AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR032404

Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy

Dapeng Feng[1], **Jiangtao Liu**[1], **Kathryn Lawson**[1], **and Chaopeng Shen**[1]

[1]Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA, USA

Key Points:

- Differentiable (δ) hydrologic models show that process clarity and a performance approaching deep learning (DL) can be both attained
- Unlike DL, δ models can output untrained physical variables, which agreed well with alternative estimates
- Dynamic parameterization has a moderate impact but is needed to narrow the gap to DL, suggesting that current model states are inadequate

Correspondence to:
C. Shen,
cshen@engr.psu.edu

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

Received 24 MAR 2022
Accepted 11 SEP 2022

Author Contributions:
Conceptualization: Chaopeng Shen
Data curation: Jiangtao Liu
Formal analysis: Dapeng Feng
Funding acquisition: Chaopeng Shen
Investigation: Dapeng Feng
Methodology: Dapeng Feng
Project Administration: Chaopeng Shen
Resources: Jiangtao Liu
Software: Dapeng Feng, Jiangtao Liu
Supervision: Chaopeng Shen
Validation: Dapeng Feng
Visualization: Dapeng Feng
Writing – original draft: Dapeng Feng, Chaopeng Shen
Writing – review & editing: Kathryn Lawson, Chaopeng Shen

Abstract Predictions of hydrologic variables across the entire water cycle have significant value for water resources management as well as downstream applications such as ecosystem and water quality modeling. Recently, purely data-driven deep learning models like long short-term memory (LSTM) showed seemingly insurmountable performance in modeling rainfall runoff and other geoscientific variables, yet they cannot predict untrained physical variables and remain challenging to interpret. Here, we show that differentiable, learnable, process-based models (called δ models here) can approach the performance level of LSTM for the intensively observed variable (streamflow) with regionalized parameterization. We use a simple hydrologic model HBV as the backbone and use embedded neural networks, which can only be trained in a differentiable programming framework, to parameterize, enhance, or replace the process-based model's modules. Without using an ensemble or post-processor, δ models can obtain a median Nash-Sutcliffe efficiency of 0.732 for 671 basins across the USA for the Daymet forcing data set, compared to 0.748 from a state-of-the-art LSTM model with the same setup. For another forcing data set, the difference is even smaller: 0.715 versus 0.722. Meanwhile, the resulting learnable process-based models can output a full set of untrained variables, for example, soil and groundwater storage, snowpack, evapotranspiration, and baseflow, and can later be constrained by their observations. Both simulated evapotranspiration and fraction of discharge from baseflow agreed decently with alternative estimates. The general framework can work with models with various process complexity and opens up the path for learning physics from big data.

Plain Language Summary Recently, deep neural networks like long short-term memory (LSTM) have received a lot of attention for producing high-accuracy simulations in hydrology, but they do not respect physical laws and remain difficult to understand. However, what if you can have a model with similar accuracy, but with clarity about physical processes? What if at the same time the model respects physical laws like mass conservation and produces interpretable outputs (like soil moisture, groundwater storage, evapotranspiration and baseflow), with which you can tell a whole story to stakeholders? What if the same framework allows you to ask precise questions about different parts of hydrology and re-examine your understanding of some parts of the physical system or check if your past equations are correct? This paper delivers a system that achieves these grand goals and opens many avenues for further exploration.

1. Introduction

Regional hydrologic models have been widely deployed for operational flood forecasting (Johnson et al., 2019; Maidment, 2017), future change projection (Hagemann et al., 2013), and water resources management (Beck et al., 2020; Guo et al., 2021; Mizukami et al., 2017). They also need to support downstream applications such as crop yield prediction (Ines et al., 2013), pest control (Piou et al., 2019), and water quality modeling (Dick et al., 2016; Strauch et al., 2017). The accuracy of these models has important implications for relevant government agencies and public stakeholders that place trust in them. The demand for accurate modeling capabilities will likely be on the rise due to increased risks of floods and droughts because of climate change (IPCC, 2021).

Traditionally, regional hydrologic models describe not only streamflow but also other water stores in the hydrologic cycle (snow, surface ponding, soil moisture, and groundwater), as well as fluxes (evapotranspiration, surface runoff, subsurface runoff, and baseflow), whereas newer, data-driven machine learning approaches tend to focus on prediction of the variable on which it has been trained. The physical states (stores) and fluxes in traditional models help to provide a full narrative of the event, for example, high antecedent soil moisture or thawing snow primed the watershed for floods, which are important for communication with stakeholders.

They allow us to pose specific scientific questions like *what are the types of floods that have occurred—soil moisture dependent precipitation excess, snowmelt, or rain-on-snow?* (Berghuijs et al., 2016). Some process granularity (meaning the ability to describe the level of details that discern different processes) is also critically important for downstream applications. For example, agricultural models require knowledge of soil moisture and evapotranspiration demand, while water temperature models require a separation between surface runoff and baseflow. Through calibrating the models to streamflow, the hope was that the other fluxes and states would also be constrained via their physical linkages. However, due to the dreaded issue of parameter nonuniqueness or equifinality (Beven, 2006) (where calibrated parameters can take vastly different positions in the parameter space and obtain essentially equivalent calibrated outcomes), calibration sometimes results in distorted physics, that is, multiple studies reported poor results for ET when the model was solely calibrated on streamflow (Rientjes et al., 2013; Tsai et al., 2021).

Recently, purely data-driven deep learning (DL) models (LeCun et al., 2015; Shen, 2018a, 2018b) showed surprisingly strong performance in hydrologic modeling, but they do not resolve internal hydrologic dynamics. These models have very generic internal structures that allow them to learn directly from big data without invoking problem-specific theories and assumptions. A particularly popular architecture in hydrology is long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997). LSTM's accuracy has been demonstrated for many hydrologic variables on large data sets including soil moisture (Fang & Shen, 2020; Fang et al., 2017, 2019; Liu et al., 2022), streamflow (Feng et al., 2020; Konapala et al., 2020; Kratzert, Klotz, Shalev, et al., 2019; Sun et al., 2021; Xiang & Demir, 2020; Xiang et al., 2020), dissolved oxygen (Zhi et al., 2021), groundwater (Solgi et al., 2021; Wunsch et al., 2021), and water temperature (Rahmani, Lawson, et al., 2021; Rahmani, Shen, et al., 2021), covering every part of the hydrologic cycle (Shen et al., 2021). It is widely publicized that LSTM represented a "step-change" in performance which also suggests our traditional models were far from optimality (Nearing et al., 2021). The growth from two LSTM papers in hydrology in 2017 to 300+ papers in 2021 (Shen & Lawson, 2021) demonstrated its appeal and popularity.

Nevertheless, many variables of interest are not adequately observed, so pure DL models do not apply to them. For hydrologic modeling, it remains challenging to interpret what is being learned by LSTM, in part because it does not output physical states or fluxes. While it is possible to correlate LSTM's cell states to some physical states with regression approaches (Lees et al., 2021), the physical meaning of these cell states cannot be guaranteed explicitly, and we cannot run such regression tests before having access to the observations of those physical variables. The regression approach also does not allow us to freely ask questions about how the system functions. The hydrologic community appears to be at a crossroads: seemingly, they cannot have both predictive power and scientific understanding at the same time, yet both are crucial for projecting the impacts of our future climate.

For streamflow prediction, the current best performance with LSTM without using an ensemble and with forcing data from the North American Land Data Assimilation System (NLDAS; different forcing data sets have a moderate influence on results) (Xia et al., 2012) shows a median Nash-Sutcliffe model efficiency coefficient (NSE) of 0.72, reported for the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set (Addor et al., 2017; Newman et al., 2014) with 671 basins across the USA (Feng et al., 2020; Kratzert et al., 2020). In a more interpretable modeling framework, Jiang et al. (2020) showed the possibility to connect a convolutional neural network (CNN) as a post-processor (or correction) layer to a conceptual process-based model encoded as a recurrent neural network. Their model achieved median NSE values of 0.48 and 0.71 without and with the CNN post-processor, respectively, for 591 basins (a subset of the 671 CAMELS basins). While this work is encouraging and is an important step forward, a workflow without the use of a post-processing layer could potentially provide better physical significance and interpretability.

Our earlier work showed that a framework called differentiable parameter learning (dPL) could employ big data and machine learning approaches to find parameter sets for process-based geoscientific models (Tsai et al., 2021). dPL provided similar performance compared to evolutionary algorithms for the main calibration variable, and presented better results for spatial generalization and uncalibrated variables, while also using orders of magnitude lower computational effort than traditional calibration. dPL superseded earlier regionalization schemes in PUB tests. Such strengths are partly because dPL leverages a beneficial data scaling curve where the inclusion of many sites allows them to synergize with each other and make the whole procedure markedly more efficient and well-constrained. dPL is named as such because it employs differentiable programming (Baydin et al., 2018), which tracks the gradients of the outputs of process-based models with respect to input parameters or neural

network weights, for the models' parameterization. Differentiability enables large-scale neural networks to work with process-based models. While dPL gives an initial glimpse at the power of applying differentiable programming to hydrology, it alone does not penetrate model process descriptions, and thus the performance is still limited by the structure of the existing model. In terms of accuracy, dPL can find close to optimal parameters for process-based models, but the end outcomes still lag far behind pure DL models.

Expanding from the advances of dPL, here we demonstrate a new hydrologic modeling paradigm where learnable, differentiable process-based models with embedded neural networks (NN) to provide parameterization or module replacements can achieve similar predictive performance as LSTM models. We call them δ models because they fully leverage differentiable programming so the models are "learnable" and "evolvable." Further, such learnable models (with a physical model as the backbone) respect mass balances, can output important internal physical fluxes, and allow us to ask and answer new scientific questions. We imposed several additional requirements on our framework: (a) each step of the main model calculation either has physical logic associated with physical terms, or uses a neural network but with its effects quantified, (b) mass balances are observed, and (c) multiple internal fluxes and states are described (e.g., groundwater and surface water contributions should be distinguished in our test case). This paper is intended to be a concise report on the methodology, performance, and implications, with myriad research questions set in motion for the future.

In hydrologic modeling, there is a large distinction between locally calibrated models and regionalized models (meaning parameters are related to autocorrelated, widely applicable features, not just calibrated on a gauge) (Hrachowitz et al., 2013). Only regionalized models are applicable to ungauged basins, though they will always have poorer performance than in gauged basins with site-by-site calibration, sometimes substantially (Beck et al., 2020; Hogue et al., 2005; Kumar et al., 2013; Rosero et al., 2010). LSTM can be regionalized when given basin-averaged attributes like soil composition or slope as inputs to distinguish between basins (in some studies, when the LSTM network is trained on a basin-by-basin basis, it essentially is a locally calibrated model). LSTM is also not immune to performance degradation when applied to ungauged basins, but it has been shown to score higher than locally calibrated hydrologic models even when tested out of sample (Feng et al., 2021; Kratzert, Klotz, Herrnegger, et al., 2019). In fact, regionalized LSTM remained competitive even in the scenario of being applied in large, contiguous regions with no data if some ensemble methods were used (Feng et al., 2021). Therefore, while locally calibrated models are certainly useful, in this work we focus only on regionalized frameworks for wider applicability in the future.

2. Methods and Data Sets

As an overview, we use an existing process-based model as a backbone, which can be based on either discrete (modifying from a backbone with discrete formulation) or continuous (written in a differential equation format) formulations. Only the discrete version is demonstrated in this paper. In this example, we selected the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Aghakouchak & Habib, 2010; Beck et al., 2020; Bergström, 1976, 1992; Seibert & Vis, 2012) as the backbone. HBV is a simple bucket-type hydrologic model that simulates hydrologic variables including snow water equivalent, soil water, groundwater storage, evapotranspiration, quick flow, baseflow, and total streamflow. We then altered parts of the backbone model structure, coupled it with differentiable parameter learning (dPL) as a regionalized parameterization scheme, and replaced parts of the model with neural networks. Different from Bennett and Nijssen (2021), the key here is that the whole framework is differentiable, so we can enable "end-to-end" training and no intermediate ground truth data for the output of the neural networks is required. We then trained and tested the model on the CAMELS data set.

2.1. dPL and dPL + HBV for Comparison

Differentiable parameter learning (dPL) only concerns the parameter space and is used to support parameterization of the evolved HBV model, as well as to serve as a comparison case when coupled with the unmodified HBV model. The dPL framework can be described concisely as $\theta = g_A(A,x)$ where θ represents HBV parameters, A contains some static attributes such as topography, soil texture, land cover, and geology, x is the meteorological forcings (Section 2.3.1), and g_A is the parameter estimation neural network. We tried treating parts of θ as being either static or dynamic. If a parameter is treated as static, the same value is used throughout the HBV simulation. If it is treated as dynamic, the model gets a new value for this parameter every day, which we call dynamic parameterization (DP), denoted by super-

Table 1
Major Equations for the Original and Modified Time-Discrete HBV Models

Module	Default HBV equations	Equation modifications
Parameterization	[The parameter set $\{\theta^r, \gamma, \beta\}$ is to be calibrated individually for each basin or regionalized]. θ^r represents parameters other than γ and β.	[We employed regionalized training using dPL] Static parameters: $\{\theta^r, \gamma, \beta\} = g_A(A, x)$ Dynamic parameters: $\{\theta^r, \gamma^t, \beta^t\} = g_A(A, x)$ (or only one of γ and β is dynamic)
Snow (S_p and S_{liq})	$\mathbf{\Delta S_p = P_s + R_{fz} - s_{melt}}$: Solid snowpack P_s: Precipitation as snow $R_{fz} = (\theta_{TT} - T)\theta_{DD}\theta_{rfz}$: Refreeze of ponding water $s_{melt} = (T - \theta_{TT})\theta_{DD}$: Snowmelt $\mathbf{\Delta S_{liq} = s_{melt} - R_{fz} - I_{snow}}$: Liquid in snow $I_{snow} = S_{liq} - \theta_{CWH}{*}S_p$: Snowmelt infiltration	Unchanged
Surface soil water (S_s)	$\mathbf{\Delta S_s = I_{snow} + P_r - P_{eff} - E_x - E_T}$ P_r: Precipitation as rain $P_{eff} = W(P_r + I_{snow})$: Effective rainfall to produce runoff $W = \min((S_s/\theta_{FC})^{\beta}, 1)$: Soil wetness factor; θ_{FC} is maximum soil moisture $E_x = (S_s - \theta_{FC})$: Excess $E_T = \eta{*}E_p$: Actual ET, E_p is potential ET $\eta = \min(S_s/(\theta_{FC}\theta_{LP}),1)$: ET efficiency; θ_{LP} is soil moisture threshold of evaporation	Dynamic parameters: $W = \min((S_s/\theta_{FC})\text{^}\beta^t, 1)$ $\eta = \min((S_s/(\theta_{FC}\theta_{LP}))\text{^}\gamma^t, 1)$ NN replacement option: $P_{eff} = NN_r(\theta_{FC}, \beta, S_s, S_s/\theta_{FC}, P_r + I_{snow})$
Upper subsurface zone (S_{uz})	$\mathbf{\Delta S_{uz} = P_{eff} + E_x - P_{erc} - Q_0 - Q_1}$ $P_{erc} = \min(\theta_{perc}, S_{uz})$: Percolation $Q_0 = \theta_{K0}(S_{uz} - \theta_{uzl})$: Fast flow $Q_1 = \theta_{K1}S_{uz}$: subsurface stormflow	Unchanged
Lower subsurface zone (S_{lz})	$\mathbf{\Delta S_{lz} = P_{erc} - Q_2}$ $Q_2 = \theta_{K2}S_{lz}$: baseflow	Unchanged
Discharge	$\mathbf{Q = Q_0 + Q_1 + Q_1}$	n parallel components: $Q = \sum_{i=1}^{n} f_i\,(Q_{0,i} + Q_{1,i} + Q_{2,i})$
Routing	$Q^*(t) = \int_o^{t\max} \xi\,(s : \theta_a, \theta_\tau) * Q(t - s)ds$ $\xi\,(t : \theta_a, \theta_\tau) = \frac{1}{\Gamma(\theta_a)\theta_\tau^{\theta_a}} t^{\theta_a - 1} e^{-\frac{t}{\theta_\tau}}$	

Note. The main balance equations are bolded and the supporting equations and explanations are indented below them. For simplicity, some minor equations and thresholding functions (for preserving mass positivity) are omitted here. Δ indicates changes in a daily time step; S terms with subscripts indicate state variables; θ terms in the HBV equations represent physical parameters, and θ^r represents physical parameters other than γ and β.

script ^{t}. To support both static and dynamic parameterization, the network g_A is in fact an LSTM model. For either static or dynamic parameterization, the same network g_A is used to parameterize all basins so we can achieve regionalized predictions, but the parameters will be different because the inputs to g_A are different between basins. We further note that the parameters of the physical models themselves are not what was directly tuned during training; we only tuned the weights of the g_A neural networks. A main feature of the dPL framework is that θ is subsequently provided to the HBV model and g_A is trained together with the rest of the model by the global loss function involving all sites. In machine learning language, this framework is called "end-to-end" in that there are no intermediate ground-truth data as supervising data for θ. This avoids two issues: first is that we do not normally have ground-truth data for θ (while some soil parameters may be available, most other parameters like groundwater recession parameters or runoff factors are not); second is that we minimize the chance for error with the intermediate steps.

2.2. The Differentiable, Learnable Process-Based Model

The time-discrete HBV model is described succinctly in Table 1 and illustrated in Figure 1. This model has five state variables. We implemented the time-discrete HBV model on PyTorch, a platform supporting automatic differentiation (Paszke et al., 2017). Other platforms could work similarly.

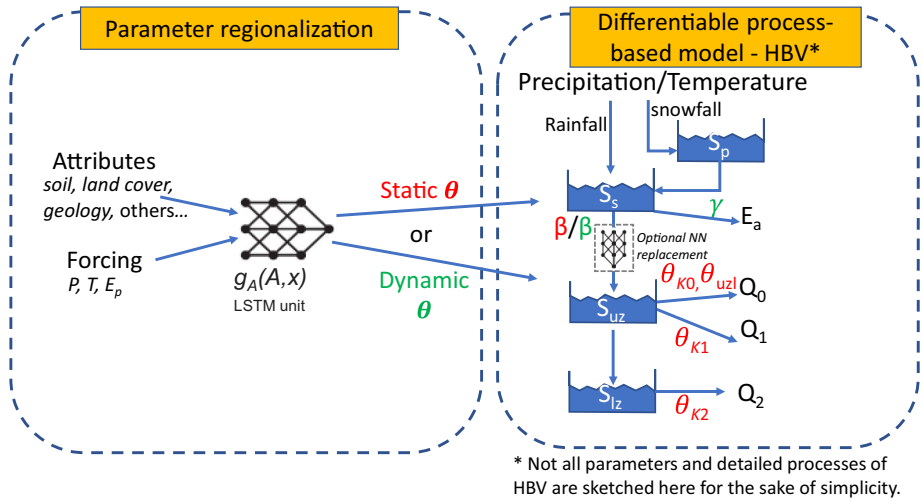

Figure 1. The sketch of the differentiable model framework. A neural network g_A (in this case a long short-term memory [LSTM] unit) outputs the physical parameters for a process-based model (here, the HBV hydrologic model). Each physical parameter can be set as either static or time-dynamic, and either way, there is no feedback from δ models' states to these LSTM-estimated parameters during a forward run. The process-based model supports differentiable programming as it is implemented on a machine learning platform. In the process-based models, certain parts of the model can be replaced by neural networks and the structure can be updated. Here, we also increased the number of storage components to represent sub-basin-scale heterogeneity. Notably, the framework is trained in an "end-to-end" fashion and there are no intermediate supervising data (or labels) for the outputs of the neural networks (either g_A or the optional NN replacements). The main loss function is calculated between the process-based model output and observations—in our example, discharge—and the gradients are backpropagated all the way to these neural networks. Red symbols are static parameters and green symbols are time-dependent parameters.

Here are the explanations and rationales for the modifications:

1. We added a parameter to the ET equation (γ in Table 1's "Equation modifications" column for surface soil water) to allow finer controls on ET efficiency by landscape characteristics and vegetation. Previously, the related parameters (maximum soil moisture and soil moisture threshold for evaporation reduction) were intended to only reflect the impacts of soil. The idea of the power-law formulation came from Lai and Katul (2000). We only use this added γ parameter in the models with dynamic parameterization (DP), as to be described in the next point 2.
2. We tested setting one or both of two parameters as being time dynamic. The first is setting γ to be dynamic (γ^t) to mitigate the errors from the potential ET equation and to mimic the impacts of vegetation. For the latter, the impacts would be season-dependent due to the phenological cycle of vegetation, for example, in the summer, the vegetation roots will be more active; after drought, vegetation may need time to recover its water use efficiency. Because a model like LSTM can automatically capture such effects while the original HBV cannot, HBV may never reach the performance of LSTM without considering seasonally varying parameters. The second parameter that can be set to dynamic is β. This parameter characterizes the relationship between surface soil moisture and effective rainfall (the amount of water available for runoff). The curve roughly characterizes how soil wetness translates into effective rainfall, and β is inversely related to runoff. Having a dynamic β^t can allow the forcing history to influence runoff production.
3. To represent spatial heterogeneity, we conceptualized each basin as being composed of multiple parallel components, each of which can be described by an HBV model. The total fluxes and states, for example, discharge, ET, and water storage, are weighted average fluxes and states of these components. The parameters of all components are simultaneously estimated by g_A. As an early exploration, we assigned a uniform weight to all components, but future efforts can examine relating the weights to soil and vegetation fractions as done in some traditional hydrologic models.
4. We tested replacing the effective rainfall part of the model with a neural network (NN_r) ($P_{eff} = NN_r(\theta_{FC}, \beta, S_s, S_s/\theta_{FC}, P_r + I_{snow})$) to examine if there were more suitable relationships to describe the moisture-runoff relationship than the original one ($P_{eff} = W(P_r + I_{snow})$; $W = \min((S_s/\theta_{FC})^{\beta}, 1)$ in the surface soil water (S_s) row in Table 1). The NN replacement method can be similarly applied to other modules like groundwater and

ET, but here we only applied it to the effective rainfall component to constrain the model's flexibility. Other replacements can be investigated in future work.

5. As in Tsai et al. (2021) and Mizukami et al. (2017), we added a routing module to the hydrologic model with two parameters (Table 1). This module assumes a gamma function for the unit hydrograph and convolves the unit hydrograph with the runoff to produce the final streamflow output, which is compared to the observed streamflow. This routing module is employed in all of our HBV-based simulations in this paper.

We use the symbology $\delta_n(\ )$ to represent different dPL + evolved HBV models for easy reference. Here, n is the number of multiple components. δ_1 represents the single-component (original) HBV model; δ_n represents the evolved HBV model with multiple components; $\delta_n(\beta^t)$ represents the multi-component model with dynamic parameter β; and $\delta_n(\mathrm{NN}_r)$ represents the multi-component model with the soil-runoff relation replaced by a simple neural network (NN_r). Here, only the $\delta_n(\mathrm{NN}_r)$ model employs the optional neural network replacement for the process inside the HBV calculation. Many other backbones other than HBV can be used, but all of the models in this paper used HBV as the backbone. All the δ models used are summarized in Table 2.

2.2.1. Model Training, Hyperparameters, and Other Details

We trained the models on 15 years' worth of data from 1 October 1980 to 30 September 1995 and evaluated the performance on another 15 years' worth of data from 1 October 1995 to 30 September 2010. The 5 years' worth of data from 1 October 1980 to 30 September 1985 were used as validation data to select the hyperparameters. For the hyperparameters of the LSTM unit used as g_A in dPL + HBV (Figure 1), we chose 256 hidden states. A mini-batch size of 100 and a length of 365 days were used to form the training instances of time series in one mini-batch, calculate the loss function, and train the dPL + HBV framework. These two hyperparameters were the same as our previous LSTM-based streamflow model trained on the CAMELS data set (Feng et al., 2020). An alternative is to use an automatic hyperparameter tuning package, but we did not explore that option here. For each training instance, we used additional 1 year of meteorological forcings as the warm-up period for initializing the state variables of HBV during training. For the simulation and performance evaluation on the testing period, we used the forcings of the whole training period to first initialize the state variables. The number of components n was set as 16 for all the multi-component HBV models. In addition to the dPL models, we also ran the purely data-driven LSTM streamflow model with the same training configuration and on the same time periods for performance comparison. The hyperparameters of the LSTM streamflow model were the same as in our previous study (Feng et al., 2020). Additionally, to be comparable with previous regionalized modeling studies (Rakovec et al., 2019), we also ran some dPL experiments on the same training/testing periods and meteorological forcings as theirs.

2.2.2. Loss Function and Evaluation Metrics

Our loss function (Equation 1) was defined as a weighted combination of two parts based on root-mean-square error (RMSE) calculations on all basins (in practice, a mini-batch of basins during training). The first part was the RMSE calculated on the predicted and observed streamflow, while the second part was the RMSE calculated on the transformed streamflow (Equation 2). The transformation in the second part of the loss aims at improving low flow representation. We used a parameter α to assign weights to different parts of the loss function. Different values of α were manually tuned using validation data and generally larger values would lead to better base flow characterization but lower peak flow performance. We chose the value 0.25 to balance the performance on low and peak flow and used it to train all the HBV models. For the benchmark LSTM streamflow model, we applied the transformation (Equation 2) and normalization as preprocessing steps, which were found to give slightly better NSEs in our previous study (Feng et al., 2020). The loss function of the LSTM streamflow model was the RMSE of the normalized data.

$$\mathrm{Loss} = (1.0-\alpha)\sqrt{\frac{\sum_{b=1}^{B}\sum_{t=1}^{T}\left(Q_s^{t,b}-Q_o^{t,b}\right)^2}{B*T}} + \alpha\sqrt{\frac{\sum_{b=1}^{B}\sum_{t=1}^{T}\left(\hat{Q}_s^{t,b}-\hat{Q}_o^{t,b}\right)^2}{B*T}} \tag{1}$$

$$\hat{Q}^{t,b} = \mathrm{Log}_{10}\left(\sqrt{Q^{t,b}+\varepsilon}+0.1\right) \tag{2}$$

Here, $Q^{t,b}$ represents the streamflow on day t and basin b; the subscripts s and o represent simulations and observations, respectively; $\hat{Q}^{t,b}$ is the transformed streamflow in order to improve low flow representations; ε is a small positive value (here, using 1E−6) for stabilizing the gradient calculation of the square root at zero; B and T respectively represent the number of basins (same as the batch size, 100 here) and total days (same as the length

of training instance, 365 here) in a training minibatch; and α is a weighted parameter which is 0.25. For model evaluation, we computed Nash-Sutcliffe model efficiency coefficients (NSE; Nash & Sutcliffe, 1970). NSE is one minus the ratio of the error variance of the modeled time series divided by the variance of the observed time series. It is 1 for a perfect model and 0 for the long-term mean value used as the prediction. The Kling-Gupta model efficiency coefficient (KGE; Gupta et al., 2009) is another popular metric that considers correlation, bias, and flow variability error. KGE, similar to NSE, is 1 for a perfect simulation.

2.3. Input and Observation Data Sets

2.3.1. Forcing, Streamflow, and Attribute Data

We used the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set (Addor et al., 2017; Newman et al., 2014) which contains meteorological forcings, streamflow observations, and attributes for 671 basins in the conterminous United States (CONUS). We tested multiple sets of forcings (one at a time), including Daily Surface Weather Data on a 1-km Grid for North America (Daymet) (Thornton et al., 2020) and the North American Land Data Assimilation System (NLDAS) (Xia et al., 2012) meteorological forcings from CAMELS with the daily minimum and maximum NLDAS temperature data obtained from Kratzert (2019a). This was both to test the suitability of the forcing data sets as well as to see if the differentiable models responded in the same way as LSTM. The HBV model only used three forcing variables due to its formulations: precipitation (P), temperature (T), and potential evapotranspiration (E_p). Describing the total evaporative demand, E_p was estimated using the Hargreaves (1994) method, which considers mean, maximum, and minimum temperatures and latitudes. For g_A, $x = \{P, T, E_p\}$ was used as the dynamic forcings. For the comparison LSTM streamflow model, we used $\{P, T$, solar radiation, vapor pressure, day lengths$\}$ as forcings. The inputs to HBV were predetermined by its structure, while they can include any relevant variable for LSTM. While this seems to give LSTM an advantage as it would have access to more data, the main point is to show how much progress can be made by making HBV differentiable and learnable, and we want the comparison LSTM to represent the highest bar.

We used the streamflow data compiled by the CAMELS data set, which was, in turn, obtained from the streamflow network of the United States Geological Survey (USGS). Static attribute data from CAMELS including 35 topography, climate, land cover, soil, and geology variables in total (Table A1 in Appendix A) were included as inputs to gA in dPL to train regionalized models. We also acquired the simulations of other regionalized models from previous studies as comparisons for our dPL models (Kratzert, Klotz, Shalev, et al., 2019; Rakovec et al., 2019).

2.3.2. Baseflow Index and Evapotranspiration for Evaluation

The baseflow index (BFI^{L13}) was derived from applying Lyne and Hollick filters with warmup periods to streamflow hydrographs by Ladson et al. (2013), which was compiled and included in CAMELS. Moderate Resolution Imaging Spectroradiometer (MODIS), part of the National Aeronautics and Space Administration Earth Observing System (NASA/EOS) project, uses satellite data to estimate terrestrial surface evapotranspiration. The improved ET algorithm of the MOD16A2 data set is based on the Penman-Monteith equation (Monteith, 1965; Mu et al., 2011; Running et al., 2017). The input data include meteorological reanalysis data of daily surface downward solar radiation and air temperature, and MODIS products such as albedo, land cover, leaf area index, and fraction of photosynthetically active radiation. The data cover all basins in CAMELS. We composited the outputs from our models as 8-day data following the same composite methodology as MOD16A2, and compared them.

3. Results and Discussion

We first briefly showcase the surprising performance of the evolved differentiable HBV models and put them in the context of previous literature. Then, we compare simulated internal variables to other estimates. We will mainly use the results forced by Daymet for discussion while also providing NLDAS-forced results for comparison.

3.1. Streamflow Metrics

On the CAMELS data set, strikingly, some of the δ models, namely $\delta_n(\gamma^t, \beta^t)$, $\delta_n(NN_r)$, δ_n, $\delta_n(\beta^t)$, and $\delta_n(\gamma^t)$ in Table 2, achieved median NSE values of >0.71. $\delta_n(\gamma^t, \beta^t)$, in particular, had a median NSE of 0.732, which

Table 2
Model Performances for the Test Period With Daymet Forcing Data and Without the Use of Ensemble

Model name	Median streamflow NSE	Spatial correlation of BFI	Median ET correlation
$\delta_n(\gamma^t, \beta^t)$	0.732	0.760	0.844
$\delta_n(NN_r)$	0.723	0.774	0.817
δ_n	0.714	0.739	0.779
$\delta_n(\beta^t)$	0.729	0.725	0.801
$\delta_n(\gamma^t)$	0.716	0.731	0.842
dPL + HBV (=δ_1)	0.640	0.560	0.770
LSTM (ours)	0.748	-	-
LSTM (Table A1 in Kratzert et al. [2020])	0.74	-	-

Note. δ refers to the differentiable process-based model built on HBV as the backbone; subscript n ($n = 16$ here) indicates multiple components were used. Superscript t (as in γ^t or β^t) indicates where dynamic parameterization is applied (otherwise, parameters are treated as constants).

approached that of LSTM (median NSE = 0.748, which represents the known best value with Daymet forcing and without an ensemble–this result is slightly higher than the corresponding value of 0.74 reported in Kratzert et al. [2020]) (Figure 2 and Table 2). They were both clearly ahead of dPL + HBV (representing the unmodified model with optimized, regionalized parameters, median NSE = 0.640), and MPR + mHM (representing another well-established traditional model with a different regionalization scheme, median NSE = 0.53). Other tests showed similar patterns. For the KGE metric, $\delta_n(\gamma^t, \beta^t)$ and δ_n were both very close to LSTM (Figure 2b). In addition, with the NLDAS forcing, the median NSEs of both LSTM and $\delta_n(\gamma^t, \beta^t)$ declined to 0.719 and 0.711, respectively, which means the gap between them was even smaller with NLDAS (Figure A1 in Appendix A). This suggests Daymet has a better forcing quality, and also suggests LSTM and δ models have been able to extract value from a better forcing dataset.

These strengths in metrics have matching real-world implications as shown by time series comparisons for several example basins (Figure 3). The HBV model underestimated flood peaks for basin A in east Texas, while the

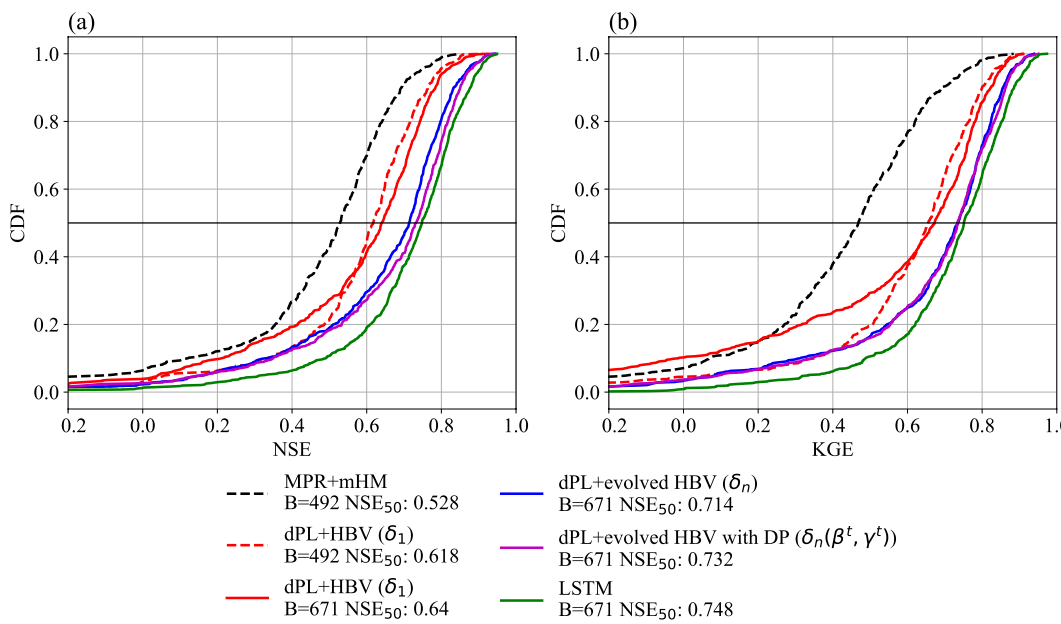

Figure 2. NSE comparison of different models on the CAMELS data set for the testing period. The dashed lines represent models with the training/testing periods of hydrologic years 1999–2008/1989–1999 using Maurer et al. (2002) forcing, while the solid lines represent models with the training/testing periods of 1980–1995/1995–2010 using Daymet forcing. MPR + mHM is from Rakovec et al. (2019). The letter B here represents the number of CAMELS basins used for the evaluation. NSE_{50} is the median NSE value.

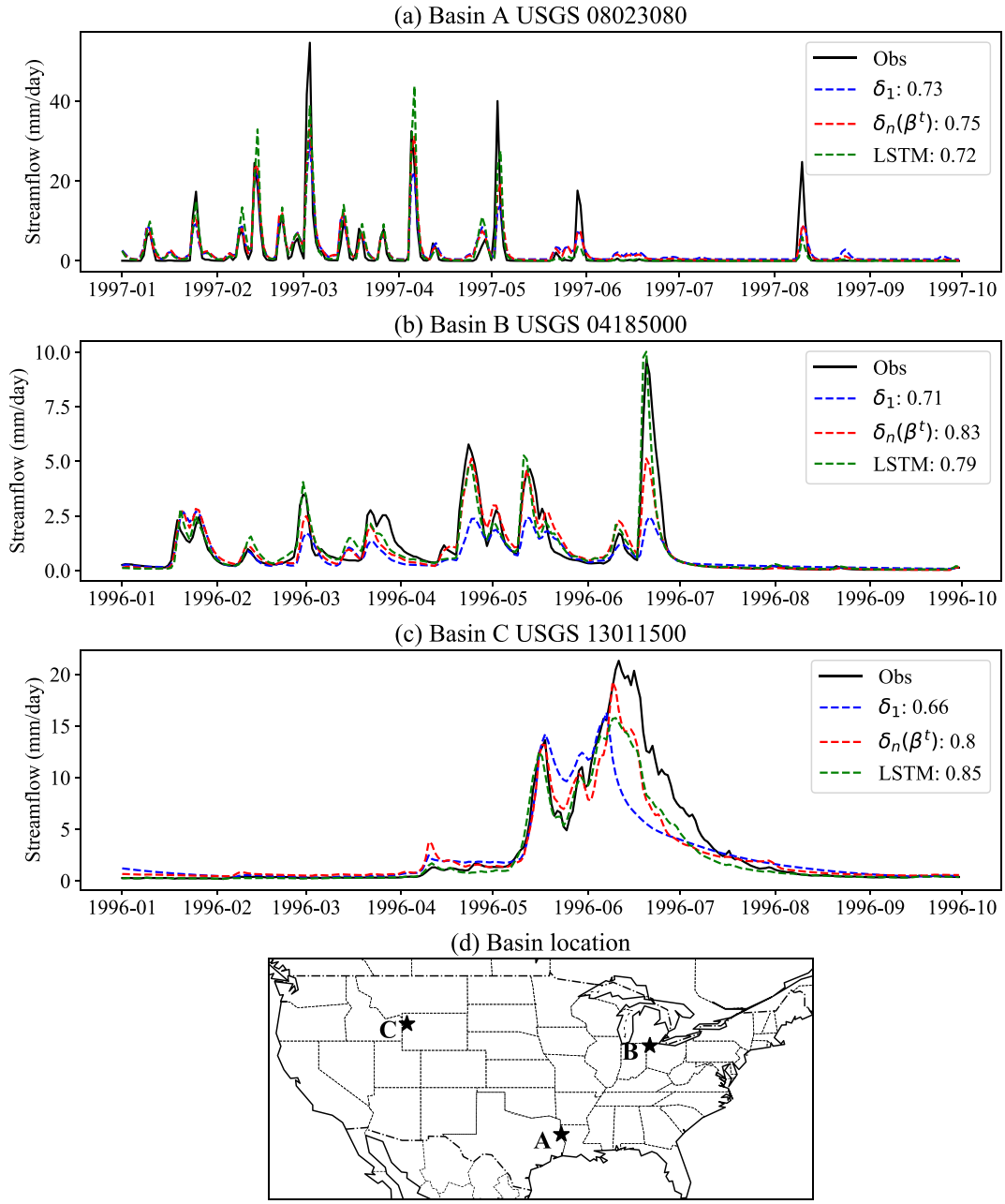

Figure 3. (a–c) Time series comparisons for several basins with NSE values close to the median—dPL+ original HBV (δ_1) versus LSTM versus evolved HBV with DP, $\delta_n(\beta^t)$. The numbers in the legends show the NSE metric for the whole testing period (1995–2010). (d) Locations of basins in (a–c).

evolved structure could moderately mitigate this problem (Figure 3a). For basin B, close to the Great Lakes, the evolved model corrected the peak magnitude from the original HBV model (Figure 3b). We can clearly see improved flow dynamics from basin C located in the Rocky Mountains. The evolved model improved both the peak and the recession limb here (Figure 3c).

The fact that we can approach LSTM performance using a process-based, mass-conservative model is refreshing because in previous studies, the gap between LSTM and process-based models with a wide variety of structures appeared to be uncrossable. These results indicated that it is difficult for expert-derived formulas to match the ability of trained neural networks, a pattern we have seen repeatedly in recent studies across various disciplines. In this work, however, we observe that the gaps in performance between LSTM and learnable models (δ models)

are not so large. This highlights the power of having learnable structures and adaptivity in the model. We note that while the median is close, LSTM is better for the low-performing basins (lower 20th percentile). This could be due to LSTM's ability to correct bias in precipitation, which frequently causes low performance but cannot be fixed by δ models due to mass balance limitations. Future effort with δ models could include specific terms to learn precipitation bias, for example, as a function of mountainous elevation or snowfall (Adam & Lettenmaier, 2003).

Considering the nuances of the configurations and comparing them laterally with the literature, the performance of optimization in our differentiable framework, dPL (median NSE = 0.732) should be close to optimum for this backbone, HBV. First, we have not employed a neural network as a post-processor. Previous results showed that using an LSTM as a post-processor to a hydrologic model produces the same results as LSTM itself (Frame et al., 2021). In this case, post-processors can strongly alter the simulation so that the contributions from the front-end model are no longer clear. Second, as mentioned earlier, previous results with traditional hydrologic models almost always showed regionalized parameterization to be weaker than site-by-site calibration (Beck et al., 2020)—previous best results with regionalized hydrologic models had a median of 0.53 in the CONUS (Rakovec et al., 2019). Lastly, LSTM used more forcing variables in this study, and thus had more information. The result of our regionalized formulation is close to the record-holding LSTM models. Presumably, like LSTM, it also benefits from the efficiency and data synergy of learning from big data (Fang et al., 2022).

Adding parallel units appeared to have a large benefit (Table 2), judging by the differences between the δ_n (n denoting multi-component) and δ_1 (1 denoting single-component) models. With just this change alone, we were able to elevate the median NSE from ~0.640 to 0.714 for δ_n, or 0.723 for $\delta_n(\mathrm{NN_r})$. As described earlier, the point of adding parallel components is to mimic landscape heterogeneity resulting from land cover, slope, and soil combinations in a basin. The results suggest that heterogeneity plays an important role in the basin and must be resolved, although it may not need the explicit overlay of geospatial data sets as commonly done in physics-based hydrologic models. The idea of multiple components is spiritually similar to LSTM's large number of hidden and cell states. However, while LSTM can be modified to have a notion of approximate mass conservation with physical meaning of the stores (Hoedt et al., 2021), in our framework, the stores and fluxes are naturally linked to familiar physical concepts. These results were obtained by applying a uniform weight fraction to all the components, but we hypothesize that using geographically meaningful areal fractions (which would require substantially more data preprocessing effort and is out of the scope of this work) could slightly improve the results.

The models that do not utilize dynamic parameterization (DP) already presented substantial improvements over traditional models, but DP further pulled up the performance. The δ_n and $\delta_n(\mathrm{NN_r})$ models (without DP) were at median NSE values of 0.714 and 0.723, respectively (Table 2), already significantly higher than the regionalized original HBV (median NSE = 0.640) or mHM (median NSE = 0.53, using Maurer forcing). We also ran a case with dPL+ unmodified HBV using Maurer forcing for comparison and obtained the red dashed line with median NSE 0.618 in Figure 2. This shows that DP is not necessarily needed for the model to be operational. However, we should not let go of trying to further narrow the gap to LSTM (adding DP raised median NSE from 0.714 to 0.732) because it is symptomatic of some more fundamental structural issues to be explored below. After various attempts (Table 2 and some undocumented efforts), we were never able to get a median NSE above 0.725 without DP.

The estimated dynamic runoff factor β^t, which is negatively related to runoff, has a dominant seasonal cycle that resembles seasonal water storage, with spikes that sometimes correspond to storms (Figure 4, obtained from the $\delta_n(\beta^t)$ model). β^t starts at the highest level in September, reaches bottom in March or April, and starts to rise in May or June. It corresponds to increasing runoff (other conditions being equal) from October to March, and reducing runoff from May to September, which matches the rhythm of the seasonal water storage cycle. In the original HBV, only the surface layers (S_s and S_{uz}) influence quick runoff, and the deeper groundwater reservoirs have no feedback to surface runoff—that is, quick flow generation is not influenced by how much water there is in deeper groundwater layers in the model. In reality, this feedback can be important, as large water storage in the basin can prime the basin for large runoffs. Hence, an explanation is that the dynamic parameterization captured this problem and introduced a mechanism for high water storage accumulated over months to increase runoff. This is consistent with the observations that β^t in the hot Texas basins (Figures 3a and 4a) has flashier and less seasonal responses—runoff in these basins tend to be flash infiltration excess, and the effect of water storage is limited. β^t for Basin B in the Midwest has a clear seasonal cycle (on a side note, the high-frequency

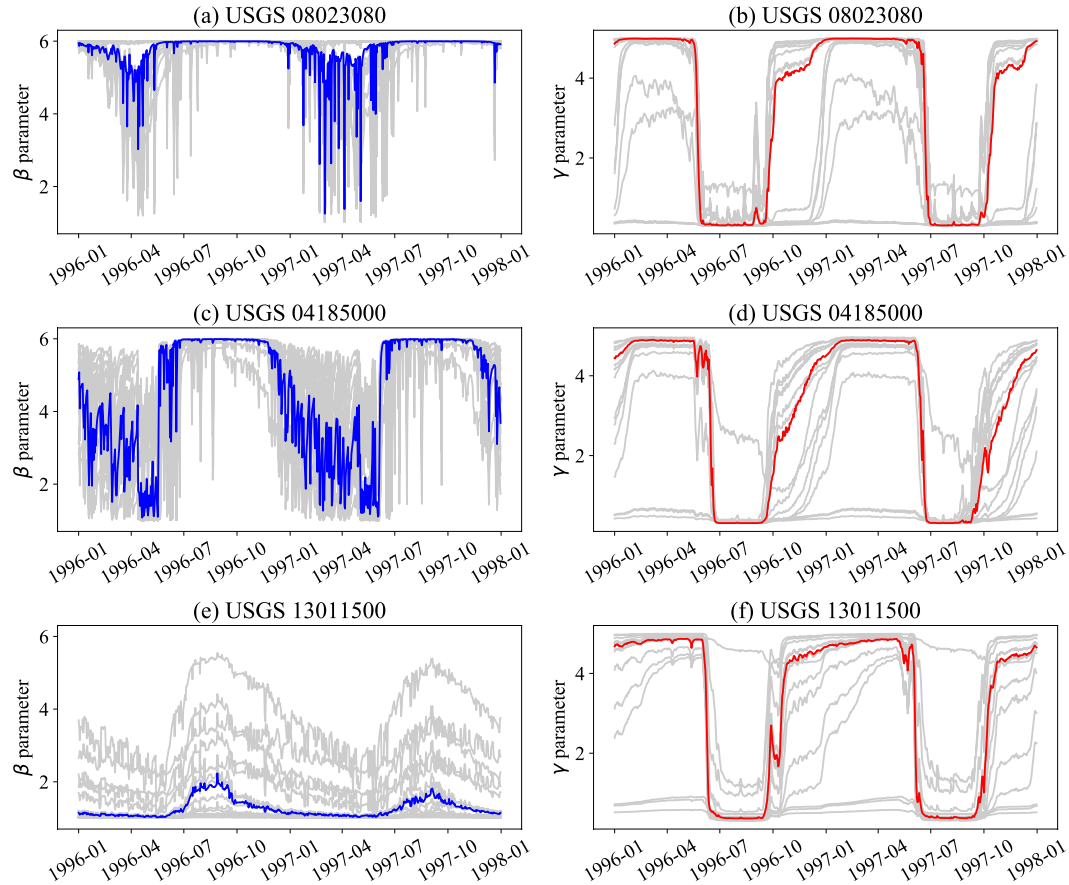

Figure 4. Time series of dynamic parameters β and γ estimated from the models $\delta_n(\beta^t)$ and $\delta_n(\gamma^t)$, respectively, for the three basins shown in Figure 3. The gray lines represent the parameter values from all 16 components while the solid colored (blue or red) lines represent the component which has the median temporal average parameter value among all the components.

fluctuations may be suppressed in the future via a smoothness condition), consistent with the hydrology of the region (Niu et al., 2014; Shen et al., 2013). β^t for Basin C is much drier and has a more muted seasonal cycle. Another concurrent possibility is that β^t also reflects a more established canopy in warm seasons to increase abstraction and reduce runoff.

For γ^t and β^t, their states in DP reflect the impacts of accumulated forcings at the monthly scale. Models like LSTM could implicitly capture such seasonal effects. In contrast, the original HBV, as well as many conceptual models, simply does not have these processes and memory. Other parameters must be distorted for the original model to compensate for their absence. If our theory is true, we expect that no process-based models, learnable or not, can achieve optimal performance without adding these states.

The value of replacing the soil moisture-runoff formula with an NN was moderate: $\delta_n(\mathrm{NN}_r)$ is between δ_n and $\delta_n(\gamma^t, \beta^t)$ in terms of both NSE and ET correlation (Table 2). This means the formulation of the soil moisture-runoff equation in HBV may not be inadequate but the NN may have found a better relationship. This NN also provided an opportunity to study more suitable relationships in the model, which can be pursued in the immediate next stage of the work. We caution that the effect of NN_r may be conditional on the other setup of the model such as forcings (due to their different bias characteristics) and other modules. However, the state-of-the-art performance likely results from the accumulation of many small steps.

3.2. Comparisons of the Baseflow Index and ET

As opposed to the LSTM model which can only predict trained variables like total streamflow, the evolved HBV (δ) models can elucidate the different sources of streamflow and ET. Overall, we found that the baseflow and ET

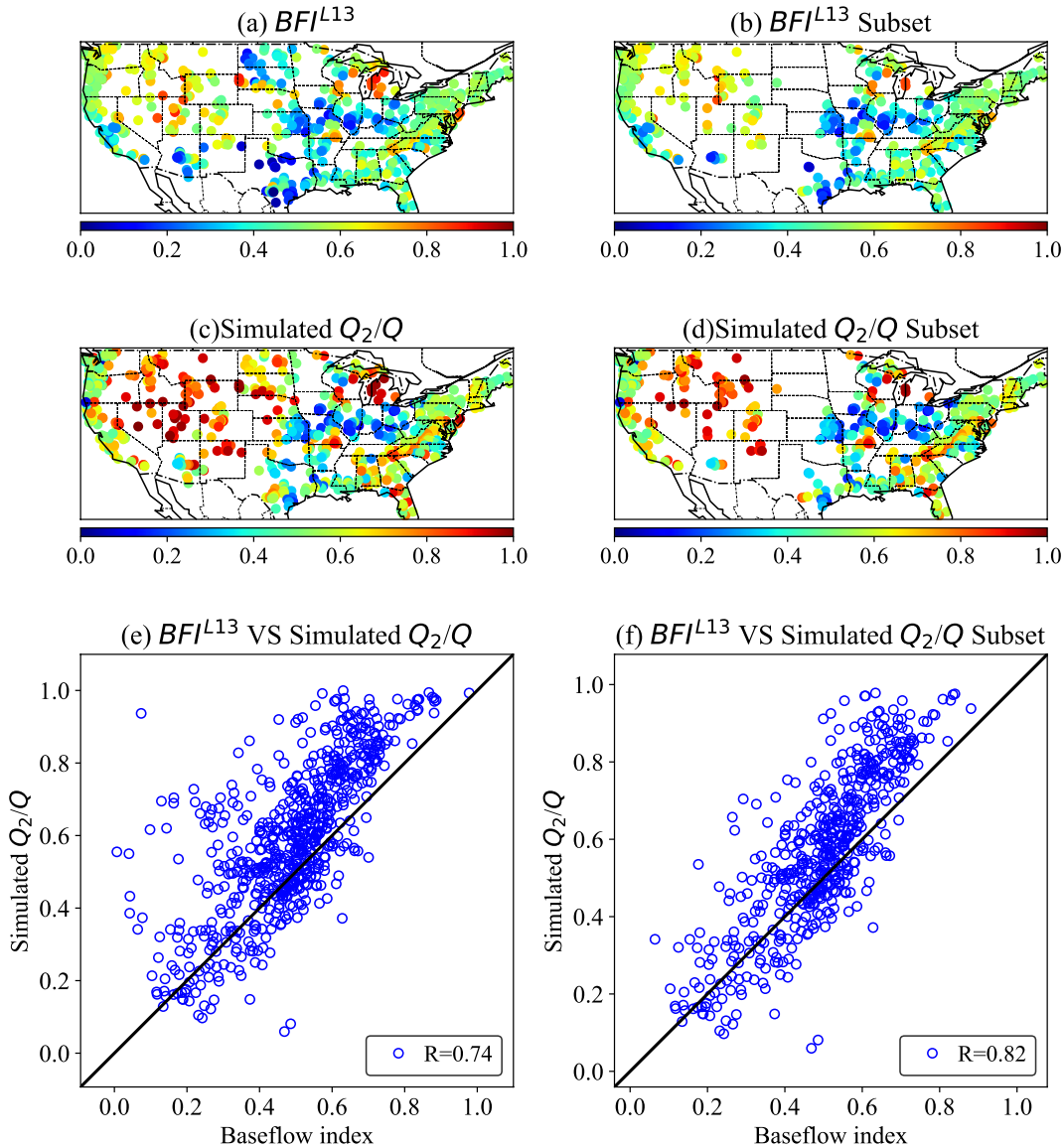

Figure 5. (a) Map of baseflow index (BFI) according to USGS baseflow separation analysis; (c) Baseflow predicted by the trained model (long-term average Q_2/Q from dPL + evolved HBV, δ_n); and (e) the correlation between USGS BFI and Q_2/Q. (b, d, and f) Same as (a, c, and e) but for the subset of basins with NSE > 0.5.

metrics tended to be consistent with NSE—that is, models with better NSE also tended to have better correlations of baseflow and ET (Table 2), although some exceptions did exist. The $\delta_n(\gamma^t, \beta^t)$ model which had the highest median NSE also ranked the second highest in baseflow correlation with BFI^{L13} and the highest in median ET correlation with MODIS. dPL+ unmodified HBV (δ_1) had the lowest streamflow NSE, and at the same time, lowest BFI and ET correlations with alternative estimates.

There is a decent correlation ($R = 0.760$ for $\delta_n(\gamma^t, \beta^t)$ and 0.739 for δ_n) between the simulated BFI (Q_2/Q) and BFI derived from Ladson et al. (2013), BFI^{L13} (Figure 5e), showing the subsurface module in evolved HBV has captured the rough baseflow patterns across the CONUS (Figures 5c and 5e). It should be noted that there is no ground truth in the baseflow index and the results of baseflow recession analysis can vary significantly based on the procedure and assumptions employed and do not represent ground truth. Hence, we place more emphasis on the general spatial pattern rather than the absolute values in the BFI. We notice high BFI associated with thick and permeable soil in the southeast and western US, and in the Upper Colorado River basin.

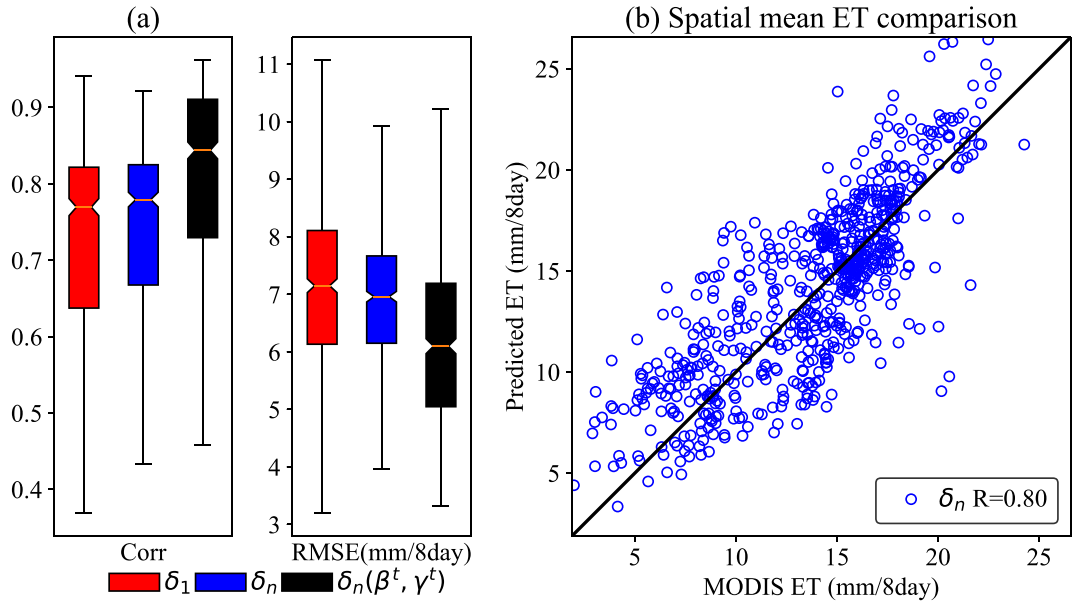

Figure 6. (a) Temporal correlation between simulated ET and the estimates from MOD16A2 8-day product for the CAMELS basins (each box summarizes 671 basins), from three evolved HBV models. (b) Spatial correlation of the long-term mean ET (each circle represents a basin).

We also notice low BFI values along the Appalachian Plateau extending into the Central Plains, in line with previous analysis of terrestrial water storage data and streamflow (Fang & Shen, 2017). The HBV-predicted BFI (Q_2/Q) becomes higher than BFI^{L13} toward the higher ranges. Q_2/Q from basins with NSE values over 0.5 has a higher correlation with BFI^{L13} than for the whole data set (Figure 5f), again shows consistency between streamflow and ET metrics. On a side note, the basins with NSE < 0.5 (the basins present in Figure 5a but missing in Figure 5b) concentrate on the upper Great Plains, which has been known for a long time to cause difficulties for various kinds of models (Feng et al., 2020; Newman et al., 2015). We suspect this is due to incorrect watershed boundaries, highly concentrated rainfall, and the existence of cross-basin water transfer, that is, large-scale groundwater flow and springs.

It is not surprising that the simulated Q_2/Q agrees to a certain extent with the values derived from baseflow separation analysis, which is based on streamflow hydrographs. Passing water through the system and releasing it as Q_2 is the primary way for the model to mathematically introduce slowly varying base flows. However, the point is that we can now directly diagnose different parts of streamflow contributions by learning from data, and further link them to downstream applications such as water temperature predictions.

For δ_1, δ_n, and $\delta_n(\gamma^t, \beta^t)$ models, the median temporal correlation of ET was respectively 0.770, 0.779, and 0.844 (Table 2), and the median RMSE was respectively 7.15, 6.96, and 6.10 mm/8 days against the 8-day integrated MOD16A2 ET product on CAMELS basins. As mentioned above, correlation with MOD16A2 improved from the original HBV to $\delta_n(\gamma^t, \beta^t)$, which was consistent with the improvement of streamflow NSE (Figure 6a, Table 2). These ET metrics appear better than the accuracy levels reported for other estimates evaluated against MOD16A2. To give an example, when evaluated on a monthly scale, Velpuri et al. (2013) reported R^2 of ~0.56 and an RMSE between 26 and 32 mm/month between MOD16A2 and FLUXNET stations. Even without training on ET, the differentiable process-based models can predict reasonable daily ET, which LSTM cannot do at all.

We also note that the model with dynamic parameterization has a higher median correlation with MOD16A2. γ^t had a minor benefit for NSE but a more noticeable benefit on ET, even though the model did not train on ET. The ET correlation of $\delta_n(\gamma^t)$ is better than δ_n, and $\delta_n(\gamma^t, \beta^t)$ is better than $\delta_n(\beta^t)$. This suggests the system calls for memory mechanisms (referring to state variables that hold information and a way to update them) associated with ET. All of these observations suggest ET can be better aided by using dynamic parameterization for γ. Regarding why it is useful to enable dynamic parameterization for γ^t, which is related to ET, we hypothesize

that two factors are at play: (a) the potential ET formula (Hargreaves) could not adequately capture the ET demand but DP (with γ^t) could correct it, and (b) there are some missing memory mechanisms (or states) in the design of the HBV model, in terms of representing seasonally varying ecosystem states like vegetation density, leaf area index, rooting depth, and xylem states (Mackay et al., 2015). These states reflect the impacts of accumulated forcings at the monthly scale. Models like LSTM could implicitly capture such seasonal effects. In contrast, the original HBV model, as well as many other conceptual models, simply does not have a functional vegetation module.

It is difficult to provide a lateral comparison based on daily ET data to the literature regarding the value of learning from streamflow. There is a lack of evaluation of ET on the CAMELS basins in the literature, while the FLUXNET data, which were often used to evaluate ET products, were not coincident with the CAMELS basins to examine the effects of learning from streamflow data. MOD16A2, estimated using a modified Penman-Monteith equation (Mu et al., 2011), should not be interpreted as ground truth. The point here is not that this model produces more accurate ET (which will be studied in future work), but that the differentiable process-based model can now be constrained and evaluated by multi-source, multifaceted observations.

3.3. Further Discussion

In this work, we limited ourselves to frameworks that support prediction in ungauged basins, even though we leave the rigorous PUB benchmark to future work. Our parameterization and learning schemes are regionalized and based on widely available inputs, and are thus applicable to large scales. Due to the lack of physical laws, LSTM may not learn the true causal relationship between static inputs and outputs. Therefore, we hypothesize the decline in performance from training basins to ungauged basins will be less significant for our learnable physical model than LSTM. Our future work will rigorously evaluate this hypothesis.

It is an expert's choice as to which model (each with varying degree of complexity) we use as a backbone, and how much model structure we retain, but that choice impacts what kind of question can be asked. Here, we chose a backbone that discerns soil moisture, groundwater, quickflow, and baseflow such that we can support downstream applications like stream temperature and ecosystem modeling. It is possible to use an even simpler hydrologic model with fewer storage components and less process granularity, but such a model would have limited support for downstream applications. On the flip side, one could use a much more complex model as a backbone, but generally, adding too many structural constraints may degrade model accuracy. The effect of the backbone should be investigated in the future with alternative backbones. In the end, we may have to make a conscious choice between interpretability and accuracy. We recommend backbone models with a process granularity that enables providing a narrative to stakeholders.

Having process granularity and physical fluxes and variables gives us another advantage—we can now use multiple observations to constrain the model or inform unobserved variables. For example, since soil moisture observations are more widely available from satellite missions, we could use these observations, apart from streamflow, to constrain the model and update model states using the DL equivalent of data assimilation. Including additional observations to constrain different parts of the hydrologic model could reduce uncertainty and make the model more robust (Dembélé et al., 2020; Efstratiadis & Koutsoyiannis, 2010). This would not be possible for an alternative model whose backbone does not include soil moisture as an output.

Here, the NNs are embedded into a time-discrete model which explicitly integrates over the time steps. Strictly speaking, the NNs in this framework learn the time-integrated operator which varies based on the time step size. This could be interpreted as using a forward Euler scheme to integrate over time. It is also possible to place the NN on the differential equation and use more accurate numerical solvers for time integration (Rackauckas et al., 2021), which would force the NN to learn the continuous operator. Both approaches are valid, but there may be performance or computational efficiency differences. The discrete version would require little change to the existing models (the code could be translated verbatim), which already have well-understood numerical schemes. The differential equation version would need stable integrator schemes in place, but it would also be a valuable avenue to explore.

While we showed the necessity of employing dynamic parameterization (DP) to reduce the gap to LSTM, we caution against an overuse of DP in the model. For one, many parameters, such as groundwater recession parameters, should not be dynamic. Additionally, because LSTM is so powerful, if we apply DP throughout the system, it has a high chance of achieving high performance, but we may start to lose physical significance and the system may revert to being an LSTM variant. The effects of DP should be quantified, its role justified by physical mechanisms, and its importance should be limited. In our case, DP only had a minor impact (median NSE 0.714 improved to 0.732). In other words, the variance to be explained by DP should not overwhelm the physical part. For future differentiable modeling methods, one important research direction would be ways to limit or constrain the roles of neural networks and retain physical significance.

We think the groundwater component still has significant room for improvement because the current groundwater formulations are too simple. The inter-bucket exchange terms in conceptual hydrologic models are typically unidirectional and do not consider the feedbacks and balances between multiple layers. Future efforts should pay special attention to improving the groundwater component, even though groundwater improvement does not necessarily get reflected in the NSE values.

4. Conclusion

The strong performance of the evolved HBV model shows that, with the aid of learnable units (both for parameterization and process description), the process-based modeling paradigm can be elevated to nearly state-of-the-art performance. The differentiable, learnable modeling we proposed here seeks to reduce the architectural elements of DL while coupling its fundamental elements (backpropagation, regularization, gradient descent, etc.) to domain process descriptions. The success of this framework means we can now use differentiable programming to ask mechanistic questions. The above statements are not to say that pure DL models are not useful—they are an inherent component of the differentiable modeling framework, remain highly valuable because they can be quickly set up to gauge the information content in data sets, and can provide a wealth of diagnostic signals.

We have imposed upon ourselves some stringent conditions (mass conservation, physical calculations, and interpretable physical outputs). The unchanged parts of the numerical model serve as physical constraints, allowing our model to output physical states and fluxes that are valuable for downstream applications. Utilizing a backbone with a sufficient level of process granularity is important to help with this mission. In the near future, the impacts of different backbones and their implications on internal variables need to be investigated. As processes continue to be improved and new backbones are tested, it would not be impossible for differentiable models to completely equal or exceed the performance of LSTM or other deep networks like transformers.

The strengths of the differentiable models and LSTM models over traditional hydrologic models highlight the power of adaptive learning capacity. There seems to be a chasm between the performance of models with and without learnable components, but the gaps are minor between those alternatives that do have them. However, to enable learning complex functions, differentiable programming will be required because traditional optimization capabilities can barely handle more than dozens of parameters. Numerically approximating the derivatives is prohibitively expensive for large neural networks. Thus, we see differentiable programming as an inevitable and promising path toward substantial growth.

Since the evolved HBV model has better results than dPL + HBV by replacing specific parts of the model, it is now easier (compared to using LSTM) to answer questions, for example, *What should have been the moisture-runoff equation if the original one was not good enough*? *How should we parameterize a groundwater recession parameter*? These questions themselves were beyond the scope of this work, but can be answered by replacing the physical descriptions of differentiable HBV with neural networks. Since we showed a well-performing parameterization scheme and that NN replacements for certain modules can be learned, it logically follows that this flexibility allows us to ask novel questions and help answer some of hydrologists' nemesis questions in the future.

Appendix A

Figure A1 provides the benchmark results (same experiments as Figure 2) with NLDAS forcings. Table A1 describes the inputs to the parameter estimation network.

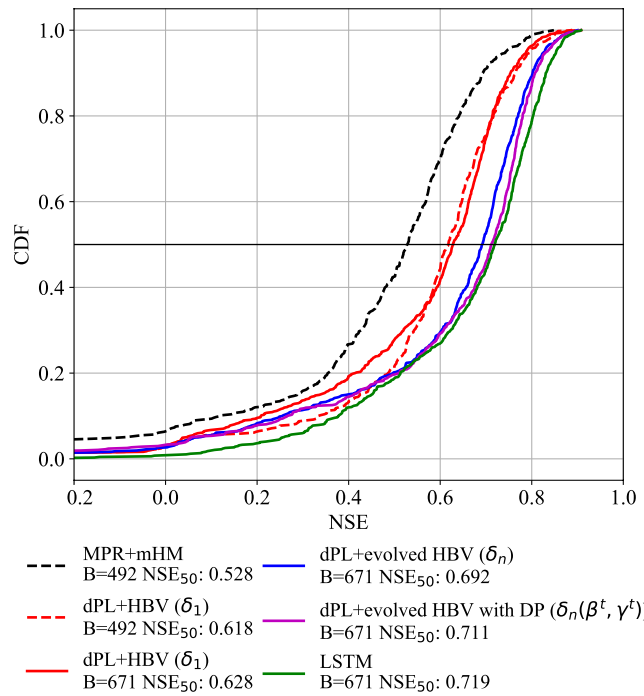

Figure A1. The comparison of NSE for the NLDAS forcing. The items are the same as Figure 2a but the differentiable models and the LSTM were forced with NLDAS. The dashed lines were still run with the Maurer forcing to enable comparison.

Table A1
The Summary of Input Variables to the Parameter Learning Neural Network g_A Including Three Dynamic Forcings and 35 Static Attributes

Variables		Descriptions	Units
Forcing	P	Precipitation	mm/day
	T	Daily mean temperature	°C
	E_p	Potential evapotranspiration	mm/day
Attributes	p_mean	Mean daily precipitation	mm/day
	pet_mean	Mean daily PET	mm/day
	p_seasonality	Seasonality and timing of precipitation	-
	frac_snow	Fraction of precipitation falling as snow	-
	Aridity	PET/P	-
	high_prec_freq	Frequency of high precipitation days	days/year
	high_prec_dur	Average duration of high precipitation events	days
	low_prec_freq	Frequency of dry days	days/year
	low_prec_dur	Average duration of dry periods	days
	elev_mean	Catchment mean elevation	m
	slope_mean	Catchment mean slope	m/km
	area_gages2	Catchment area (GAGESII estimate)	km^2
	frac_forest	Forest fraction	-

Table *A1*
Continued

Variables	Descriptions	Units
lai_max	Maximum monthly mean of the leaf area index	-
lai_diff	Difference between the maximum and minimum monthly mean of the leaf area index	-
gvf_max	Maximum monthly mean of the green vegetation	-
gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction	-
dom_land_cover_frac	Fraction of the catchment area associated with the dominant land cover	-
dom_land_cover	Dominant land cover type	-
root_depth_50	Root depth at 50th percentiles	m
soil_depth_pelletier	Depth to bedrock	m
soil_depth_statgso	Soil depth	m
soil_porosity	Volumetric soil porosity	-
soil_conductivity	Saturated hydraulic conductivity	cm/hr
max_water_content	Maximum water content	m
sand_frac	Sand fraction	%
silt_frac	Silt fraction	%
clay_frac	Clay fraction	%
geol_class_1st	Most common geologic class in the catchment	-
geol_class_1st_frac	Fraction of the catchment area associated with its most common geologic class	-
geol_class_2nd	Second most common geologic class in the catchment	-
geol_class_2nd_frac	Fraction of the catchment area associated with its 2nd most common geologic class	-
carbonate_rocks_frac	Fraction of the catchment area as carbonate sedimentary rocks	-
geol_porosity	Subsurface porosity	-
geol_permeability	Subsurface permeability	m^2

Data Availability Statement

The code for the differentiable model can be downloaded at https://doi.org/10.5281/zenodo.7091334. CAMELS data can be downloaded at https://dx.doi.org/10.5065/D6MW2F4D (Addor et al., 2017; Newman et al., 2014). The extended NLDAS and Maurer forcing data for CAMELS can be downloaded at https://doi.org/10.4211/hs.0a68bfd7ddf642a8be9041d60f40868c (Kratzert, 2019a) and https://doi.org/10.4211/hs.17c896843cf940339c-3c3496d0c1c077 (Kratzert, 2019b). MODIS ET data can be downloaded at https://modis.gsfc.nasa.gov/data/dataprod/mod16.php (Running et al., 2017).

Acknowledgments
DF was supported by US National Science Foundation Awards EAR #1832294 and EAR #2221880. CS was supported by US National Science Foundation Award OAC #1940190. JL and KL were supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under contract DE-SC0016605. Computing was partially supported by US National Science Foundation Award PHY #2018280. We thank the three anonymous reviewers whose comments helped to improve the quality of the manuscript.

References

Adam, J. C., & Lettenmaier, D. P. (2003). Adjustment of global gridded precipitation for systematic bias. *Journal of Geophysical Research*, *108*(D9), 4257. https://doi.org/10.1029/2002JD002499

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). Catchment attributes for large-sample studies [Dataset]. UCAR/NCAR. https://doi.org/10.5065/D6G73C3Q

Aghakouchak, A., & Habib, E. (2010). Application of a conceptual hydrologic model in teaching hydrologic processes. *International Journal of Engineering Education*, *26*(4), 963–973.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018). Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, *18*(153), 1–43.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, *125*(17), e2019JD031485. https://doi.org/10.1029/2019JD031485

Bennett, A., & Nijssen, B. (2021). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models. *Water Resources Research*, *57*(5), e2020WR029328. https://doi.org/10.1029/2020WR029328

Berghuijs, W. R., Woods, R. A., Hutton, C. J., & Sivapalan, M. (2016). Dominant flood generating mechanisms across the United States. *Geophysical Research Letters*, *43*(9), 4382–4390. https://doi.org/10.1002/2016GL068070

Bergström, S. (1976). *Development and application of a conceptual runoff model for Scandinavian catchments* (PhD Thesis). Swedish Meteorological and Hydrological Institute (SMHI). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:smhi:diva-5738

Bergström, S. (1992). *The HBV model—Its structure and applications (RH No. 4; SMHI Reports)*. Swedish Meteorological and Hydrological Institute (SMHI). Retrieved from https://www.smhi.se/en/publications/the-hbv-model-its-structure-and-applications-1.83591

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. *Water Resources Research*, *56*(1), e2019WR026085. https://doi.org/10.1029/2019WR026085

Dick, J. J., Soulsby, C., Birkel, C., Malcolm, I., & Tetzlaff, D. (2016). Continuous dissolved oxygen measurements and modelling metabolism in Peatland Streams. *PLoS One*, *11*(8), e0161363. https://doi.org/10.1371/journal.pone.0161363

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The data synergy effects of time-series deep learning models in hydrology. *Water Resources Research*, *58*(4), e2021WR029583. https://doi.org/10.1029/2021WR029583

Fang, K., Pan, M., & Shen, C. (2019). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 2221–2233. https://doi.org/10.1109/TGRS.2018.2872131

Fang, K., & Shen, C. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. *Water Resources Research*, *53*(9), 8064–8083. https://doi.org/10.1002/2016WR020283

Fang, K., & Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *Journal of Hydrometeorology*, *21*(3), 399–413. https://doi.org/10.1175/JHM-D-19-0169.1

Fang, K., Shen, C., Kifer, D., & Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophysical Research Letters*, *44*(21), 11030–11039. https://doi.org/10.1002/2017GL075619

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9), e2019WR026793. https://doi.org/10.1029/2019WR026793

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophysical Research Letters*, *48*(14), e2021GL092999. https://doi.org/10.1029/2021GL092999

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 model diagnostics. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905. https://doi.org/10.1111/1752-1688.12964

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *WIREs Water*, *8*(1), e1487. https://doi.org/10.1002/wat2.1487

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., et al. (2013). Climate change impact on available water resources obtained using multiple global climate and hydrology models. *Earth System Dynamics*, *4*(1), 129–144. https://doi.org/10.5194/esd-4-129-2013

Hargreaves, G. H. (1994). Defining and using reference evapotranspiration. *Journal of Irrigation and Drainage Engineering*, *120*(6), 1132–1139. https://doi.org/10.1061/(asce)0733-9437(1994)120:6(1132)

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., et al. (2021). MC-LSTM: Mass-conserving LSTM. *ArXiv:2101.05186 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/2101.05186

Hogue, T. S., Bastidas, L., Gupta, H., Sorooshian, S., Mitchell, K., & Emmerich, W. (2005). Evaluation and transferability of the Noah land surface model in semiarid environments. *Journal of Hydrometeorology*, *6*(1), 68–84. https://doi.org/10.1175/JHM-402.1

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Ines, A. V. M., Das, N. N., Hansen, J. W., & Njoku, E. G. (2013). Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. *Remote Sensing of Environment*, *138*, 149–164. https://doi.org/10.1016/j.rse.2013.07.018

IPCC. (2021). In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Eds.), *Climate change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

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*, *47*(13), e2020GL088229. https://doi.org/10.1029/2020GL088229

Johnson, J. M., Munasinghe, D., Eyelade, D., & Cohen, S. (2019). An integrated evaluation of the National Water Model (NWM)–Height Above Nearest Drainage (HAND) flood mapping methodology. *Natural Hazards and Earth System Sciences*, *19*(11), 2405–2420. https://doi.org/10.5194/nhess-19-2405-2019

Konapala, G., Kao, S.-C., Painter, S. L., & 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. (2019a). CAMELS extended NLDAS forcing data [Dataset]. HydroShare. https://doi.org/10.4211/hs.0a68bfd7ddf642a8be9041d60f40868c

Kratzert, F. (2019b). CAMELS Extended Maurer Forcing Data [Dataset]. HydroShare. https://doi.org/10.4211/hs.17c896843cf940339c3c3496d0c1c077

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. (2020). A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff modeling. *Hydrology and Earth System Sciences*, *2020*, 1–26. https://doi.org/10.5194/hess-2020-221

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

Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, *49*(1), 360–379. https://doi.org/10.1029/2012wr012195

Ladson, A. R., Brown, R., Neal, B., & Nathan, R. (2013). A standard approach to baseflow separation using the Lyne and Hollick filter. *Australasian Journal of Water Resources*, *17*(1), 25–34. https://doi.org/10.7158/13241583.2013.11465417

Lai, C.-T., & Katul, G. (2000). The dynamic role of root-water uptake in coupling potential to actual transpiration. *Advances in Water Resources*, *23*(4), 427–439.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539
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
Liu, J., Rahmani, F., Lawson, K., & Shen, C. (2022). A multiscale deep learning model for soil moisture integrating satellite and in situ data. *Geophysical Research Letters*, *49*(7), e2021GL096847. https://doi.org/10.1029/2021GL096847
Mackay, D. S., Roberts, D. E., Ewers, B. E., Sperry, J. S., McDowell, N. G., & Pockman, W. T. (2015). Interdependence of chronic hydraulic dysfunction and canopy processes can improve integrated models of tree response to drought. *Water Resources Research*, *51*(8), 6156–6176. https://doi.org/10.1002/2015WR017244
Maidment, D. R. (2017). Conceptual framework for the national flood interoperability experiment. *JAWRA Journal of the American Water Resources Association*, *53*(2), 245–257. https://doi.org/10.1111/1752-1688.12474
Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., & Nijssen, B. (2002). A long-term hydrologically based dataset 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
Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017WR020401
Monteith, J. L. (1965). Evaporation and environment. *Symposia of the Society for Experimental Biology*, *19*, 205–234.
Mu, Q., Zhao, M., & Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment*, *115*(8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019
Nash, J. E., & Sutcliffe, 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
Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., 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
Newman, A. J., Sampson, K., Clark, M. P., Bock, A., Viger, R. J., & Blodgett, D. (2014). A large-sample watershed-scale hydrometeorological dataset for the contiguous USA [Dataset]. UCAR/NCAR. https://doi.org/10.5065/D6MW2F4D
Niu, J., Shen, C., Li, S.-G., & Phanikumar, M. S. (2014). Quantifying storage changes in regional Great Lakes watersheds using a coupled subsurface-land surface process model and GRACE, MODIS products. *Water Resources Research*, *50*(9), 7359–7377. https://doi.org/10.1002/2014WR015589
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in PyTorch. In *Paper presented at 31st Conference on Neural Information Processing Systems (NIPS 2017)*.
Piou, C., Gay, P.-E., Benahi, A. S., Ebbe, M. A. O. B., Chihrane, J., Ghaout, S., et al. (2019). Soil moisture from remote sensing to forecast desert locust presence. *Journal of Applied Ecology*, *56*(4), 966–975. https://doi.org/10.1111/1365-2664.13323
Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., et al. (2021). Universal differential equations for scientific machine learning. Retrieved from http://arxiv.org/abs/2001.04385
Rahmani, F., Lawson, K., Ouyang, 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*(2), 024025. https://doi.org/10.1088/1748-9326/abd501
Rahmani, F., Shen, C., Oliver, S., Lawson, K., & Appling, A. (2021). Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. *Hydrological Processes*, *35*(11), e14400. https://doi.org/10.1002/hyp.14400
Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., et al. (2019). Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States. *Journal of Geophysical Research: Atmospheres*, *124*(24), 13991–14007. https://doi.org/10.1029/2019JD030767
Rientjes, T. H. M., Muthuwatta, L. P., Bos, M. G., Booij, M. J., & Bhatti, H. A. (2013). Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. *Journal of Hydrology*, *505*, 276–290. https://doi.org/10.1016/j.jhydrol.2013.10.006
Rosero, E., Yang, Z.-L., Wagener, T., Gulden, L. E., Yatheendradas, S., & Niu, G.-Y. (2010). Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season. *Journal of Geophysical Research*, *115*(D3), D03106. https://doi.org/10.1029/2009jd012035
Running, S., Mu, Q., & Zhao, M. (2017). MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006 [Dataset]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD16A2.006
Sahu, R. K., Müller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B., & Agarwal, D. (2020). Impact of input feature selection on groundwater level prediction from a multi-layer perceptron neural network. *Frontiers in Water*, *2*, 573034. https://doi.org/10.3389/frwa.2020.573034
Seibert, J., & Vis, M. J. P. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012
Shen, C. (2018a). Deep learning: A next-generation big-data approach for hydrology. *Eos Transactions American Geophysical Union*, *99*. https://doi.org/10.1029/2018eo095649
Shen, C. (2018b). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643
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
Shen, C., & Lawson, K. (2021). Applications of deep learning in hydrology. In *Deep Learning for the Earth Sciences* (pp. 283–297). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119646181.ch19
Shen, C., Niu, J., & Phanikumar, M. S. (2013). Evaluating controls on coupled hydrologic and vegetation dynamics in a humid continental climate watershed using a subsurface—Land surface processes model. *Water Resources Research*, *49*(5), 2552–2572. https://doi.org/10.1002/wrcr.20189
Solgi, R., Loáiciga, H. A., & Kram, M. (2021). Long short-term memory neural network (LSTM-NN) for aquifer level time series forecasting using in-situ piezometric observations. *Journal of Hydrology*, *601*, 126800. https://doi.org/10.1016/j.jhydrol.2021.126800

Strauch, A. M., MacKenzie, R. A., & Tingley, R. W. (2017). Base flow-driven shifts in tropical stream temperature regimes across a mean annual rainfall gradient: Tropical stream temperature. *Hydrological Processes*, *31*(9), 1678–1689. https://doi.org/10.1002/hyp.11084

Sun, A. Y., Jiang, P., Mudunuru, M. K., & Chen, X. (2021). Explore spatio-temporal learning of large sample hydrology using graph neural networks. *Water Resources Research*, *57*(12), e2021WR030394. https://doi.org/10.1029/2021WR030394

Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E., Kao, S.-C., & Wilson, B. E. (2020). *Daymet: Daily surface weather data on a 1-km grid for North America, version 4*. ORNL DAAC. https://doi.org/10.3334/ORNLDAAC/1840

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., & Verdin, J. P. (2013). A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment*, *139*, 35–49. https://doi.org/10.1016/j.rse.2013.07.013

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, *25*(3), 1671–1687. https://doi.org/10.5194/hess-25-1671-2021

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*(D3), D03109. https://doi.org/10.1029/2011JD016048

Xiang, Z., & Demir, I. (2020). Distributed long-term hourly streamflow predictions using deep learning – A case study for State of Iowa. *Environmental Modelling & Software*, *131*, 104761. https://doi.org/10.1016/j.envsoft.2020.104761

Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resources Research*, *56*(1), e2019WR025326. https://doi.org/10.1029/2019WR025326

Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., et al. (2019). Physics-constrained machine learning of evapotranspiration. *Geophysical Research Letters*, *46*(24), 14496–14507. https://doi.org/10.1029/2019gl085291

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