https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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

Sungwook Wi[1], **Scott Steinschneider**[1]

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

Correspondence to: Sungwook Wi (sw2275@cornell.edu)

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

Abstract

Deep learning 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 models can produce physically plausible projections of streamflow under significant amounts of climate change. We investigate this question here, focusing specifically on modeled responses to increases in temperature and potential evapotranspiration (PET). Previous research has shown that temperature-based methods to estimate PET lead to overestimates of water loss in rainfall-runoff models under warming, as compared to energy budget-based PET methods. 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 should exhibit less water loss when forced with smaller (energy budget-based) projections of future PET. We conduct this assessment using three process-based rainfall-runoff models and three deep learning models, trained and tested across 212 watersheds in the Great Lakes basin. The deep learning 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 water loss (MC-LSTM-PET). We first compare historical streamflow 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 for changes in average flow, as well as low flows, high flows, and streamflow timing. Finally, we also explore similar projections using a National LSTM fit to a broader set of 531 watersheds across the contiguous United States. The main results of this study are as follows:

1. The three Great Lakes deep learning models significantly outperform all process models in streamflow estimation under spatiotemporal validation, with only small differences between the

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

DL models. The MC-LSTM-PET also matches the best process models and outperforms the MC-LSTM in estimating evapotranspiration under spatiotemporal validation.

2. All process models show a downward shift in average flows under warming, but this shift is significantly larger under temperature-based PET estimates than energy budget-based PET. The MC-LSTM-PET model exhibits similar differences in water loss across the different PET forcings, consistent with the process models. However, the LSTM exhibits unrealistically large water losses under warming as compared to the process models using Priestley-Taylor PET, while the MC-LSTM is relatively insensitive to PET method.
3. All deep learning models exhibit smaller changes in high flows and streamflow timing as compared to the process models, while deep learning projections of low flows are all very consistent and within the range projected by process models.
4. Like the Great Lakes LSTM, the National LSTM also shows unrealistically large water losses under warming. However, when compared to the Great Lakes deep learning models, projections from the National LSTM were more stable when many inputs were changed under warming and better aligned with process model projections for streamflow timing. This suggests that the addition of more, diverse watersheds in training does help improve climate change projections from deep learning models, but this strategy alone may not guarantee reliable projections under unprecedented climate change.

Ultimately, the results of this work 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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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 state-of-the-science in rainfall-runoff modeling.

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; 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. 2021), suggesting

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

that DL hydrologic models can learn fundamental hydrologic processes. A corollary to 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 compare. Recently, Wi and Steinschneider (2022) (hereafter WS22) addressed this challenge directly, forwarding an experimental design 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 physically plausible hydrologic projections should show a decline in total annual average streamflow compared to a baseline historical simulation, due to increases in potential evapotranspiration (PET) with warming (and no changes in precipitation). Results showed that the 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 model trained to 531 catchments.

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 this depended heavily on the accuracy of the process-model ET.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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. This architecture slightly underperformed a standard LSTM when predicting out-of-sample extreme events (Frame et al., 2021b), and 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 this mass conserving architecture have not been tested when employed under previously unobserved climate change.

For all models considered in WS22, a major focus was evaluating the direction of annual total runoff change in the presence of warming and no change in precipitation. However, that study did not consider the magnitude of runoff change and how it relates to projected changes in PET. As we argue below, this comparison provides a unique way to assess the physical plausibility of future hydrologic projections. Several studies have investigated the effects of different PET estimation methods on the magnitude of PET and runoff change in a warming climate (Lofgren et al., 2011; Shaw and Riha, 2011; Lofgren and Rouhana, 2016; Milly and Dunne, 2017; Lemaitre-Basset et al. 2022). Broadly, this work has 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 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 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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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.

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 physical plausibility of hydrologic projections under climate change. In this study, we adopt this strategy for DL rainfall-runoff models and forward an experimental design in which both 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 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 larger 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 similar and large 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 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 an outcome would indicate that some degree of PIML is 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).

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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 that a standard LSTM will produce unrealistic hydrologic projections because it relies on historical and geographically pervasive correlations between temperature and PET to project streamflow losses under warming. We also hypothesize that PIML-based DL models will be better able to relate future projections of temperature and PET to streamflow change, especially those PIML approaches that directly map PET to evaporative water loss in their architecture.

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

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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

benchmarking datasets for DL rainfall-runoff models (e.g., CAMELS-GB; Coxon et al., 2020). For direct comparability to previous results from the 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 2001 and December 2017. All streamflow gauging stations have a drainage area greater than or equal to 200 km^2 and less than 5% missing data in the study period. The watersheds are evenly distributed across the five lake basins and the Ottawa River basin, and they represent a range of land use/land cover types and degrees of hydrologic alteration from human activity. In the experiments described further below, 141 of the watersheds are designated as training sites, and the remaining 71 watersheds are used for testing (see Figure 1). In addition, the period between January 2001 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).

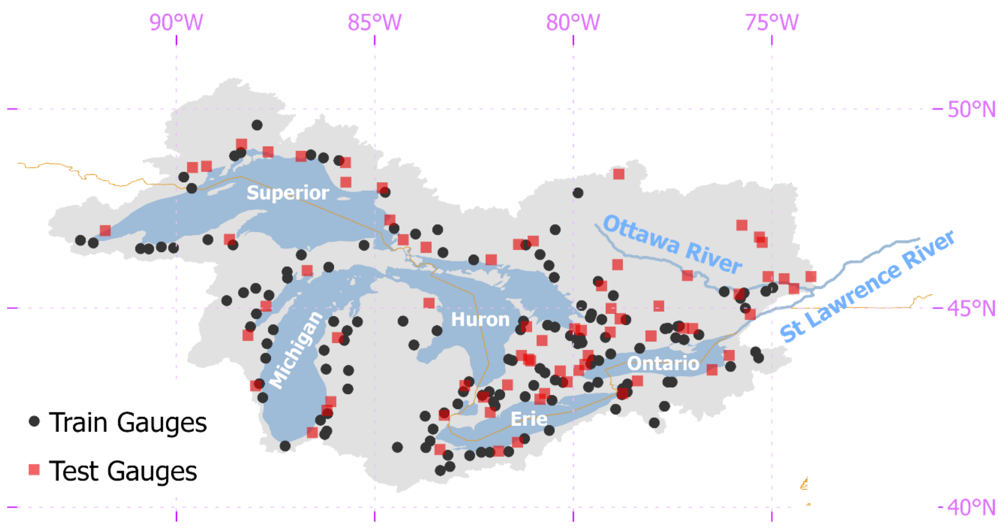

Figure 1. Great Lakes domain, with training and testing streamflow gauges used throughout this study.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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

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

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

Attribute	Description
p_mean	Mean daily precipitation
pet_mean	Mean daily potential evapotranspiration
aridity	Ratio of mean PET to mean precipitation
t_mean	Mean of daily maximum and daily minimum temperature
frac_snow	Fraction of precipitation falling on days with mean daily temperatures below 0°C

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

high_prec_freq	Fraction of high-precipitation days (= 5 times mean daily precipitation)
high_prec_dur	Average duration of high-precipitation events (number of consecutive days with = 5 times mean daily precipitation)
low_prec_freq	Fraction of dry days (< 1 mm d-1 daily precipitation)
low_prec_dur	Average duration of dry periods (number of consecutive days with daily precipitation < 1 mm d-1)
mean_elev	Catchment mean elevation
std_elev	Standard deviation of catchment elevation
mean_slope	Catchment mean slope
std_slope	Standard deviation of catchment slope
area_km2	Catchment area
Temperate-or-sub-polar-needleleaf-forest	Fraction of land covered by “Temperate-or-sub-polar-needleleaf-forest”
Temperate-or-sub-polar-broadleaf-forest	Fraction of land covered by “Temperate-or-sub-polar-broadleaf-forest”
Temperate-or-sub-polar-shrubland	Fraction of land covered by “Temperate-or-sub-polar-shrubland”
Temperate-or-sub-polar-grassland	Fraction of land covered by “Temperate-or-sub-polar-grassland”
Mixed-Forest	Fraction of land covered by “Mixed-Forest”
Wetland	Fraction of land covered by “Wetland”
Cropland	Fraction of land covered by “Cropland”
Barren-Lands	Fraction of land covered by “Barren-Lands”
Urban-and-Built-up	Fraction of land covered by “Urban-and-Built-up”
Water	Fraction of land covered by “Water”
BD	Soil bulk density (g cm-3)
CLAY	Soil clay content (% of weight)
GRAV	Soil gravel content (% of volume)
OC	Soil organic carbon (% of weight)
SAND	Soil sand content (% of weight)
SILT	Soil silt content (% of weight)

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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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.

3. Methods

We design an experiment to test the two primary hypotheses of this study, namely that a standard LSTM will overestimate hydrologic 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 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 force all models with climate change scenarios composed of historical precipitation and historical but warmed temperatures, as well as PET based on those warmed temperatures. 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 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 contiguous United States (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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

warming than an LSTM fit only to locations in the Great Lakes basin. Figure 2 presents an overview of our experimental design.

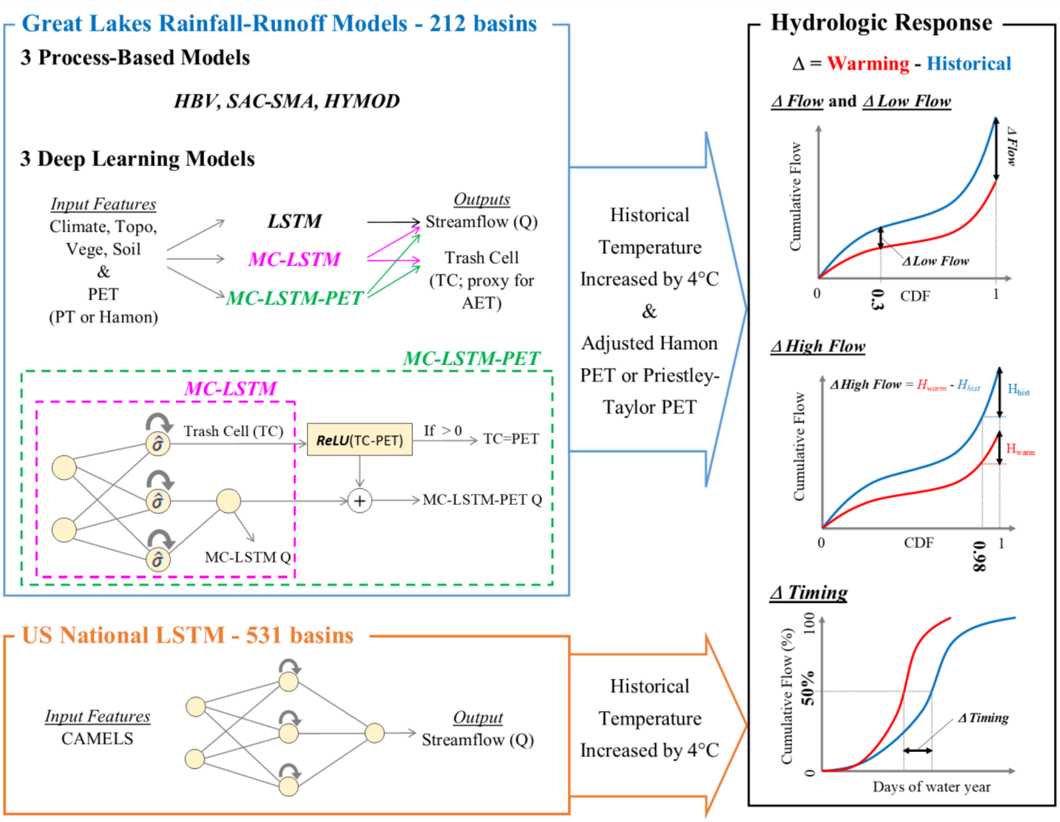

Figure 2. Overview of experiment design.

3.1. Models

3.1.1. Benchmark Conceptual Models

We develop three 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. We

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

calibrate the models with the genetic algorithm from Wang et al. (1991) to maximize the Kling-Gupta Efficiency (KGE; Gupta et al. 2009), 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. Benchmark models are developed for the 71 testing sites in two ways: 1) separate models are trained for the testing sites during the training period; and 2) each testing site is assigned a donor from among the 141 training sites, and the calibrated parameters from that donor site are transferred to the testing site. The first of these approaches enables a comparison between DL models fit only to the training sites to benchmark models developed for the testing sites, i.e., a spatial out-of-sample versus in-sample comparison. The second of these approaches enables a more direct spatial out-of-sample comparison between DL and benchmark models. We note that donor sites were used to assign model parameters to testing sites in the benchmarking study of Mai et al. (2022), and to retain direct comparability to the results of that work we use the same donor sites for each testing site. Donor sites were selected based on spatial proximity, while also prioritizing donor sites that were nested within the watershed of the testing site.

3.1.2. LSTM

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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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

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

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

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

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

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

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

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

The LSTM takes K=39 input features: 9 dynamic and 30 static. The dynamic input features are basin-averaged climate, including daily precipitation, maximum temperature, minimum temperature, net incoming shortwave radiation, specific humidity, surface air pressure, zonal and meridional components of wind, and PET. The static features represent catchment attributes (see Table 1) and are repeated for all time steps in the input sequences $\boldsymbol{x}$. All input features are standardized before training (by subtracting the mean and dividing by the standard deviation for data across all training sites in the training period). Note that we

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

do not standardize the observed streamflow, besides dividing my drainage area to represent streamflow in units of millimeters.

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

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

where N is the number of training watersheds and T_n is the number samples in the n^{th} watershed. $\hat{Q}_{n,t}$ and $Q_{n,t}$ are, respectively, the streamflow prediction and observation for basin n and day t. To estimate $\hat{Q}_{n,t}$, we feed into the network an input sequence for the past T=365 days. The model was developed with 1 hidden layer composed of D=256 nodes, a mini-batch size of 256, a learning rate of 0.0005, and a drop-out rate of 0.4, and it was trained across 30 epochs. All hyperparameters (number of hidden layer nodes, mini-batch size, learning rate, dropout rate, and number of epochs) were selected in a 5-fold cross-validation on the training sites. 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 under climate change, 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.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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:

$$\boldsymbol{i}[t] = \hat{\sigma}(\boldsymbol{W}_i\boldsymbol{x}[t] + \boldsymbol{U}_i\boldsymbol{c}[t-1] + \boldsymbol{V}_i\boldsymbol{a}[t] + \boldsymbol{b}_i) \quad \text{(Eq. 3.1)}$$

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

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

$$\boldsymbol{m}[t] = \boldsymbol{R}[t]\boldsymbol{c}[t-1] + \boldsymbol{i}[t]\boldsymbol{x}[t] \quad \text{(Eq. 3.4)}$$

$$\boldsymbol{c}[t] = (1-\boldsymbol{o}[t]) \odot \boldsymbol{m}[t] \quad \text{(Eq. 3.5)}$$

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

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

The mass in $\boldsymbol{m}[t]$, which is stored across *D* elements in the vector, is then distributed to the output of the hidden layer, $\boldsymbol{h}[t]$, or the next cell state, $\boldsymbol{c}[t]$. To account for water losses from evapotranspiration or other sinks, one element of the *D*-dimensional vector $\boldsymbol{h}[t]$ is considered a 'trash cell', and the output of this cell

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

is ignored when calculating the final streamflow prediction, which at time T is given by the sum of outgoing water mass:

$$y[T] = \sum_{d=1}^{D-1} h_d[T] \quad \text{(Eq. 4)}$$

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

3.1.4. MC-LSTM-PET

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

$$y[t] = \sum_{d=1}^{D-1} h_d[t] + ReLU(h_D[t] - PET[t]) \quad \text{(Eq. 5)}$$

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. 4) 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., 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 2001 -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.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

Several metrics are considered for model evaluation, including percent bias (PBIAS), the Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), Kling-Gupta Efficiency (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 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). PET (in mm/day) under the Hamon method is calculated as follows (Shaw and Riha, 2011):

$$PET_H = \alpha_H \times 29.8 \times Hr_{day} \frac{e_{sat}}{T_a + 273.2} \quad \text{(Eq. 6)}$$

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

$$e_{sat} = 0.611 \times exp\left(\frac{17.27 \times T_a}{237.3 + T_a}\right) \quad \text{(Eq. 7)}$$

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

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

$$PET_{PT} = \alpha_{PT}\left(\frac{\Delta(T_a) \times (R_n - G)}{\lambda(\Delta(T_a) + \gamma)}\right) \times 1000 \quad \text{(Eq. 8)}$$

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

We then develop a simple climate change scenario 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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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 a similar climate change scenario for 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 (e.g., pet_mean, aridity, t_mean, frac_snow; see Table 1) and run all models using two settings: 1) with climate changes only to the dynamic features, and 2) with climate changes to both dynamic and static features.

Ultimately, for each model we compare hydrologic projections 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. We examine four different metrics for this comparison, including:

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

- AVG.Q: the average 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 years, where the center of mass is defined as the day of the 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 but similar warming while process-based and PIML models can, we would expect that under the LSTM using both PET projections, average 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 more closely across the two different PET projections, at least in terms of the difference in magnitude of average streamflow declines. For comparison, we also explore the differences in low flow (FLV), high flow (FHV), and timing (COM) metrics across all model versions, where we have less reason to anticipate how DL and process models will differ in their projections and across PET formulations.

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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

Figure S1). For the process-based models, we show results for models fit to the training sites and then used as donors at the testing sites, as well as models fit to the testing sites directly. We denote the latter with the suffix "-test" and note that performance metrics at the training sites are not available for process models fit to the testing sites.

Several insights emerge from Figure 3. First, for the training sites during the training period, all models perform very well (Figure 3a). Across the three process models, the median KGE is 0.82, 0.83, and 0.81 for HBV, SAC-SMA, and HYMOD, respectfully. However, unsurprisingly, the DL models perform better for the training data, with median KGE values all equal or above 0.88. The LSTM performs best in this case. Under temporal validation (training sites during the testing period), performance degrades somewhat across all models, and the differences in KGE between all process-based models and between all DL models shrink considerably (Figure 3c). Larger performance declines are seen at the testing sites during the training period (Figure 3b) and testing period (Figure 3d). Here, the median KGE for all process models falls to between 0.56-0.57 when streamflow at the testing sites is estimated with donor models from nearby gauged watersheds. In contrast, process models fit to the testing sites (denoted "-test") exhibit performance similar to that seen in Figure 3a,c. All three DL models perform quite well for the testing sites, with median KGE values above 0.71 in both time periods. This is only modestly below the median KGE for the process models fit to the testing sites, which is quite impressive given that this represents the spatial out-of-sample performance of the DL models. We even see that for approximately 10% of testing sites during the training period, the DL models outperform the process models fit to those locations in that period.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

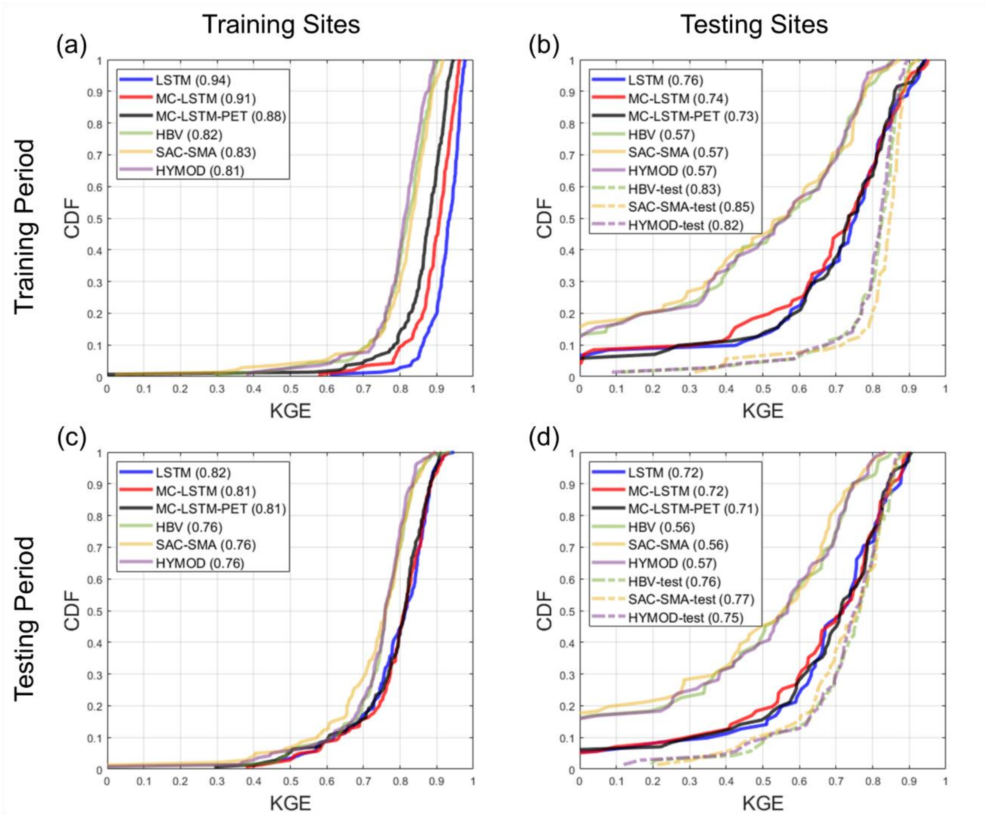

Figure 3. The distribution of Kling-Gupta efficiency (KGE) for streamflow estimates across sites from each model at the (a) 141 training sites and (b) 71 testing sites for the training period. Similar results for the testing period are shown in panels (c) and (d), respectively. For the process models 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 2 shows the median KGE, NSE, PBIAS, FHV, and FHL across testing sites for all models, excluding the process models fit to the testing sites. Similar to Figure 3, all three DL models outperform the donor-based process models at the testing sites for all metrics, with the exception of PBIAS during the training period. The performance across the three different DL models is similar, although there are some notable differences. In particular, the LSTM outperforms the MC-LSTM and MC-LSTM-PET for KGE, NSE, and FLV, 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. We 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.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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

	Testing Sites: Training Period					Testing Sites: Testing Period				
Model	KGE	NSE	PBIAS	FHV	FLV	KGE	NSE	PBIAS	FHV	FLV
LSTM	**0.76**	**0.77**	9.66	17.58	**30.98**	**0.72**	**0.68**	12.15	26.01	**27.32**
MC-LSTM	0.74	0.72	9.48	**15.52**	41.46	0.72	0.65	12.13	22.82	35.80
MC-LSTM-PET	0.73	0.72	8.63	18.80	48.10	0.71	0.66	**10.22**	**22.49**	44.43
HBV	0.57	0.42	**8.41**	32.61	50.41	0.56	0.45	11.24	36.29	46.67
SAC-SMA	0.57	0.43	11.03	34.54	42.08	0.56	0.41	12.13	36.74	49.29
HYMOD	0.57	0.41	9.58	32.70	52.24	0.57	0.45	11.16	36.34	53.62

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 0.77-0.80). 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.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

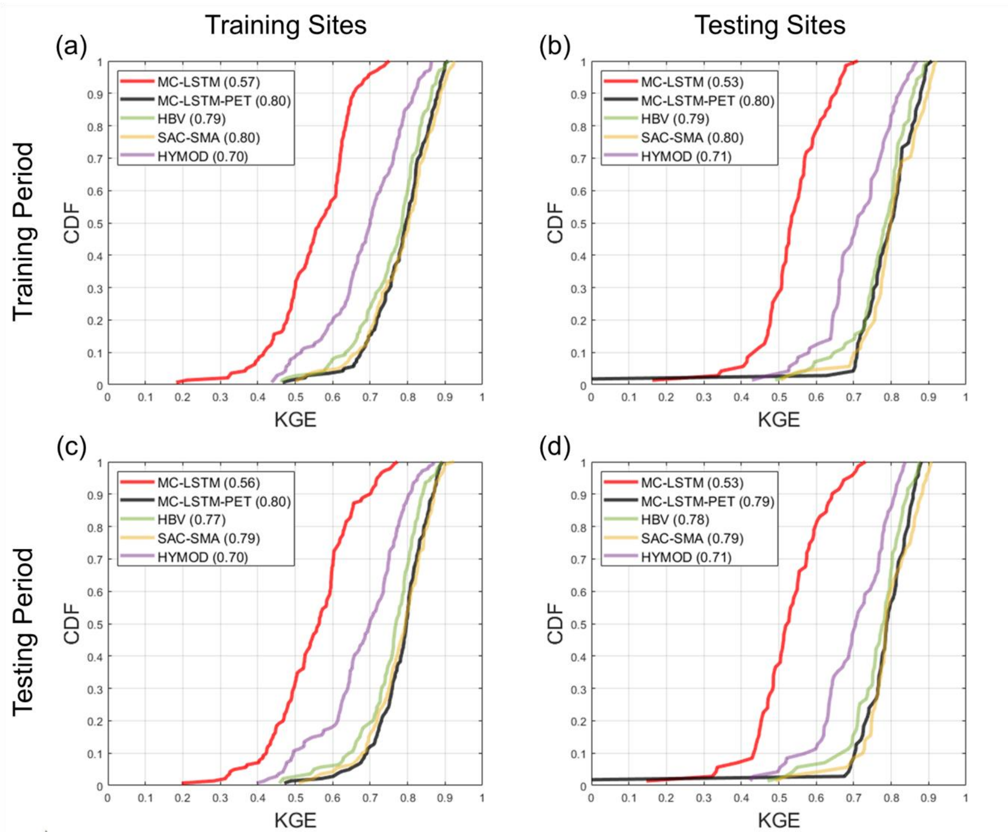

Figure 4. The Kling-Gupta efficiency (KGE) for AET estimated from each model at the (a) 141 training sites and (b) 71 testing sites for the training period. Similar results for the testing period are shown in panels (c) and (d), respectively. The LSTM is not included in this comparison. All models are 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 AET versus water allocations to the trash cell for the two models from five randomly sampled testing sites across both training and testing periods. Trash cell water from the MC-LSTM is not only more scattered around observed AET compared to the MC-LSTM-PET, but it also exhibits many outlier values that are two to five times larger than observed AET. The MC-LSTM-PET follows the variability of AET much more closely, with virtually no outliers

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

that exceed 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 in the DL model.

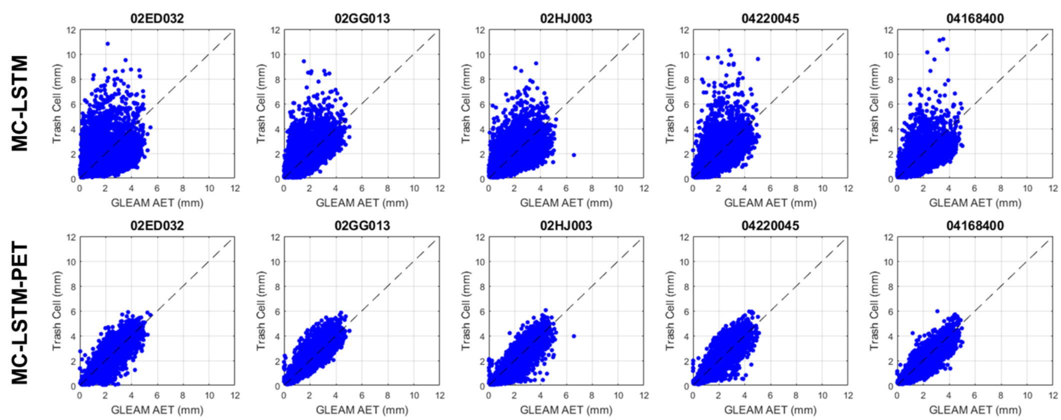

Figure 5. Scatterplots of daily AET versus trash cell water for the (top) MC-LSTM and (bottom) MC-LSTM-PET at five randomly selected testing sites across both training and testing periods. All models are trained using Priestley-Taylor PET.

4.2. Evaluating Hydrologic Response under Warming

Next, we evaluate streamflow projections 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. First, we show the differences in historic and projected 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 (shown at one sample location for demonstration; Figure 6a). However, under the scenario with 4 °C of warming, Hamon-based PET is significantly larger than Priestley-Taylor based PET (Figure 6b). On average, this difference reaches ~16% across all training sites and exhibits very little variability across locations (Figure 6c). The primary reason for the difference in projected change in PET is that the Hamon method attributes PET entirely to temperature, while only a portion of PET is based on temperature in the Priestley-Taylor method, with the rest based on R_n. It is worthwhile to note that R_n does change with temperature through its effects on net outgoing longwave radiation, but these changes are small.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

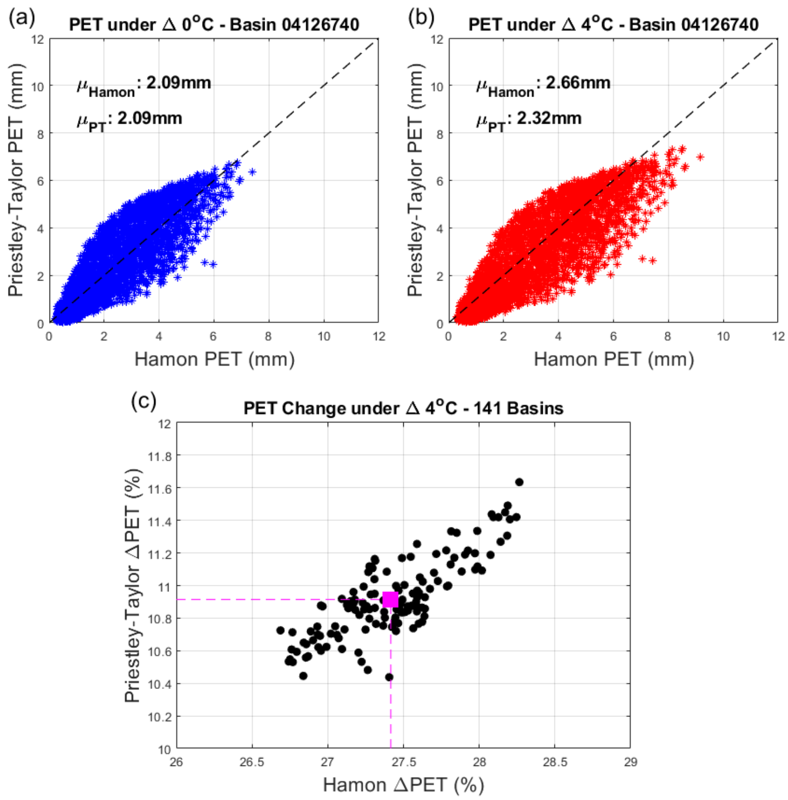

Figure 6. (a) Daily PET estimated using the Hamon and Priestley-Taylor method for one sample watershed, under historic climate conditions in the training period. (b) Same as (a), but under the climate change scenario with 4 °C of warming. (c) Percent change in average PET with 4 °C of warming across all training sites using the Hamon and Priestley-Taylor methods.

Figure 7 shows how these differences in PET under warming propagate into changes in different attributes of streamflow across training sites in the training period. The left and right columns of Figure 7 show projections 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.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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 streamflow to the process-based models, with the median ΔAVG.Q across sites ranging between -23% and -17% 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 -9% to -5%, 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 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 -25% to -15% and little variability across sites (Figure 7c,d). The process models disagree significantly on changes to FLV and 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.

The differences between process-based and DL simulated changes for high flows (FHV; Figure 7e,f) and streamflow timing (COM; Figure 7g,h) are relatively consist, with the process models exhibiting larger declines in high flows and earlier shifts in streamflow timing compared to the DL models. The choice of

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

PET method has a 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 a week less than the median change across the process models. The method of PET estimation has relatively little impact on both process model and DL based estimates of change in COM.

We 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 S4). These results suggest that changing the static watershed properties associated with long-term climate characteristics can degrade the quality of the projections, at least when the climate changes 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).

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

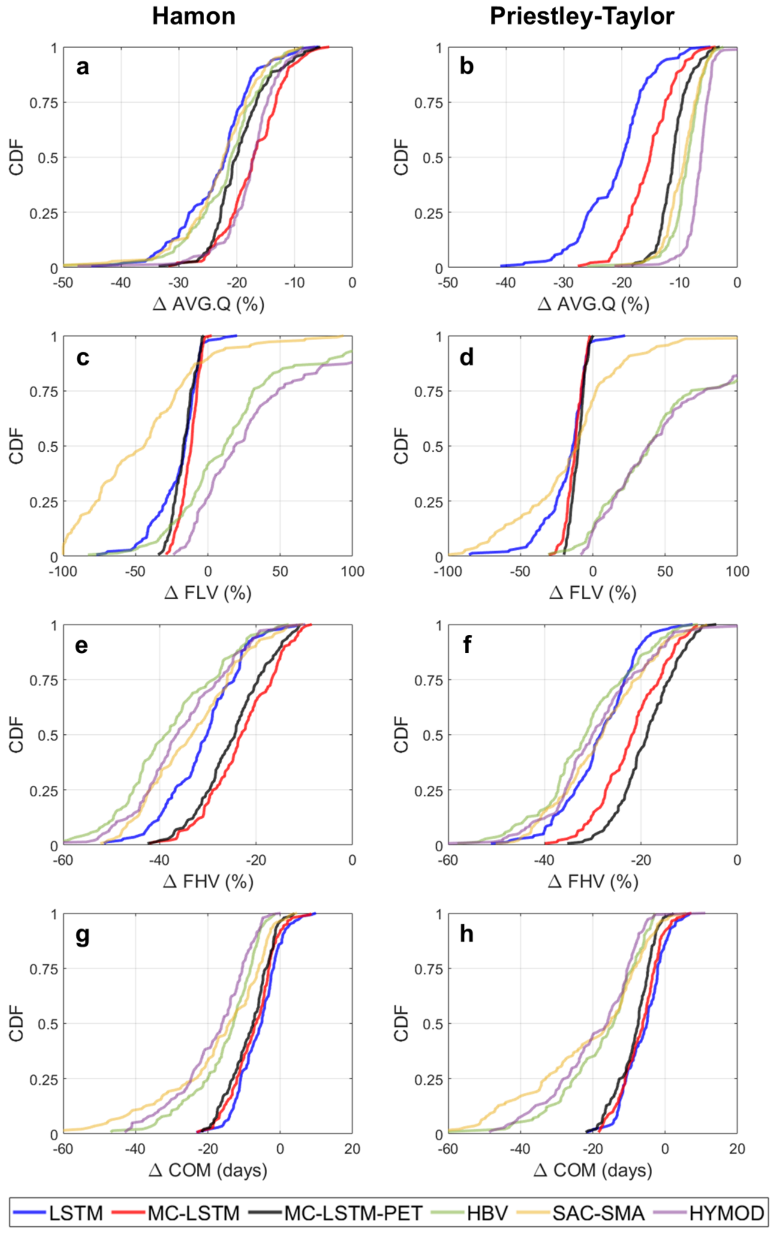

Figure 7. The distribution of change in (a,b) AVG.Q, (c,d) FLV, (e,f) FHV, and (g,h) COM across the 141 training sites and all models under a scenario of 4°C warming using (a,c,e,g) Hamon PET and (b,d,f,h) Priestley-Taylor PET. For the DL models, changes were only made to the dynamic inputs (i.e., no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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

One reason why the Great Lakes LSTM exhibits excessive hydrologic 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. To evaluate this hypothesis, we examine changes to AVG.Q, FLV, FHV, and COM under 4°C warming at the 29 CAMELS watersheds within the Great Lakes basin using the National LSTM (Figure 8). For comparison, we also examine similar changes under all six Great Lakes DL and process models at 17 of those 29 CAMELS basins that were used in the training and testing sets for the Great Lakes models, and 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 contiguous United States (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 moderately larger than the median projections of loss under the process models using Hamon PET and much 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 similar to the process models, and for COM, the projections are closer to the process models than for any Great Lakes DL model. 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 S6), in contrast to the Great Lakes DL models. Therefore, the use of more watersheds in training that span a more diverse set of climate conditions likely benefit the model when

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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.

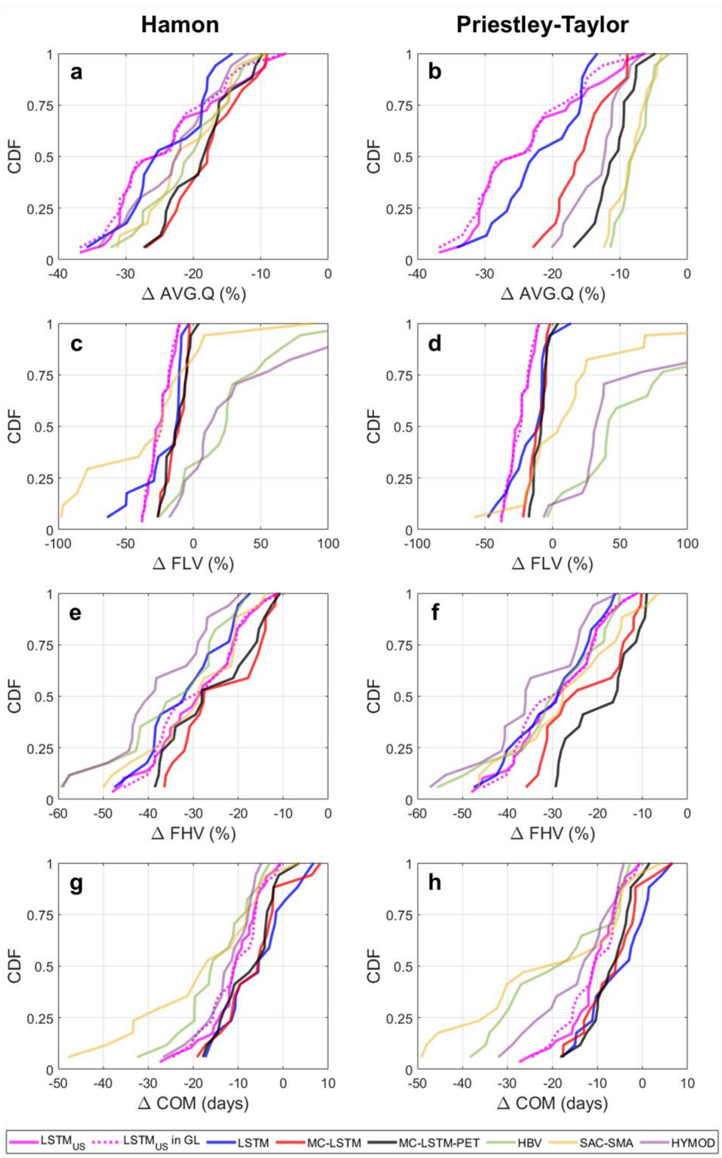

Figure 8. The distribution of change in (a,b) AVG.Q, (c,d) FLV, (e,f) FHV, and (g,h) COM across 29 CAMELS sites within the Great Lakes basin under the National LSTM (solid pink), as well as for 17 of those 29 sites from the Great Lakes DL and process models, under a scenario of 4°C warming. Results from the National LSTM for those 17 sites are also highlighted (dashed pink). For the Great Lakes

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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 streamflow, (Priestly-Taylor estimated) PET, and R_s 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 the difference in long-term (1980-2014) average streamflow (Figure 9a), PET (Figure 9b), and R_s (Figure 9c) across all pairs of basins in the CAMELS dataset with average long-term precipitation within 1% of each other, and plot these differences against the differences in average temperature across each pair. The results show that the difference in average 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 median decline in average streamflow seen for the National LSTM in Figure 8. We also note that average PET increases by approximately 20% between watersheds that differ in average temperature by 4°C (Figure 9b). However, higher PET in warmer watersheds is related both to the direct effect of temperature on vapor pressure deficit, as well as to the fact that higher incoming solar radiation co-occurs in warmer watersheds (R_s is approximately 9% higher across watershed pairs that differ by 4°C; Figure 9c). Using the Priestley-Taylor method, we estimate that average PET would only increase by between 9-14% (median of 11.5%) if temperatures warm by 4°C and R_s is held at historic values, while R_n is increased slightly due to declines in net outgoing longwave radiation with warming (Figure 9d). However, the National LSTM appears to convolute the effects of temperature and R_s and cannot separate out their effects on ET-based water loss, leading to larger projected 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).

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

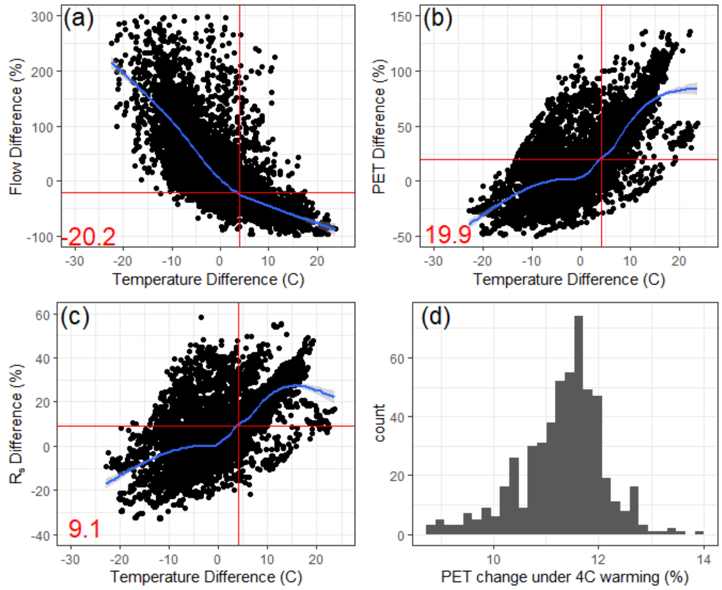

Figure 9. The percent difference in long-term (1980-2014) average (a) streamflow, (b) Priestley-Taylor based PET, and (c) downward shortwave radiation (R_s) for all pairs of CAMELS basins with average precipitation within 1% of each other, plotted against differences in average temperature for each pair. 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 an analysis that evaluates the physical plausibility of future streamflow projections under climate change using DL rainfall-runoff models. The basis for this evaluation is anchored to the assumption that differences in streamflow projections should emerge under very different projections of future PET, and that realistic projections 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 should be able to respond to lower energy-budget based PET projections under warming and, all else equal, project 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 of future PET under warming. This

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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.

Results from this work also suggest that PIML-based DL models can capture physically plausible streamflow responses under climate change 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 (the MC-LSTM-PET) was able to project changes in average 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 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 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 timing that were more like the process models than projections from the Great Lakes DL models. While it is challenging to know which set of projections are correct for these streamflow properties, these result overall favor projections from the National LSTM and highlight the benefits of DL rainfall-runoff models trained to a larger set of diverse watersheds for climate change analysis.

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

One important limitation of this study is how we constructed the climate change scenarios, with 4°C warming but no change to net incoming shortwave radiation and slight decreases in net outgoing longwave radiation with warming (i.e., slight increases in R_n). 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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

(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.

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.

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

Acknowledgements

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

Competing Interests

The authors declare no competing interests.

Data and Code 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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

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

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

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

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

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

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

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

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

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

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

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

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

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

https://doi.org/10.5194/egusphere-2023-1744
Preprint. Discussion started: 9 August 2023

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

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

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

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

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

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

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

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

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

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