Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR032123

Assessing the Physical Realism of Deep Learning Hydrologic Model Projections Under Climate Change

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

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

Key Points:

- A long short-term memory network trained to 15 watersheds can produce misleading increases in annual runoff under significant warming
- When also trained with outputs from process models, the regional network can produce more reliable runoff projections, but not always
- A network trained to over 500 basins mostly produces realistic runoff projections with warming, but also runoff increases in glacial areas

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Wi,
sw2275@cornell.edu

Citation:

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

Received 1 FEB 2022
Accepted 26 AUG 2022

Author Contributions:

Conceptualization: Sungwook Wi, Scott Steinschneider
Data curation: Sungwook Wi, Scott Steinschneider
Formal analysis: Sungwook Wi, Scott Steinschneider
Funding acquisition: Scott Steinschneider
Investigation: Sungwook Wi, Scott Steinschneider
Methodology: Sungwook Wi, Scott Steinschneider
Project Administration: Sungwook Wi, Scott Steinschneider
Resources: Sungwook Wi, Scott Steinschneider
Supervision: Scott Steinschneider
Validation: Sungwook Wi, Scott Steinschneider
Visualization: Sungwook Wi

Abstract This study examines whether deep learning models can produce reliable future projections of streamflow under warming. We train a regional long short-term memory network (LSTM) to daily streamflow in 15 watersheds in California and develop three process models (HYMOD, SAC-SMA, and VIC) as benchmarks. We force all models with scenarios of warming and assess their hydrologic response, including shifts in the hydrograph and total runoff ratio. All process models show a shift to more winter runoff, reduced summer runoff, and a decline in the runoff ratio due to increased evapotranspiration. The LSTM predicts similar hydrograph shifts but in some watersheds predicts an unrealistic increase in the runoff ratio. We then test two alternative versions of the LSTM in which process model outputs are used as either additional training targets (i.e., multi-output LSTM) or input features. Results indicate that the multi-output LSTM does not correct the unrealistic streamflow projections under warming. The hybrid LSTM using estimates of evapotranspiration from SAC-SMA as an additional input feature produces more realistic streamflow projections, but this does not hold for VIC or HYMOD. This suggests that the hybrid method depends on the fidelity of the process model. Finally, we test climate change responses under an LSTM trained to over 500 watersheds across the United States and find more realistic streamflow projections under warming. Ultimately, this work suggests that hybrid modeling may support the use of LSTMs for hydrologic projections under climate change, but so may training LSTMs to a large, diverse set of watersheds.

Plain Language Summary Recent research has shown that deep learning models can outperform process models in hydrologic prediction and forecasting, but it is unclear whether they can be used to project streamflow response under climate change. The concern is that deep learning models will be unable to reliably extrapolate beyond the range of historical climate, whereas process models can leverage physics to make such projections. To test this question, this study trained a deep learning hydrologic model (termed a long short-term memory network, LSTM) to data from 15 watersheds in California and also trained three process-based hydrologic models to the same watersheds for comparison. We also developed two other versions of the LSTM that use the output from the process models during training. We forced all models with the same scenarios of warming and compared their hydrologic response. The results suggested that the LSTM trained using process model data as input can improve the realism of streamflow projections under warming, but this is not guaranteed. We also conducted a similar experiment with an LSTM trained with data from over 500 watersheds and found more realistic hydrologic responses, suggesting deep learning models may provide more reliable projections when trained with more diverse data.

1. Introduction

Deep learning (DL) models currently represent the state-of-the-art in hydrologic prediction (Nearing et al., 2021; Shen et al., 2021), as evidenced by their unmatched accuracy in predicting streamflow (Liu et al., 2020), soil moisture (Li et al., 2021), evapotranspiration (Ahmed et al., 2021), stream temperature (Rahmani et al., 2021), and water quality indicators (Aldhyani et al., 2020; Zhi et al., 2021). For streamflow prediction, long short-term msemory networks (LSTMs, Hochreiter & Schmidhuber, 1997) have proven particularly effective due to their strong inductive bias toward storing information over time (Hoedt et al., 2021). This property is well suited for capturing hydrologic dynamics driven by multi-scale memory effects within a watershed, such as the persistence and release of water from soil moisture and snowpack. LSTMs trained across a large number of watersheds have been shown to outperform process-based hydrologic models by a substantial margin (Kratzert et al., 2018), even at hourly timescales (Gauch, Kratzert, et al., 2021) and for watersheds treated as unseen by the LSTM (Kratzert, Klotz, Herrnegger, et al., 2019). Given their high degree of performance, LSTMs are being considered for a range of hydrologic applications, including forecasting (Cheng et al., 2020; Sharma et al., 2021), streamflow estimation

Writing – original draft: Sungwook Wi, Scott Steinschneider
Writing – review & editing: Sungwook Wi, Scott Steinschneider

for ungauged sites (Yin et al., 2021), and post-processing of process-based hydrologic models (Frame, Kratzert, Raney, et al., 2021).

One application of DL models that remains underexplored is their use in hydrologic projections under climate change. Very few studies have utilized DL or machine learning (ML) models to predict time series of streamflow under projections of future climate, and of those that have (Anaraki et al., 2021; Das & Nanduri, 2018; Ghosh & Mujumdar, 2008; Lee et al., 2020; Zhu et al., 2019), few if any have assessed the credibility of these projections. The reticence to use DL and ML models for climate change projections may be in part due to concerns that data-driven models will be unable to extrapolate beyond their training data to unfamiliar circumstances where system dynamics change considerably (de Silva et al., 2020), such as would be experienced under significant climate change. Because DL models represent physical systems without accounting for underlying laws such as conservation of mass and energy, spurious predictions are possible under extrapolation (Read et al., 2019). In contrast, process-based models are rooted in scientific theory, arguably making them more suitable for projecting non-stationarity in hydrologic systems (Fatichi et al., 2016; Paniconi & Putti, 2015).

However, the assumption that DL models cannot reliably extrapolate to previously untested conditions deserves additional scrutiny, especially for models trained to a large and diverse set of basins (Nearing et al., 2021). Recent work has shown that LSTMs trained to hundreds of basins can predict extreme events with more accuracy than process models, even when the largest extremes are withheld during training (Frame, Kratzert, Klotz, et al., 2021). In addition, Lees et al. (2021) showed that the cell states of globally trained LSTMs used to predict streamflow strongly correlated with latent hydrologic states, such as soil moisture storage and snowpack. These results suggest that DL hydrologic models are learning fundamental processes from the data, a concept further bolstered by the fact that LSTMs trained to many basins produce better out-of-sample streamflow predictions in a given basin than an LSTM only trained to that basin or a smaller subset (Gauch, Mai, & Lin, 2021; Kratzert, Klotz, Shalev, et al., 2019). If DL hydrologic models are in fact learning underlying process through the analysis of large volumes of data, they may be capable of extrapolating predictions under new boundary forcing.

Physics-informed machine learning (PIML; Faghmous & Kumar, 2014; Karniadakis et al., 2021; Karpatne et al., 2017; Jiang et al., 2020; Reichstein et al., 2019; Willard et al., 2022) presents another avenue that could further bolster the use of DL hydrologic models for climate change projections. PIML imbues data-driven techniques with process-knowledge constructs through, for example, the combination of data-driven and physically based models, physics-guided loss functions, model architectures that conserve mass and energy, or by selecting feature variables that maximize information content (Hanson et al., 2020; Jia et al., 2018; Kashinath et al., 2021). Several of these techniques have been tested for streamflow prediction. For instance, Xie et al. (2021) used synthetic events and additional penalties within the loss function to impose constraints of physical consistency between precipitation and streamflow into an LSTM for basins across the contiguous United States. Alternatively, mass conservation can be directly embedded into DL model architecture to ensure cumulative streamflow predictions do not exceed precipitation inputs (Hoedt et al., 2021; Nearing et al., 2021). This architecture slightly underperformed a standard LSTM when predicting out-of-sample extreme events (Frame, Kratzert, Klotz, et al., 2021), but still outperformed process-based models despite being subject to the same closure constraints (Frame et al., 2022). DL models have also been used as emulators of more complex hydrologic process models (Liang et al., 2019), and in hybrid modeling approaches in which some aspects of streamflow prediction are handled by theory-based models while others are modeled using ML or DL (Bhasme et al., 2021). As an alternative hybrid modeling strategy, several studies have used process-based model outputs (e.g., discharge, soil moisture, and snowpack) as inputs for DL (Frame, Kratzert, Raney, et al., 2021; Konapala et al., 2020; Lu et al., 2021).

From the developments above, a question emerges as to whether DL hydrologic models, with or without a physics-informed component, can in fact be used for long-term hydrologic projections under climate change (Nearing et al., 2019). To date, this question remains unanswered. In this work, we probe this question with a series of relatively simple numerical experiments in which we compare the projections of DL hydrologic models against those of process-based hydrologic models under scenarios of atmospheric warming. We consider standard and PIML-based LSTMs of daily streamflow, fit regionally to 15 watersheds in a case study in California, and

compare them against process-based benchmark models. We also conduct a similar experiment using a national DL hydrologic model fit to hundreds of watersheds across the US. We hypothesize that:

- H1: An LSTM fit regionally to a small set of watersheds will not produce reliable streamflow projections under warming, because the historical data set does not provide sufficient information for the model to extrapolate hydrologic response under unprecedented warming.
- H2: PIML-based LSTMs fit regionally will be able to improve the physical realism of model projections under warming by leveraging information about hydrologic response to climate change from process-based models.
- H3: An LSTM fit to a diverse set of hundreds of watersheds will also produce more reliable hydrologic projections under warming, even without a PIML component, because this model can learn the observed hydrologic responses in historically warm basins and transfer those responses to historically cooler basins exposed to climate change induced warming.

To assess the reliability of modeled streamflow projections, we compare them to well-established expectations of hydrologic response under warming. In particular, we assume that hydrologic responses to warming (without any change in precipitation) will lead to: (a) shifts in the monthly hydrograph in snow-dominated regions, where cold season runoff increases and warm season runoff declines due to more precipitation falling as rain in the cold season, earlier snowmelt, and less accumulated snow pack; and (b) declines in the total runoff ratio (cumulative runoff divided by cumulative precipitation) driven by higher evapotranspiration under warming. These two responses are consistent across a large body of literature for the contiguous United States and globally (Cayan et al., 2001; Dierauer et al., 2018; Gordon et al., 2022; Kapnick & Hall, 2010; Lehner et al., 2017; Liu et al., 2022; Martin et al., 2020; McCabe et al., 2017; Milly & Dunne, 2020; Mote et al., 2018; Rungee et al., 2021; Stewart et al., 2005; Woodhouse & Pederson, 2018), with the only exception found in glacially fed watersheds where significant perennial ice and snow melt can lead to increases in runoff with warming (Hugonnet et al., 2021; Huss & Hock, 2018; Pritchard, 2019). By assuming these two responses hold in any physically realistic representation of non-glacial hydrologic systems in our study domain, we are able to evaluate the plausibility of future hydrologic projections from the models tested in this work.

The goal of these experiments is not a comprehensive analysis of the physical consistency of DL hydrologic projections under climate change, but rather a first-order assessment to determine whether such projections are physically plausible and whether the use of PIML strategies or larger training data sets improves physical consistency under warming. Ultimately, we aim to contribute a better understanding of the potential uses and limitations of DL in applications throughout the hydrologic sciences.

2. Data and Methods

To assess DL hydrologic model projections under climate change, we build a regional LSTM that predicts daily, naturalized streamflow in 15 watersheds within the Central Valley of California (Section 2.1), and develop three regional, spatially distributed process-based models (HYMOD, SAC-SMA, and VIC) as benchmarks (Section 2.2). We force all models under climate scenarios with historical precipitation but warmer temperatures and then assess their hydrologic response. In our assessment, we focus on shifts in the monthly hydrograph and the runoff ratio, with the former revealing seasonal responses (i.e., changes in peak timing and streamflow volume throughout the year) and the latter highlighting the change in mass balance between precipitation and streamflow over a multi-year period (Section 2.3). We also test two alternative PIML versions of the LSTM (Section 2.4) in which process model outputs are used as additional: (a) training targets (multi-output LSTM, or LSTM-MO) and (b) input features (LSTM-IN). The LSTM-MO networks are trained against process model evapotranspiration (ET) and soil moisture in addition to streamflow, while the LSTM-IN uses process model ET as an additional input because this variable captures water loss under warming. Lastly, we test climate change responses using the national LSTM of Kratzert et al. (2021), trained to a much larger (>500) set of watersheds across the United States, to assess the implication of a large, diverse set of watersheds for streamflow projections under warming (Section 2.5). Figure 1 provides a conceptual overview of our experimental design.

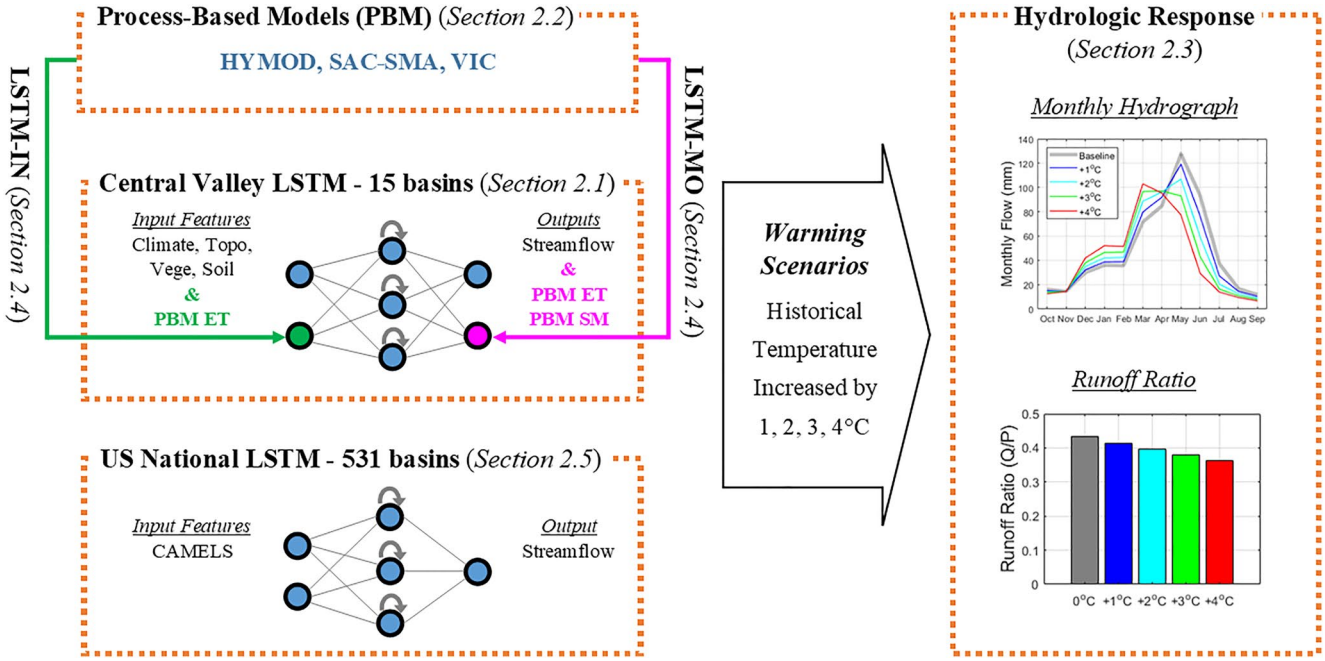

Figure 1. Overview of experiment design.

2.1. Central Valley LSTM

We develop a regional LSTM (hereafter CV-LSTM) for predicting daily streamflow in 15 major headwater watersheds in the Central Valley of California (Figure 2). The CV-LSTM architecture consists of three layers: an input layer, one hidden layer, and an output layer. Cells within the hidden layer feature gates, or activated linear functions, and cell states that enable learning long-term dependencies between input and output time series. Input gates control how information from inputs and previous cell states flow to the current state value; forget gates enable attenuation of information within the cell state over time; and output gates control information flow from current state values to hidden layer outputs. Memory within cell states enables the LSTM to model multi-time scale hydrological processes. We refer the reader to Kratzert et al. (2018) for a detailed description of the LSTM structure in the context of hydrologic systems.

The CV-LSTM takes 38 input features: 4 dynamic and 34 static. The first three dynamic input features are basin-averaged climate, including daily precipitation, maximum temperature, and minimum temperature. These data are derived from the 1/16-degree climate product of Livneh et al. (2015). The fourth dynamic feature is day length computed as a function of latitude and day of year (Forsythe et al., 1995). The static features represent catchment attributes, which have been shown to improve the generalizability of LSTM networks across multiple locations (Kratzert, Klotz, Herrnegger, et al., 2019). We include static catchment attributes related to topography, soil, and land cover, derived from the Shuttle Radar Topographic Mission (SRTM) 90-m digital elevation model (Jarvis et al., 2008), 1-km-resolution Advanced Very High Resolution Radiometer (AVHRR) global land cover data (Hansen et al., 2010), and 1-km-resolution State Soil Geographic (STATSGO) data set (Miller & White, 1998), respectively. In total, we define nine topographic features, 12 land covers, and 13 soil characteristics. All input features are standardized before training (by subtracting the mean and dividing by the standard deviation). We provide the full list of CV-LSTM input features in Table S1 in Supporting Information S1.

The output layer of the CV-LSTM has a single neuron targeting observed daily streamflow in millimeters. Note that we do not standardize the observed streamflow. We ensure nonnegative streamflow predictions using the rectified linear unit (ReLU) activation function for the output neuron, expressed as $\mathrm{ReLU}(x) = \max(0,x)$. The California Data Exchange Center (CDEC) archive provides estimates of unimpaired flow (Full Natural Flow, FNF) that represents the natural water production of a river basin unaltered by upstream human modifications (e.g., diversions, storage, and export/import of water between watersheds). We use the daily FNF data for each of the 15 watersheds. Before training the CV-LSTM, we conducted a quality check of the FNF data and identified

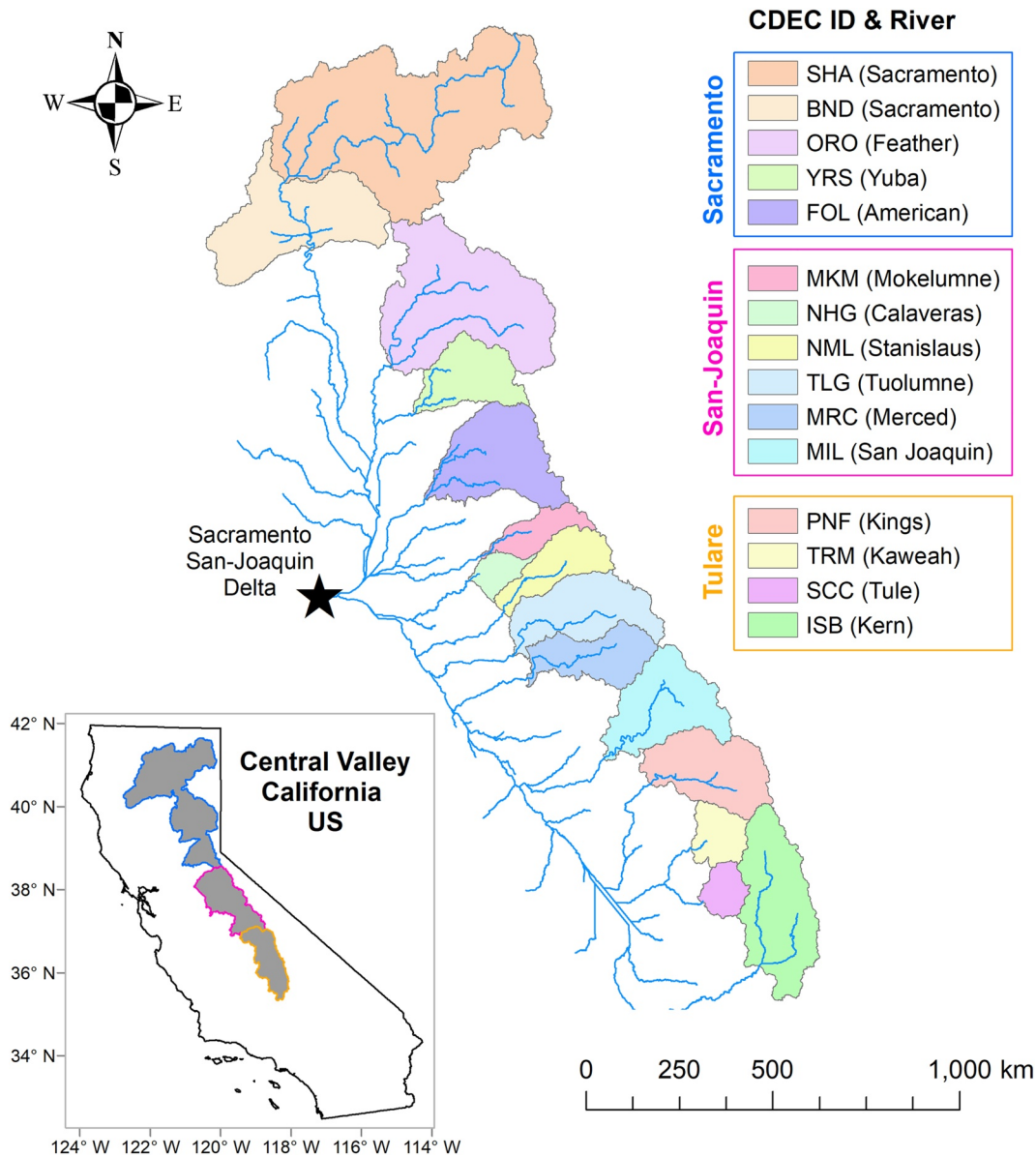

Figure 2. Fifteen watersheds draining into the Central Valley of California, US. Watersheds are labeled based on their ID in the California Data Exchange Center.

suspicious peak flow events, many of which occurred during the mid-to-late summer (i.e., the dry season) and were inconsistent with streamflow measurements at nearby USGS gauges. As a result, we discarded 21 peaks in total (approximately 1.5 days per basin on average). Details describing the quality control are given in Text S1, Tables S2–S4, and Figure S1 in Supporting Information S1.

We train the CV-LSTM by minimizing the mean-squared error averaged over the 15 CDEC watersheds:

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

where N is the number of watersheds and T_n is the number samples in the *nth* basin. $\hat{Q}_{n,t}$ 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 270 days (including day t), which was determined by a grid search over a range of parameter values. Likewise, we found optimal values of other hyperparameters (epochs, dropout rate, mini-batch size, and hidden layer size) by running a 4-fold cross validation grid search. We provide the details of the grid

search and optimal hyperparameters in Text S2, Table S5, and Figures S2–S3 in Supporting Information S1. Network weights are tuned using the ADAM optimizer (Kingma & Ba, 2015). The model is trained with data between 1989 and 2003 and evaluated over the test period of 2004–2013. Here, all years represent water years, that is, October 1 through September 30. The model is trained 10 separate times with different random initializations to account for uncertainty in the training process.

It is worth noting that we also tested the CV-LSTM trained with the Nash Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970) averaged across all basins, which is the objective function used to calibrate our process-based models (see Section 2.2). There are only marginal differences in LSTM streamflow predictions when using either mean-squared error or NSE as the loss function. We chose mean-squared error over NSE as the loss function of the LSTM for consistency with LSTM variants described in Section 2.4.

2.2. Process-Based Hydrologic Models

We develop three process-based hydrologic models as benchmarks, including HYMOD (Boyle, 2001), SAC-SMA (Burnash, 1995) coupled with SNOW-17 (Anderson, 1976), and VIC (Liang et al., 1994). These models are spatially distributed, built using over 6,000 hydrologic response units (HRUs) with Lohmann routing model (Lohmann et al., 1998) that traces the runoff from HRUs through the river channel. We determine the HRUs based on climate and soil information, that is, HRUs are defined by segregating each 1/16° Livneh climate grid cell into mutually exclusive polygons based on different soil classes from the 1-km STATSGO soil data set within that climate grid cell. Figure S4 in Supporting Information S1 shows HRUs across all 15 watersheds, as well as HRU elevation and vegetation that has been upscaled from the original SRTM 90-m DEM (by taking the mean) and 1-km AVHRR vegetation (by taking the majority class).

We employ a large number of HRUs to make the greatest use of spatial gradients in climate and land cover data. However, we parameterized the process models in a parsimonious way to avoid overfitting, such that hydrologic model parameters for HYMOD, SAC-SMA, and VIC are spatially distributed based on the main drivers of each hydrologic process. For instance, the STATSGO soil information (i.e., 13 soil types across the watersheds) determines the distribution of parameters for soil moisture accounting processes. Each HRU with one soil type is assigned one set of parameters that control soil moisture accounting, while another HRU with a different soil type is assigned a different set of soil moisture accounting parameters. In this way, HRUs in different watersheds but with the same soil type will have the same soil moisture accounting parameters. Similarly, the AVHRR vegetation cover (12 cover types) determines the parameter distributions of potential evapotranspiration (PET) modules. Thus, the 15 watersheds share parameters for PET and soil moisture accounting processes, and the total number of parameters for HYMOD, SAC-SMA, and VIC are limited to 100, 204, and 65 across the 15 watersheds (i.e., about 6.7, 13.6, and 4.3 parameters per watershed on average). Note that the number of parameters for each model includes one set of snow and routing parameters that are shared across all 15 watersheds. The complete list of parameters for each model is provided in Tables S6-S8 in Supporting Information S1. Other VIC parameters not listed in Table S8 in Supporting Information S1 were adopted from Livneh et al. (2013). The unit hydrograph parameterization of Wi et al. (2017) added two additional parameters in the Lohmann routing model.

We calibrate the models with the genetic algorithm (Wang, 1991) based on a pooled calibration approach (Wi et al., 2015), in which the algorithm seeks optimal parameters that maximize the average NSE for the 15 FNF series simultaneously. We employ the same training and testing periods as used for the CV-LSTM. The pooled calibration approach has proven effective in reducing streamflow prediction uncertainty and improving predictions in ungauged basins, particularly when using a parsimonious model structure to avoid over-parameterization. Similar to the CV-LSTM, we conduct 10 separate calibrations for each process-based model using different random initializations to account for parameter uncertainty.

The process-based models predict daily streamflow using the same meteorological forcings (Livneh et al., 2015) as the CV-LSTM, including daily precipitation and maximum and minimum temperature. The VIC model also uses wind speed from the same database. For temperature, we further process the data so that temperatures for each HRU are scaled to account for differences between the average elevation of each climate grid cell and the average elevation of HRUs within that climate grid. Rescaling is based on lapse rates inferred from the Moderate Resolution Imaging Spectroradiometer (MODIS) land surface temperature product (MOD11A1 LST, Wan, 2014), which exhibits strong correlation with air temperature near the surface (e.g., Zhang et al., 2018).

Here, we process the MOD11A1 data by month for the period 2001–2019 to adjust HRU temperatures with monthly lapse rates, which are averaged across all 15 watersheds (i.e., we apply one monthly lapse rate for all watersheds). The results are presented in Figure S5 in Supporting Information S1. Monthly lapse rates range from −4.4°C/km (November) to −6.5°C/km (March). With the monthly lapse rates, each process-based model can better represent spatial variations in temperature across HRUs, which can improve model predictions in snow dominated regions (Immerzeel et al., 2014). We did not employ temperature rescaling for the CV-LSTM because watersheds are represented as lumped (rather than spatially distributed) units within the LSTM.

We emphasize here that significant attention was paid to the formulation and calibration of the process-based models (as described above) to ensure a fair comparison with the DL hydrologic model, and to justify their use in PIML approaches described in Section 2.4.

2.3. Evaluating Hydrologic Response Under Warming

To probe the question of whether the CV-LSTM can produce reliable streamflow projections under warming, we increase the historical temperature time series by 1°, 2°, 3°, and 4°C and use them as inputs to the network, while keeping all other inputs the same. We also force the process models with the same warming scenarios and use their projections as benchmarks. To evaluate the changes in hydrologic response projected by each model, we first visually assess shifts in the monthly hydrograph under warming and then we analyze the total runoff ratio (defined as cumulative streamflow divided by cumulative precipitation over the entire test period). The monthly hydrograph reflects seasonal responses of the 15 watersheds to warming. In this snow-dominated region, we anticipate the well-established response of reduced summer runoff and increases in winter runoff due to more precipitation falling as rain in the winter, earlier snowmelt, and increased ET (Gordon et al., 2022).

We assess changes in mass balance using the total runoff ratio. Here, we posit that the runoff ratio should decline across the warming scenarios because temperatures and thus ET increase while precipitation remains unchanged (Rungee et al., 2021). To test this, we evaluate if projections of the total runoff ratio for each model monotonically declines with increasing temperature, that is, whether $R_{n+1} \leq R_n$ for all $n \in [0,1,2,3,4]$ holds for the sequence of total runoff ratios R_n corresponding to a degree of warming n. We note that all 15 California basins have negligible glacial cover, with an average (maximum) of 0.02% (0.12%) of the watershed area covered by perennial snow and ice, based on the land cover data in Falcone et al. (2010). For context, Comeau et al. (2009), Jost et al. (2012), and Moore et al. (2020) found that glacial melt contributes non-negligible percentages of annual streamflow for perennial snow and ice cover around 1%–2% of watershed area. In addition, Schaner et al. (2012) found negligible glacial melt contributions to annual flows throughout the Sacramento-San Joaquin basin in California.

2.4. CV-LSTM Variants: LSTM-MO and Hybrid-LSTM

Here, we redesign the CV-LSTM as a PIML-based version of the model. We consider two LSTM variants, both of which aim to integrate the network with process-based model outputs so that they can learn internal hydrologic processes predicted by the process-based models. The goal of this exercise is to determine whether such PIML approaches can improve how the LSTM projects streamflow under warming, as compared to a standard LSTM that does not leverage process-model output. The two alternative versions of the CV-LSTM tested here utilize process model outputs as either additional training targets (LSTM-MO) or input features (LSTM-IN).

The LSTM-MO uses process-based model predictions of basin-averaged ET and soil moisture. We develop two versions of the LSTM-MO: the first (LSTM-MO1) uses an additional output neuron targeting ET simulated by a process model. The second (LSTM-MO2) has two additional output neurons for ET and soil moisture, respectively. We use loss functions of L_{MO1} and L_{MO2} to train the LSTM-MO1 and LSTM-MO2, respectively:

$$L_{MO1} = \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\hat{Q}_{n,t} - Q_{n,t}\right)^2 + \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\widehat{ET}_{n,t} - ET_{n,t}\right)^2 \tag{2}$$

$$L_{MO2} = \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\hat{Q}_{n,t} - Q_{n,t}\right)^2 + \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\widehat{ET}_{n,t} - ET_{n,t}\right)^2 + \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\widehat{SM}_{n,t} - SM_{n,t}\right)^2 \tag{3}$$

where $\widehat{ET}_{n,t}$ ($\widehat{SM}_{n,t}$) and $ET_{n,t}$ ($SM_{n,t}$) are the evapotranspiration (soil moisture) predictions by the LSTM and a process model for basin n and day t, respectively. We note that weights can be placed on each term in the loss functions above to emphasize one component over another. We experimented with several different weighting schemes and found very little difference in the final models. We therefore adopt an equal weighting scheme in Equations 2 and 3.

The LSTM-IN uses ET from a process-based model as an additional input feature to the network, rather than a target variable. We also experimented with using streamflow and soil moisture as additional inputs, but it had little impact on the final model. Therefore, we omit including these versions of the LSTM-IN, so that the results from the LSTM-IN can be clearly attributed to the use of process-based model ET as input.

The LSTM-MO and LSTM-IN models are trained in the same way as the CV-LSTM (i.e., the same training and test sets and hyperparameters), including 10 different random initializations to account for uncertainty in the training process.

2.5. US National LSTM

Previous work has suggested that LSTM streamflow predictions become more generalizable when the model is trained across a large, diverse set of basins (Gauch, Mai, & Lin, 2021; Kratzert, Klotz, Shalev, et al., 2019). We hypothesize that LSTMs trained to many basins can also better extrapolate streamflow predictions under new climate forcing. To test this, we utilize a national scale LSTM pretrained by Kratzert et al. (2021) on 531 basins across the contiguous United States (hereafter National-LSTM) with the meteorological forcing from Maurer et al. (2002). The National-LSTM was trained using a different set of data compared to our CV-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) data set (Newman et al., 2015). Similar to the models above, the National-LSTM was trained with 10 random initializations to capture training uncertainty. We take the National-LSTM trained over 9 years (2000–2008) directly from Kratzert et al. (2021) without any adjustments. We conduct the same warming scenario experiments for the 531 basins in the National-LSTM (we increase both minimum and maximum temperatures) and evaluate the change in total runoff ratios across basins for the test period of 1990–1999.

The US National-LSTM uses five dynamic input features from the CAMELS data set, including daily precipitation, maximum temperature, minimum temperature, radiation, and water vapor pressure. There is a strong correlation between vapor pressure and minimum temperature in the CAMELS data set, since minimum temperature is used to estimate the water vapor pressure (Newman et al., 2015; Thornton et al., 2021). Thus, to run the National-LSTM under warming scenarios, we also adjust the vapor pressure input based on the changes imposed to minimum temperature. We provide details of the process of adjusting water vapor pressure under warming in the Supporting Information (Text S3, Figures S6–S7 in Supporting Information S1).

3. Results

3.1. Comparison of CV-LSTM and Process-Based Models

Figure 3 shows the prediction skill (NSE) of the CV-LSTM and three process-based models on the test set (2004–2013) for each watershed, estimated from the ensemble mean from the 10 separate training trials. Overall, the CV-LSTM exhibits better performance across the watersheds than all three process models, with an average NSE of 0.86 over the 15 watersheds. The SAC-SMA model performs best among the process models, with an average NSE of 0.79. In all watersheds except NML, the CV-LSTM performs better than SAC-SMA. Both HYMOD and VIC perform significantly worse than the other two models, with VIC performing the worst (average NSE of 0.64). Both VIC and HYMOD struggle with basins further south in the San-Joaquin and Tulare basins (NML, MIL, PNF, and SCC). For all four models, we found the lowest NSE in the Tule River basin (SCC). The variance of NSEs in the Tule basin across the 10 training trials (not shown) is also significantly higher than other watersheds for all models. We also note that the LSTM performance is robust to alternative training and testing sets. Similar to the approach taken in Salvi et al. (2016), we retrained the CV-LSTM on the 15 coldest years in the record and tested the model on the 10 warmest years, and testing set performance remained very strong (average NSE of 0.79 across the 15 basins).

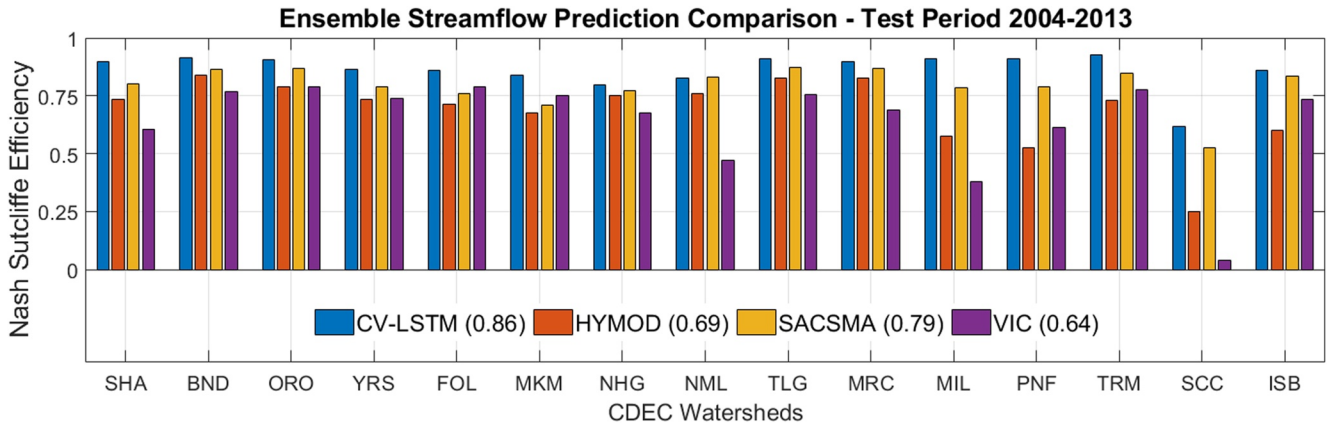

Figure 3. Comparison of streamflow predictions in the test set between the CV-LSTM and three process-based models for the 15 CDEC watersheds. Nash Sutcliffe Efficiency (NSE) estimated from the ensemble mean from the 10 separate training trials is used as a performance metric. An average NSE over the 15 watersheds is shown in parentheses for each model.

Next, we evaluate streamflow projections under 1°, 2°, 3°, and 4°C warming scenarios over the test period for all four models, with precipitation unchanged from its historical values. Figure 4 shows the monthly streamflow climatology of all models for the Feather River watershed at Oroville Dam (ORO). The monthly streamflow climatology for each model was calculated from the ensemble mean daily streamflow prediction (i.e., average over the 10 training trials). Three major insights emerge from Figure 4. First, all models project monthly streamflow to change under warming in ways that conform to expectations, that is, increases in winter runoff and decreases in late spring and summer runoff. An analysis of state variables from the process models (not shown) confirms this is due to more precipitation as rain in the cold season, earlier snowmelt, and increased evapotranspiration triggered by warming. The CV-LSTM captures the same basic pattern despite not having explicit representations of these processes in its architecture; these processes are instead represented by the internal states of the network (Lees et al., 2021).

Second, there is a substantial amount of variability in the projected monthly hydrographs across the four different models. The CV-LSTM projects larger increases in wintertime runoff and smaller declines in summer runoff compared to the SAC-SMA model. HYMOD predicts even smaller increases in winter runoff compared to large declines in spring and summer runoff (especially under the warmest scenario), whereas VIC exhibits the largest summertime declines and a more noticeable shift in peak flow timing. The variability in response is striking given that all models performed similarly in the ORO basin, with HYMOD and VIC having only slightly lower out-of-sample performance compared to CV-LSTM and SAC-SMA (see Figure 3). The differences in monthly hydrograph response across the four (skillful) models highlight the difficulty in determining the true response of the watershed to warming.

Finally, Figure 4 shows the annual net changes of streamflow relative to the baseline for each projection. Because of the increases in ET under warming, all process models project that annual total streamflow decreases, approximately at a rate of 26 mm (HYMOD), 16 mm (SAC-SMA), and 18 mm (VIC) per degree Celsius warming. These annual declines are also visible in the monthly hydrographs for all three process models: the area between the baseline and warming scenarios during the summer (i.e., water losses) outweighs the area between the baseline and warming scenarios in the cold season (i.e., water gains). Importantly, the pattern of net change in annual streamflow projected by the CV-LSTM does not match that of the process models. Under all levels of warming, the CV-LSTM projects slight increases in annual streamflow over the baseline, that is, it projects annual net changes in streamflow that differ in both sign and magnitude compared to the process models. The annual water balance results seen in Figure 4 are not unique to the Feather River basin.

Figure 5 shows the total runoff ratios, calculated from the ensemble mean daily streamflow from the 10 training trials, for all 15 watersheds under the different warming scenarios for the CV-LSTM (Figure 5a) and SAC-SMA (Figure 5b) models. For each watershed, we tested whether the sequence of runoff ratios increases with increasing temperature, that is, whether the model violates the assumption that total streamflow should decline with

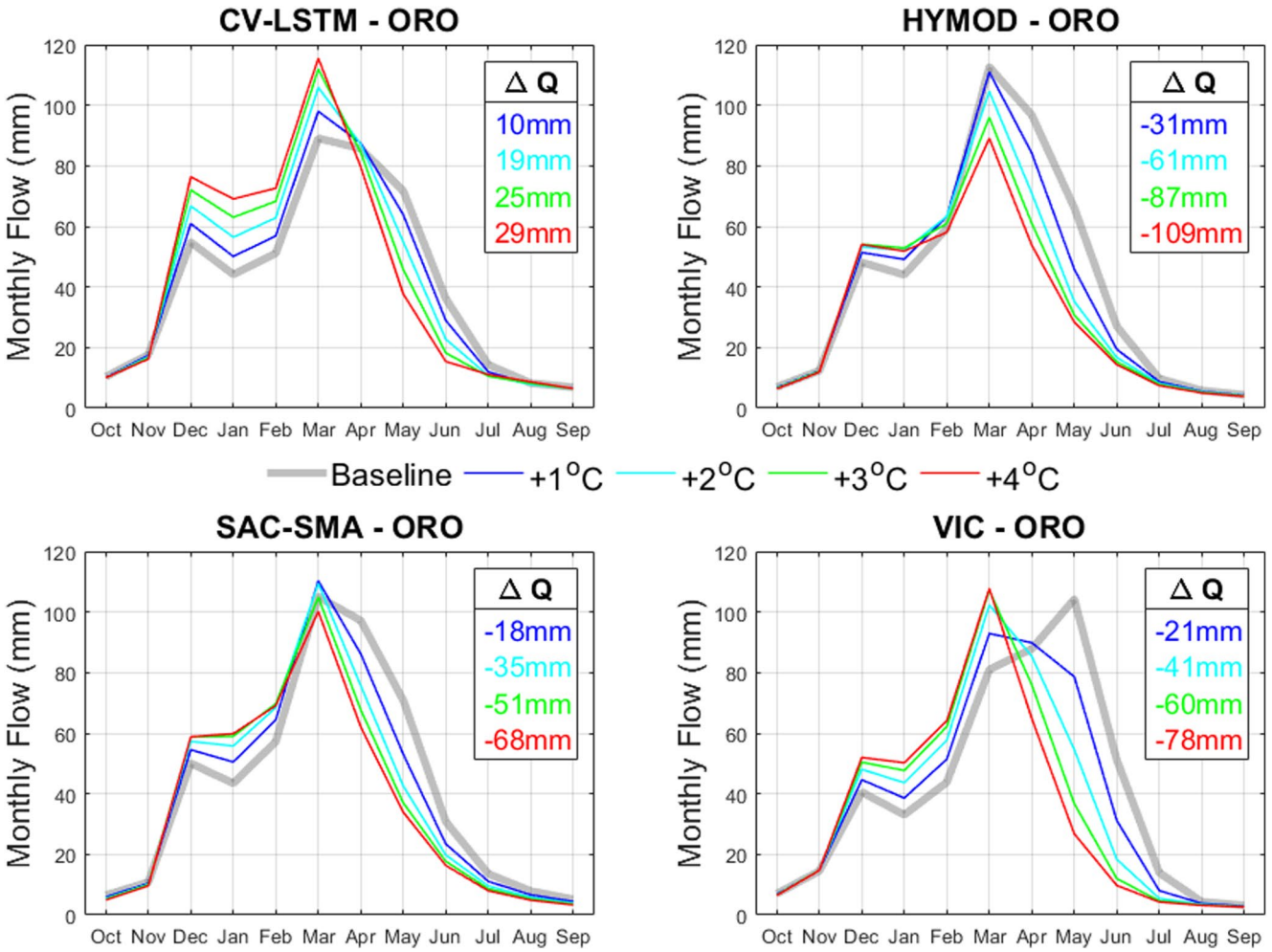

Figure 4. Monthly hydrograph under the warming scenarios for ORO (the Feather River watershed at Oroville Dam). Monthly streamflow predictions are averaged over the test period of 2004–2013. Baseline simulations (i.e., no warming) are specific to each model. The net increases of annual streamflow relative to the baseline annual streamflow are shown in the inserted panels.

warming if precipitation is held constant. We report the watersheds that violate this assumption with an asterisk appended to the watershed CDEC ID.

The CV-LSTM exhibits an increasing runoff ratio with warming for 6 watersheds (Figure 5a). In contrast, the SAC-SMA model always projects that the total runoff ratio will decline with warming, across all basins (Figure 5b). This is also the case for the other two process models (see Figures S8 and S9 in Supporting Information S1). These results show that the CV-LSTM, while more skillful in daily out-of-sample prediction, may produce misleading streamflow projections under warming for a large number of watersheds in the domain.

3.2. Evaluation of CV-LSTM Variants

Next, we evaluate whether the two PIML variants of the CV-LSTM (LSTM-MO and LSTM-IN) help reduce the number of instances where the runoff coefficient increases with warming. We first focus on LSTM-MO, for which there are two versions: a two-node output network that predicts streamflow and process-model ET (LSTM-MO1) and a three-node output network that predicts streamflow, process-model ET, and process-model soil moisture (LSTM-MO2). Both LSTM-MO1 and LSTM-MO2 were fit separately using output from each of the three process models (i.e., their ensemble mean predictions of daily ET and soil moisture), for a total of six models. The networks effectively capture both ET and soil moisture for all three process models. For instance, LSTM-MO1 predicts SAC-SMA ET with an average test-period NSE of 0.98 over the 15 watersheds, while

10.1029/2022WR032123

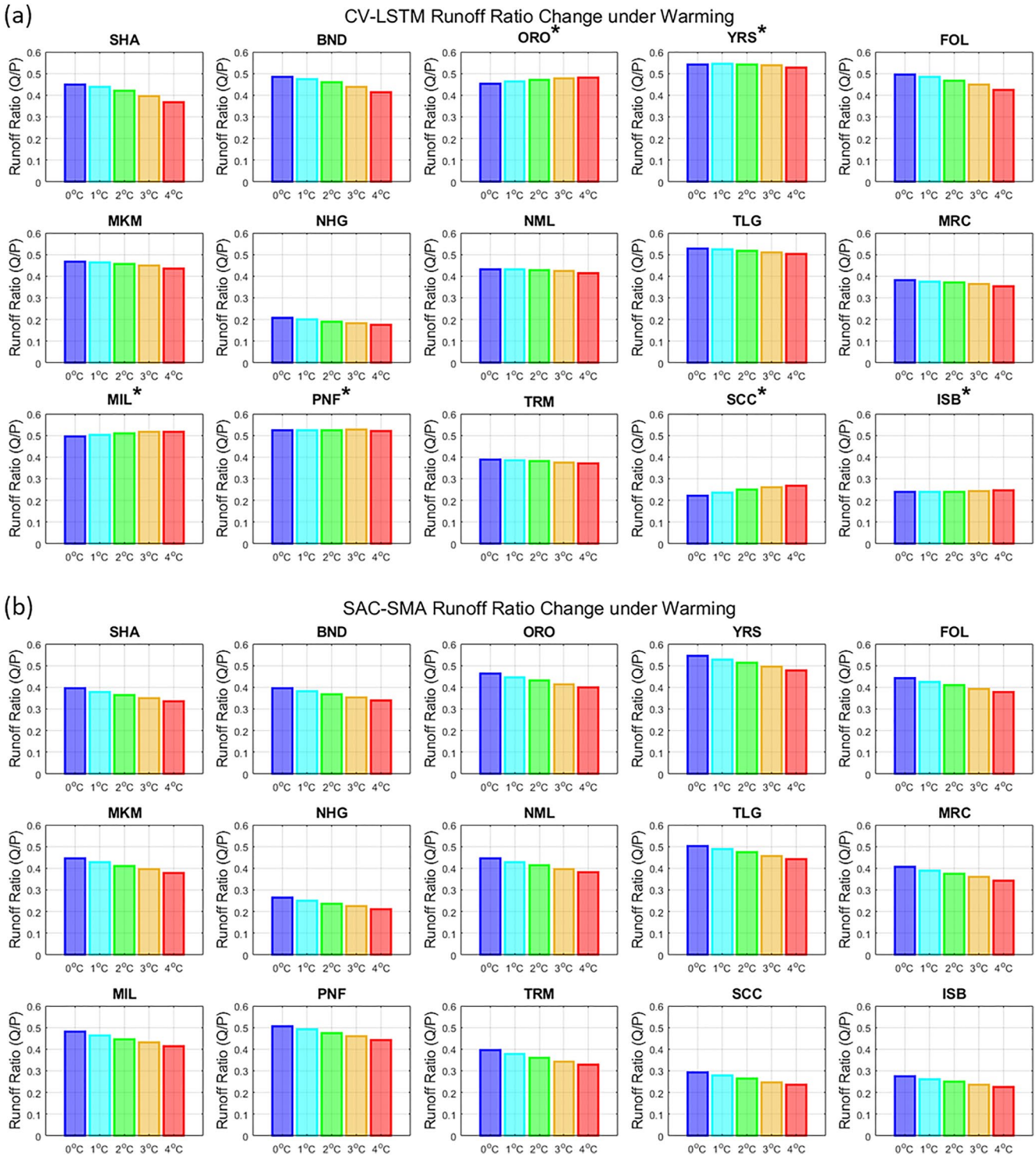

Figure 5. Total runoff ratio changes for the (a) CV-LSTM and (b) SAC-SMA for each watershed under the scenarios of warming. The bars represent the total runoff ratios across warming scenarios for each watershed, calculated from the ensemble mean daily streamflow from the 10 training trials. The asterisk indicates the watersheds violating the assumption that the sequence of total runoff ratios should decrease with increasing temperature.

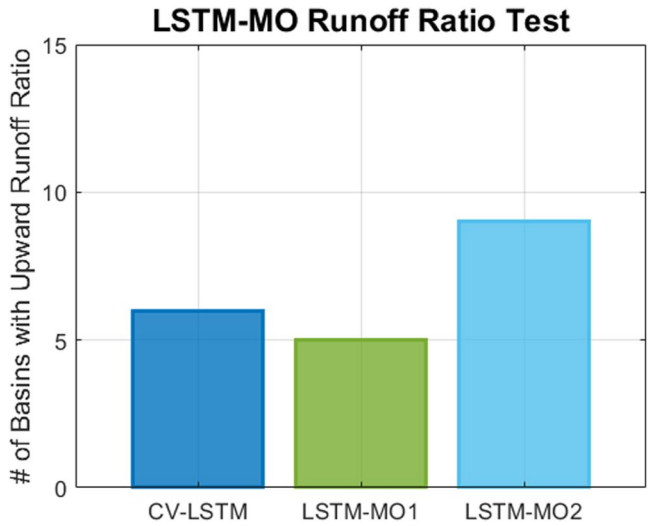

Figure 6. The total number of violations of the monotonic decreasing sequence test across 15 watersheds for the CV-LSTM and two multi-output LSTMs (LSTM-MO1 and LSTM-MO2 using the outputs from SAC-SMA).

LSTM-MO2 predicts test-period ET and soil moisture from the SAC-SMA model with an average NSE of 0.98 and 0.97, respectively. Importantly, there was no drop in out-of-sample skill in streamflow prediction for either of the two versions of LSTM-MO compared to the original CV-LSTM (see Figure S10 in Supporting Information S1).

Figure 6 summarizes the number of watersheds where the runoff coefficient increases with warming across all 15 watersheds. We show results for the original CV-LSTM and the two versions of LSTM-MO fit to the ensemble mean predictions of SAC-SMA for ET and soil moisture, but results are similar for the other process models. The comparison between CV-LSTM and LSTM-MO1 suggests only a small change in the number of basins where the runoff coefficient increases with warming, whereas for LSTM-MO2, the number of these violations increases significantly. Overall, these results show that the addition of process model ET and soil moisture as training targets does not help shape the LSTM network in a way to better preserve total water budgets under warming.

Results shift significantly when process-model output are used as an input feature, as was done for the LSTM-IN variant. Figure 7 shows the total runoff ratio under warming for all 15 CDEC watersheds projected by LSTM-IN fit with SAC-SMA ET as input. Here all watersheds show no increases in the runoff coefficient with warming. For many basins, the rate of decline in the runoff coefficient with warming under the LSTM-IN is similar to that of the SAC-SMA model (Figure 5b), even though the magnitude of the coefficient can differ between the two models. In addition, the LSTM-IN exhibits the same out-of-sample streamflow prediction skill as the original CV-LSTM (see Figure S10 in Supporting Information S1).

This result suggests that a form of PIML that uses process model output as input features can promote more realistic streamflow behavior under warming. However, further investigation shows that this result depends on

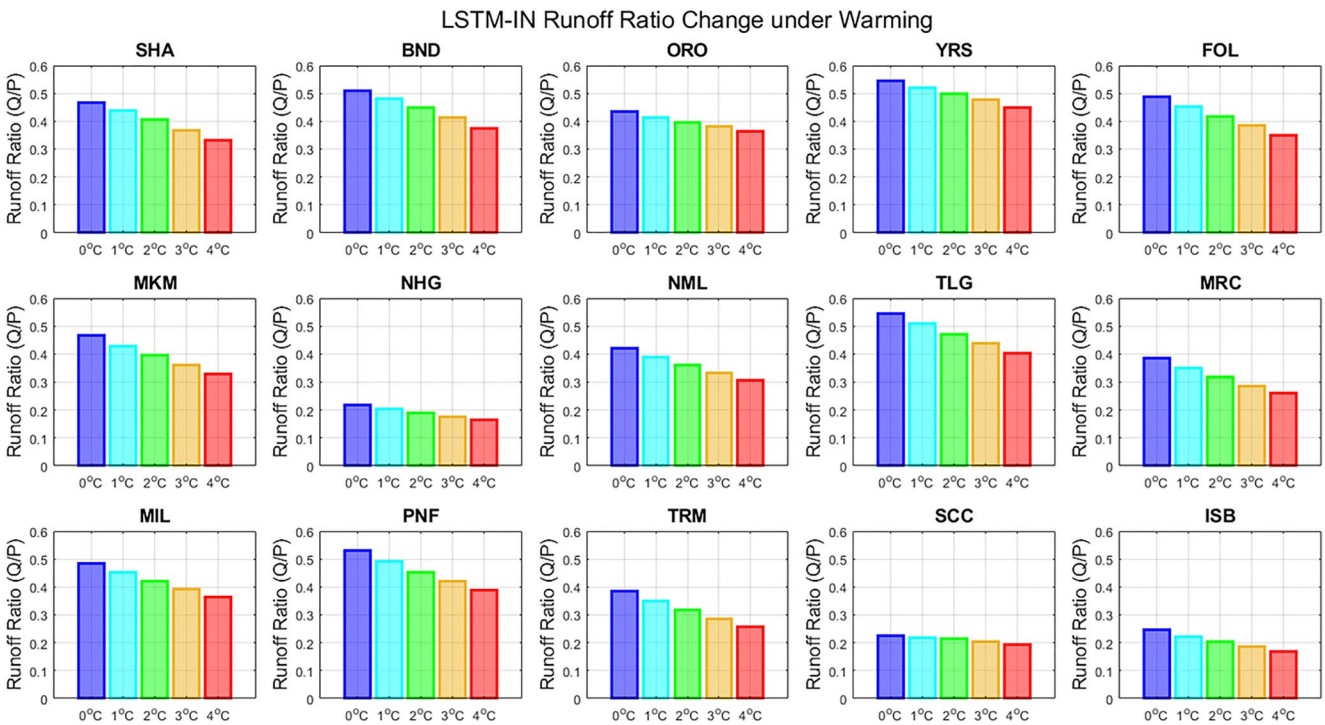

Figure 7. Same as Figure 5 but for the LSTM-IN fit using the ensemble mean ET from the SAC-SMA model as an input feature.

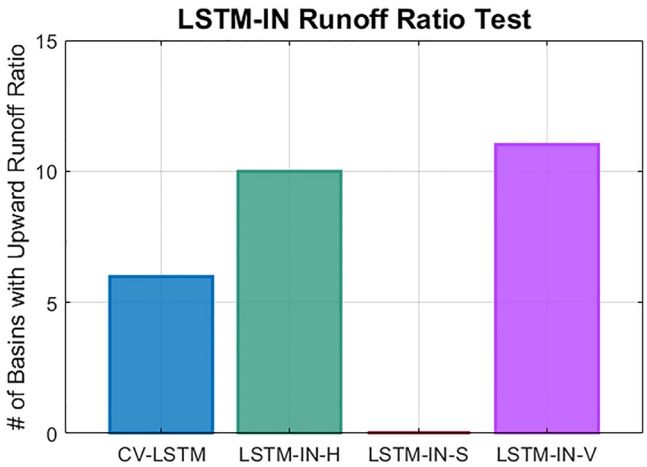

Figure 8. Same as Figure 6 but for the LSTM-IN using the ensemble mean ET from HYMOD (LSTM-IN-H), SAC-SMA (LSTM-IN-S), and VIC (LSTM-IN-V). Results for the original CV-LSTM are also shown for comparison.

the process model being used. Figure 8 shows the total number of basins in which the runoff coefficient increases with warming for the LSTM-IN model fit using ET from each of the process models. Like Figure 6, we show the total number of basins where the total runoff ratio increases with warming across all 15 basins. Unlike the LSTM-IN fit using ET from SAC-SMA, the other two LSTM-IN models fit using ET from HYMOD and VIC do not eliminate instances of increasing total runoff with warming. The number of violations actually grows considerably when the LSTM-IN is fit with ET from HYMOD and VIC.

To better understand the differences between LSTM-IN seen in Figure 8, Figure 9 shows the monthly evapotranspiration climatology for each of the process models for the Feather River basin at Oroville Dam. The discrepancies in monthly ET among the models are significant. HYMOD and VIC estimate significantly higher spring and summer ET compared to SAC-SMA, with VIC showing the largest values in the late summer. VIC underestimates winter ET compared to the other two models and exhibits a later peak compared to HYMOD. The intermonthly variance of ET in SAC-SMA is much smaller compared to the other models. As a reference, we also provide the monthly ET climatology obtained from the MODIS product of MOD16A2 V6 (Running et al., 2017) that computes ET based on the Penman-Monteith equation with inputs of meteorological reanalysis data and MODIS remotely sensed vegetation properties. MODIS estimates higher winter ET and lower summer ET than all process models but the ET seasonal variation of MODIS is closest to that of SAC-SMA. The patterns shown in Figure 9 are similar across the other 14 watersheds. Overall, the results from Figures 7–9 imply that the LSTM-IN can improve the realism of streamflow predictions under warming, but this depends on the fidelity of the process model selected.

3.3. Evaluation of National-LSTM

The results presented up to this point have all been based on variants of an LSTM fit to 15 watersheds in California. Here, we test whether an LSTM fit to many more watersheds, and therefore exposed to a larger diversity of hydrologic response across a range of climates, can produce physically plausible streamflow projections under warming. Figure 10 shows the watersheds for which the ensemble mean streamflow of the National-LSTM predicts an increasing total runoff ratio under warming scenarios at 531 CAMELS watersheds across the US. There are 27 basins (~5%) that violate the assumption of a monotonic decline in runoff ratio with warming for 531 CAMELS basins. The frequency of violations in the National-LSTM is smaller than that seen for the CV-LSTM (40% of the California basins showed a violation), although the sample size for the latter is much smaller.

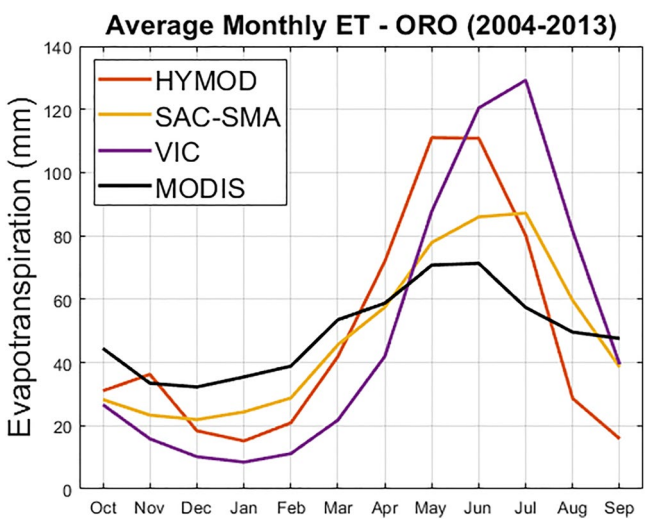

Figure 9. Comparison of monthly evapotranspiration climatology from the ensemble mean predictions of three process models and the MODIS ET product (MOD16A2).

Importantly, the spatial distribution of violations in the National-LSTM is not uniform across the US. Rather, all instances where the runoff coefficient increases with warming are limited to the Western US, and primarily in snow-dominated mountainous regions. There are no violations in the Eastern US. We note that five CAMELS basins are nested within the larger basins used in the CV-LSTM. The total runoff ratio increases with warming in some of these nested basins and decreases in others, with no clear consistency with the pattern of total runoff change projected by the CV-LSTM for the larger basins. Furthermore, some (but not all) of the locations in the Western US where the National-LSTM predicts increasing total runoff with warming are situated in glacier-fed regions (shown at a coarse resolution in the blue hatched regions in Figure 10; see Bidlake et al. (2007) for the source of glacial areas). In total, 17 of the 27 CAMELS watersheds that exhibit an increasing total runoff ratio with warming have at least some perennial snow or ice cover based on the land cover data in Falcone et al. (2010), although

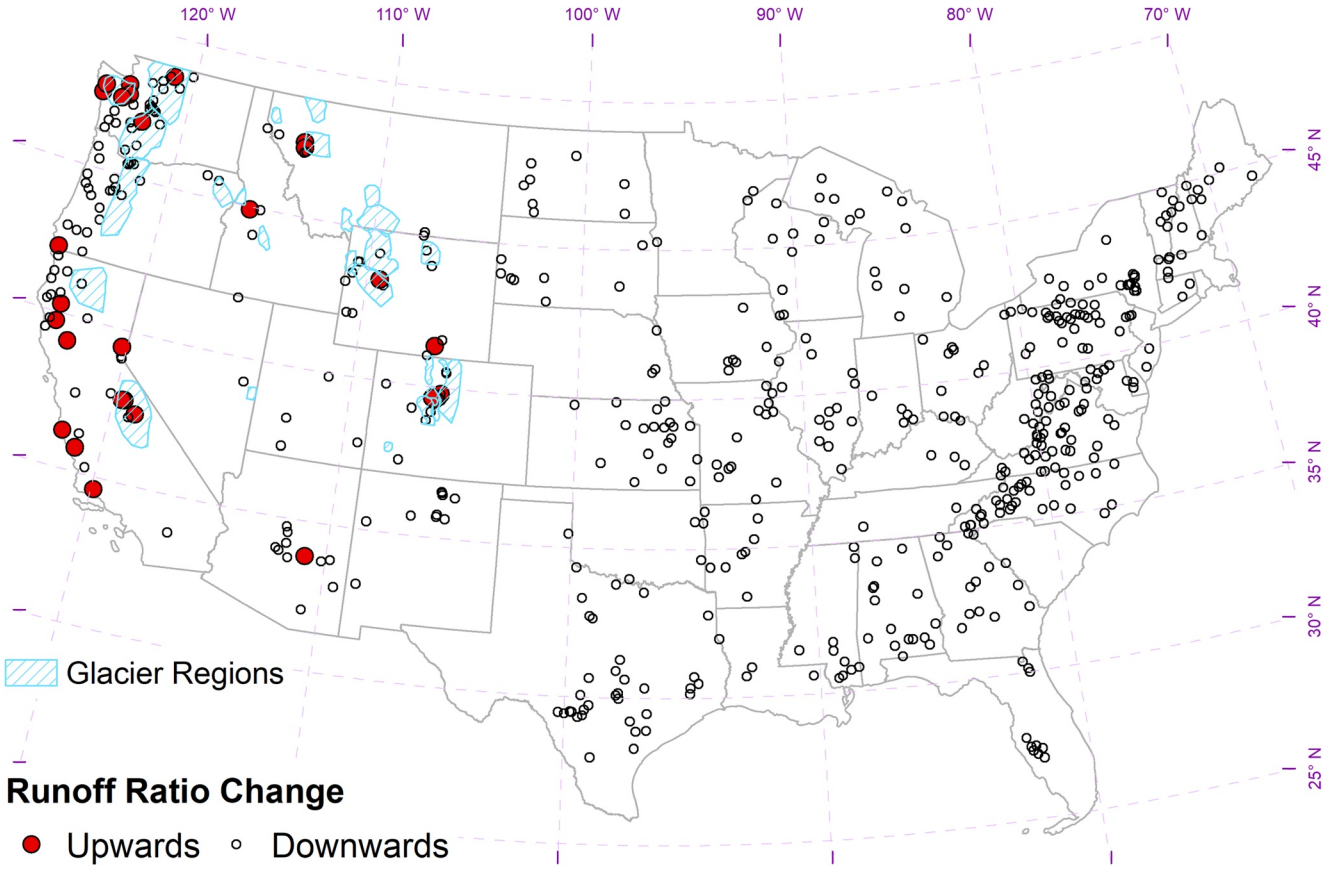

Figure 10. Evaluation of the total runoff ratio under warming scenarios for the 531 basins in the National-LSTM. Red circles represent the CAMELS watersheds in which the total runoff ratio increases with warming. Glacial regions, taken from Bidlake et al. (2007), are highlighted in blue hatching.

only nine of those 27 sites have over 1% of their drainage area classified as perennial snow or ice. These 9 locations are all situated in the Pacific Northwest and Rocky Mountains of Montana, Wyoming, and Colorado, where Schaner et al. (2012) estimated that at least 5% of annual flow is derived from glacial meltwater. Those same areas also exhibit high historical runoff ratios (even greater than 1 in some basins), as derived from the CAMELS observations as well as the National-LSTM predictions (Figure S11 in Supporting Information S1). This result is important because, at least initially, it is reasonable to expect an increase in total runoff in glacierized regions under warming as the higher temperatures can release excess water from the glacial reservoir. This response cannot continue indefinitely, as the glacier or over-year snowpack would eventually be depleted, but the initial response of runoff totals to warming is physically plausible (at least for basins with a non-negligible percentage of glacial coverage by area). This result, when taken together with the overall percent reduction in basins exhibiting mass balance violations from the National-LSTM compared to the CV-LSTM, suggests that an LSTM trained to many diverse basins may also improve streamflow projections under warming, even without explicit inclusion of a PIML-component.

4. Discussion and Conclusion

In this study, we address the question of whether deep learning models can produce reliable future projections of streamflow under climate change. Using 15 watersheds in the California Central Valley as a case study, we assess monthly and decadal scale hydrologic responses under simple warming scenarios based on a regional LSTM (CV-LSTM), three process-based benchmark models (HYMOD, SAC-SMA, and VIC), and two PIML variants of the LSTM that leverage process-based model output (LSTM-MO and LSTM-IN). We perform a simi-

lar experiment with a national-scale LSTM (National-LSTM; Kratzert et al., 2021) fit to 531 basins across the contiguous US. The significant findings from our experiments are as follows:

- The CV-LSTM projects similar shifts in the monthly hydrograph under warming to the process-based models, with seasonal differences between the LSTM and process models well within the variability exhibited across the process models.
- However, the CV-LSTM projects an increasing trend in the total runoff ratio under warming in 6 of 15 basins, while all process models project a monotonic decrease in total runoff ratio across all basins.
- The multi-output LSTM networks (LSTM-MO1 and LSTM-MO2) do not help rectify the unrealistic increases in the total runoff ratio.
- The LSTM-IN using SAC-SMA ET estimates as an input variable corrects unrealistic increases in the total runoff ratio for the 15 California Central Valley watersheds. However, similar corrections are not observed for the LSTM-IN models using ET estimates from HYMOD or VIC.
- Similar to the CV-LSTM, the National-LSTM simulates an increasing runoff ratio under warming in a subset of basins. However, the percentage of basins where this response occurs is lower, and it is more common in glaciated regions where higher runoff ratios with warming might be expected (at least initially).

From the findings above, we conclude that hybrid modeling using process model outputs as additional input features can support the use of LSTMs for hydrologic projections under a changing climate, although this is not guaranteed (i.e., it depends on the fidelity of the process-based hydrologic model). We also find some evidence that training LSTMs to a large, diverse set of watersheds may also help improve the realism of hydrologic projections under climate change.

Previous work has documented the effect of warming on the hydrologic response in US mountainous regions: a decreased fraction of precipitation falling as snow, a shift in snowmelt season to earlier in the spring, increased winter runoff, and reduced warm season runoff (Dierauer et al., 2018; Mote et al., 2018; Rungee et al., 2021). The streamflow projections from the CV-LSTM and all three process models conform to these expectations (see Figure 4). However, unlike the process model projections, the CV-LSTM projections can show increases in total streamflow over the baseline under sufficient warming. Throughout this work, we assumed such a response was unrealistic under the assumption that total runoff should decline as temperature rises if precipitation is held constant. This assumption formed the basis for many of the model evaluations presented in this work, and therefore deserves scrutiny. Specifically, we could entertain an alternative assumption that warming translates runoff from the warm season (when ET is high) to the cold season (when ET is low) via reduced snowpack, thus reducing the opportunity for summer evapotranspiration and increasing the total runoff across the year. Under this alternative assumption, the climate change responses predicted by the CV-LSTM could be considered plausible.

However, we see this alternative assumption as highly unlikely for two reasons. First, several observation-based studies have linked warming temperatures to long-term declines in runoff efficiency in the US, caused by increased evapotranspiration and diminished soil moisture (Lehner et al., 2017; Martin et al., 2020; Overpeck & Udall, 2020; Woodhouse & Pederson, 2018). Only one study did not detect significant downward trends in runoff efficiency in locations across the West (McCabe et al., 2018), but the effects of trending temperature and precipitation were not clearly separated in the analysis. In addition, others have found that reduced snow due to warming and the consequent decline in albedo has increased evapotranspiration, leading to further declines in river flows (Liu et al., 2022; McCabe et al., 2017; Milly & Dunne, 2020). In aggregate, these empirical results raise significant doubts about any model-based projections of increasing runoff efficiency under warming. Second, when closely examining the results of the National-LSTM, some basins with minimal or no snow influence can also exhibit increasing runoff coefficients with warming (see southern California and Arizona in Figure 10). We can think of no causal mechanisms that would drive this response in these locations, and so conclude that in at least some locations the LSTM-projected responses under warming are questionable.

The unrealistic climate change responses of the CV-LSTM motivated our development of two PIML-based model variants. For the multi-output LSTM variants (LSTM-MO1 and LSTM-MO2), we hypothesized that predicting intermediate hydrologic states (i.e., process model ET and soil moisture) in addition to streamflow would improve physical consistency in the network (Khandelwal et al., 2020) and therefore help fix mass balance violations. This was not the case, even though all multi-output LSTMs successfully predicted ET and soil moisture from the process models (average NSE above 0.9 across the 15 basins, 3 models, and for both variables)

while maintaining very similar skill in streamflow prediction as compared to the CV-LSTM. We note that all multi-output LSTM models project ET to increase under warming, much like the ET patterns projected by the process models (see Figure S12 in Supporting Information S1). This means that the multi-output LSTMs are capable of learning process model ET responses under warming, but such learned responses do not translate into the predicted streamflow response. Taken together, these results suggest that the multi-output LSTM structure as implemented in this work is not a promising PIML strategy to help learn physically consistent hydrologic processes under climate change. However, alternative multi-output strategies may prove more effective, such as introducing a loss function to force the outputs to interact with each other in a way to respect mass conservation (Ruckstuhl et al., 2021). This effort is left for future work.

In the literature, results have been mixed in terms of improved hydrologic prediction when using hybrid DL models that use process-based model outputs as additional inputs. Hybrid DL streamflow models have shown noticeable improvement over individual process and DL models when using process model predictions of streamflow (Lu et al., 2021), predictions of soil moisture (Humphrey et al., 2016), and prediction errors for streamflow (Konapala et al., 2020; Tian et al., 2018). Conversely, Frame, Kratzert, Raney, et al. (2021) trained a hybrid LSTM on 531 CAMELS basins across the contiguous US using 20 National Water Model (Gochis et al., 2015) outputs as additional inputs and showed that the hybrid DL performed better than the process model but not the standalone LSTM.

In the present study, we tested various hybrid DL models that integrate the CV-LSTM with process model predictions of evapotranspiration, soil moisture, and streamflow separately from three different models (i.e., HYMOD, SAC-SMA, and VIC). Overall, we developed 9 hybrid LSTM models that use process model output as additional input to the network, and for all, we found no drop in streamflow prediction skill compared to the original CV-LSTM (in terms of NSE). This suggests that the hybrid model skill for streamflow prediction is not affected by the performance of the process model, which varied considerably across HYMOD, SAC-SMA, and VIC. We also found very little difference between hybrid LSTMs that use only process-model ET as input and those that also use process model soil moisture and streamflow predictions as inputs. This motivated our focus on the ET-only hybrid-LSTMs, in order to help isolate the effects of process-model ET as an input.

Most importantly, we found potential for the LSTM-IN to extrapolate streamflow predictions under warming, but this result depended on the process model ET used as input. The LSTM-IN using ET from SAC-SMA was the only hybrid model that helped eliminate all instances of increasing total runoff with warming, while the number of unrealistic projections actually increased when using ET from the other two process models. This suggests that process model credibility plays a large role in the success of hybrid DL approaches in the context of climate change. In the Feather River basin, where streamflow prediction skill for all process models is high and comparable to that of the CV-LSTM (Figure 3), we showed that discrepancies in monthly modeled ET are significant (Figure 9). Estimates of ET from SAC-SMA were closest to independent ET estimates based on remotely sensed data from MODIS (MOD16A2), which are considered a reliable baseline (He et al., 2019). Thus, our results indicate that a hybrid DL approach to improve streamflow projections under warming requires a credible process model of ET. In addition, ET estimates from HYMOD and VIC were very high in the warm season (Figure 9), yet when used in the LSTM-IN, they led to an increase in the runoff ratio. The reasons for this are unclear, but may be related to limitations in the physics representation or parameterization of ET or other processes in these models. For instance, other work has found that VIC can overestimate ET due to incorrect representations of soil moisture fluxes (Dash et al., 2021), and that assimilation of remotely sensed data can help improve VIC-based soil moisture prediction (Gavahi et al., 2020). Overall, these results present a significant challenge to the hybrid approach because assessing the credibility of ET at basin scales is difficult and highly uncertain (Aguilar et al., 2018; Liu et al., 2016). Therefore, while this study highlights the potential for hybrid DL approaches to improve hydrologic projections under climate change, future research is needed to explore methods that ensure added value under this strategy.

Even without hybrid methods, deep learning models can advance hydrologic projection through extensive training with vast quantities of data (Shen et al., 2021). For instance, LSTMs trained on a large number of watersheds can improve streamflow prediction in ungauged basins (Kratzert, Klotz, Herrnegger, et al., 2019), even in data-sparse regions (Ma et al., 2021). Our study adds some support to these findings by demonstrating physically consistent streamflow projections under warming at a decadal scale using the National-LSTM, at least over the eastern half of the US. However, challenges remain. As shown in Figure 10, the number of basins showing increasing runoff

ratios under warming using the National-LSTM is not insignificant, with ~5% of basins exhibiting this behavior. Yet these projections from the National-LSTM—and their credibility or lack thereof—should be interpreted with care. Some (but not all) locations in the National-LSTM that exhibited increasing runoff ratios with warming were located in glacial regions (particularly the Pacific Northwest and Rocky Mountain Range), where higher temperatures could trigger rapid glacial melt and a transitory increase in runoff efficiency. In these locations, the historical total runoff ratios derived from the CAMELS data tend to be high (see Figure S11 in Supporting Information S1), possibly due to glacial melt (Hugonnet et al., 2021). Thus, for several sites it is not clear if the projections from the National-LSTM are unrealistic, at least initially (before glacier volume depletion). Future work could consider the addition of glacier-related inputs (e.g., glacier volume as an initial boundary condition) to help the LSTM better isolate glacial impacts on runoff under warming. Alternatively, glacially fed basins could be removed altogether from national-scale LSTMs meant to predict hydrologic responses under climate change.

Recent literature clearly shows that DL hydrologic models outperform process-based models in hydrologic prediction, but the work in this study suggests that the lack of explicit process representation may still lead to unrealistic hydrologic projections under significant climate change. We have identified promising avenues to help resolve these issues, including hybrid approaches and extensive training on a large, diverse set of watersheds, but basins showing increasing runoff with rising temperature are still possible under both techniques. Thus, we cannot yet definitively claim that either a hybrid approach or a DL model trained to a large set of watersheds always lead to physically consistent streamflow projections under warming. Still, when comparing the National-LSTM and CV-LSTM results of this study, we conclude that future work advancing DL-based climate change projections should focus on models trained to a large and diverse set of data with many basins whenever possible. Consistent with previous work (Gauch, Mai, & Lin, 2021; Kratzert, Klotz, Shalev, et al., 2019), our results show that this approach will likely improve the reliability of climate change projections, given the larger set of data available to learn how hydrologic systems respond to previously unobserved climate conditions. Future work could also explore the utility of explicit mass balance constraints (Frame et al., 2021b, 2022; Hoedt et al., 2021; Nearing et al., 2021) or additional fine-tuning for specific regions, but regional LSTMs trained only to a small set of sites (like the CV-LSTM of this study) should only be employed if data limitations prevent the use of a model trained to many sites. In either case, climate change projections could benefit from ET or other flux estimates from process models, or possibly pre-turning on both historical and future projected process model output (e.g., Read et al., 2019), but significant care is needed to ensure the quality of these estimates. This will be challenging, especially if applied to DL models trained and tested on a large number of watersheds, but this strategy presents a promising avenue to explore in future work.

Data Availability Statement

The code and data used for this project are available at https://doi.org/10.5281/zenodo.6998882. The pre-trained US national models are available at https://doi.org/10.5281/zenodo.4670268 (Kratzert et al., 2021). The CAMELS data are available from NCAR at https://ral.ucar.edu/solutions/products/camels.

Acknowledgments
This research was supported by the U.S. National Science Foundation (Grant No. OIA-2040613). We thank Drs. Grey Nearing and Frederik Kratzert for their comments and feedback, as well as the comments of three anonymous reviewers, which significantly improved the quality of this work.

References

Aguilar, A. L., Flores, H., Crespo, G., Martin, M. I., Campos, I., & Calera, A. (2018). Performance assessment of MOD16 in evapotranspiration evaluation in Northwestern Mexico. *Water*, *10*(7), 901. https://doi.org/10.3390/w10070901

Ahmed, A. A. M., Deo, R. C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., & Yang, L. (2021). Hybrid deep learning method for a week-ahead evapotranspiration forecasting. *Stochastic Environmental Research and Risk Assessment*, *36*(3), 831–849. https://doi.org/10.1007/s00477-021-02078-x

Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water quality prediction using artificial intelligence algorithms. *Applied Bionics and Biomechanics*, *2020*, 6659314. https://doi.org/10.1155/2020/6659314

Anaraki, M. V., Farzin, S., Mousavi, S. F., & Karami, H. (2021). Uncertainty analysis of climate change impacts on flood frequency by using hybrid machine learning methods. *Water Resources Management*, *35*(1), 199–223. https://doi.org/10.1007/s11269-020-02719-w

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

Bhasme, P., Vagadiya, J., & Bhatia, U. (2021). Enhancing predictive skills in physically-consistent way: Physics informed machine learning for hydrologic processes. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2104.11009

Bidlake, W. R., Josberger, E. G., & Savoca, M. E. (2007). *Water, ice, and meteorological measurements at south cascade glacier, Washington, balance years 2004 and 2005 (scientific investigation report 2007-5055)*. U.S. Geological Survey.

Boyle, D. P. (2001). *Multicriteria calibration of hydrologic models*, Doctoral dissertation, The University of Arizona. Retrieved From http://hdl.handle.net/10150/290657

Burnash, R. J. (1995). The NWS river forecast system–Catchment modeling. In V. Singh (Ed.), *Computer models of watershed hydrology* (pp. 311–366). Water Resources Publication.
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
Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology*, *590*, 125376. https://doi.org/10.1016/j.jhydrol.2020.125376
Comeau, L. E. L., Pietroniro, A., & Demuth, M. N. (2009). Glacier contribution to the North and South Saskatchewan rivers. *Hydrological Processes*, *23*(18), 2640–2653. https://doi.org/10.1002/hyp.7409
Das, J., & Nanduri, U. V. (2018). Assessment and evaluation of potential climate change impact on monsoon flows using machine learning technique over Wainganga River basin, India. *Hydrological Sciences Journal*, *63*(7), 1020–1046. https://doi.org/10.1080/02626667.2018.1469757
Dash, S. S., Sahoo, B., & Raghuwanshi, N. S. (2021). How reliable are the evapotranspiration estimates by Soil and Water Assessment Tool (SWAT) and Variable Infiltration Capacity (VIC) models for catchment-scale drought assessment and irrigation planning? *Journal of Hydrology*, *592*, 125838. https://doi.org/10.1016/j.jhydrol.2020.125838
de Silva, B. M., Higdon, D. M., Brunton, S. L., & Kutz, J. N. (2020). Discovery of physics from data: Universal laws and discrepancies. *Frontiers in Artificial Intelligence*, *3*, 25. https://doi.org/10.3389/frai.2020.00025
Dierauer, J. R., Whitfield, P. H., & Allen, D. M. (2018). Climate controls on runoff and low flows in mountain catchments of Western North America. *Water Resources Research*, *54*(10), 7495–7510. https://doi.org/10.1029/2018WR023087
Faghmous, J. H., & Kumar, V. (2014). A big data guide to understanding climate change: The case for theory-guided data science. *Big Data*, *2*(3), 155–163. https://doi.org/10.1089/big.2014.0026
Falcone, J., Carlisle, D., Wolock, D., & Meador, M. (2010). GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. *Ecology*, *91*(2), 621. https://doi.org/10.1890/09-0889.1
Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., et al. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology*, *537*, 45–60. https://doi.org/10.1016/j.jhydrol.2016.03.026
Forsythe, W. C., Rykiel, E. J., Jr., Stahl, R. S., Wu, H., & Schoolfield, R. M. (1995). A model comparison for daylength as a function of latitude and day of year. *Ecological Modelling*, *80*(1), 87–95. https://doi.org/10.1016/0304-3800(94)00034-F
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 in Review. https://doi.org/10.31223/X5BH0P
Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., et al. (2021). Deep learning rainfall-runoff predictions of extreme events. Hydrology and Earth System Sciences Discussions, [preprint]. https://doi.org/10.5194/hess-2021-423
Frame, J. M., Kratzert, F., Raney, A., II, Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-processing the national water model with long short-term memory networks for streamflow predictions and diagnostics. *Journal of the American Water Resources Association*, *57*(6), 1–12. https://doi.org/10.1111/1752-1688.12964
Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall-runoff prediction at multiple timescales with a single long short-term memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021
Gauch, M., Mai, J., & Lin, J. (2021). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, *135*, 104926. https://doi.org/10.1016/j.envsoft.2020.104926
Gavahi, K., Abbaszadeh, P., Moradkhani, H., Zhan, X., & Hain, C. (2020). Multivariate assimilation of remotely sensed soil moisture and evapotranspiration for drought monitoring. *Journal of Hydrometeorology*, *21*(10), 2293–2308. https://doi.org/10.1175/JHM-D-20-0057.1
Ghosh, S., & Mujumdar, P. P. (2008). Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in Water Resources*, *31*(1), 132–146. https://doi.org/10.1016/j.advwatres.2007.07.005
Gochis, D. J., Yu, W., & Yates, D. N. (2015). *The WRF-Hydro model technical description and user's guide*, version 3.0. NCAR Technical Document, (p. 120). Retrieved From http://www.ral.ucar.edu/projects/wrf_hydro/
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
Hansen, M. C., Defries, R. S., Townshend, J. R. G., & Sohlberg, R. (2010). Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, *21*(6–7), 1331–1364. https://doi.org/10.1080/014311600210209
Hanson, P. C., Stillman, A. B., Jia, X., Karpatne, A., Dugan, H. A., Carey, C. C., et al. (2020). Predicting lake surface water phosphorus dynamics using process-guided machine learning. *Ecological Modelling*, *430*, 109136. https://doi.org/10.1016/j.ecolmodel.2020.109136
He, M., Kimball, J. S., Yi, Y., Running, S. W., Guan, K., Moreno, A., et al. (2019). Satellite data-driven modeling of field scale evapotranspiration in croplands using the MOD16 algorithm framework. *Remote Sensing of Environment*, *230*, 111201. https://doi.org/10.1016/j.rse.2019.05.020
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., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., et al. (2021). MC-LSTM: Mass-conserving LSTM. *International Conference on Machine Learning*, (pp. 4275–4286). PMLR. Retrieved from https://arxiv.org/abs/2101.05186
Hugonnet, R., McNabb, R., Berthier, E., Menoounos, B., Nuth, C., Girod, L., et al. (2021). Accelerated global glacier mass loss in the early twenty-first century. *Nature*, *592*(7856), 726–731. https://doi.org/10.1038/s41586-021-03436-z
Humphrey, G., Gibbs, M. S., Dandy, G. C., & Maier, H. R. (2016). A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology*, *540*, 623–640. https://doi.org/10.1016/j.jhydrol.2016.06.026
Huss, M., & Hock, R. (2018). Global-scale hydrological response to future glacier mass loss. *Nature Climate Change*, *8*(2), 135–140. https://doi.org/10.1038/s41558-017-0049-x
Immerzeel, W. W., Petersen, L., Ragettli, S., & Pellicciotti, F. (2014). The importance of observed gradients of air temperature and precipitation for modeling runoff from a glacierized watershed in the Nepalese Himalayas. *Water Resources Research*, *50*(3), 2212–2226. https://doi.org/10.1002/2013WR014506
Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. (2008). Hole-filled SRTM for the globe version 4. *CGIAR-CSI SRMT 90m Database*. Retrieved From http://srtm.csi.cgiar.org
Jia, X., Karpatne, A., Willard, J., Steinbach, M., Read, J., Hanson, P. C., et al. (2018). Physics guided recurrent neural networks for modeling dynamical systems: Application to monitoring water temperature and quality in lakes. Retrieved from https://arxiv.org/abs/1810.02880
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*(13), e2020GL088229. https://doi.org/10.1029/2020GL088229
Jost, G., Moore, R. D., Menounos, B., & Wheate, R. (2012). Quantifying the contribution of glacier runoff to streamflow in the upper Columbia River Basin, Canada. *Hydrology and Earth System Sciences*, *16*(3), 849–860. https://doi.org/10.5194/hess-16-849-2012

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

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Review Physics*, *3*(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

Karpatne, 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

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., et al. (2021). Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200093. https://doi.org/10.1098/rsta.2020.0093

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., et al. (2020). Physics guided machine learning methods for hydrology. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2012.02854

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., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

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

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. S. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample data sets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

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 Discussions, *26*(12), 3079–3101. https://doi.org/10.5194/hess-2021-566

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*(9), 4124–4133. https://doi.org/10.1002/2017GL073253

Li, Q., Wang, Z., Shangguan, W., Li, L., Yao, Y., & Yu, F. (2021). Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. *Journal of Hydrology*, *600*, 126698. https://doi.org/10.1016/j.jhydrol.2021.126698

Liang, J., Li, W., Bradford, S. A., & Šimůnek, J. (2019). Physics-informed data-driven models to predict surface runoff water quantity and quality in agricultural fields. *Water*, *11*(2), 200. https://doi.org/10.3390/w11020200

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land-surface water and energy fluxes for general-circulation models. *Journal of Geophysical Research*, *99*(D7), 14415–14428. https://doi.org/10.1029/94jd00483

Liu, D., Jiang, W., Mu, L., & Wang, S. (2020). Streamflow prediction using deep learning neural network: Case study of Yangtze river. *IEEE Access*, *8*, 90069–90086. https://doi.org/10.1109/ACCESS.2020.2993874

Liu, W., Zhou, J., Li, Y., Sun, F., Fu, G., Li, X., & Sang, Y. F. (2016). A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. *Journal of Hydrology*, *538*, 82–95. https://doi.org/10.1016/j.jhydrol.2016.04.006

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

Livneh, B., Bohn, T., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., et al. (2015). A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950-2013. *Scientific Data*, *2*(1), 150042. https://doi.org/10.1038/sdata.2015.42

Livneh, B., Rosenberg, E. A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K., et al. (2013). A long-term hydrologically based data set of land surface fluxes and states for the conterminous United States: Update and extensions. *Journal of Climate*, *26*(23), 9384–9392. https://doi.org/10.1175/JCLI-D-12-00508.1

Lohmann, D., Raschke, E., Nijssen, G., & Lettenmaier, D. (1998). Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model. *Hydrological Sciences Journal*, *43*(1), 131–141. https://doi.org/10.1080/02626669809492107

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

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*(5), e2020WR028600. https://doi.org/10.1029/2020WR028600

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 of the United States of America*, *117*(21), 11328–11336. https://doi.org/10.1073/pnas.1916208117

Maurer, E. P., Wood, A., Adam, J., Lettenmaier, D. P., & Nijssen, B. (2002). A long-term hydrologically based data set of land surface fluxes and states for the conterminous United States. *Journal of Climate*, *15*(22), 3237–3251. https://doi.org/10.1175/1520-0442(2002)015<3237:althbd>2.0.co;2

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

McCabe, G. J., Wolock, D. M., & Valentin, M. (2018). Warming is driving decrease in snow fractions while runoff efficiency remains mostly unchanged in snow-covered areas of the Western United States. *Journal of Hydrometeorology*, *19*(5), 803–814. https://doi.org/10.1175/JHM-D-17-0227.1

Miller, D., & White, R. A. (1998). A conterminous United States multilayer soil characteristics data set for regional climate and hydrology modeling. *Earth Interactions*, *2*(2), 1–26. https://doi.org/10.1175/1087-3562(1998)002<0002:cusms>2.0.co;2

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

Moore, R. D., Pelto, B., Menounos, B., & Hutchinson, D. (2020). Detecting the effects of sustained glacier wastage on streamflow in variably glacierized catchments. *Frontiers of Earth Science*, *12*. https://doi.org/10.3389/feart.2020.00136

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

Nash, J. E., & Sutcliff, J. V. (1970). River flow forecasting through conceptual models part I–A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

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*(3), e2020WR028091. https://doi.org/10.1029/2020WR028091

Nearing, G. S., Pelissier, C. S., Kratzert, F., Klotz, D., Gupta, H. V., Frame, J. M., & Sampson, A. K. (2019). Physically informed machine learning for hydrological modeling under climate nonstationarity. In *Science and technology infusion climate bulletin* (pp. 22–24). NOAA's National Weather Service 44th NOAA Annual Climate Diagnostics and Prediction Workshop. Retrieved from https://www.nws.noaa.gov/ost/climate/STIP/44CDPW/44cdpw-GNearing.pdf

Newman, A., Clark, M. P., Sampson, K., Wood, A., Hay, L., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set 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

Overpeck, J. T., & Udall, B. (2020). Climate change and the aridification of North America. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(22), 11856–11858. https://doi.org/10.1073/pnas.2006323117

Paniconi, C., & Putti, M. (2015). Physically based modeling in catchment hydrology at 50: Survey and outlook. *Water Resources Research*, *51*(9), 7090–7129. https://doi.org/10.1002/2015WR017780

Pritchard, H. D. (2019). Asia's shrinking glaciers protect large populations from drought stress. *Nature*, *569*(7758), 649–654. https://doi.org/10.1038/s41586-019-1240-1

Rahmani, F., Lawson, K., Ouynag, W., Appling, A., Oliver, S., & Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*, *16*, 024025. https://doi.org/10.1088/1748-9326/abd501

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, *55*(11), 9173–9190. https://doi.org/10.1029/2019WR024922

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Ruckstuhl, Y., Janjic, T., & Rasp, S. (2021). Training a convolutional neural network to conserve mass in data assimilation. *Nonlinear Processes in Geophysics*, *28*(1), 111–119. https://doi.org/10.5194/npg-28-111-2021

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

Running, S., Mu, Q., & Zhao, M. (2017). MOD16A2 MODIS/terra net evapotranspiration 8-day L4 global 500m SIN grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD16A2.006

Salvi, K., Ghosh, S., & Ganguly, A. R. (2016). Credibility of statistical downscaling under nonstationary climate. *Climate Dynamics*, *46*(5–6), 1991–2023. https://doi.org/10.1007/s00382-015-2688-9

Schaner, N., Voisin, N., Nijssen, B., & Lettenmaier, D. P. (2012). The contribution of glacier melt to streamflow. *Environmental Research Letters*, *7*(3), 034029. https://doi.org/10.1088/1748-9326/7/3/034029

Sharma, S., Ghimire, G. R., & Siddique, R. (2021). Machine learning for postprocessing ensemble streamflow forecasts. *arXiv preprint arXiv:2106.09547*.

Shen, C., Chen, X., & Laloy, E. (2021). Editorial: Broadening the use of machine learning in hydrology. *Frontiers in Water*, *3*, 681023. https://doi.org/10.3389/frwa.2021.681023

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

Thornton, P. E., Shrestha, R., Thornton, M., Kao, S. C., Wei, Y., & Wilson, B. E. (2021). Gridded daily weather data for North America with comprehensive uncertainty quantification. *Scientific Data*, *8*(1), 190. https://doi.org/10.1038/s41597-021-00973-0

Tian, Y., Xu, Y. P., Yang, Z., Wang, G., & Zhu, Q. (2018). Integration of a parsimonious hydrological model with recurrent neural networks for improved streamflow forecasting. *Water*, *10*(11), 1655. https://doi.org/10.3390/w10111655

Wan, Z. (2014). New refinements and validation of the collection-6 MODIS land-surface temperature/emissivity product. *Remote Sensing of Environment*, *140*, 36–45. https://doi.org/10.1016/j.rse.2013.08.027

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., Ray, P., Demaria, E. M. C., Steinschneider, S., & Brown, C. (2017). A user-friendly software package for VIC hydrologic model development. *Environmental Modelling & Software*, *98*, 35–53. https://doi.org/10.1016/j.envsoft.2017.09.006

Wi, S., Yang, Y. C. E., Steinschneider, S., Khalil, A., & Brown, C. M. (2015). Calibration approaches for distributed hydrologic models in poorly gaged basin: Implication for streamflow predictions under climate change. *Hydrology and Earth System Sciences*, *19*(2), 857–876. https://doi.org/10.5194/hess-19-857-2015

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2003.04919

Woodhouse, C. A., & Pederson, G. T. (2018). Investigating runoff efficiency in upper Colorado river streamflow over past centuries. *Water Resources Research*, *54*(1), 286–300. https://doi.org/10.1002/2017WR021663

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., & Shen, C. (2021). Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. *Journal of Hydrology*, *603*, 127043. https://doi.org/10.1016/j.jhydrol.2021.127043

Yin, H., Guo, Z., Zhang, X., Chen, J., & Zhang, Y. (2021). Runoff predictions in ungauged basins using sequence-to-sequence models. *Journal of Hydrology*, *603*, 126975. https://doi.org/10.1016/j.jhydrol.2021.126975

Zhang, H., Zhang, F., Zhang, G., Che, t., & Yan, W. (2018). How accurately can the air temperature lapse rate over the Tibetan Plateau be estimated from MODIS LSTs? *Journal of Geophysical Research: Atmospheres*, *123*(8), 3943–3960. https://doi.org/10.1002/2017JD028243

Zhi, W., Feng, D., Tsai, W. P., Sterle, G., Harpold, A., Shen, C., & Li, L. (2021). From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environmental Science & Technology*, *55*(4), 2357–2368. https://doi.org/10.1021/acs.est.0c06783

Zhu, R., Yang, L., Liu, T., Wen, X., Zhang, L., & Chang, Y. (2019). Hydrological responses to the future climate change in a data scarce region, northwest China: Application of machine learning models. *Water*, *11*(8), 1588. https://doi.org/10.3390/w11081588