{\sum_j \sigma(z_j)} \quad \text{(Eq. 4)}$$

Here, $\sigma(\cdot)$ is the sigmoid activation function, while $\hat{\sigma}(\cdot)$ is a normalized sigmoid activation that produces a vector of fractions that sum to unity.

The mass in $\boldsymbol{m}[t]$, which is stored across *D* elements in the vector, is then distributed to the output of the hidden layer, $\boldsymbol{h}[t]$, or the next cell state, $\boldsymbol{c}[t]$. To account for water losses from evapotranspiration or other sinks, one element of the *D*-dimensional vector $\boldsymbol{h}[t]$ is considered a 'trash cell', and the output of this cell is ignored when calculating the final streamflow prediction, which at time *T* is given by the sum of outgoing water mass:

$$y[T] = \sum_{d=1}^{D-1} h_d[T] \quad \text{(Eq. 5)}$$

Here, the D^{th} cell of $\boldsymbol{h}$ (h_D) is set as the trash cell, and water allocated to this cell at each time step t=1,..,T is lost from the system. We note that the MC-LSTM was trained in the same way as the LSTM (i.e., same inputs, loss function, training and test sets, hyperparameter selection process, number of ensemble members with random initialization).

3.1.4. MC-LSTM-PET

We also propose a novel variant of the MC-LSTM that requires water lost from the system to not exceed PET (hereafter referred to as the MC-LSTM-PET). In the original MC-LSTM, any amount of water can be delegated to the trash cell h_D. Therefore, while water is conserved in the MC-LSTM, the model has the freedom to transfer any amount of water from $\boldsymbol{m}[t]$ to the trash cell (and out of the hydrologic system) as it seeks to improve the loss function during training. This has the benefit of handling biased data, e.g., cases where the precipitation input to the system is systematically too high compared to the measured outflow. However, this structure also has the drawback of potentially removing more water from the system than is physically plausible. To address this issue, we propose a small change to the architecture of the MC-LSTM, where any water relegated to the trash cell that exceeds PET at time t is directed back to the stream:

$$y[t] = \sum_{d=1}^{D-1} h_d[t] + ReLU(h_D[t] - PET[t]) \quad \text{(Eq. 6)}$$

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. ~~This approach assumes that the maximum allowable water lost from the system cannot exceed PET, and therefore ignores other potential terminal sinks (e.g., deep groundwater percolation that remains disconnected from the stream; lateral groundwater flows out of the watershed; human diversions). However, given that evapotranspiration accounts for the vast majority of water lost in most hydrologic systems, this assumption is likely reasonable in most cases.~~ The MC-LSTM-PET was trained in the same way as the LSTM (i.e., same inputs, loss function, training and test sets, hyperparameter selection process, number of ensemble members with random initialization).

3.2. Model Performance Evaluation

As noted previously, 141 of the watersheds are designated as training sites, and the remaining 71 watersheds are used for testing. In addition, the training and testing periods were restricted to January 2000 -December 2010 and January 2011 – December 2017, respectively. This provides three separate ways to evaluate model performance:

- Temporal validation - Performance across models is evaluated at training sites during the testing period.
- Spatial validation - Performance across models is evaluated at testing sites during the training period.
- Spatiotemporal validation - Performance across models is evaluated at testing sites during the testing period.

All three evaluation strategies are utilized. For benchmark process-based models that are calibrated locally on a site-by-site basis, we consider model versions that are transferred to testing sites from training sites, as well as models that are trained to the testing sites directly (see Section 3.1.1). The former can be used for all three evaluation strategies above, while the latter can only be used for temporal validation at the testing sites.

Following other intercomparison studies (Frame et al., 2022; Gauch et al., 2021a; Klotz et al., 2022; Kratzert et al., 2021), ~~Several~~ several metrics are considered for model evaluation, including percent bias (PBIAS), the Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), Kling-Gupta Efficient (KGE; Gupta et al. 2009), top 2% peak flow bias (FHV; Yilmaz et al. 2008), and bottom 30% low flow bias (FLV; Yilmaz et al. 2008). Each metric is calculated separately for training and testing periods for each site. For all~~the DL~~ models, all results are estimated from the ensemble mean from 10 separate training trials.

For the process models, the MC-LSTM, and the MC-LSTM-PET, we also compare simulations of AET to ~~observations of~~ AET from the GLEAM database. We note that AET data were not used to train any of the models. For the process models, AET is a direct output of the model and so can immediately be extracted for comparison, but AET is not directly simulated by the MC-LSTM or MC-LSTM-PET. Instead, we assume water delegated to the trash cell permanently leaves the system because of evapotranspiration. Several metrics are used to compare model ~~-~~based AET to GLEAM AET, including KGE, correlation, and PBIAS, and the comparison is conducted for training sites during the training period and under temporal, spatial, and spatiotemporal validation (as described above). Similar to streamflow, all AET results ~~for the MC-LSTM and MC-LSTM-PET~~ are based on the ensemble mean ~~of water delegated to the trash cell~~ from the 10 separate training trials.

3.3. Evaluating Hydrologic Response under Warming

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).

PET (in mm/day) under the Hamon method is calculated as follows (Shaw and Riha, 2011):

$$PET_H = \alpha_H \times 29.8 \times Hr_{day} \frac{e_{sat}}{T_a + 273.2} \quad \text{(Eq. 7)}$$

$$e_{sat} = 0.611 \times exp\left(\frac{17.27 \times T_a}{237.3 + T_a}\right) \quad \text{(Eq. 8)}$$

where Hr_{day} is the number of daylight hours, T_a is the average daily temperature (°C) calculated from daily minimum and maximum temperature, e_{sat} is the saturation vapor pressure (kPa), and α_H is a calibration coefficient set to 1.2 for all models in this study (similar to Lu et al., 2005).

PET under the Priestley-Taylor method is calculated as follows:

$$PET_{PT} = \alpha_{PT}\left(\frac{\Delta(T_a) \times (R_n - G)}{\lambda(\Delta(T_a) + \gamma)}\right) \times 1000 \quad \text{(Eq. 9)}$$

Here, $\Delta(T_a)$ is the slope of the saturation vapor pressure temperature curve (kPa/°C) and is a function of T_a, γ is the psychrometric constant (kPa/°C), λ is the volumetric latent heat of vaporization (MJ/m^3), R_n is the net radiation (MJ/m^2-day) equal to the difference between net incoming shortwave (R_{ns}) and net outgoing longwave (R_{nl}) radiation, G is the heat flux to the ground (MJ/m^2-day), and α_{PT} is a dimensionless coefficient set to 1.1 for all models in this study (similar to Szilagyi et al., 2017). Details on how to calculate γ, $\Delta(T_a)$, and R_{nl} are available in Allen et al. (1998), and we assume G=0. Net shortwave radiation is given by $R_{ns} = (1 - \zeta)R_s$, with $\zeta = .23$ the assumed albedo and R_s the incoming shorwave radiation. We note that net outgoing longwave radiation R_{nl} is a function of maximum and minimum temperature, actual vapor pressure, and R_s (see Eq. 39 in Allen et al. 1998). All exogenous meteorological inputs for the two methods are derived from the Regional Deterministic Reanalysis System v2 ~~RDRS-v2~~ (see Section 2). We note that using $\alpha_H = 1.2$ and $\alpha_{PT} = 1.1$ leads to very similar long-term average PET estimates between the Hamon and Priestley-Taylor methods under baseline climate conditions, helping to ensure their comparability. We also note that both PET series are highly correlated with daily average temperatures (average Pearson correlations across sites of 0.94 and 0.83 for Hamon and Priestley-Taylor PET, respectively).

We then ~~develop a simple climate change scenario~~conduct a sensitivity analysis of model response in which the historical minimum and maximum temperature time series are increased uniformly by 4 °C, and the two PET estimates are updated using these warmed temperatures. We focus the ~~climate change~~ assessment on training period data at the training sites, so that any differences in ~~climate change projections~~responses that emerge between the DL and process models are due to model structural differences and not the effects of spatiotemporal regionalization. In the Priestly-Taylor method, we maintain historical values for R_s to isolate how changes in temperature and its effect on $\Delta(T_a)$ and R_{nl} influence changes in PET. The use of historical R_s is supported by the results from CMIP5 projections presented in Lai et al. (2022), but this assumption is discussed further in Section 5.

We also ~~develop~~ conduct a similar ~~climate change scenario~~sensitivity analysis ~~for~~ on the National LSTM, which uses five dynamic input features from the CAMELS dataset (daily precipitation, maximum temperature, minimum temperature, R_s, and water vapor pressure). Here, temperatures are warmed by 4 °C, while precipitation and R_s are held at historical values. There is a strong correlation between vapor pressure and minimum temperature in the CAMELS dataset, since minimum temperature is used to estimate the water vapor pressure (Newman et al., 2015). Thus, to run the National LSTM under warming, we also adjust the vapor pressure input based on the change imposed to minimum temperature. This procedure is detailed in SW22.

For both the Great Lakes DL models and the National LSTM, the dynamic inputs are adjusted based on the warming scenarios above. 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 ~~Table 1~~Supporting Information S1 and Table S1 for details on adjustments to these features), and then run all models using two settings: 1) with ~~climate~~ changes only to the dynamic features, and 2) with ~~climate~~ changes to both dynamic features and to static features that depend on those dynamic features. In total, there are six scenarios run in this work, which are shown in Table 2.

Table 2. Overview of the setup for the different scenarios run in this analysis. All models are driven with temperatures warmed by 4°C. The Great Lakes models include the HBV, SAC-SMA, HYMOD, LSTM, MC-LSTM, and MC-LSTM models that are trained and tested to the 212 sites across the Great Lakes basin.

Scenario	Model	PET method adjusted with warmer temperatures	Are static features also changed along with dynamic features?
1	Great Lakes models	Hamon	Yes
2	Great Lakes models	Priestley-Taylor	Yes
3	Great Lakes models	Hamon	No
4	Great Lakes models	Priestley-Taylor	No
5	National LSTM	NA	Yes
6	National LSTM	NA	No

Ultimately, for each model we compare hydrologic ~~projections~~ responses under the warmed scenario to their values under the baseline scenario with no warming. For the National LSTM, we only consider basins in the CAMELS dataset within the Great Lakes Basin. For the process models, we also evaluate the uncertainty in hydrologic response based on the range predicted across the 10 different training trials, as a simple means to evaluate how parametric uncertainty influences the predictions. We examine four different metrics for this comparison, including:

- AVG.Q: the long-term ~~average~~ mean of daily streamflow ~~runoff~~ across the entire series.
- FHV: the average of the top 2% peak flows.
- FLV: the average of the bottom 30% low flows.
- COM: the median center of mass across all water years, where the center of mass is defined as the day of the water year by which half of the total annual flow has passed.

If our hypothesis is correct that the LSTM cannot distinguish water loss differences with different PET ~~projections~~ series but similar warming while process-based and PIML models can, we would expect that under the LSTM using both PET ~~projections~~series, ~~average~~ long-term mean flow will decline ~~significantly~~ substantially and with similar magnitude to the process models using the temperature-based PET method but not the energy budget-based PET method. We would also expect the National LSTM to exhibit similar behavior, even though it was able to learn from a larger set of watersheds across a more diverse range of climate conditions. Finally, if our hypothesis is correct, we would expect the PIML models (MC-LSTM, MC-LSTM-PET) to follow the process model ~~projections~~ responses more closely across the two different PET ~~projections~~series, at least in terms of the difference in magnitude of ~~average~~ long-term mean streamflow declines. ~~For~~ To facilitate a broader ~~comparison~~inter-model comparison of DL and process-based models under warming (which is largely absent from the literature), we also explore the differences in low flow (FLV), high flow (FHV), and seasonal timing (COM) metrics across all model versions, where we have less reason to anticipate how DL and process models will differ in their ~~projections~~ responses and across PET formulations. However, for responses like seasonal streamflow timing (COM), we do anticipate that realistic responses should show a shift towards more streamflow earlier in the year, as warmer temperatures lead to more precipitation falling as rain rather than snow and drive snowmelt earlier in the spring.

4. Results

4.1. Model Performance Evaluation

Figure 3 shows the distribution of KGE values across sites for streamflow from the LSTM, MC-LSTM, MC-LSTM-PET, and the three process-based models for both the training and testing sites during both the training and testing periods. All results here and elsewhere in Section 4.1 are shown for the models fit with Priestley-Taylor PET, but there is little difference in performance for the models fit with Hamon PET (see Figure S1). For the process-based models, we show results for models fit to the training sites and then used as donors at the testing sites, as well as models fit to the testing sites directly. We denote the latter with the suffix "-test" and note that performance metrics at the training sites are not available for process models fit to the testing sites.

Several insights emerge from Figure 3. First, for the training sites during the training period, all models perform very well (Figure 3a). Across the three process models, the median KGE is ~~0.82~~0.79, ~~0.83~~0.78, and ~~0.81~~0.77 for HBV, SAC-SMA, and HYMOD, respectfully. However, unsurprisingly, the DL models perform better for the training data, with median KGE values all equal or above 0.88. The LSTM performs best in this case. Under temporal validation (training sites during the testing period), performance degrades somewhat across all models, and the differences in KGE between all process-based models and between all DL models shrink considerably (Figure 3c). Larger performance declines are seen at the testing sites during the training period (Figure 3b) and testing period (Figure 3d). Here, the median KGE for all process models falls to between 0.~~56~~54-0.58~~7~~ when streamflow at the testing sites is estimated with donor models from nearby gauged watersheds. In contrast, process models fit to the testing sites (denoted "-test") exhibit performance similar to that seen in Figure 3a,c. All three DL models perform quite well for the testing sites, with median KGE values above 0.71 in both time periods. This is only modestly below the median KGE for the process models fit to the testing sites, which is quite impressive given that this represents the spatial out-of-sample performance of the DL models. We even see that for approximately ~~10~~20% of testing sites during the training period, the DL models outperform the process models fit to those locations in that period.

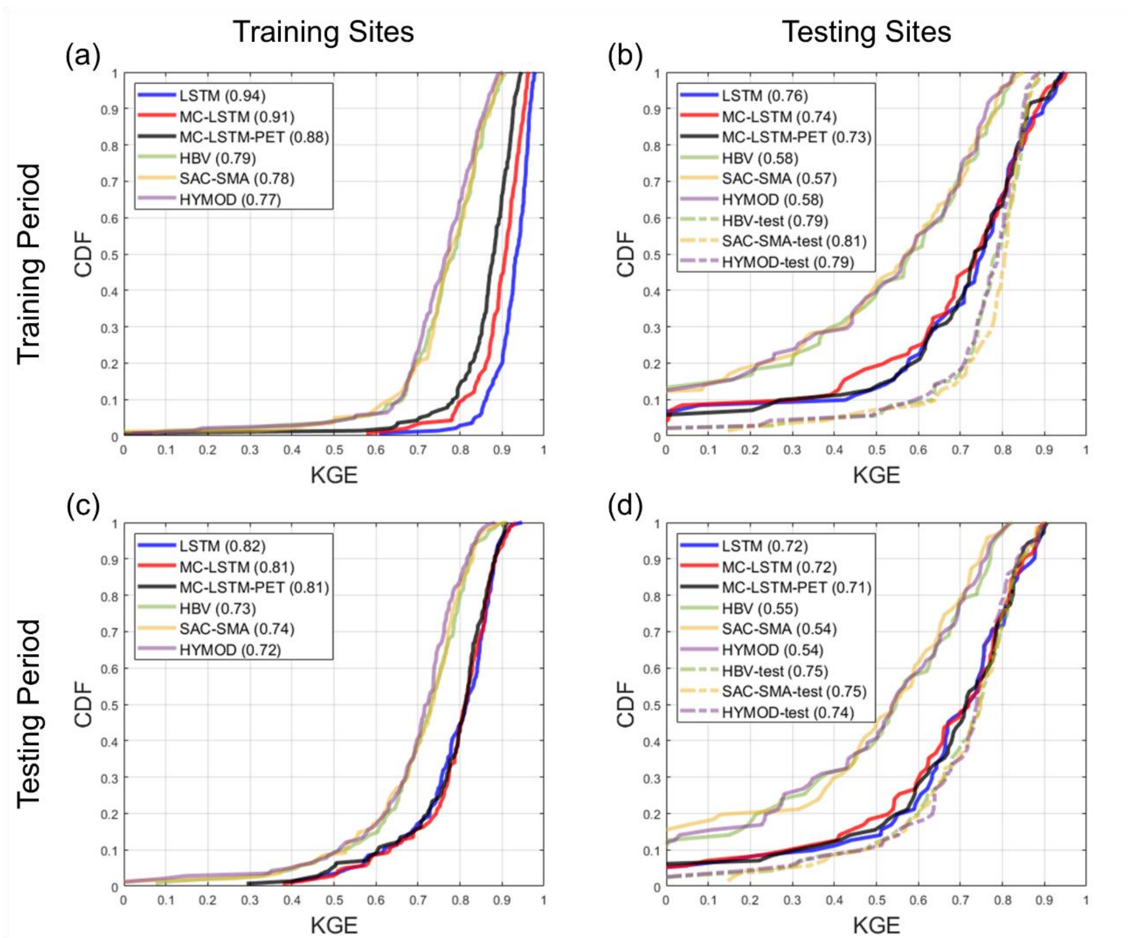


Figure 3. The distribution of Kling-Gupta efficiency (KGE) for streamflow estimates across sites from each model at the (a) the 141 training sites and (b) 71 testing sites for the training period. Similar results for the testing period are shown in panels (c) and (d), respectively. For the process models fit to the testing sites (denoted "-test"), no performance results are available at the training sites. All models are trained using Priestley-Taylor PET.

Table 3~~2~~ shows the median KGE, NSE, PBIAS, FHV, and FHL across testing sites for all models, excluding the process models fit to the testing sites. Similar to Figure 3, all three DL models outperform the donor-based process models at the testing sites for all metrics~~, with the exception of PBIAS during the training period~~. The performance across the three different DL models is similar, although there are some notable differences. In particular, the LSTM outperforms the MC-LSTM and MC-LSTM-PET for ~~KGE, ~~NSE and~~,~~

~~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. The fact that the MC-LSTM-PET performs best for PBIAS of all models suggests that the PET constraint imposed in that model improves the overall accounting of water entering and existing the watershed on a long-term basis. We also note that percent biases for FLV are high because the absolute magnitude of low flows is small, so small absolute biases still lead to large percent biases.

Table 3~~2~~. The median KGE, NSE, PBIAS, FHV, and FLV for streamflow across testing sites for the training and testing periods for all models (excluding the process models fit to the testing sites). The metric from the best performing model in each period is bolded. All models are trained using Priestley-Taylor PET.

	Testing Sites: Training Period					Testing Sites: Testing Period				
Model	KGE	NSE	PBIAS	FHV	FLV	KGE	NSE	PBIAS	FHV	FLV
LSTM	**0.76**	**0.77**	9.66	17.58	**30.98**	**0.72**	**0.68**	12.15	26.01	**27.32**
MC-LSTM	0.74	0.72	9.48	**15.52**	41.46	**0.72**	0.65	12.13	22.82	35.80
MC-LSTM-PET	0.73	0.72	**8.63**	18.80	48.10	0.71	0.66	**10.22**	**22.49**	44.43
HBV	0.58	0.50	9.99	32.22	63.96	0.55	0.50	12.68	34.76	57.20
SAC-SMA	0.57	0.48	11.74	34.72	45.17	0.54	0.47	12.24	40.45	46.78
HYMOD	0.58	0.48	10.07	33.68	58.06	0.54	0.48	12.52	36.07	60.32

Figure 4 shows similar results as Figure 3, but for the KGE based on estimates of AET. Also, only donor process models are shown for the testing sites. Results for correlation and PBIAS are available in the Supplemental Information (Figures S2-S3). Here, the LSTM is not included because estimates of AET are unavailable, while AET from the MC-LSTM and MC-LSTM-PET is based on water relegated to the trash cell. Note that none of the models were trained for AET, and so results at training sites during the training period also provide a form of model validation. Figure 4 shows that SAC-SMA and HBV predict AET with relatively high degrees of accuracy for both training and testing sites in both periods (median KGE between 0.79~~9~~-0.80). Performance is slightly worse for HYMOD. Notably, the MC-LSTM-PET exhibits very similar, strong performance for all sites and periods as compared to SAC-SMA and HBV, except for one testing site. In contrast, the MC-LSTM performs the worst of all models, with median KGE values ranging between 0.53-0.57.

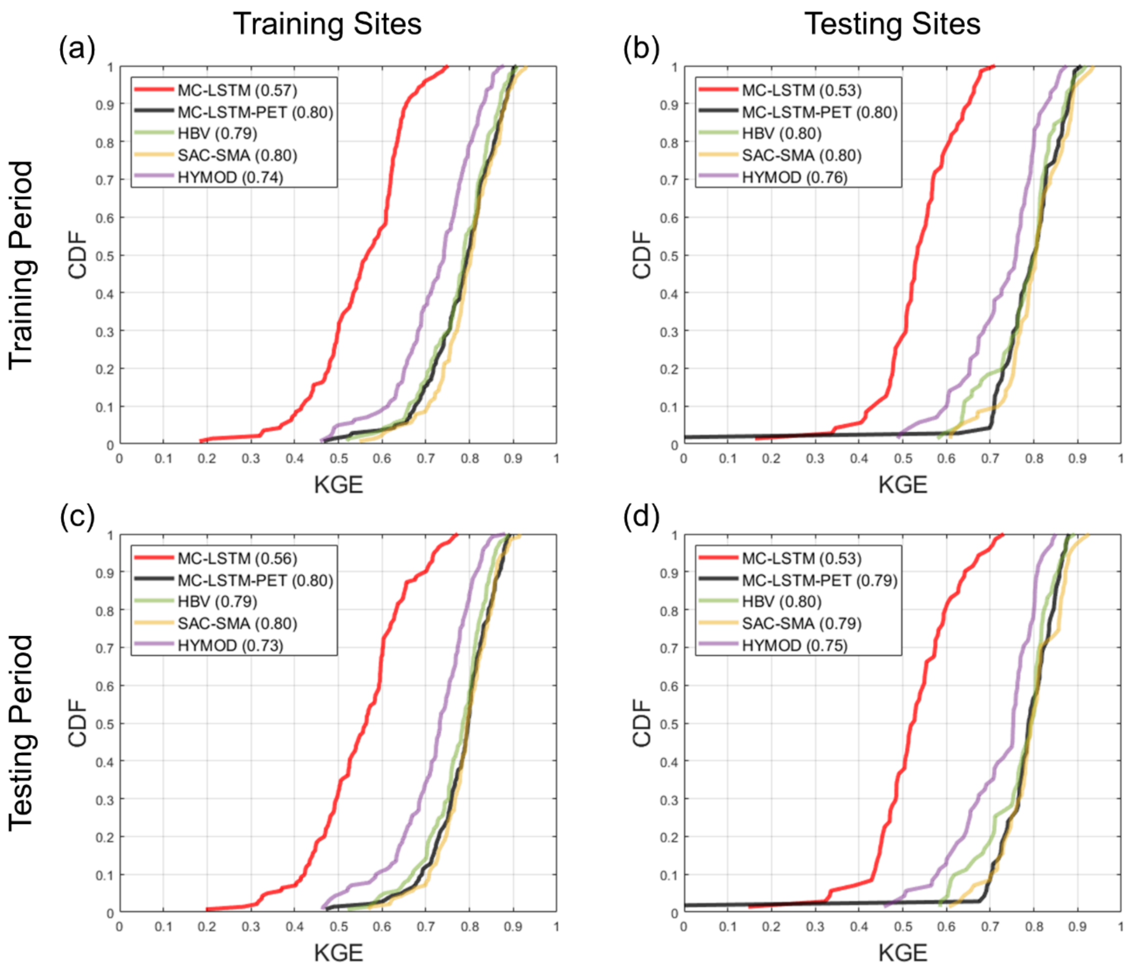


Figure 4. The Kling-Gupta efficiency (KGE) for AET estimated from each model at the (a) the 141 training sites and (b) 71 testing sites for the training period. Similar results for the testing period are shown in panels (c) and (d), respectively. The LSTM is not included in this comparison. All models are trained using Priestley-Taylor PET.

Further investigation reveals that the differences in KGE between the MC-LSTM and MC-LSTM-PET models for AET are largely driven by differences in correlation (see Figure S2). We examine this difference in more detail in Figure 5, which presents scatterplots of ~~observed~~ GLEAM AET versus water allocations to the trash cell for the two models from five randomly sampled testing sites across both training and testing periods (see ~~Table S1 for details on each site~~Figure 1; also Table S3). Trash cell water from the MC-LSTM is not only more scattered around ~~observed~~ GLEAM AET compared to the MC-LSTM-PET, but it also exhibits many outlier values that are two to five times larger than ~~observed~~ GLEAM AET. The MC-LSTM-PET follows the variability of GLEAM AET much more closely, with virtually no outliers that exceed GLEAM AET by large margins. This suggests that the PET constraint on the trash cell in the MC-LSTM-PET helps water allocated to that cell more faithfully represent ~~an ET sink~~evaporative water loss in the DL model.

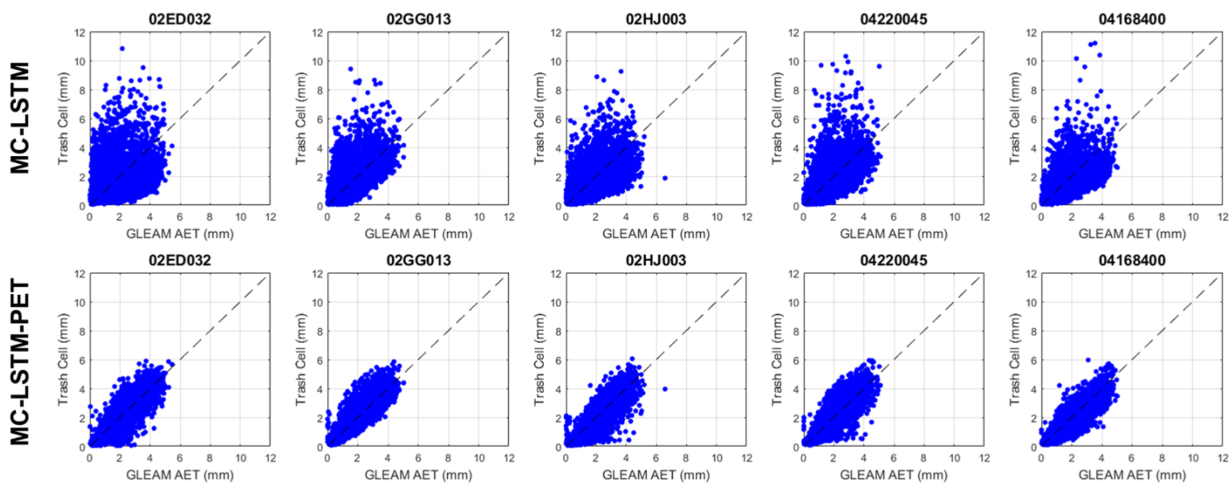


Figure 5. Scatterplots of daily AET versus trash cell water for the (top) MC-LSTM and (bottom) MC-LSTM-PET at five randomly selected testing sites across both training and testing periods. All models are trained using Priestley-Taylor PET.

4.2. Evaluating Hydrologic Response under Warming

Next, we evaluate streamflow ~~projections~~ responses under a 4 °C warming scenario. We focus on training sites during the training period, so that any differences that emerge between DL and process models are only related to model structure and not spatiotemporal regionalization. However, our results are largely unchanged if based on responses for testing sites in the testing period (see Figure S4). First, we show the differences in historic and warming-~~projected~~ adjusted PET when using the Hamon and Priestley-Taylor methods (Figure 6). For the training period without any temperature change, PET estimated from the two methods is very similar (Figure 6a; shown at one sample location for demonstration, see ~~Table S1~~Figure 1 and Table S3~~; Figure 6a~~). However, under the scenario with 4 °C of warming, Hamon-based PET is ~~significantly~~ substantially larger than Priestley-Taylor based PET (Figure 6b). On average, this difference reaches ~16% across all training sites and exhibits very little variability across locations (Figure 6c). The primary reason for the difference in ~~projected~~ the estimated change in PET is that the Hamon method attributes PET entirely to temperature, while only a portion of PET is based on temperature in the Priestley-Taylor method, with the rest based on R_n. It is worthwhile to note that R_n does ~~change~~ increase with temperature through its effects on net outgoing longwave radiation, but these changes ~~are small~~are generally less than 5% across all sites (Allen et al. 1998).

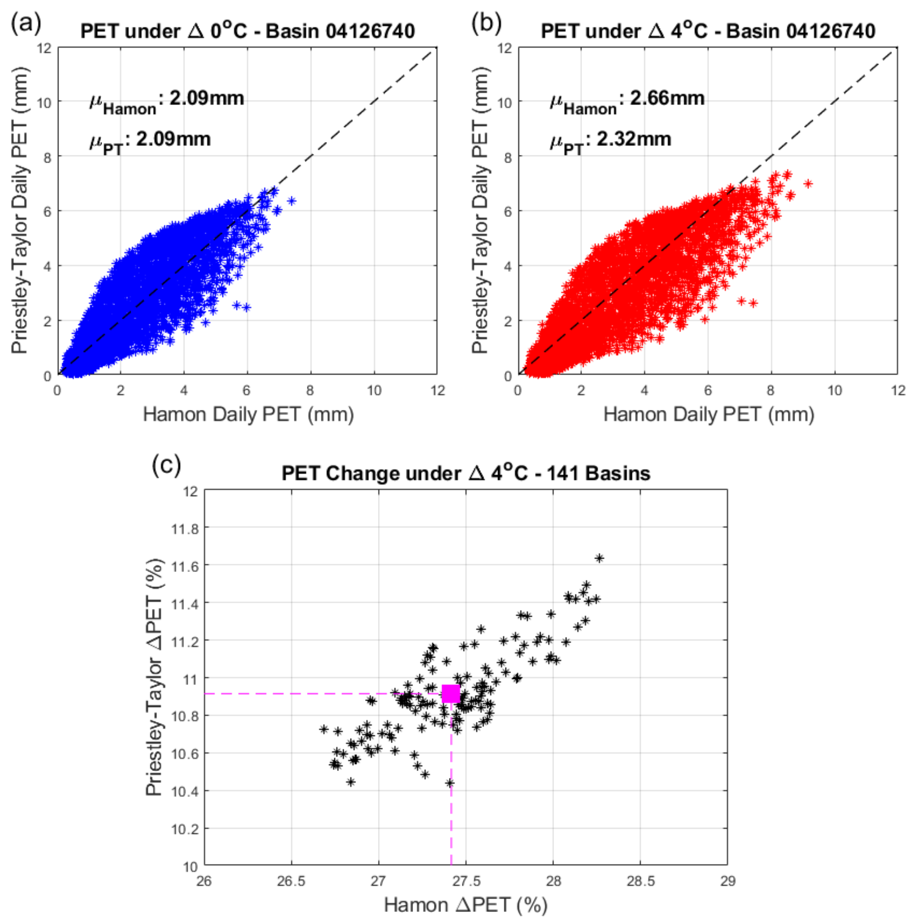


Figure 6. (a) Daily PET estimated using the Hamon and Priestley-Taylor method for one sample watershed, under historic climate conditions in the training period. (b) Same as (a), but under the ~~climate~~

~~change~~ scenario with 4 °C of warming. (c) Percent change in average PET with 4 °C of warming across all training sites using the Hamon and Priestley-Taylor methods.

Figure 7 shows how these differences in PET under warming propagate into changes in different attributes of streamflow across training sites in the training period. The left and right columns of Figure 7 show ~~projections~~ streamflow responses using Hamon and Priestley-Taylor PET, respectively, while the rows of Figure 7 show the distribution of changes ~~(as a percentage)~~ in different streamflow attributes (AVG.Q, FLV, FHV, COM) across models. Figure 7 shows results for DL models where only the dynamic inputs are changed under warming.~~,~~ ~~while Figure S4 show the same results when both the dynamic and the static climate properties are updated with warming.~~

Starting with changes in AVG.Q, Figure 7a,b shows that under the Hamon method for PET, the DL models exhibit similar changes in ~~average~~ long-term mean streamflow to the process-based models, with the median ΔAVG.Q across sites ranging between -17% and -~~23~~25% across all models. However, when using Priestley-Taylor PET, larger differences in the distribution of ΔAVG.Q emerge. Across all three process models, the median ΔAVG.Q is between -~~5~~6% to -~~10~~9%, and very few locations exhibit ΔAVG.Q less than -20%. Conversely, the LSTM shows a median water loss of -20% under Priestley-Taylor PET and a very similar distribution of water losses regardless of whether Hamon or Priestley-Taylor PET was used. The MC-LSTM is also relatively insensitive to PET, and as compared to the process models, the MC-LSTM tends to predict smaller absolute changes to AVG.Q for Hamon PET and larger changes under Priestley-Taylor PET. Only the MC-LSTM-PET model achieves water loss that is ~~significantly~~ considerably smaller under Priestley-Taylor PET than Hamon PET and closely follows the process models in both cases.

The overall pattern of change in low flows (FLV) is very similar across all three DL models, with median declines between -15% to -25% and little variability across sites (Figure 7c,d). The process models disagree ~~significantly~~ on the sign of change~~s~~ ~~to~~ for FLV, and also bound the changes predicted by the DL models.

HBV and HYMOD show mostly increases to FLV under warming and Priestley-Taylor PET, and a mix of increases and decreases across sites for Hamon PET. SAC-SMA exhibits large declines in FLV under warming and Hamon PET, and shows a median change that is similar to the DL models under Priestley-Taylor PET. The percent changes in FLV across models tend to be large because the absolute magnitude of FLV is small, and so small changes in millimeters of flow lead to large percent changes. This can be seen in sample daily hydrographs for two sites (see Figure S5), where visually the changes in low flows are difficult to discern because they are all near zero for all models, but the change in the FLV statistic varies significantly across the six models and two sites (-56% to +40%).

The differences between process-based and DL simulated changes for high flows (FHV; Figure 7e,f) and ~~streamflow~~ seasonal timing (COM; Figure 7g,h) are relatively consistent, with the process models exhibiting ~~larger~~ more substantial declines in high flows and earlier shifts in ~~streamflow~~ seasonal timing compared to the DL models. The choice of PET method has an ~~moderate~~ impact on process-model based changes in FHV, with larger declines under Hamon PET. A similar signal is also seen for the MC-LSTM-PET but not the MC-LSTM or LSTM, although the LSTM predicts changes in FHV closest to the process models.

For COM, the process models show a wide range of variability in projected change across sites, from no change to 60 days earlier. For the DL models the range of change is much narrower, and the median change in COM is ~~almost~~ approximately a week less than the median change across the process models. The earlier shift in COM across all models is consistent with anticipated changes to snow accumulation and melt dynamics under warming, with more water entering the stream during the winter and early spring as precipitation shifts more towards rainfall and ~~existing~~ snowpack melts off earlier in the year (Byun and Hamlet, 2018; Mote et al., 2018; Kayastha et al., 2022~~REFERENCES~~). However, this effect is seen more dramatically in the process models, as evidenced by more prominent changes to their daily and monthly hydrographs under warming during the winter and early spring as compared to the DL models (see ~~see~~

Figures S5 and S6~~X~~). The method of PET estimation has relatively little impact on both process model and DL based estimates of change in COM.

We note that the results above do not change even when considering the parametric uncertainty in the process models, although for some metrics (FLV), uncertainty in process model estimated changes due to parametric uncertainty is large (see Figure S7). We also note that if the static watershed properties (pet_mean, aridity, t_mean, frac_snow; see Table 1) are ~~also~~ changed to reflect warmer temperatures and higher PET, all three DL models exhibit unrealistic water gains for between 15%-40% of locations depending on the model and PET method, with the most water gains occurring under the LSTM (Figure S8~~4~~). These results suggest that changing the static watershed properties associated with long-term climate characteristics can degrade the quality of the ~~projections~~estimated responses, at least when the ~~climate~~ temperature ~~changes~~ shifts are large and the range of average temperature and PET in the training set is limited. ~~We also note that the results in Figure 7 are largely unchanged if based on projections for testing sites in the testing period (Figure S5).~~

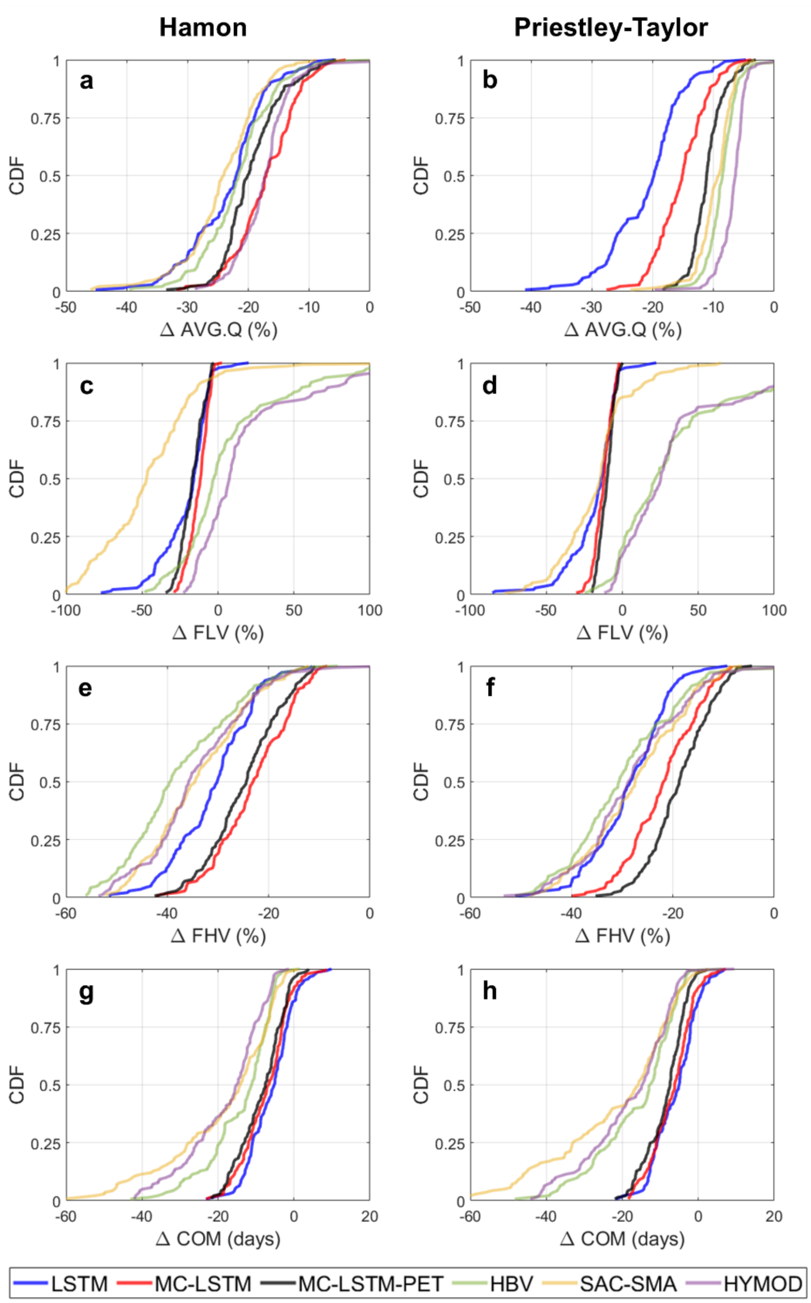


Figure 7. The distribution of change in (a,b) long term mean daily flow (AVG.Q), (c,d) low flows (FLV), (e,f) high flows (FHV), and (g,h) seasonal streamflow timing (COM) across the 141 training sites and all models under a scenario of 4°C warming using (a,c,e,g) Hamon PET and (b,d,f,h) Priestley-Taylor PET. For the ~~DL~~ deep learning models, changes were only made to the dynamic inputs (i.e., no changes to static inputs).

One reason why the Great Lakes LSTM exhibits excessive ~~hydrologic~~ water losses under warming could be that the model was trained using sites that are confined to a limited range of temperature and PET values found in the Great Lakes basin (spanning approximately 40.5°-50°N), and so is ill-suited to extrapolate hydrologic response under warming conditions that extend beyond~~-~~ this ~~range~~temperature and PET range. To evaluate this hypothesis, we examine changes to AVG.Q, FLV, FHV, and COM under 4°C warming at the 29 CAMELS watersheds within the Great Lakes basin using the National LSTM (Figure 8). For comparison, we also examine similar changes under all six Great Lakes DL and process models at 17 of those 29 CAMELS basins that were used in the training and testing sets for the Great Lakes models. ~~,~~ ~~and~~We also ~~separate out~~highlight the National LSTM ~~projections~~ predictions for those 17 sites. Note that in Figure 8, the National LSTM ~~projections~~ predictions do not differ between Hamon and Priestley Taylor PET, because PET is not an input to that model.

The National LSTM was trained to watersheds across the CONUS (spanning approximately 26°-49°N), and so was exposed to watersheds with much warmer conditions and higher PET during training. However, we find that the National LSTM still ~~projects~~ predicts very large declines in AVG.Q. For the 29 CAMELS watersheds in the Great Lakes basin, the median decline in AVG.Q under the National LSTM is approximately 25%, which is only 0-6~~moderately XX~~% larger than the median ~~projections~~ predictions of loss under the process models using Hamon PET ~~and~~ but ~~much~~ ~~XX~~16-19% larger than the process model losses under Priestley-Taylor PET (Figure 8a,b). We also see larger declines in FLV under the National LSTM as compared to the other Great Lakes DL models (Figure 8c,d). The National LSTM ~~projects~~ predicts changes in FHV (Figure 8e,f) and COM (Figure 8g,h) that are relatively similar to the process models.~~,~~ ~~and~~ ~~f~~For COM, the ~~projections~~ predictions of change are ~~closer~~ still smaller than~~to~~ the process models but closer to the process models than ~~for~~ any Great Lakes DL model,~~.~~ suggesting that the National LSTM predicts shifting snow accumulation and melt dynamics more consistently with the process models than regionally fit DL models. In addition, the hydrologic ~~projections~~ predictions are stable under the National LSTM regardless of whether only dynamic inputs or both dynamic and static inputs are changed under warming (see Figure S9~~6~~), in contrast to the Great Lakes DL models. Therefore, the use of more watersheds in training than span a more diverse set of climate conditions likely benefit the model when inputs are shifted ~~significantly~~ to reflect new climate conditions. However, as shown in Figure 8a,b, this benefit does not mitigate the tendency for the National LSTM to overestimate water loss under warming.

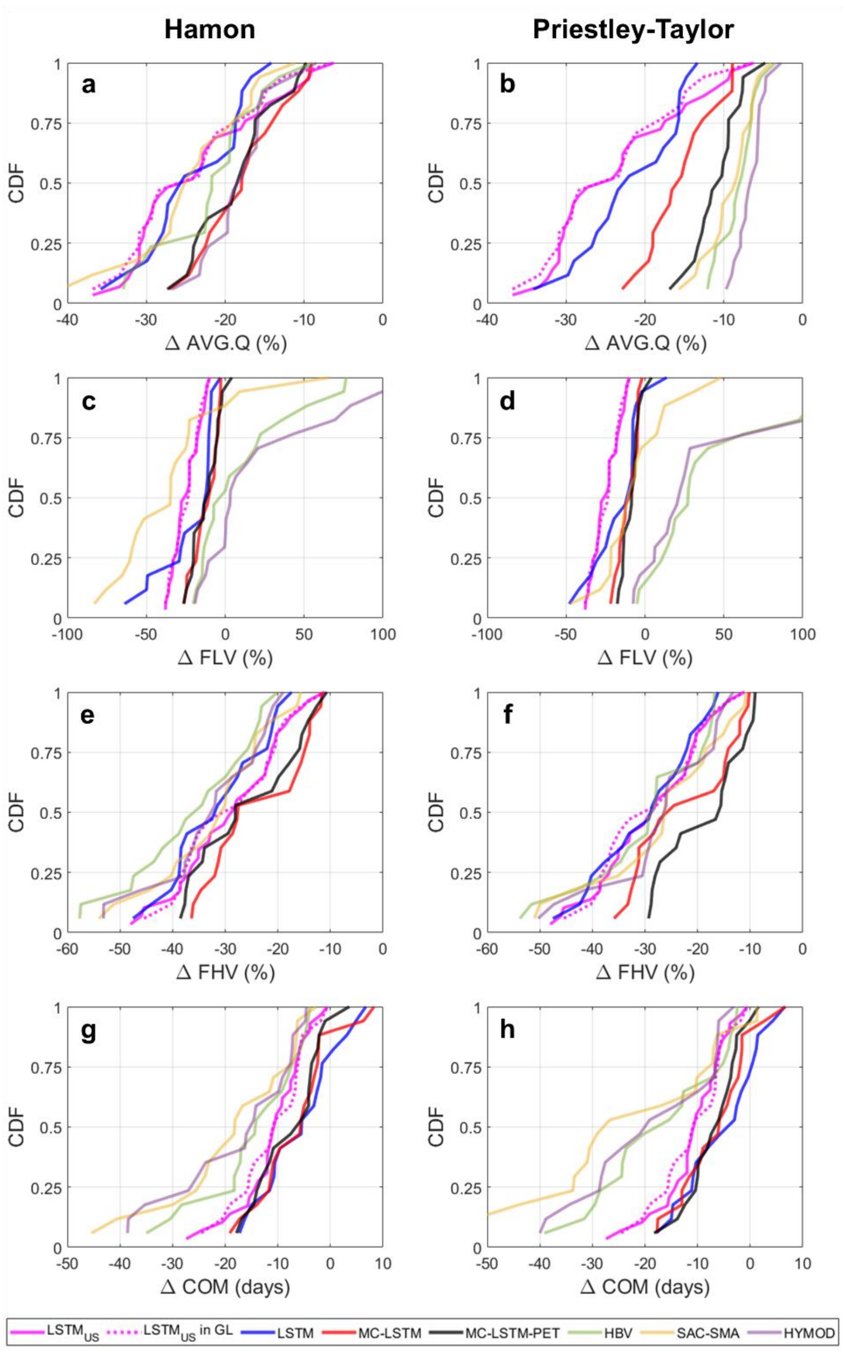


Figure 8. The distribution of change in (a,b) long term mean daily flow (AVG.Q), (c,d) low flows (FLV), (e,f) high flows (FHV), and (g,h) seasonal streamflow timing (COM) across 29 CAMELS sites within the Great Lakes basin under the National LSTM (solid pink), as well as for 17 of those 29 sites from the Great Lakes ~~DL~~ deep learning and process models, under a scenario of 4°C warming. Results from the

National LSTM for those 17 sites are also highlighted (dashed pink). For the Great Lakes models only, results differ when using (a,c,e,f) Hamon PET and (b,d,f,h) Priestley-Taylor PET. For the National LSTM, changes were made only to the dynamic inputs.

To better understand why the National LSTM predicts large water losses under warming, it is instructive to examine how ~~average~~ long-term mean streamflow, (Priestly-Taylor estimated) PET, and R_s vary across all 531 CAMELS watersheds of different average temperatures, and compare this variability to ~~projected~~ predicted changes in PET at each site under warming. Specifically, we ~~compare~~ calculate the difference in long-term (1980-2014) ~~average~~ mean streamflow (Figure 9a), PET (Figure 9b), and R_s (Figure 9c) across all pairs of basins in the CAMELS dataset with average long-term precipitation within 1% of each other (i.e., we only examine pairs of basins with very similar long-term mean precipitation). Then, for each basin pair, we ~~, and~~ plot the~~se~~ difference~~s~~ in long-term mean streamflow, PET, and R_s against the difference~~s~~ in long-term average temperature ~~across~~ for ~~each~~ that pair. The results show that the difference in ~~average~~ long-term mean streamflow across watersheds with similar precipitation becomes negative when the difference in temperature is positive (i.e., warmer watersheds have less flow on average), and that when the difference in average temperature reaches 4°C, flows differ by about 20% on average (Figure 9a). This is very similar to the ~~projected~~ predicted median decline in ~~average~~ long-term mean streamflow seen for the National LSTM in Figure 8. We also note that average PET increases by approximately 20% between watersheds that differ in average temperature by 4°C (Figure 9b). However, higher PET in warmer watersheds is related both to the direct effect of temperature on vapor pressure deficit, as well as to the fact that higher incoming solar radiation co-occurs in warmer watersheds (R_s is approximately 9% higher across watershed pairs that differ by 4°C; Figure 9c). Using the Priestley-Taylor method, we estimate that average PET would only increase by between 9-14% (median of 11.5%) if temperatures warm by 4°C and R_s is held at historic values, while R_n is increased slightly due to declines in net outgoing longwave radiation with warming (Figure 9d). However, the National LSTM appears to convolute the effects of temperature and R_s and cannot separate out their effects on ~~ET-based~~evaporative water loss, leading to larger ~~projected~~ predicted streamflow losses under 4°C warming than changes in PET would warrant. This is possibly because of the very strong correlation between at-site daily temperature and R_s historically (median correlation of 0.85 across all CAMELS watersheds).

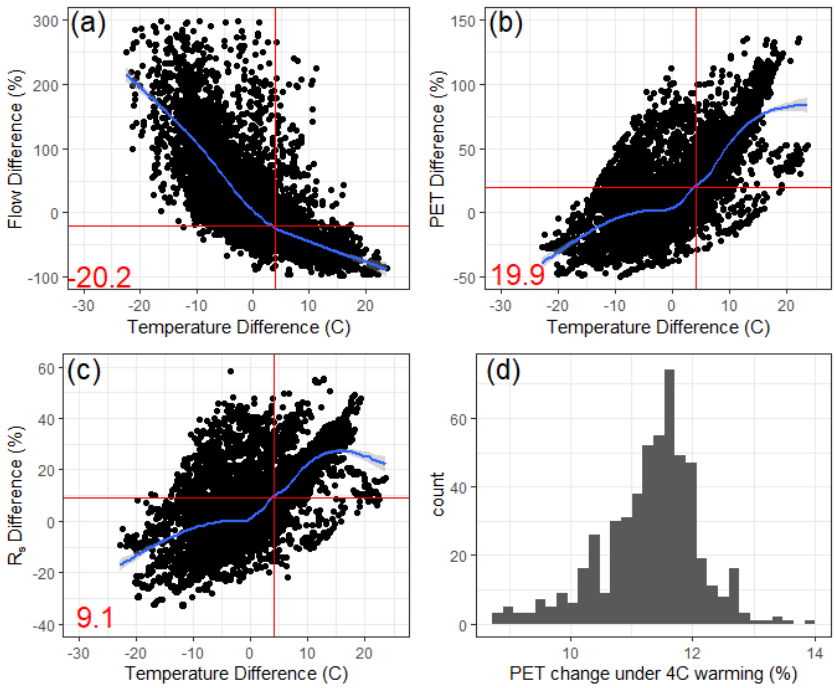


Figure 9. The percent difference in long-term (1980-2014) ~~average~~ mean (a) streamflow, (b) Priestley-Taylor based PET, and (c) downward shortwave radiation (R_s) for all pairs of CAMELS basins with average precipitation within 1% of each other, plotted against differences in average temperature for each pair. A loess smooth is provided for each scatter (blue), along with the changes in variable estimated at a 4°C temperature difference between pairs of sites (red). (d) The projected change in Priestley-Taylor based PET (as a percentage) for each CAMELS basin under 4°C warming, assuming no change in R_s.

5. Discussion and Conclusion

In this study, we contribute a sensitivity~~n~~ analysis that evaluates the physical plausibility of ~~future~~ streamflow ~~projections~~ responses under ~~climate change~~warming using DL rainfall-runoff models. The basis for this evaluation is anchored to the assumption that differences in estimated streamflow ~~projections~~ responses should emerge under very different ~~projections~~ scenarios of ~~future~~ PET under warming, and that realistic ~~projections~~ predictions of ~~future~~ PET and water loss under warming tend to be much lower than those estimated by temperature-based PET methods. Accordingly, we assume that physically plausible

~~future~~ streamflow ~~projections~~ predictions should be able to respond to lower energy-budget based PET projections under warming and, all else equal, ~~project~~ estimate smaller streamflow losses.

The results of this study show that a standard LSTM is not able to predict physically realistic differences in streamflow response across substantially different ~~projections~~ estimates of ~~future~~ PET under warming. This discrepancy ~~in future projections~~ emerged despite the fact that the standard LSTM was a far better model for streamflow estimation in ungauged basins compared to three process-based models under historic climate conditions. In addition, the National LSTM trained to a much larger set of watersheds (531 basins across 23° of latitude) using temperature, vapor pressure, and R_s directly (rather than PET) also estimated water loss under warming that far exceeded the losses estimated with process models forced with energy budget-based PET. Since water losses estimated using energy budget-based PET are generally considered more realistic (Lofgren et al., 2011; Shaw and Riha, 2011; Lofgren and Rouhana, 2016; Milly and Dunne, 2017; Lemaitre-Basset et al. 2022), this result casts doubt over the physical plausibility of the LSTM ~~projection~~predictions.

Results from this work also suggest that PIML-based DL models can capture physically plausible streamflow responses under ~~climate change~~warming while still maintaining superior prediction skill compared to process models, at least in some cases. In particular, a mass conserving LSTM that also respected the limits of water loss due to ~~ET~~ evapotranspiration (the MC-LSTM-PET) was able to ~~project~~ predict changes in ~~average~~ long-term mean streamflow that much more closely aligned with process-model based estimates, while also providing competitive out-of-sample performance across all models considered (including the other DL models). A more conventional MC-LSTM that did not limit water losses by PET was less consistent with process-based estimates of change in ~~average~~ long-term mean streamflow. These results highlight the potential for PIML-based DL models to help achieve similar performance improvements over process-based models as documented in recent work on DL rainfall-runoff models

(Kratzert et al., 2019a,b; Feng et al., 2020; Nearing et al., 2021) while also producing projections under climate change that are more consistent with theory than non-PIML DL models.

An interesting result from this study was the disagreement in the change in high flows and ~~streamflow~~ seasonal streamflow timing between all Great Lakes DL models and process models, the latter which estimated greater reductions in high flows and larger shifts of water towards earlier in the year. ~~Projections~~ Predictions from the Great Lakes DL models were also unstable if static climate properties of each watershed were changed under warming. In contrast, the National LSTM was more stable if static properties were changed, and it predicted changes to high flows and ~~streamflow~~ seasonal timing that were more like the process models than ~~projections~~ predictions from the Great Lakes DL models. The results for COM in particular suggest that the National LSTM ~~is~~may be more consistent with the process models in terms of its representation of warming effects on snow accumulation and melt processes and the resulting shifts in the seasonal hydrograph, although differences with the process model predictions were still notable. Still, these results are consistent with past work showing that large-sample LSTMs can learn to represent snow processes internally from meteorological and streamflow data (Lees et al., 2022). While its challenging to know which set of ~~projections~~ predictions are correct for these streamflow properties, these results overall favor ~~projections~~ predictions from the National LSTM over the regional LSTMs and highlight the benefits of DL rainfall-runoff models trained to a larger set of diverse watersheds for climate change analysis.

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.

~~The MC-LSTM-PET model proposed in this work represents one (relatively simple) PIML-based architectural change to an existing DL model in the hydrologic literature that can help better capture physical constraints on water loss from hydrologic systems. However, other possibilities exist. For example, the hard constraint in the MC-LSTM-PET could instead be imposed as a soft constraint through adjustments to the loss function, where water losses in the trash cell that exceed PET are penalized. The MC-LSTM-PET model could also be adjusted further to allow additional water losses in the trash cell related to human water extractions from the watershed or other terminal sinks. A different approach would be to use learnable, differentiable, process-based models with embedded neural networks (Jiang et al., 2020; Feng et al., 2022; Feng et al., 2023), which can achieve similar performance to LSTMs but can also represent and output different internal hydrologic fluxes. Further work is needed to evaluate the benefits and drawbacks of these different PIML-based approaches, preferably on large benchmarking datasets such as CAMELS.~~

~~One~~Another important limitation ~~of this study~~ is how we constructed the ~~climate change~~warming scenarios, with 4°C warming and shifts to PET but ~~but~~ no changes ~~to net incoming shortwave radiation and slight decreases in net outgoing longwave radiation with warming (i.e., slight increases in R_n)~~to other meteorological variables (net incoming shortwave radiation, precipitation, humidity, air pressure, wind speeds). These scenarios and associated sensitivity analyses were constructed in the style of other metamorphic tests for hydrologic models (Yang and Chui, 2021; Razavi, 2021; Reichert et al., 2023), where we define input changes with expected responses and test whether model behavior is consistent with these expectations. However, for DL and other machine learning ~~(ML)~~ models, the results of such sensitivity analyses may be unreliable because of distributional shifts between the training and testing data and poor out-of-distribution generalization (see Shen et al., 2021, Wang et al., 2023, and references within). When trained, conventional machine learning ~~ML~~ models try to leverage all of the correlations within the training set to minimize training errors, which is effective in out-of-sample performance only if those same patterns of correlation persistent into the testing data (Liu et al., 2021). In our experimental design, we impose a distinct shift in the joint distribution of the inputs (i.e., a covariate shift) by increasing temperatures and PET but leaving unchanged other meteorological inputs, thereby altering the correlation among inputs. Therefore, one might expect some degradation in the DL model-based predictions of streamflow under these scenarios.

~~While outside the scope of the present study, we~~The challenge of out-of-distribution generalization and its application to DL rainfall-runoff model testing under climate change highlights several important avenues for future work. First, additional efforts are needed to evaluate the ~~argue more work is needed to further explore the~~ physical plausibility of DL-based hydrologic projections under climate change~~with more standard~~ while ensuring that ~~LSTMs, with greater attention paid to~~ the joint distribution of all ~~meteorologic~~meteorological inputs used in future scenarios is realistic. For example, there are physical relationships between changes in temperature and net radiation (Nordling et al., 2021), as well as temperature, humidity, and extreme precipitation (Ali et al., 2018; Najibi et al., 2022), that should all be preserved in future climate scenarios. The use of climate model output may be well suited for such tests, although care is needed to avoid ~~significant~~ statistical bias correction and downscaling (i.e., post-processing) of multiple climate fields that could cause shifts in the joint distribution across inputs (Maraun, 2016). High-resolution convective-permitting models may be helpful in this regard, given their improved accuracy for key climate fields like precipitation ~~(~~(Kendon et al. 2017).

~~the model under historical and future climate conditions. We did not consider any changes in net incoming shortwave radiation because there is significant uncertainty in this term at local scales and its relationship to local temperature change. Projections of net incoming shortwave radiation are highly variable across~~

~~space and can even differ in the direction of change, largely because of uncertainty in the representation of clouds in climate models, future projections of aerosols, and the representation of cloud-aerosol interactions (Chen, 2021; Coppola et al., 2021; Taranu et al., 2023). The relationship between local net radiation change and local temperature change further depends on horizontal energy transport from other regions (Nordling et al., 2021). In addition, the approximation we used for changes to net outgoing longwave radiation was not designed to resolve all land-atmosphere energy balance feedbacks with changing atmospheric composition under climate change. These uncertainties, along with uncertainties in energy-budget based methods used to estimate PET (Greve et al. 2019; Liu et al., 2022), complicate future projections of atmospheric drying power under warming. Regardless, the main finding of this work remains, namely that DL models struggle to propagate different hypotheses of future PET scenarios into hydrologic projections unless explicitly directed to do so.~~

There are also several emerging techniques in machine learning ~~ML~~ to address out-of-distribution generalization directly ~~(Shen et al., 2021)~~. One ~~family~~set of promising methods ~~for the challenge of DL hydrologic modeling under climate change~~ is causal learning, defined broadly as methods ~~that~~ aimed ~~to~~at identifying input variables that have a causal relationship with the target variable and to leverage those inputs for prediction (Shen et al., 2021). PIML ~~One approach for this is to~~approaches, such as the MC-LSTM-PET model proposed in this work, fall into this category (Vasudevan et al., 2021). Here, prior scientific knowledge on casual structures can be embedded into the DL model through tailored loss functions or, as in the case of the MC-LSTM-PET model, through ~~The MC-LSTM-PET model proposed in this work represents one (relatively simple) PIML-based~~ architectural adjustments or constraints (for other examples outside of hydrology, see Lin et al., 2017; Ma et al., 2018)~~change to an existing DL model in the hydrologic literature that can help better capture physical constraints on water loss from hydrologic systems.~~ The MC-LSTM-PET model can be viewed as a specific, limited case of a broader class of ~~However, other possibilities exist. For example, the hard constraint in the MC-LSTM-PET could instead be imposed as a soft constraint through adjustments to the loss function, where water losses in the trash cell that exceed PET are penalized. The MC-LSTM-PET model could also be adjusted further to allow additional water~~

~~losses in the trash cell related to human water extractions from the watershed or other terminal sinks. A different approach would be to use~~ learnable, differentiable, process-based models ~~with embedded neural networks~~ (also referred to as hybrid differentiable models; Jiang et al., 2020; Feng et al., 2022; Feng et al., 2023a). These models use process model architectures as a backbone for model structure, which is then enhanced through flexible, data-driven learning for a subset of processes. Recent work has shown that these models ~~, which~~ can achieve similar performance to LSTMs but can also represent and output different internal hydrologic fluxes (Feng et al., 2022; Feng et al., 2023a).

However, challenges can arise when imposing architectural constraints in PIML models. For example, the MC-LSTM-PET model makes the assumption that all water loss in the system is due to evapotranspiration, and therefore cannot exceed PET. However, other terminal sinks are possible, such as human water extractions and inter-basin transfers (Siddik et al. 2023) or water lost to aquifer recharge and inter-basin groundwater fluxes (Safeeq et al., 2021; Jasechko et al., 2021). It is difficult to know the magnitude of these alternative sinks given unknown systematic errors in other inputs (e.g., underestimation of precipitation from under-catch) that confound water balance closure analyses. Still, recent techniques and datasets to help quantify these sinks (Gordon et al., 2022; Siddik et al. 2023) provide an avenue to integrate them into the MC-LSTM-PET ~~model~~ constraints ~~to improve generalizability~~. ~~However,~~Yet as constraints are added to the model architecture ~~(i.e., more assumptions are inherited from a process model backbone)~~, the potential grows for inductive bias that negatively impacts generalizability. For instance, a recent evaluation of hybrid differentiable models showed that they underperformed relative to a standard LSTM due to structural deficiencies in cold regions, arid regions, and basins with considerable anthropogenic impacts (Feng et al., 2023b). Some of these challenges may be difficult to address because only differentiable process models can be considered in this hybrid framework, limiting the process model structures that could be adapted with this approach. ~~Further~~Additional work is needed to evaluate the benefits and drawbacks of these different PIML-based approaches, preferably on large benchmarking datasets such as CAMELS or CAVARAN (Kratzert et al., 2023).

Given some of the potential challenges above, other DL methods that advance causality while making fewer assumptions on watershed-scale process controls are also worth pursuing. For example, a series of techniques have emerged that embed the concept and constraints of directed acyclic graphs within deep neural networks in such a way that the architecture of the neural network is inferred from the data to encode causality among variables (see Luo et al., 2020 and references within). That is, frameworks to optimize the architecture of the model can be designed not only to maximize out-of-sample predictive performance, but also to promote causality. Alternatively, domain-invariant learning attempts to promote the identification of features that are domain-specific versus domain invariant, by separating and labeling training data from different ‘domains’ or ‘environments’ (Ilse et al., 2021). In the case of DL rainfall-runoff models, this strategy could be implemented, for instance, by pairing observed climate and streamflow (one domain) with land surface model-based streamflow estimated using future projected climate model output (another domain), with the goal to learn invariant relationships between key climate inputs (e.g., net radiation or PET) and streamflow across the two domains. Here, there may be a benefit from including data from the land surface and climate models, where the correlation between temperature, net radiation, and PET may be weaker under projected climate change. These techniques offer an intriguing alternative for the next generation of DL hydrologic models that can generalize well under climate change, and should be the focus of further exploration. ~~identify inputs where the conditional distribution of the target variable (streamflow) given that input is invariant across heterogeneous datasets. A large focus on~~

~~Finally, we note that the results of this study do not entirely preclude the possibility that a standard LSTM, fit to a sufficiently large set of diverse watersheds, could ultimately learn more physically realistic projections under climate change. Our results with the National LSTM suggest that the signals between~~

~~temperature change and R_s on water loss may be entangled, making it difficult for the model to estimate the individual effects of changes to one of those terms (temperature) on water loss. However, it is possible that the model would produce hydrologic projections that were more in line with theory if it was given 1) high quality data on all terms related to water loss; and 2) future projections of these terms that were co-developed in physically consistent ways (e.g., from physical climate models). The R_s used in the National LSTM was based on reanalysis and so may have had meaningful errors that drove the model to attribute more water loss to warmer temperatures, and the scenario of warming given to the National LSTM (4°C warming with no change in R_s) may violate the physical relationship between temperatures and R_s. While outside the scope of the present study, we argue more work is needed to further explore the physical plausibility of hydrologic projections with more standard LSTMs, with greater attention paid to the meteorologic inputs used in the model under historical and future climate conditions.~~

Acknowledgements

This research was supported by the U.S. National Science Foundation grant CBET-2144332.

Data Availability Statement

The code used for this project is available ~~at https://doi.org/10.5281/zenodo.8190287.~~at https://doi.org/10.5281/zenodo.10027355. ~~All data used to~~ All data used to train and evaluate the models are available at https://www.hydrohub.org/mips_introduction.html#grip-gl.

References

Ali, H., Fowler, H. J., & Mishra, V. (2018). Global observational evidence of strong linkage between dew point temperature and precipitation extremes. Geophysical Research Letters, 45, 12320–12330. https://doi.org/10.1029/2018gl080557

Allen, R.G., Pereira, L.S., Raes, D., et al. (1998) Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-FAO Irrigation and Drainage Paper 56. FAO, Rome, 300(9): D05109.

Anderson, E. A. (1976). A point energy and mass balance model of a snow cover (NOAA Technical Report NWS 19). Silver Spring, MD: National Oceanic and Atmosphere Administration.

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.

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.

Bergström, S. & Forsman, A. (1973) Development of a conceptual deterministic rainfall-runoff model. Nordic Hydrol. 4, 147–170.

Beven, K. (2023). Benchmarking hydrological models for an uncertain future. Hydrological Processes, 37(5), e14882. https://doi.org/10.1002/hyp.14882

Boyle, D. P. (2001). Multicriteria calibration of hydrologic models, (Doctoral dissertation). Retrieved from UA Campus Repository (http://hdl.handle.net/10150/290657), Tucson, AZ: The University of Arizona.

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.

Burnash, R. J. (1995). The NWS river forecast system - catchment modeling. In Singh, V. (Ed.), Computer Models of Watershed Hydrology (pp. 311-366). Littleton, CO: Water Resources Publication.

Byun, K. and Hamlet, A.F. (2018), Projected changes in future climate over the Midwest and Great Lakes region using downscaled CMIP5 ensembles. Int. J. Climatol, 38: e531-e553. https://doi.org/10.1002/joc.5388

Campbell, M., Cooper, M. J. P., Friedman, K., & Anderson, W. P. (2015). The economy as a driver of change in the Great Lakes - St. Lawrence basin. *Journal of Great Lakes Research*, *41*, 69–83.

~~Cayan, D. R., Kammerdiener, S. A., Dettinger, M. D., Caprio, J. M., & Peterson, D. H. (2001). Changes in the Onset of Spring in the Western United States, *Bulletin of the American Meteorological Society*, 82(3), 399-416. https://doi.org/10.1175/1520-0477(2001)082<0399:CITOOS>2.3.CO;2~~

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. (2017). 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.

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

~~Chen, L. Uncertainties in solar radiation assessment in the United States using climate models. Clim Dyn 56, 665–678 (2021). https://doi.org/10.1007/s00382-020-05498-7~~

~~Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R. (2020). CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth Syst. Sci. Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020.~~

~~Coppola, E., Nogherotto, R., Ciarlò, J. M., Giorgi, F., van Meijgaard, E., Kadygrov, N., et al. (2021). Assessment of the European Climate Projections as Simulated by the Large EURO-CORDEX Regional and Global Climate Model Ensemble. Journal of Geophysical Research: Atmospheres, 126, e2019JD032356. https://doi.org/10.1029/2019JD032356~~

Demargne, J. et al. (2014). The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. Bull. Amer. Meteor. Soc., 95, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1.

Fan, Y. (2019). Are catchments leaky? WIREs Water, 6(6). https://doi.org/10.1002/wat2.1386

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resources Research, 56, e2019WR026793. https://doi.org/ 10.1029/2019WR026793

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., Lawson, K., and Shen, C. (2023a). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment, Hydrol. Earth Syst. Sci., 27, 2357–2373, https://doi.org/10.5194/hess-27-2357-2023.

Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., Liu, J., Pan, M., Lawson, K., and Shen, C. (2023b). 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.

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.

Frame, J.M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., et al. (2021b). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26, 3377-3392, https://doi.org/10.5194/hess-26-3377-2022.

Frame, J.M., Kratzert, F., Raney II, A., Rahman, M., Salas, F.R., & Nearing, G.S. (2021a). Post-processing the National Water Model with Long Short-Term Memory networks for streamflow predictions and diagnostics. *Journal of the American Water Resources Association*, 1-12. https://doi.org/10.1111/1752-1688.12964

Fry, L. M., Hunter, T. S., Phanikumar, M. S., Fortin, V., and Gronewold, A. D. (2013), Identifying streamgage networks for maximizing the effectiveness of regional water balance modeling, Water Resour. Res., 49, 2689– 2700, doi:10.1002/wrcr.20233.

Gasset, N., Fortin, V., Dimitrijevic, M., Carrera, M., Bilodeau, B., Muncaster, R., Gaborit, É., Roy, G., Pentcheva, N., Bulat, M., Wang, X., Pavlovic, R., Lespinas, F., Khedhaouiria, D., and Mai, J.: A 10 km North American precipitation and land-surface reanalysis based on the GEM atmospheric model, Hydrol. Earth Syst. Sci., 25, 4917–4945, https://doi.org/10.5194/hess-25-4917-2021, 2021.

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

Gauch, M., Mai, J., & Lin, J. (2021b). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling and Software*, *135*, 104926. https://doi.org/10.1016/j.envsoft.2020.104926

~~Greve, P., Roderick, M.L., Ukkola, A.M., and Wada, Y. (2019), The aridity index under global warming, Environmental Research Letters, 14, 124006, https://doi.org/10.1088/1748-9326/ab5046.~~

~~Gordon, B.L., Brooks, P.D., Krogh, S.A., Boisrame, G.F.S., Carrol, R.W.H., McNamara, J.P., & Harpold, A.A. (2022), Why does snowmelt-driven streamflow response to warming vary? A data-driven review and predictive framework, *Environmental Research Letters*, 15 (5), 053004. https://doi.org/10.1088/1748-9326/ac64b4~~

Gordon, B. L., Crow, W. T., Konings, A. G., Dralle, D. N., & Harpold, A. A. (2022). Can we use the water budget to infer upland catchment behavior? The role of data set error estimation and interbasin groundwater flow. *Water Resources Research*, *58*, e2021WR030966. https://doi.org/10.1029/2021WR030966

Greve, P., Roderick, M.L., Ukkola, A.M., and Wada, Y. (2019), The aridity index under global warming, Environmental Research Letters, 14, 124006, https://doi.org/10.1088/1748-9326/ab5046.

Gronewold, A. D., and Rood, R. B. (2019). Recent water level changes across Earth's largest lake system and implications for future variability. *Journal of Great Lakes Research*, *45*(1), 1–3.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decom- position of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91.

Hamon, W. R. (1963). Estimating Potential Evapotranspiration, T. Am. Soc. Civ. Eng., 128, 324–338, https://doi.org/10.1061/TACEAT.0008673.

Hansen, C., Shafiei Shiva, J., McDonald, S., and Nabors, A. (2019). Assessing Retrospective National Water Model Streamflow with Respect to Droughts and Low Flows in the Colorado River Basin. Journal of the American Water Resources Association 964– 975. https://doi.org/10.1111/1752-1688.12784.

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoedt, P.J., F. Kratzert, D. Klotz, C. Halmich, M. Holzleitner, G. Nearing, et al. (2021). MC-LSTM: Mass-Conserving LSTM. *arXiv e-prints*, arXiv:2101.05186. Retrieved from https://arxiv.org/abs/2101.05186

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., and Fenicia, F. (2022). Improving hydrologic models for predictions and process understanding using neural ODEs, Hydrol. Earth Syst. Sci., 26, 5085–5102, https://doi.org/10.5194/hess-26-5085-2022.

Hrachowitz, M. et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58:6, 1198-1255, DOI: 10.1080/02626667.2013.803183

Ilse, M., Tomczak, J.M., and Forré, P. (2021). Selecting Data Augmentation for Simulating Interventions. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:4555-4562.

Jasechko, S., Seybold, H., Perrone, D. et al. Widespread potential loss of streamflow into underlying aquifers across the USA. Nature 591, 391–395 (2021). https://doi.org/10.1038/s41586-021-03311-x

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. Geophysical Research Letters, 46, e2020GL088229. https://doi. org/10.1029/2020GL088229

~~Kapnick, S., & Hall, A. (2010). Observed Climate–Snowpack Relationships in California and their Implications for the Future, *Journal of Climate*, 23(13), 3446-3456. https://doi.org/10.1175/2010JCLI2903.1~~

Karpantne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318-2331. https://doi.org/10.1109/TKDE.2017.2720168

Kayastha, M.B., Ye, X., Huang, C., and Xue, P. (2022), Future rise of the Great Lakes water levels under climate change, Journal of Hydrology, 612 (Part B), 128205, https://doi.org/10.1016/j.jhydrol.2022.128205.

Kendon, Elizabeth J., Nikolina Ban, Nigel M. Roberts, Hayley J. Fowler, Malcolm J. Roberts, Steven C. Chan, Jason P. Evans, Giorgia Fosser, and Jonathan M. Wilkinson. (2017). Do Convection-Permitting Regional Climate Models Improve Projections of Future Precipitation Change? *Bulletin of the American Meteorological Society* 98 (1): 79–93. https://doi.org/10.1175/BAMS-D-15-0004.1.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *arXiv e-prints*, arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980

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.

Konapala, G., Kao, S. C., Painter, S., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environmental Research Letters, 15(10), 104022. https://doi.org/10.1088/1748-9326/aba927

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019a). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*, 11,344–11,354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. S. (2019b). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*, 5089-5110. https://doi.org/10.5194/hess-23-5089-2019

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.

Kratzert, F., Nearing, G., Addor, N. et al. (2023), Caravan - A global community dataset for large-sample hydrology. Sci Data 10, 61. https://doi.org/10.1038/s41597-023-01975-w

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.

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

Lai, C., Chen, X., Zhong, R., and Wang, Z. (2022), Implication of climate variable selections on the uncertainty of reference crop evapotranspiration projections propagated from climate variables projections under climate change, Agricultural Water Management, 259(1), 107273, https://doi.org/10.1016/j.agwat.2021.107273.

Lee, D., Lee, G., Kim, S., & Jung, S. (2020). Future Runoff Analysis in the Mekong River Basin under a Climate Change Scenario Using Deep Learning. *Water*, *12*(6):1556. https://doi.org/10.3390/w12061556

~~Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., et al. (2021). Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrology and Earth System Sciences, 26 (12), https://doi.org/10.5194/hess-26-3079-2022.~~

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J. (2022). 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.

~~Lehner, F., Wahl, E., R., Wood, A. W., Blatchford, D. B., & Llewellyn, D. (2017). Assessing recent declines in Upper Rio Grande runoff efficiency from a paleoclimate perspective. *Geophysical Research Letters*, *44*, 4124-4133. https://doi.org/10.1002/2017GL073253~~

Lehner, B., Verdin, K., and Jarvis, A. (2008). New Global Hydrography Derived From Spaceborne Elevation Data, Eos T. Am. Geophys. Un., 89, 93–94.

Lemaitre-Basset, T., Oudin, L., Thirel, G., and Collet, L.: Unraveling the contribution of potential evaporation formulation to uncertainty under climate change, Hydrol. Earth Syst. Sci., 26, 2147–2159, https://doi.org/10.5194/hess-26-2147-2022, 2022.

Li, K., Huang, G., Wang, S., Razavi, S., & Zhang, X. (2022). Development of a joint probabilistic rainfall-runoff model for high-to-extreme flow projections under changing climatic conditions. Water Resources Research, 58, e2021WR031557. https://doi. org/10.1029/2021WR031557

Lin, L., Gettelman, A., Fu, Q. et al. Simulated differences in 21st century aridity due to different scenarios of greenhouse gases and aerosols. Climatic Change 146, 407–422 (2018). https://doi.org/10.1007/s10584-016-1615-3

Lin, C., Jain, S., Kim, H., Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data, Nucleic Acids Research, Volume 45, Issue 17, 29 September 2017, Page e156, https://doi.org/10.1093/nar/gkx681

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. (2021). Heterogeneous risk minimization. In ICML, PMLR. PMLR.

Liu, X., Li, C., Zhao, T., and Han, L. (2020) Future changes of global potential evapotranspiration simulated from CMIP5 to CMIP6 models, Atmospheric and Oceanic Science Letters, 13:6, 568-575, DOI: 10.1080/16742834.2020.1824983

Liu, Z., Han, J., and Yang, H. (2022), Assessing the ability of potential evaporation models to capture the sensitivity to temperature, Agricultural and Forest Meteorology, 317, 108886.

~~Liu, Z., Han, J., and Yang, H. (2022), Assessing the ability of potential evaporation models to capture the sensitivity to temperature, Agricultural and Forest Meteorology, 317, 108886.~~

~~Liu, Z., Wang T., Han, J., Yang, W., & Yang, H. (2022). Decreases in mean annual streamflow and interannual streamflow variability across snow-affected catchments under a warming climate. *Geophysical Research Letters*, *49*(3), e2021GL097442. https://doi.org/10.1029/2021GL097442~~

Lofgren, B.M., Hunter, T.S., Wilbarger, J. (2011), Effects of using air temperature as a proxy for potential evapotranspiration in climate change scenarios of Great Lakes basin hydrology, Journal of Great Lakes Research, 37 (4), 744-752.

Lofgren, B. M., and Rouhana, J. (2016) Physically Plausible Methods for Projecting Changes in Great Lakes Water Levels under Climate Change Scenarios. J. Hydrometeor., 17, 2209–2223, https://doi.org/10.1175/JHM-D-15-0220.1.

Lu, D., Konapala, G., Painter, S. L., Kao, S. C., & Gangrade, S. (2021). Streamflow simulation in data-scarce basins using Bayesian and physics-informed machine learning models. Journal of Hydrometeorology, 22(6), 1421– 1438. https://doi.org/10.1175/JHM-D-20-0082.1

Lu, J., Sun, G., McNulty, S.G. and Amatya, D.M. (2005), A comparison of six potential evapotranspiration methods for regional use in the southeastern United States. JAWRA Journal of the American Water Resources Association, 41: 621-633. https://doi.org/10.1111/j.1752-1688.2005.tb03759.x

~~Lu, J., Sun, G., McNulty, S.G. and Amatya, D.M. (2005), A comparison of six potential evapotranspiration methods for regional use in the southeastern United States. JAWRA Journal of the American Water Resources Association, 41: 621-633. https://doi.org/10.1111/j.1752-1688.2005.tb03759.x~~

Luo, Y., Peng, J. & Ma, J. (2020). When causal inference meets deep learning. Nat Mach Intell 2, 426–427. https://doi.org/10.1038/s42256-020-0218-x

Ma, J., Yu, M., Fong, S. et al. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 15, 290–298. https://doi.org/10.1038/nmeth.4627

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. Water Resources Research, 57, e2020WR028600. https://doi. org/10.1029/2020WR028600

Mai et al. (2022). The Great Lakes runoff intercomparison project phase 4: the Great Lakes (GRIP-GL), Hydrologic and Earth System Sciences, 26 (13), 3537-3573, https://doi.org/10.5194/hess-26-3537-2022.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C. (2017). GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10, 1903– 1925, https://doi.org/10.5194/gmd-10-1903-2017.

~~Martin, J. T., Pederson, G. T., Woodhouse, C. A., Cook, E. R., McCabe, G. J., Anchukaitis, K. J., et al. (2020). Increased drought severity tracks warming in the United States' largest river basin. *Proceedings of the National Academy of Sciences*, *117*(21). https://doi.org/10.1073/pnas.1916208117~~

Maraun, D. (2016). Bias Correcting Climate Change Simulations - a Critical Review. Curr Clim Change Rep 2, 211–220. https://doi.org/10.1007/s40641-016-0050-x

~~McCabe, G. J., Wolock, D. M., Pederson, G. T., Woodhouse, C. A., & McAfee, S. (2017). Evidence that recent warming is reducing upper Colorado River flows. *Earth Interactions*, *21*(10), 1-14. https://doi.org/10.1175/EI-D-17-0007.1~~

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.

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.

Milly, P.C.D. and Dunne, Krista A. (2017). A Hydrologic Drying Bias in Water-Resource Impact Analyses of Anthropogenic Climate Change. Journal of the American Water Resources Association (JAWRA) 53(4): 822– 838. https://doi.org/10.1111/1752-1688.12538

~~Milly, P. C. D., & Dunne, K. A. (2020). Colorado River flow dwindles as warming-driven loss of reflective snow energizes evaporation. *Science*, *367*(6483), 1252-1255. https://doi.org/10.1126/science.aay9187~~

~~Monteith, J. L. (1965), Evaporation and environment, in: Symposia of the society for experimental biology, volume 19, Cambridge University Press (CUP), Cambridge, UK, 205–234 pp.~~

Mote, P. W., Li, S., Lettenmaier, D. P., Xiao, M., & Engel, R. (2018). Dramatic declines in snowpack in the western US. *npj Climate and Atmospheric Science*, *1:2*. https://doi.org/10.1038/s41612-018-0012-1

NACLMS: NACLMS website, http://www.cec.org/north-american- environmental-atlas/land-cover-2010-landsat-30m/ (last access: 31 May 2023), 2017.

Najibi, N., Mukhopadhyay, S., & Steinschneider, S. (2022). Precipitation scaling with temperature in the Northeast US: Variations by weather regime, season, and precipitation intensity. Geophysical Research Letters, 49, e2021GL097100. https://doi.org/10.1029/2021GL097100

Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*, e2020WR028091. https://doi.org/10.1029/2020WR028091

Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., and Nevo, S. (2022). Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, Hydrol. Earth Syst. Sci., 26, 5493–5513, https://doi.org/10.5194/hess-26-5493-2022.

Newman, A., Clark, M. P., Sampson, K., Wood, A., Hay, L., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209-223. https://doi.org/10.5194/hess-19-209-2015

Nordling, K., Korhonen, H., Raisanen, J., Partanen, A.-I., Samset, B.H., and Merikanto, J. (2021), Understanding the surface temperature response and its uncertainty to CO_2, CH_4, black carbon, and sulfate, Atmos. Chem. Phys., 21, 14941-14958.

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

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.

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.

Priestley, C. H. B., and Taylor, R. J. (1972). On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. Mon. Wea. Rev., 100, 81–92, https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2.

Pryor, S.C., Barthelmie, R.J., Bukovsky, M.S. et al. Climate change impacts on wind power generation. Nat Rev Earth Environ 1, 627–643 (2020). https://doi.org/10.1038/s43017-020-0101-7

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling, Environmental Modelling and Software, 105159, https://doi.org/10.1016/j.envsoft.2021.105159.

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.

~~Rungee, J., Ma, Q., Goulden, M. L., & Bales, R. (2021). Evapotranspiration and runoff patterns across California's Sierra Nevada. *Frontiers in Water*, *3:655485*. https://doi.org/10.3389/frwa.2021.655485~~

Safeeq, M., Bart, R. R., Pelak, N. F., Singh, C. K., Dralle, D. N., Hartsough, P., & Wagenbrenner, J. W. (2021). How realistic are water-balance closure assumptions? A demonstration from the southern sierra critical zone observatory and kings river experimental watersheds. Hydrological Processes, 35: e14199. https://doi.org/10.1002/hyp.14199

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.

Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H. (2014). A global soil data set for earth system modeling, J. Adv. Model. Earth Sy., 6, 249–263.

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

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. (2021). Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:2108.13624.

Siddik, M.A.B., Dickson, K.E., Rising, J. et al. Interbasin water transfers in the United States and Canada. Sci Data 10, 27 (2023). https://doi.org/10.1038/s41597-023-01935-4

Steinman, A.D. et al. (2017), Ecosystem services in the Great Lakes, Journal of Great Lakes Research, 43 (3), 161-168. https://doi.org/10.1016/j.jglr.2017.02.004

~~Stewart, I. T., Cayan, D. R., & Dettinger, M. D. (2005). Changes toward Earlier Streamflow Timing across Western North America, *Journal of Climate*, 18(8), 1136-1155. https://doi.org/10.1175/JCLI3321.1~~

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

Szilagyi, J., Crago, R., and Qualls, R. (2017), A calibration-free formulation of the complementary relationship of evaporation for continental-scale hydrology, J. Geophys. Res. Atmos., 122, 264– 278, doi:10.1002/2016JD025611.

~~Taranu, I.S., Somot, S., Alias, A. et al. Mechanisms behind large-scale inconsistencies between regional and global climate model-based projections over Europe. Clim Dyn 60, 3813–3838 (2023). https://doi.org/10.1007/s00382-022-06540-6~~

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., and Zhang, Y. (2023): 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.

Vasudevan, R.K., Ziatdinov, M., Vlcek, L. et al. (2021). Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality. npj Comput Mater 7, 16. https://doi.org/10.1038/s41524-020-00487-0

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.

Wang, Q. J. (1991). The genetic algorithm and its application to calibrating conceptual rainfall-runoff models, *Water Resources Research*, *27*(9), 2467-2471. https://doi.org/10.1029/91WR01305

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P.S. (2023). Generalizing to Unseen Domains: A Survey on Domain Generalization, in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 8, pp. 8052-8072, 1 Aug. 2023, doi: 10.1109/TKDE.2022.3178128.

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

~~Wolock, D.M., McCabe, G.J. (1999). Estimates of runoff using water-balance and atmospheric general circulation models. Journal of the American Water Resources Association 35: 1341–1350.~~

~~Woodhouse, C. A., & Pederson, G. T. (2018). Investigating runoff efficiency in upper Colorado river streamflow over past centuries. *Water Resources Research*, *54*, 286-300. https://doi.org/10.1002/2017WR021663~~

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

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

Yang, Y., & Chui, T. F. M. (2021). Reliability assessment of machine learning models in hydrological predictions through metamorphic testing. Water Resources Research, 57, e2020WR029471. https://doi.org/10.1029/2020WR029471

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.

Zhong, L., Lei, H., & Gao, B. (2023). Developing a physics-informed deep learning model to simulate runoff response to climate change in Alpine catchments. Water Resources Research, 59, e2022WR034118. https://doi.org/10.1029/2022WR034118

Supplemental Material 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[1], **Scott Steinschneider**[1]

[1]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA

Summary

This supplementary material file contains one supplemental section of text, ~~six eight figures, one supplemental section of text, and~~ three ~~additional~~ tables, and ~~eight~~nine figures in support of the analysis and conclusions presented in the main article.

Text S1: Adjustments to Static Attributes

In the primary article, we describe two sets of scenarios for the deep learning models used in this work: 1) one in which changes are only made to the dynamic inputs features of each model, and 2) one with changes to both dynamic features and to static features that depend on those dynamic features. Here we describe in more detail the adjustments made to the static features for each site, which include: pet_mean, aridity, t_mean, frac_snow (see Table S1 below for the definition of these features). Importantly, these are the static features that are dependent on temperature and PET, the two dynamic inputs adjusted in our analysis.

To adjust t_mean, we use the full time series of daily maximum and minimum temperature (on which t_mean was originally based), and shift those time series upward by 4°C. Using those adjusted series, we calculate daily average temperature as the mean of maximum and minimum temperature on each day, and then calculate the long-term mean of daily average temperature to develop an updated estimate of t_mean.

To adjust frac_snow, we first calculate the adjusted time series of daily average temperature based on the time series of daily maximum and minimum temperature shifted upward by 4°C. Then, we count all days in the record when precipitation occurs and this adjusted time series of daily average temperature is below 0°C, and divide this number by the total number of days of non-zero precipitation in the record. The resulting value is the updated value for frac_snow.

We develop two versions of adjusted pet_mean, one based on Hamon PET and the other for Priestley-Taylor PET. The adjusted Hamon PET is based entirely on the series of daily maximum and minimum temperature shifted by 4°C. We use Eqs. 7-8 in the main article to calculate daily Hamon PET under warming. We then take the long-term mean of this time series to develop an updated estimate of pet_mean. Similarly, for Priestley-Taylor PET, we couple the warmed temperature time series with the unadjusted time series of net shortwave radiation, and then use the approach in Eq. 9 in the main article to calculate a daily time series of Priestley-Taylor PET. We again take the long-term mean of this time series to develop an updated estimate of pet_mean.

Finally, we develop two versions of adjusted aridity, one based on Hamon PET and the other for Priestley-Taylor PET. In both cases, we calculate adjusted aridity as the ratio of the updated values for pet_mean under warming and the unadjusted value for long-term mean precipitation (another static input to the models).

Table S1. Static watershed attributes that are adjusted in a subset of scenarios used in this analysis.

Attribute	Description
pet_mean	Mean daily potential evapotranspiration
aridity	Ratio of mean PET to mean precipitation
t_mean	Mean of daily maximum and daily minimum temperature

frac_snow	Fraction of precipitation falling on days with mean daily temperatures below 0°C

Additional Supporting Tables

Table S2. Range of values considered in the grid search during hyper-parameter tuning.

Hyper-parameter	**Values Tested**
Number of Hidden Layer Nodes	64, 96, 128, 256
Mini-Batch Size	64, 128, 256, 512
Learning Rate	0.0001, 0.0005, 0.001, 0.005
Number of Epochs	30, 50
Dropout Rate*	0, 0.2, 0.4

Table S3. Additional details for gauges highlighted in Figures 5 and 6 of main article.

Gauge ID	**Country**	**Site Name**	**Drainage Area (km^2)**
02ED032	Canada	Willow Creek near Minesing	231
02GG013	Canada	Black Creek near Bradshaw	213
02HJ003	Canada	Ouse River near Westwook	283
04126740	United States	Platte River at Honor, MI	324
04220045	United States	Oak Orchard Creek near Shelby NY	378
04168400	United States	Lower River Rouge at Dearborn, MI	236

Additional Supporting Figures

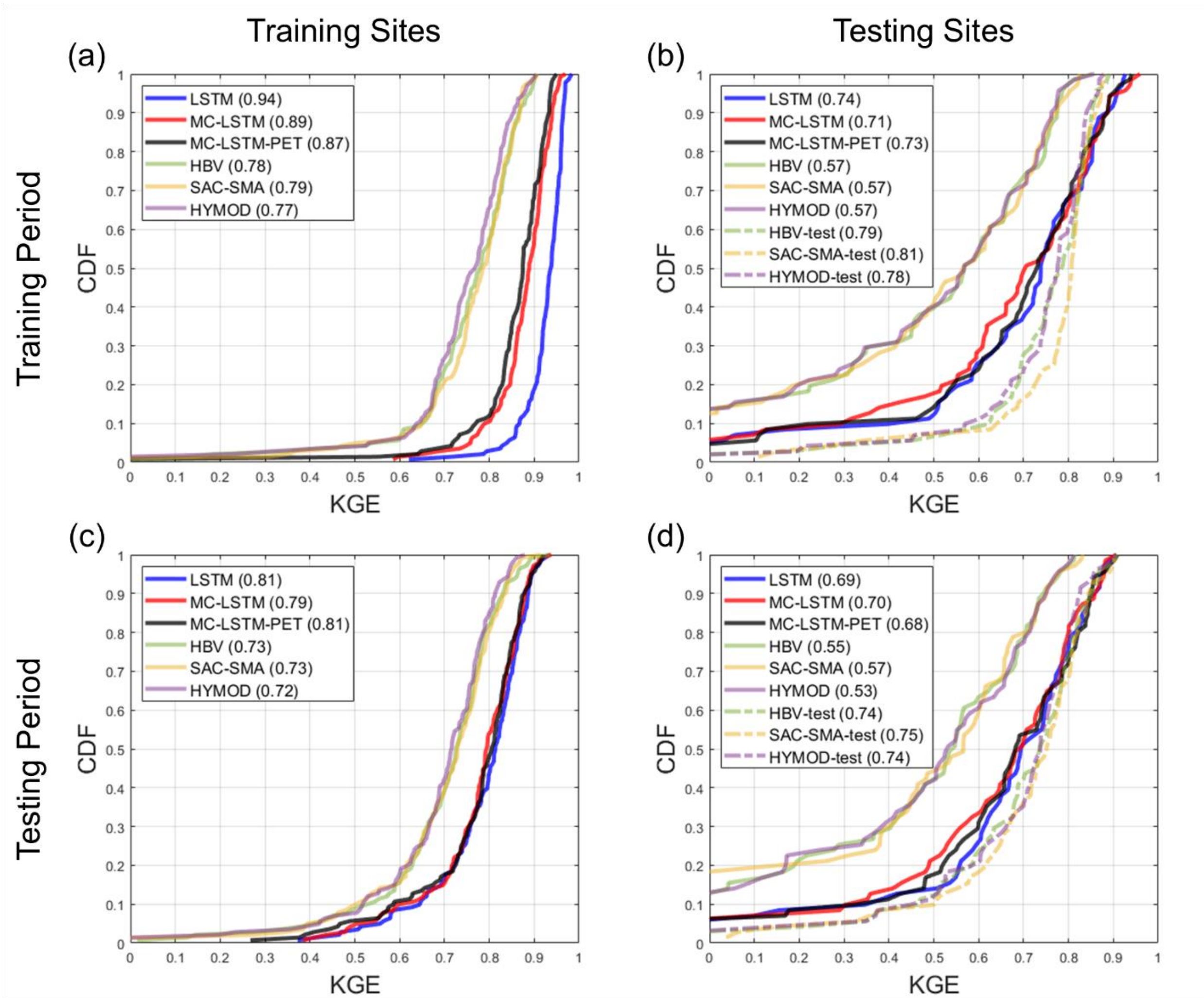


Figure S1. The distribution of Kling-Gupta efficiency (KGE) for streamflow estimates across sites from each model at the (a) the 141 training sites and (b) 71 testing sites for the training period. Similar results for the testing period are shown in panels (c) and (d), respectively. For the process models fit to the testing sites (denoted "-test"), no performance results are available at the training sites. All models are trained using Hamon PET.

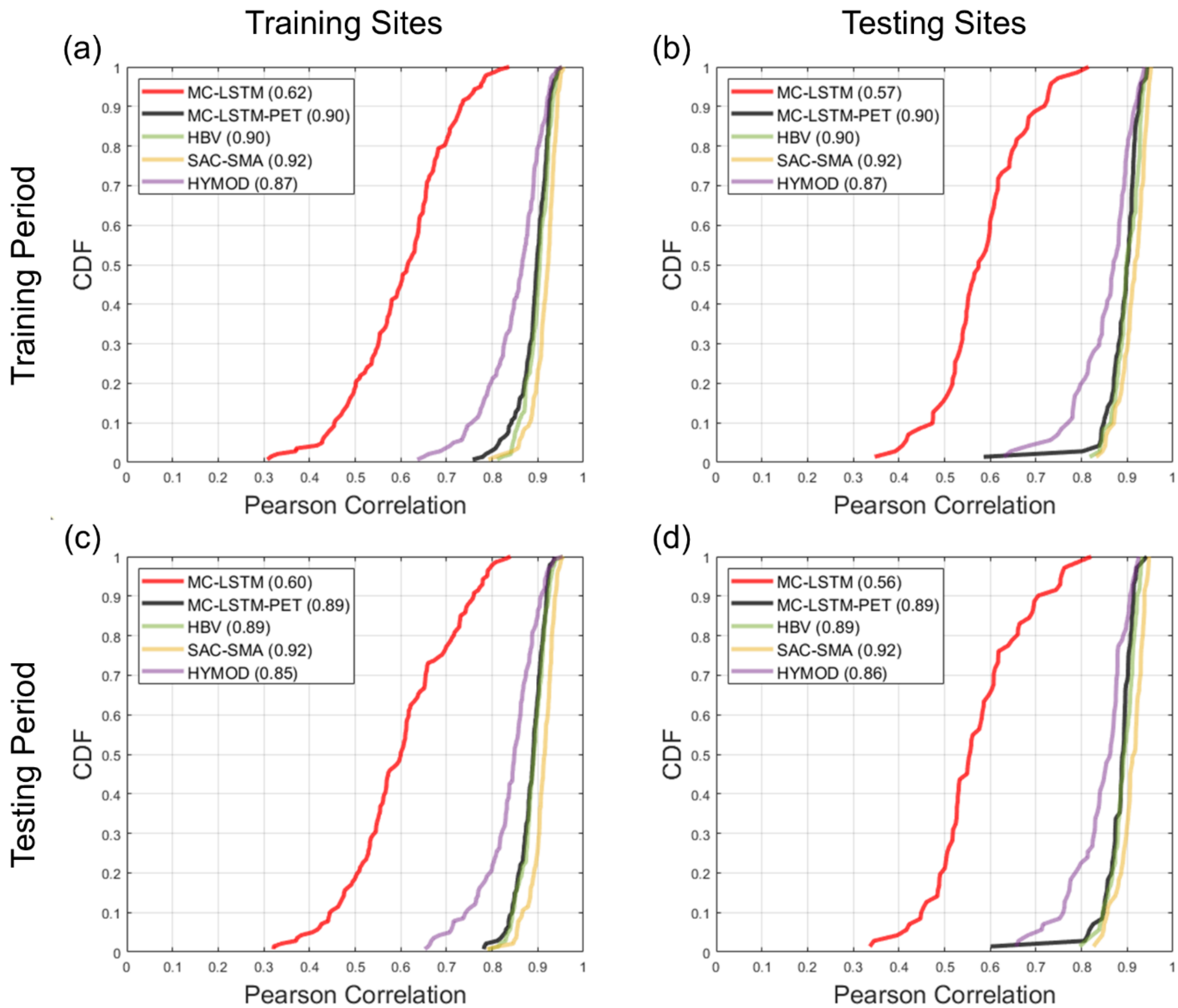


Figure S2. The correlation between model estimated and ~~observed~~ GLEAM AET from each model at the (a) the 141 training sites and (b) 71 testing sites for the training period. Similar results for the testing period are shown in panels (c) and (d), respectively. The LSTM is not included in this comparison. All models are trained using Priestley-Taylor PET.

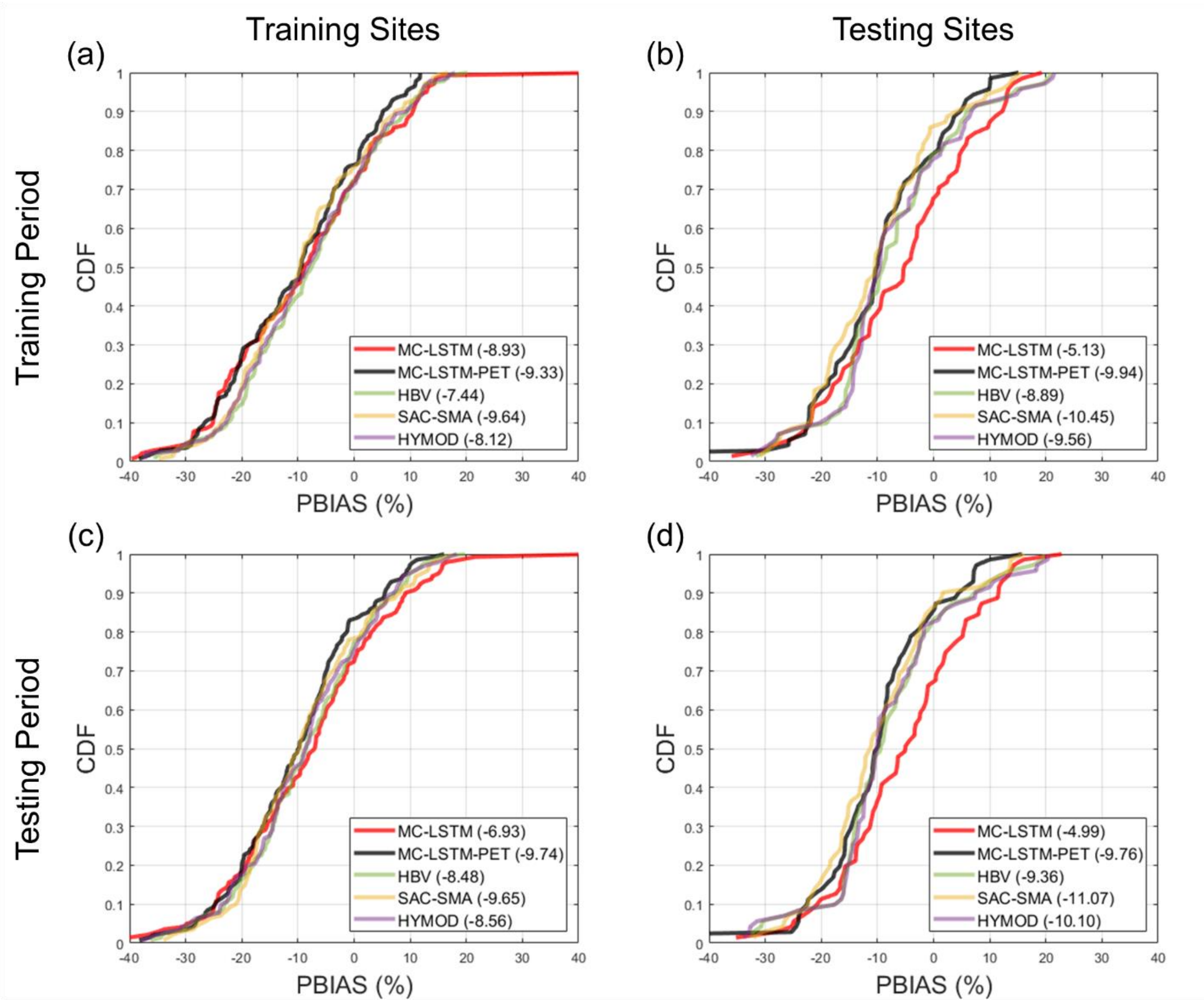


Figure S3. The PBIAS between model estimated and GLEAM ~~observed~~ AET from each model at the (a) the 141 training sites and (b) 71 testing sites for the training period. Similar results for the testing period are shown in panels (c) and (d), respectively. The LSTM is not included in this comparison. All models are trained using Priestley-Taylor PET.

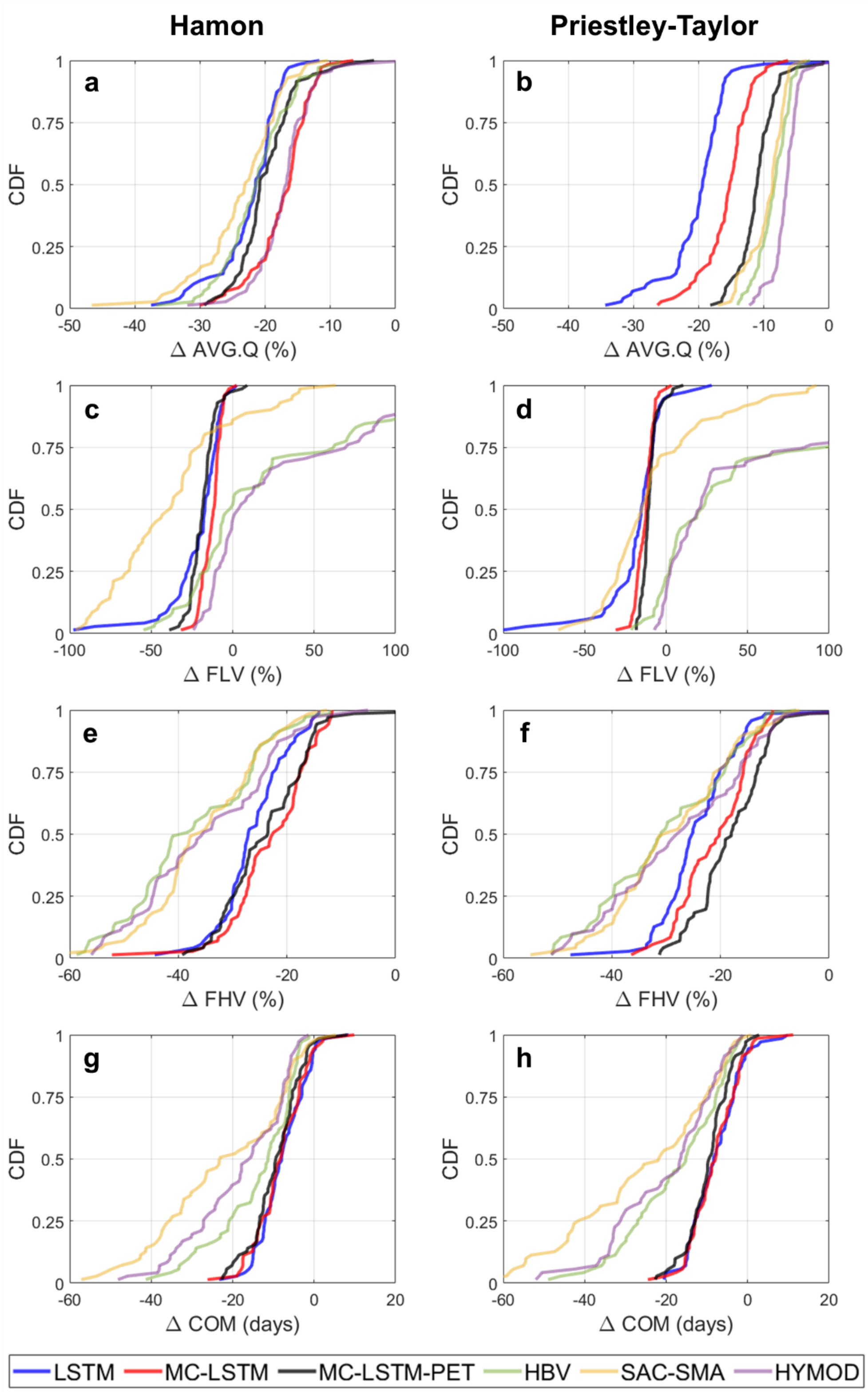


Figure S4~~5~~. The distribution of change in (a,b) AVG.Q, (c,d) FLV, (e,f) FHV, and (g,h) COM across the 71 testing sites and all models under a scenario of 4°C warming using (a,c,e,g) Hamon PET and (b,d,f,h) Priestley-Taylor PET. For the DL models, changes were only made to the dynamic inputs (i.e., no changes to static inputs).

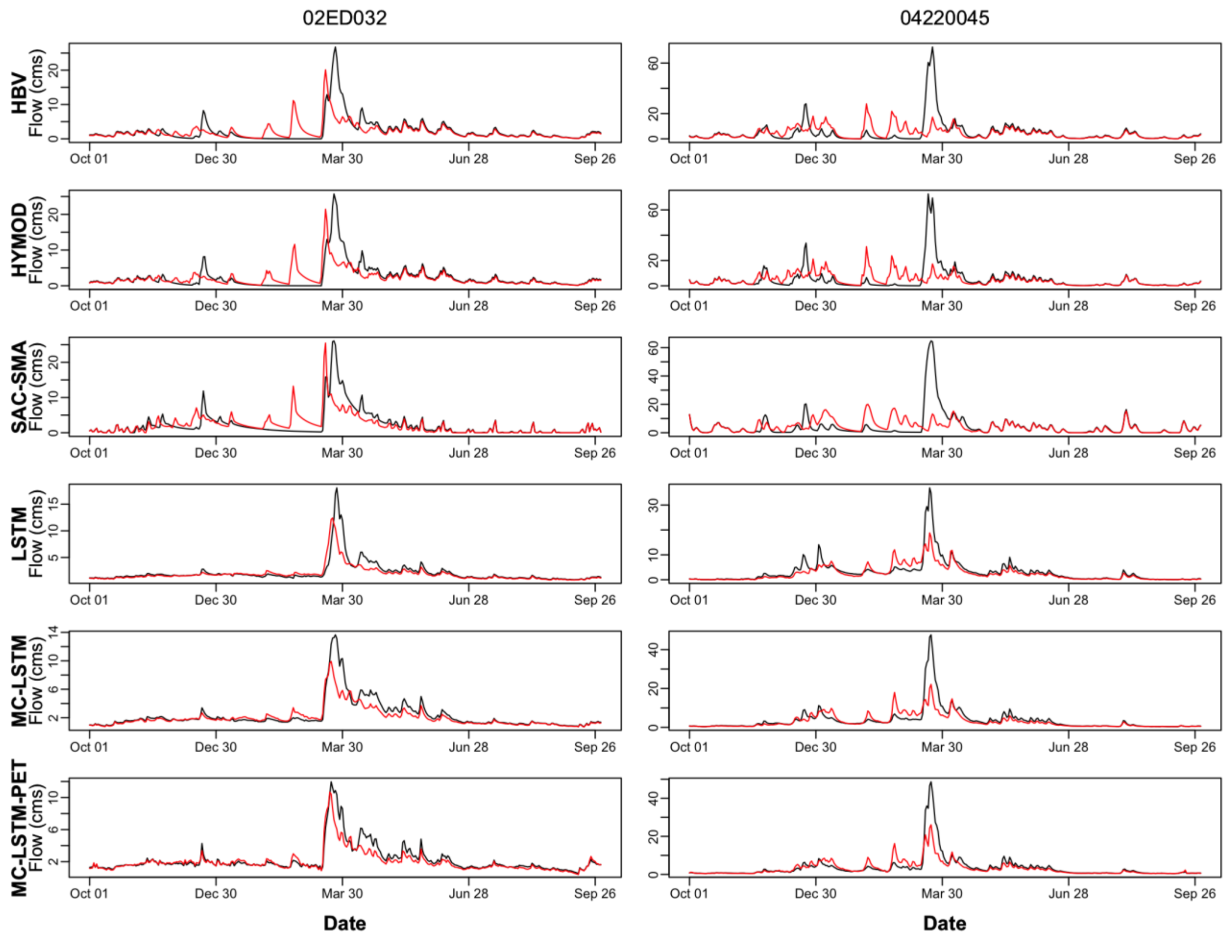


Figure S5. Daily streamflow hydrograph for one water year (2002 October- 2003 September) across the three different process-based models (HBV, HYMOD, SAC-SMA) and deep-learning models (LSTM, MC-LSTM, MC-LSTM-PET) under 0°C warming (black) and 4°C warming (red). Results are shown for two sites (highlighted in Figure 1 of the main article), and are constructed with models using Priestley-Taylor PET.

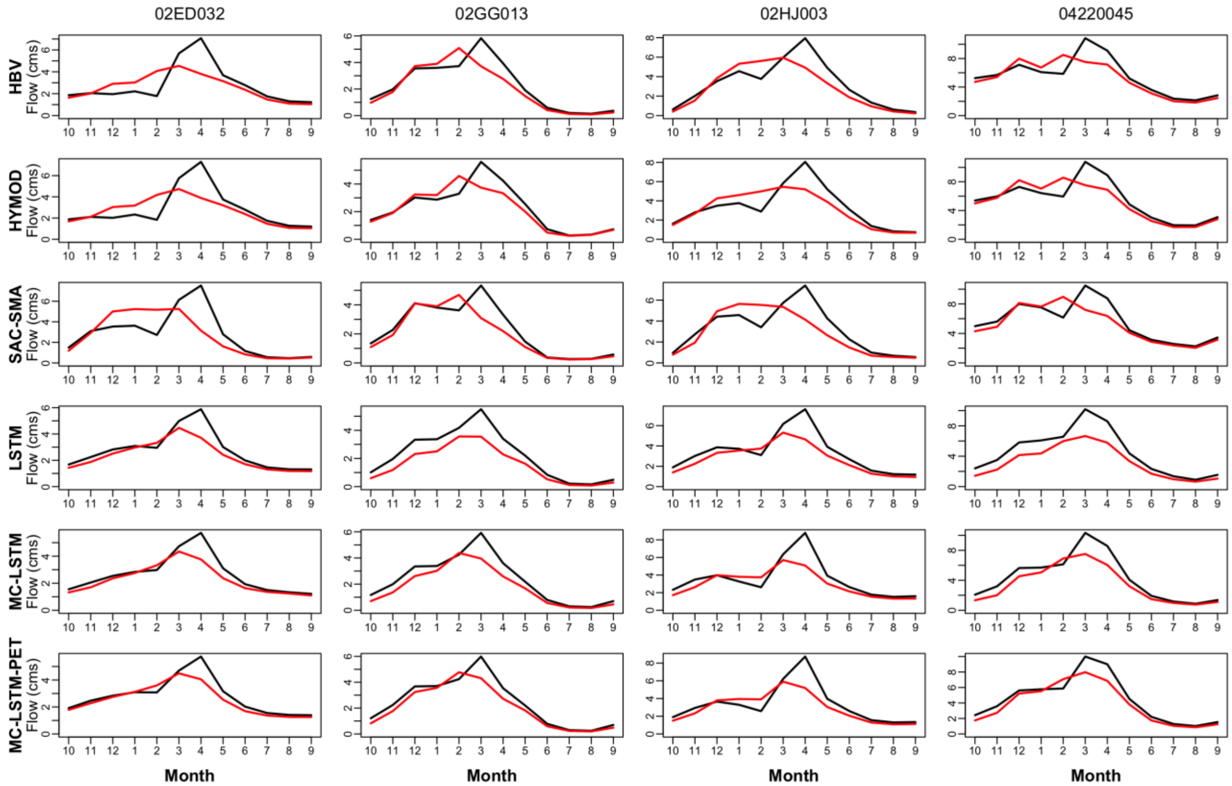


Figure S6. Mean monthly streamflow averaged across the entire record, shown ~~throughout the water year (October-September) across~~for the three different process-based models (HBV, HYMOD, SAC-SMA) and deep-learning models (LSTM, MC-LSTM, MC-LSTM-PET) under 0°C warming (black) and 4°C warming (red). Results are shown on a water year basis (October-September) for four sites (highlighted in Figure 1 of the main article), and are constructed with models using Priestley-Taylor PET.

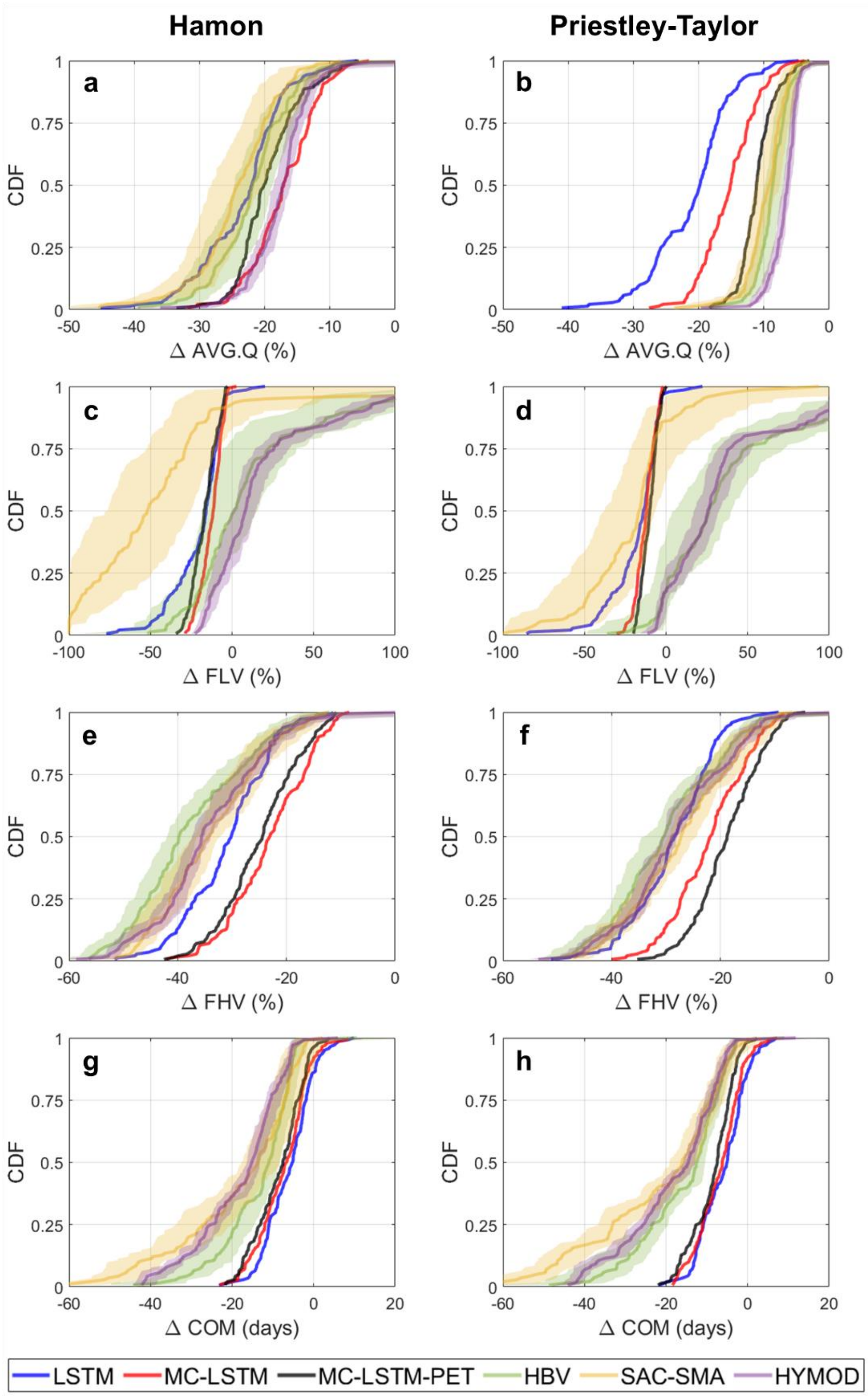


Figure S7. The distribution of change in (a,b) long term mean daily flow (AVG.Q), (c,d) low flows (FLV), (e,f) high flows (FHV), and (g,h) seasonal streamflow timing (COM) across the 141 training sites and all models under a scenario of 4°C warming using (a,c,e,g) Hamon PET and (b,d,f,h) Priestley-Taylor PET. For the deep learning models, changes were only made to the dynamic inputs (i.e., no changes to static inputs). For the process models, the uncertainty in the change in each streamflow attribute across 10 different training trails is shown as translucent shading.

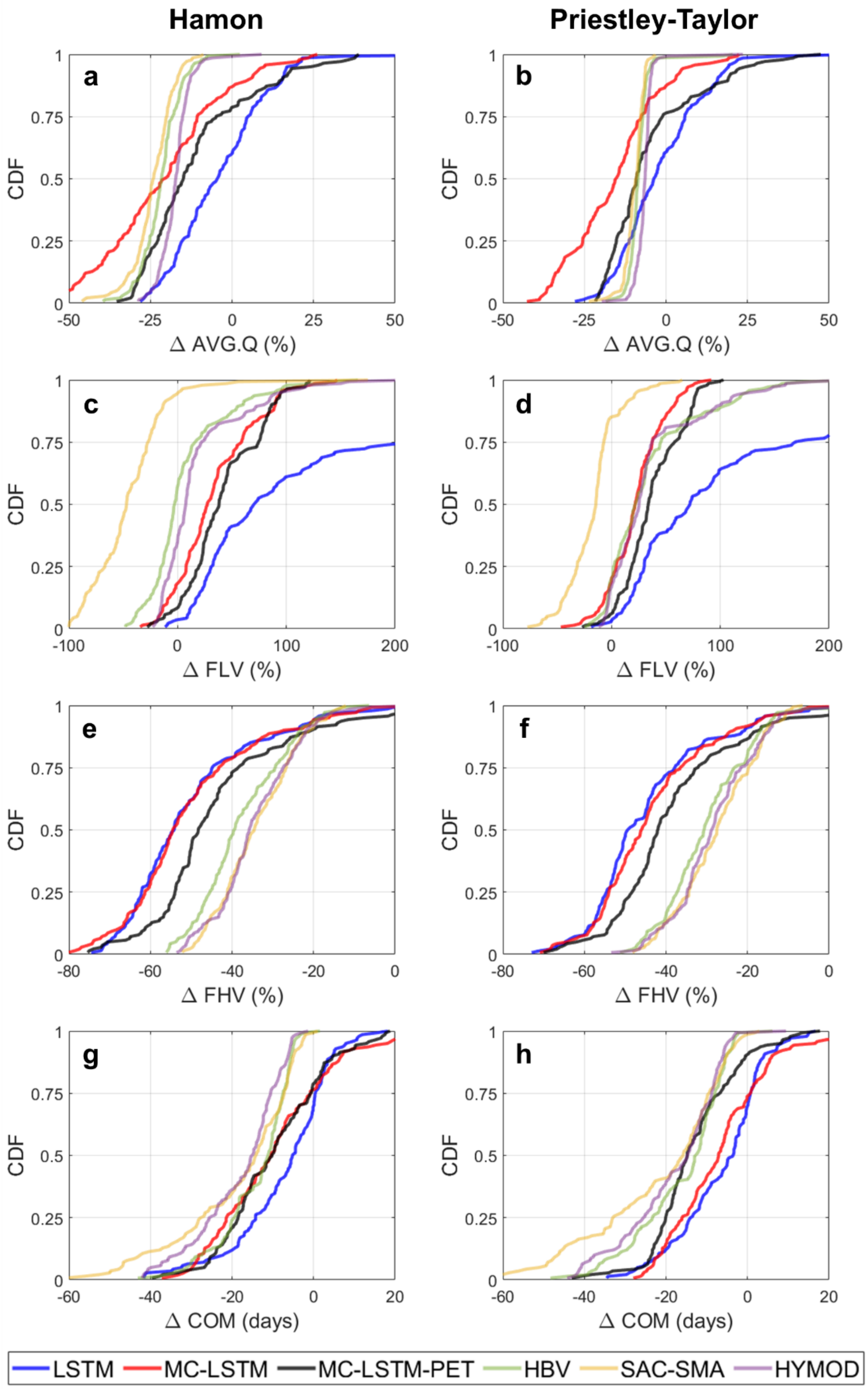


Figure S84. The distribution of change in (a,b) AVG.Q, (c,d) FLV, (e,f) FHV, and (g,h) COM across the 141 training sites and all models under a scenario of 4°C warming using (a,c,e,g) Hamon PET and (b,d,f,h) Priestley-Taylor PET. For the DL models, changes were made to both the dynamic and static inputs.

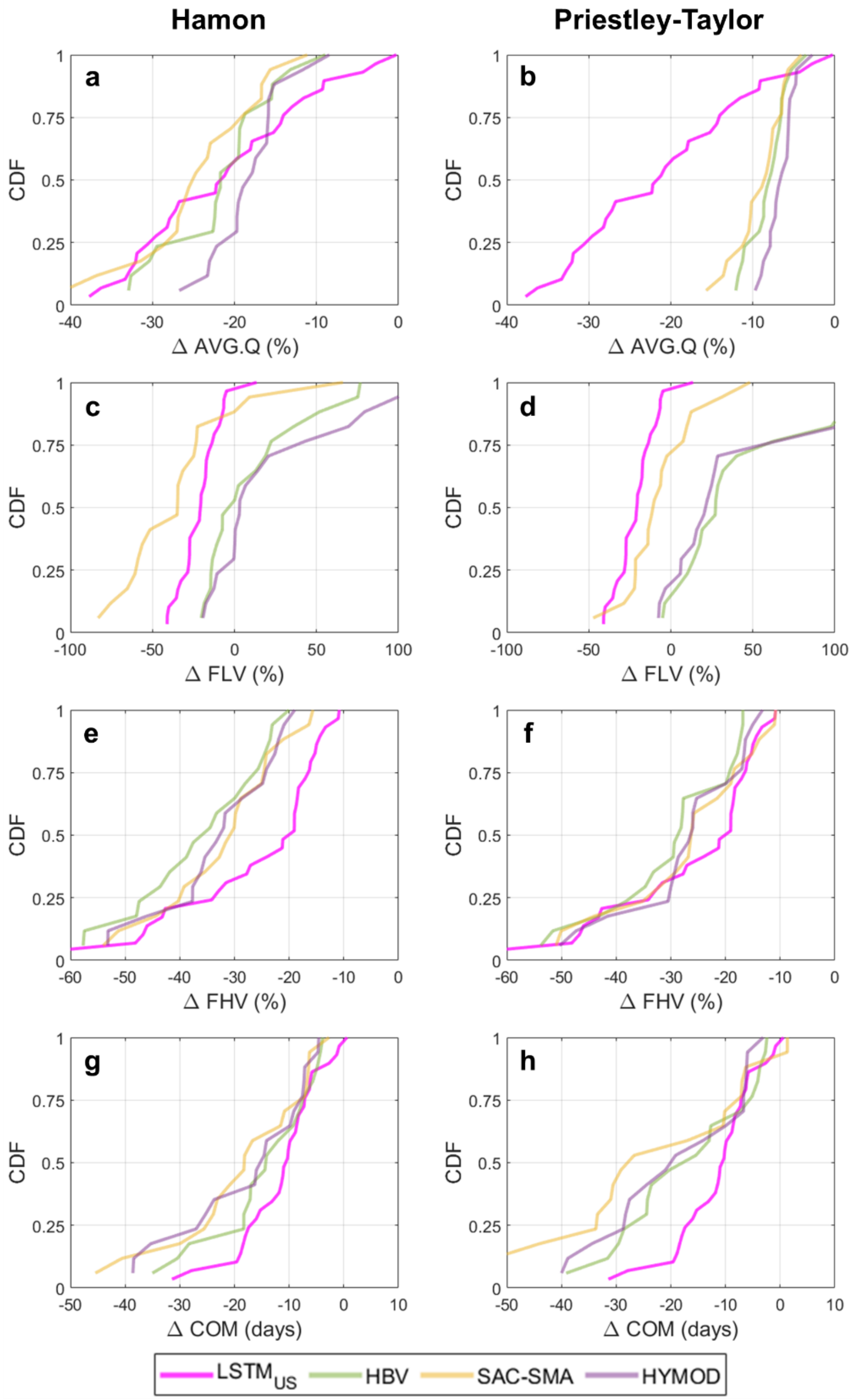


Figure S9~~6~~. The distribution of change in (a,b) AVG.Q, (c,d) FLV, (e,f) FHV, and (g,h) COM across 29 CAMELS sites within the Great Lakes basin under the National LSTM, as well as for 17 of those 29 sites from the Great Lakes process models, under a scenario of 4°C warming. For the process models only, results differ when using (a,c,e,f) Hamon PET and (b,d,f,h) Priestley-Taylor PET. For the National LSTM, changes were made to both the dynamic and static inputs.