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 #2

General comments:

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

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

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

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

The reviewer is entirely correct that snowmelt dynamics in regions like the Great Lakes will change under warming. Our focus on PET and long-term average water loss was motivated by the need for a strategy to prove whether or not the hydrologic responses to warming from the DL models are credible. In many cases, it is difficult to know a priori the true magnitude of hydrologic responses to different types of climate change. Therefore, unless the DL models estimate a hydrologic change that disagrees in sign with estimated (and well-accepted) changes from a process model, it is difficult to draw firm conclusions on the credibility of DL-based estimates of hydrologic change.

In our previous work (SW22), we focused on the sign of streamflow change to determine if the DL models were estimating reasonable hydrologic changes under strict warming. In this work, we tried to add nuance to this assessment by focusing on the magnitude of average water loss under warming, leveraging the distinct responses that should be expected under temperature-based and energy budget-based PET estimates to evaluate the reasonableness of estimated responses.
When designing our experiment, we carefully weighed which attributes of streamflow change to include, trying to balance two factors. On the one hand, we wanted to provide a broad set of statistics that captured different aspects of streamflow change, since we thought this would be of interest to a wider readership, especially since these types of assessments of DL models under warming are very rare in the literature. On the other hand, we wanted to limit the number of streamflow attributes to only those that were directly relevant to our core arguments, namely that DL models struggled to estimate long-term average water loss under warming due to their inability to separate historical temperature and PET correlations that will likely change in the future.

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

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

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

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

1) First, we have added language in the Methods, Results, and the Discussion and Conclusion section making a more direct connection between the COM statistic and model estimated changes to snow dynamics. We highlight in these additions that the DL models do not represent snow processes (or any hydrologic processes) explicitly, but rather learn these processes from the data. We note here (and also in our revisions) that others (e.g., Lees et al., 2022) have correlated internal states from LSTMs to independent measures of snow and find very strong relationships, suggesting that these models do learn the patterns of snow accumulation and melt directly from precipitation, temperature, and streamflow data. However, we note in our Discussion and Conclusion that it seems the National LSTM may do this better than the Great Lakes DL models, given the stronger consistency with the process models in terms of changes to the COM statistic under warming (although differences were still notable).

2) Second, when discussing the results in Figure 7 regarding the COM statistic, we also now refer to a new figure in the Supporting Information that shows the shift in the monthly hydrology between the baseline and 4C warming scenario across all models. This figure more clearly shows the shifts in flow during the winter and spring, and how this shift differs between the process and DL models, in order to more clearly demonstrate how these models are simulating hydrologic response under changing snow conditions with warming.


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

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

- The Title, which now reads: On the need for physical constraints in deep leaning rainfall-runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration
In the Abstract, Methods, Results, and Discussion and Conclusions, we have made numerous changes to emphasize the fact that this study is conducting a sensitivity analysis rather than a formal set of projections under internally consistent climate change scenarios. Where appropriate, we removed reference to the phrases “climate change”, “projections”, and “future predictions” to avoid convoluting our analysis with actual future projections under climate change, when we are really referring to a sensitivity analysis of model responses to warmed historical temperatures and associated PET changes.

In addition, we have added significant discussion and additional literature review to our Discussion and Conclusion section to address the concerns expressed here (and by Reviewer 1). Specifically, we now discuss the machine learning literature on causality and distribution shifts, specifically in the context of our experimental design and its limitations to address the full spectrum of climate change, and the potential for that design to influence the results of this work.

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

Please see our two responses to the major comments above.

Specific comments follow,

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

We recognize that the abstract for this paper is longer than most conventional papers. The journal does provide guidelines that “abstracts should be short, clear, concise…”, as the reviewer notes. However, they also promote more comprehensive abstracts so that readers can get a fully sense of the article’s content from the abstract, and then don’t impose page limits on the full manuscript, which together is their strategy to promote both conciseness and completeness (see here: https://www.hydrology-and-earth-system-sciences.net/about/faqs.html):

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

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

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

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

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

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

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

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

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

We have added several of the in-line citations we previously cited in this paragraph to the end of the line in question, as all those articles show convincingly that temperature-based PET methods overestimate PET and water loss under warming.
L170-172, as noted in my previous comments, the hypotheses may suggest that PET is the sole or overwhelming cause of declining streamflow.

While we agree that other factors (like snow accumulation and melt dynamics) will significantly impact streamflow behavior under warming, we do still assert that changes in PET (which could include changes in plant transpiration under changes in CO$_2$, as mentioned by the reviewer earlier) are the dominant cause of *long term average* streamflow decline under warming, absent any precipitation changes. We emphasize ‘long term average’ to distinguish our focus from declines in streamflow in certain seasons, which could be balanced by increases in streamflow in other seasons. This is exactly what one might expect due to changing snow accumulation and melt dynamics, where warmer temperatures lead to more water running off into the stream during the winter and early spring months, which is then balanced by less runoff from snowmelt later in the spring and early summer. In this situation, precipitation entering the watershed is being redistributed in terms of the timing of streamflow, but the long-term average streamflow is not significantly impacted. The major way that climate can drive a long-term average decline in streamflow across multiple years is by either changing the long-term total amount of water entering the system (i.e., precipitation) or by changing the long-term total amount of water that leaves the system before it can reach the stream (i.e., evapotranspiration).

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

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

In the figure below, we show the distribution of Pearson correlation values between daily average temperature (the mean of daily tmax and tmin) and PET calculated using either the Hamon method or the Priestley Taylor method, shown across all basins in the Great Lakes region. The average correlation between temperature and Hamon PET and Priestley Taylor PET is 0.94 and 0.83, respectively. We think the most natural place to present this information is when discussing the calculation of Hamon and Priestley Taylor PET in Section 3.3, after the other meteorological data and sites in the study domain have been introduced. We have now included the mean correlation values in this section in the revised manuscript.
Fig. 2. I am confused by the presentation of the timing, as it seems to suggest that the cumulative flux of warming stops increasing after a certain day of the water year.

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

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

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

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

For a broader discussion of this point, please refer to our response to the third general comment from Reviewer #3.
L412, I think it depends on how many catchments of your study are natural without human intervention.

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

Here, the ReLU activation ensures that any water in the trash cell \((h_D)\) which exceeds PET at time \(t\) is added to the streamflow prediction \(y[t]\), but the streamflow prediction is the same as the original MC-LSTM (Eq. 5) if water in the trash cell is less than PET. This approach assumes that the maximum allowable water lost from the system cannot exceed PET, and therefore ignores other potential terminal sinks (e.g., inter-basin lateral groundwater flows; human diversions and inter-basin transfers). This assumption is more strongly supported in moderately-sized (> 200 km²), low-gradient, non-arid watersheds where inter-basin groundwater flows are less impactful (Fan 2019; Gordon et al., 2022), such as the Great Lakes basins examined in this work. However, we discuss the potential to relax the assumptions of the MC-LSTM-PET model in Section 5.

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

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

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

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

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

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

We also consider changes to some of the static input features that depend on temperature and PET in their calculation (e.g., pet_mean, aridity, t_mean, frac_snow; see Table 1 for...
L510, does the "year" here indicate calendar year or water year, it seems water year based on Figure 2.

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

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

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

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

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

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

L663-664, as I mentioned in the general comments, this may be due to the role of snow dynamics being treated differently between the process-based and deep learning models. Unfortunately, analysis on snowmelt was not included in the study.
Please see our response to the reviewer’s major comment above regarding how we now better address hydrologic responses to changing snow accumulation and melt in our analysis of model sensitivity to warming.

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

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

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

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

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

Agreed. Consistent with the reviewer’s second major comment, we have avoided using the term “climate change scenario” throughout the manuscript when referring to the scenarios used in our sensitivity analysis. In this instance, we changed our wording to reference the “warming scenarios” used in this work. More broadly, we now discuss more thoroughly the climate change signals not addressed in this work in the last paragraph of the Conclusion, including Clausius-Clapeyron scaling and its potential impact on extreme precipitation under warming.
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 Wi¹, Scott Steinschneider¹

¹Department 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 overestimates 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, assuming 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, 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 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, we also explore similar projections
responses using a National LSTM fit 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 learningDL 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 learningDL models exhibit smaller changes in high flows and streamflow-seasonal timing of flows as compared to the process models while deep learningDL 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 wereit is more stable when many inputs were-arc changed under warming and better aligned 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 sensitivity analysis suggest that physical considerations regarding model architecture and input variables are 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 to argue that DL models represent the 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), suggesting that DL hydrologic models can learn fundamental hydrologic processes. A corollary potential implication to of this finding is that these models 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 different climate conditions, because there are no future observations against which to
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 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 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 on 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 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 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, these architectural changes can 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). 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, the benefits of 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, these studies have shown that temperature-based PET estimation methods (e.g., Hamon, Thornthwaite) significantly 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 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, 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 produce 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 (Grönewold 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 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 km² 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 (Rs), and temperature are aggregated into a basin-wide daily precipitation average, daily Rs 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).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_mean</td>
<td>Mean daily precipitation</td>
</tr>
<tr>
<td>pet_mean</td>
<td>Mean daily potential evapotranspiration</td>
</tr>
<tr>
<td>aridity</td>
<td>Ratio of mean PET to mean precipitation</td>
</tr>
<tr>
<td>t_mean</td>
<td>Mean of daily maximum and daily minimum temperature</td>
</tr>
<tr>
<td>frac_snow</td>
<td>Fraction of precipitation falling on days with mean daily temperatures below 0°C</td>
</tr>
<tr>
<td>high_prec_freq</td>
<td>Fraction of high-precipitation days (= 5 times mean daily precipitation)</td>
</tr>
<tr>
<td>high_prec_dur</td>
<td>Average duration of high-precipitation events (number of consecutive days with = 5 times mean daily precipitation)</td>
</tr>
<tr>
<td>low_prec_freq</td>
<td>Fraction of dry days (&lt; 1 mm d-1 daily precipitation)</td>
</tr>
<tr>
<td>low_prec_dur</td>
<td>Average duration of dry periods (number of consecutive days with daily precipitation &lt; 1 mm d-1)</td>
</tr>
<tr>
<td>mean_elev</td>
<td>Catchment mean elevation</td>
</tr>
<tr>
<td>std_elev</td>
<td>Standard deviation of catchment elevation</td>
</tr>
</tbody>
</table>
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 $c[t]$ that is updated over a sequence of $t=1,...,T$ time steps based on a sequence of inputs $x = x[1], ..., x[T]$, where each input $x[t]$ 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 $h[t]$, which form the output of the hidden layer in the model. The structure of the LSTM is given as follows:

\[ i[t] = \sigma(W_i x[t] + U_i h[t-1] + b_i) \]  
(Eq. 1.1)

\[ f[t] = \sigma(W_f x[t] + U_f h[t-1] + b_f) \]  
(Eq. 1.2)

\[ g[t] = \tanh(W_g x[t] + U_g h[t-1] + b_g) \]  
(Eq. 1.3)

\[ o[t] = \sigma(W_o x[t] + U_o h[t-1] + b_o) \]  
(Eq. 1.4)

\[ c[t] = f[t] \odot c[t-1] + i[t] \odot g[t] \]  
(Eq. 1.5)

\[ h[t] = o[t] \odot \tanh(c[t]) \]  
(Eq. 1.6)

\[ y[T] = \text{ReLU}(W_y h[T] + b_y) \]  
(Eq. 1.7)

Here, the input gate ($i[t]$) controls how candidate information ($g[t]$) from inputs and previous hidden states flows to the current cell state ($c[t]$); the forget gate ($f[t]$) enables removal of information within the cell state over time; and the output gate ($o[t]$) 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, $h[T]$ 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 $\text{ReLU}(x) = \max(0,x)$. 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 $x$. 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 basin-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:

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

(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:

\[
\hat{c}[t-1] = \frac{c[t-1]}{||c[t-1]||_1} \quad \text{(Eq. 3.1)}
\]
\[
i[t] = \sigma(W_i x[t] + U_i \hat{c}[t-1] + V_i a[t] + b_i) \quad \text{(Eq. 3.2)}
\]
\[
a[t] = \sigma(W_o x[t] + U_o \hat{c}[t-1] + V_o a[t] + b_o) \quad \text{(Eq. 3.3)}
\]
\[
R[t] = \hat{\sigma}(W_R x[t] + U_R \hat{c}[t-1] + V_R a[t] + b_R) \quad \text{(Eq. 3.4)}
\]
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 $c[t-1]$ and $x[t]$ to $m[t]$, by basing these quantities on a normalized sigmoid activation function that sums to unity:

$$\hat{\sigma}(z_j) = \frac{\sigma(z_j)}{\sum_j \sigma(z_j)}$$

(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 $m[t]$, which is stored across $D$ elements in the vector, is then distributed to the output of the hidden layer, $h[t]$, or the next cell state, $c[t]$. To account for water losses from evapotranspiration or other sinks, one element of the $D$-dimensional vector $h[t]$ 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:

$$y[T] = \sum_{d=1}^{D-1} h_d[T]$$

(Eq. 5)
Here, the $D^{th}$ cell of $h (h_D)$ is set as the trash cell, and water allocated to this cell at each time step $t=1,\ldots,T$ 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 $m[t]$ 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:

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

(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 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):

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

\[
e_{sat} = 0.611 \times \exp \left( \frac{17.27 \times T_a}{237.3+T_a} \right)
\]  
(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 \(\alpha_H\) 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:

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

Here, \(\Delta(T_a)\) is the slope of the saturation vapor pressure temperature curve (kPa/°C) and is a function of \(T_a\). \(\gamma\) is the psychrometric constant (kPa/°C), \(\lambda\) is the volumetric latent heat of vaporization (MJ/m³), \(R_n\) is the net radiation (MJ/m²-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/m²-day), and \(\alpha_{PT}\) 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 \(\gamma\), \(\Delta(T_a)\), 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 shortwave 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 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 S1 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.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Model</th>
<th>PET method adjusted with warmer temperatures</th>
<th>Are static features also changed along with dynamic features?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Great Lakes models</td>
<td>Hamon</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Great Lakes models</td>
<td>Priestley-Taylor</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Great Lakes models</td>
<td>Hamon</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Great Lakes models</td>
<td>Priestley-Taylor</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>National LSTM</td>
<td>NA</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>National LSTM</td>
<td>NA</td>
<td>No</td>
</tr>
</tbody>
</table>

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 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 projection-series, average long-term mean flow will decline significantly 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 projection-series, at least in terms of the difference in magnitude of average long-term mean streamflow declines. ForTo 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 0.82079, 0.83078, and 0.84077 for HBV, SAC-SMA, and HYMOD, respectively. 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 0.5604-0.587 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 40% 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 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 32. 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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td><strong>0.76</strong></td>
<td><strong>0.77</strong></td>
<td>9.66</td>
<td>17.58</td>
<td><strong>30.98</strong></td>
<td><strong>0.72</strong></td>
<td><strong>0.68</strong></td>
<td>12.15</td>
<td>26.01</td>
<td><strong>27.32</strong></td>
</tr>
<tr>
<td>MC-LSTM</td>
<td>0.74</td>
<td>0.72</td>
<td>9.48</td>
<td>15.52</td>
<td>41.46</td>
<td><strong>0.72</strong></td>
<td>0.65</td>
<td>12.13</td>
<td>22.82</td>
<td>35.80</td>
</tr>
<tr>
<td>MC-LSTM-PET</td>
<td>0.73</td>
<td>0.72</td>
<td><strong>8.63</strong></td>
<td>18.80</td>
<td>48.10</td>
<td>0.71</td>
<td>0.66</td>
<td><strong>10.22</strong></td>
<td>22.49</td>
<td>44.43</td>
</tr>
<tr>
<td>HBV</td>
<td>0.58</td>
<td>0.50</td>
<td>9.99</td>
<td>32.22</td>
<td>63.96</td>
<td>0.55</td>
<td>0.50</td>
<td>12.68</td>
<td>34.76</td>
<td>57.20</td>
</tr>
<tr>
<td>SAC-SMA</td>
<td>0.57</td>
<td>0.48</td>
<td>11.74</td>
<td>34.72</td>
<td>45.17</td>
<td>0.54</td>
<td>0.47</td>
<td>12.24</td>
<td>40.45</td>
<td>46.78</td>
</tr>
<tr>
<td>HYMOD</td>
<td>0.58</td>
<td>0.48</td>
<td>10.07</td>
<td>33.68</td>
<td>58.06</td>
<td>0.54</td>
<td>0.48</td>
<td>12.52</td>
<td>36.07</td>
<td>60.32</td>
</tr>
</tbody>
</table>

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 0.799-0.80). 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; 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 S4; Figure 1.
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 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 not generally change with temperature through its effects on net outgoing longwave radiation, but these changes are small (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
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 of 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 -56% to -109%, 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 changes 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
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 S84). These results suggest that changing the static watershed properties associated with long-term climate characteristics can degrade the quality of the projected 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 of temperature and PET. 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. We also separate out the National LSTM projections for those 17 sites. Note that in Figure 8, the National LSTM projections 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 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% larger than the median projections of loss under the process models using Hamon PET and but much 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 changes in FHV (Figure 8e,f) and COM (Figure 8g,h) that are relatively similar to the process models, and for COM, the projections 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 are stable under the National LSTM regardless of whether only dynamic inputs or both dynamic and static inputs are changed under warming (see Figure S96), 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 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 Rs vary across all 531 CAMELS watersheds of different average temperatures, and compare this variability to projected 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 Rs (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 calculate these differences in long-term mean streamflow, PET, and Rs across the 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 (Rs 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 Rs is held at historic values, while Rs 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 Rs, 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](image)

**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 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 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 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 are correct for these streamflow properties, these results overall favor projections 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 a 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.

Another important limitation of this study is how we constructed the climate change warming scenarios, with 4°C warming and shifts to PET 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, 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 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 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).

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 of promising methods for the challenge of DL hydrologic modeling under climate change is causal learning, defined broadly as methods that aimed to identify 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 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 adjustments or constraints (for other examples outside of hydrology, see Lin et al., 2017; Ma et al., 2018). 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) to change 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 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. All data used to train and evaluate the models are available at https://www.hydrohub.org/mips_introduction.html#grip-gl.

References


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 Wi¹, Scott Steinschneider¹

¹Department 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.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pet_mean</td>
<td>Mean daily potential evapotranspiration</td>
</tr>
<tr>
<td>aridity</td>
<td>Ratio of mean PET to mean precipitation</td>
</tr>
<tr>
<td>t_mean</td>
<td>Mean of daily maximum and daily minimum temperature</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Values Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Hidden Layer Nodes</td>
<td>64, 96, 128, 256</td>
</tr>
<tr>
<td>Mini-Batch Size</td>
<td>64, 128, 256, 512</td>
</tr>
<tr>
<td>Learning Rate*</td>
<td>0.0001, 0.0005, 0.001, 0.005</td>
</tr>
<tr>
<td>Number of Epochs</td>
<td>30, 50</td>
</tr>
<tr>
<td>Dropout Rate*</td>
<td>0, 0.2, 0.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Gauge ID</th>
<th>Country</th>
<th>Site Name</th>
<th>Drainage Area (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>02ED032</td>
<td>Canada</td>
<td>Willow Creek near Minesing</td>
<td>231</td>
</tr>
<tr>
<td>02GG013</td>
<td>Canada</td>
<td>Black Creek near Bradshaw</td>
<td>213</td>
</tr>
<tr>
<td>02HJ003</td>
<td>Canada</td>
<td>Ouse River near Westwook</td>
<td>283</td>
</tr>
<tr>
<td>04126740</td>
<td>United States</td>
<td>Platte River at Honor, MI</td>
<td>324</td>
</tr>
<tr>
<td>04220045</td>
<td>United States</td>
<td>Oak Orchard Creek near Shelby NY</td>
<td>378</td>
</tr>
<tr>
<td>04168400</td>
<td>United States</td>
<td>Lower River Rouge at Dearborn, MI</td>
<td>236</td>
</tr>
</tbody>
</table>
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 S45. 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 S96. 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.