https://doi.org/10.5194/hess-2022-245
Preprint. Discussion started: 8 August 2022

Hydrology and Earth System Sciences
Open Access
Discussions

The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment

Dapeng Feng[1], Hylke Beck[2], Kathryn Lawson[1] and Chaopeng Shen*,[1]

[1] Civil and Environmental Engineering, The Pennsylvania State University

[2] Joint Research Centre of the European Commission, Ispra, Italy

Correspondence to: Chaopeng Shen (cshen@engr.psu.edu)

Abstract. Differentiable, learnable process-based hydrologic models (abbreviated as δ or delta models) with regionalized parameterization pipelines were recently shown to provide daily streamflow prediction performance that closely approach state-of-the-art long short-term memory (LSTM) deep networks. Meanwhile, δ models provide a full suite of diagnostic physical variables and guaranteed mass conservation. Due to their physical constraints, we hypothesize that they are suitable for making extrapolated predictions. Here, we ran experiments to test (1) their ability to extrapolate to regions far from streamflow gauges; and (2) their ability to make credible projections of long-term (decadal-scale) change trends. We evaluated the models based on daily hydrograph metrics (Nash-Sutcliffe model efficiency coefficient, etc.), as well as projected decadal streamflow trends. The results show that, for spatial interpolation (test in randomly sampled ungauged basins, or PUB), δ models had mixed comparisons with LSTM, presenting better trends for annual mean flow and high flow but slightly worse for low flow. For spatial extrapolation (test in regionally held out basins, or PUR, representing a highly data-scarce scenario), δ models started to surpass LSTM in daily hydrograph metrics, and its advantages in mean and high flow trends became more prominent. In addition, an untrained variable, evapotranspiration, retained good seasonality even for extrapolated cases. δ models' parameterization pipeline produced parameter fields that maintain remarkably stable spatial patterns even in highly data-scarce scenarios, which explains their robustness. Combined with their interpretability and ability to assimilate multi-source observations, δ models are strong candidates for regional and global scale hydrologic simulations for climate change impact assessment.

1. Introduction

Hydrologic models are essential tools to quantify the spatio-temporal dynamics of water resources in both data-dense and data-sparse regions (Hrachowitz et al., 2013). The parameters of hydrologic models are typically calibrated or regionalized for regional applications (Beck et al., 2016), which requires streamflow data, while for global-scale applications models are often uncalibrated (Hattermann et al., 2017; Zaherpour et al., 2018), leading to large predictive uncertainty. Many regions across the world, e.g., parts of South America, Africa, and Asia, suffer from a paucity of publicly available streamflow data (Hannah et al., 2011), which precludes calibration. Yet the water resources in many of these regions face severe pressures due to, among others, population expansion, environmental degradation, climate change (Boretti & Rosa, 2019), and extreme-weather-related disasters, e.g., floods (Ray et al., 2019) and heatwaves in India, and droughts in East Africa. Therefore it is important to better quantify the impacts of these pressures in these regions (Sivapalan, 2003) and estimate the future water cycle.

There has been a surge of interest in deep learning (DL) models such as long short-term memory (LSTM) networks in hydrology due to their high predictive performance, yet DL is not without limitations. LSTMs have made tremendous progress in predicting a wide variety of variables including soil moisture (Fang et al., 2017; Liu et al., 2022; O & Orth, 2021), streamflow (Feng et al., 2020, 2021; Kratzert, Klotz, Herrnegger, et al., 2019), stream temperature (Qiu et al., 2021; Rahmani et al., 2021), and dissolved oxygen (Kim et al., 2021; Zhi et al., 2021), among others (Shen, 2018; Shen & Lawson, 2021). DL is able to harness the synergy between data points and thus thrives in a big data environment (Fang et al., 2022; Kratzert, Klotz, Herrnegger, et al., 2019; Tsai et al., 2021). However, DL models are still difficult to interpret and do not predict variables without extensive observations. In addition, it is challenging to answer specific scientific questions using DL models, e.g., "*what is the relationship between variable soil moisture and runoff?"*, as LSTM's internal relationships may not be straightforwardly interpretable by humans.

Large-scale predictions for ungauged basins (PUB) (Figure 1 left) or ungauged regions (PUR) (Figure 1 right) challenge the ability of a model and its parameterization schemes to generalize in space. For both kinds of tests, regionalized LSTM models hold the performance record (Feng et al., 2021; Kratzert, Klotz, Herrnegger, et al., 2019). While no clear definition has been universally given for PUB, these PUB tests are typically conducted by randomly holding out basins for testing. As such, PUB can be considered spatial "*interpolation*", as there will always be training gauges surrounding the test basins. While LSTM's performance declines from temporal to PUB tests, it obtains better results than established process-based models calibrated on the test basins (Feng et al., 2021; Kratzert, Klotz, Herrnegger, et al., 2019). However, it is uncertain if process-based models' poorer performance is simply due to structural deficiencies and if they would experience similar declines for PUB. Stepping up in difficulty, prediction for ungauged regions (PUR) refers to tests where a large region is entirely held out for testing. As such, PUR better represents the case of spatial "*extrapolation*" encountered in global hydrologic assessment (Feng et al., 2021). For PUR, LSTM's performance further declines significantly (Feng et al., 2021). No systematic PUR tests have been done for process-based models, however, perhaps because there has been a serious underappreciation of the difference between PUB and PUR and the risk of model failures due to large data gaps.

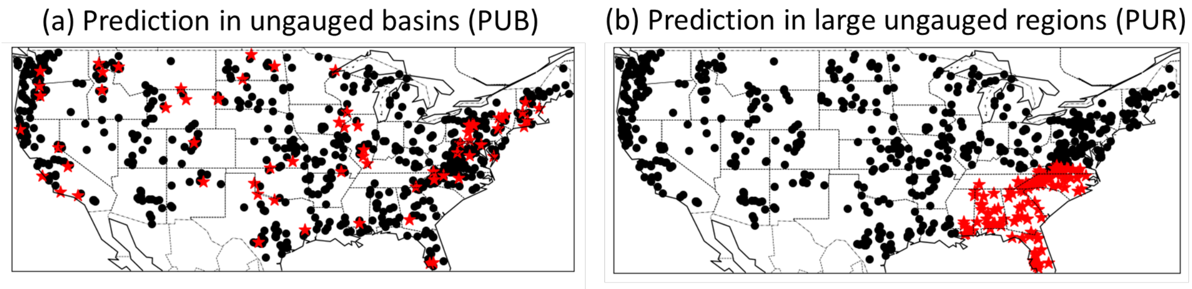

(a) Prediction in ungauged basins (PUB)
(b) Prediction in large ungauged regions (PUR)

Figure 1. A comparison of spatial generalization tests: (left) prediction in ungauged basin (PUB) and (right) prediction in ungauged region (PUR) tests. The black dots are the training basins while the red stars are the test basins for one fold. In the study we ran cross validation to obtain the spatial out-of-sample predictions for basins in the CAMELS dataset.

Recently, a new class of models adopting differentiable programming (the computing paradigm where the gradient of each operation is tracked) (Baydin et al., 2018) has shown great promise (Innes et al., 2019; Tsai et al., 2021). Regardless of the computational platforms chosen for them, differentiable models mix physical process descriptions with neural networks (NNs), which serve as learnable elements for parts of the model pipeline. The paradigm supports backpropagation and neural-network-style end-to-end training on big data so no ground-truth data is required for the direct outputs of the neural network. The first demonstration in geosciences was a method we called differentiable parameter learning (dPL), which uses NNs to provide parameterization to process-based models (or their differentiable surrogate models) (Tsai et al., 2021). Not only did the work propose a novel large-scale parameterization paradigm, it further uncovered the benefits of big data: we gain stronger optimization results, acquire parameters which are more spatially generalizable and physically coherent (in terms of uncalibrated variables), and save orders of magnitude in computational power. Only a framework that can assimilate big data, such as a differentiable one, could fully leverage these benefits. However, dPL is still limited by the presence of flawed structures in most existing process-based models, and some performance degradation is further introduced when a surrogate model is used. As a result, with a LSTM-based surrogate for the VIC hydrologic model, dPL's performance is still significantly lower than that of LSTM. One valuable avenue to boost performance is to append neural networks as a postprocessor to the physics-based model (Frame et al., 2021; Jiang et al., 2020), but this is not the path we choose here.

Strikingly, differentiable models can be elevated to approach the performance level of state-of-the-art LSTM models with postprocessors (Feng et al., 2022). We obtained a set of differentiable, learnable process-based models, which we call ***δ models***, by updating model structures based on the conceptual hydrologic model HBV. For the same CAMELS benchmark, we obtained a median NSE of 0.715 for the NLDAS forcing data, which is already very similar to LSTM (0.720). Furthermore, we can now output diagnostic physical fluxes and states such as baseflow, evapotranspiration, water storage, and soil moisture. Differentiable models can thus trade a small amount of performance metric for the full suite of physical variables, process clarity, and the possibility to learn science from data.

There are two perspectives with which we can view δ models: they can be regarded as deep networks whose learnable functional space is restricted to the subspace permitted by the process-based backbone; or they can be viewed as process-based models with learnable and adaptable components provided by NNs. The flow of information from inputs to outputs is regulated. For example, in the setup in Feng et al. (2022), the parameterization network can only influence the groundwater flow process via influencing the parameters (but not the flux calculation itself). It does not allow information mixing at all calculation steps (as opposed to LSTM, in which most steps are full matrix multiplications that mix information between different channels). For another example, because mass balance is observed, a parameter leading to larger annual mean evapotranspiration will necessarily reduce long-term streamflow output. Mass balance is the primary connective tissue between different hydrologic stores and fluxes. This important constraint can lead to tradeoffs between processes if there are errors with inputs like precipitation, but it imposes a stronger constraint on the overall behavior of the model. Nevertheless, Feng et al. (2022) was conducted only for temporal tests (training on some basins and testing on those same basins but for a different time period) but not for PUB or PUR, which may show a different picture. For these new types of models, the generalizability of these models under varied data density scenarios is highly uncertain. Before we use those models for the purpose of learning knowledge, we seek to understand their ability to generalize.

Our main research question in this paper is whether differentiable process-based models can generalize well in space and provide reliable large-scale hydrologic estimates in data-scarce regions. Our hypothesis is that, since the differentiable models have stronger structural constraints, they should exhibit some advantages in extrapolation, both in space and time, compared to LSTM and existing process-based models. An implicit hypothesis is that the relationships learned by the parameterization component are general, so they can be transferred to untrained regions. If these hypotheses are true, it would make this category of models appropriate for global hydrologic modeling, which is desirable considering they can also provide a full narrative of the hydrologic processes, fluxes and states. Since δ models have similar performance to LSTM in temporal tests, they represent a chance to truly test the value of model structures and the impact of extrapolation. In this paper, we designed both PUB and PUR experiments. Furthermore, apart from typical metrics calculated on the daily hydrographs, we also evaluated the simulated trends of mean annual flow and different flow regimes, which are critical aspects for climate change impact assessments but have not been adequately assessed before.

2. Data and Methods

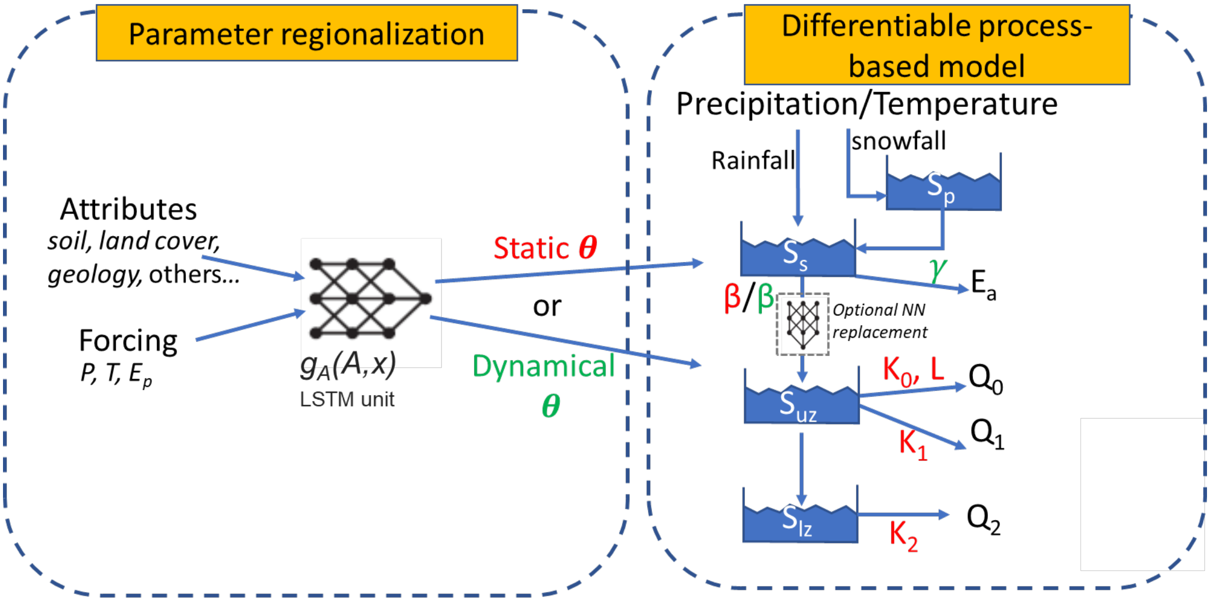
2.1. Differentiable models

As an overview, a differentiable model implements a process-based model as an evolvable backbone on a differentiable computing platform such as PyTorch, Tensorflow, or Julia, and uses intermingled neural networks (NNs) to provide parameterization (meaning a way to infer parameters for the model using raw information) or process enhancement. In our setup, the parameterization and processes are learned from all the available data using a whole-domain loss function, therefore supporting regionalized PUB applications and even out-of-training-region (PUR) applications.

For the process-based backbone, we employed the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Aghakouchak & Habib, 2010; Beck, Pan, et al., 2020; Bergström, 1976, 1992; Seibert & Vis, 2012), a simple, bucket-type conceptual hydrologic model. HBV has state variables like snow storage, soil water, and subsurface storage, and can simulate flux variables like evapotranspiration (ET), recharge, surface runoff, shallow subsurface flow, and groundwater flow. The parameters of HBV are learned from basin characteristics by a DL network just as in dPL (Figure 2). Here, we made two changes to the HBV structure. The first modification was to increase the number of parallel storage components of the HBV model (16 used here), to represent the heterogeneity within basins. The state and flux variables were calculated as the average of different components, and the parameters of all these components are learned from the neural network g_A. The second modifcation was that, for some tested versions of the model, we turned some static parameters of HBV into time-dependent parameters with a different value for each day (we call this dynamic parameterization, or DP). For example, we set the runoff curve shape coefficient parameter to be time-dependent (β^t) as explained in Appendix B. The dynamic parameters are also learned by the neural network g_A, from basin characteristics and climate forcings. More details about differentiable models can be found in our previous study (Feng et al., 2022).

2.2. The comparison models

We compared the performance of δ models with a pure LSTM streamflow model for spatially out-of-sample predictions. The regionalized LSTM model was based on Feng et al. (2020), taking meteorological forcings and basin attributes (detailed below) as inputs. The hyperparameters of both LSTM and δ models were manually tuned in the previous studies and retained in this study. The loss function was calculated as root-mean-square error (RMSE) for a minibatch of basins with a one year look-back period, but across many iterations the training will go through the entire training dataset. Same as Feng et al., (2022), the RMSE was calculated on both the unnormalized predictions and transformed predictions to improve low flow representation and a loss with weighted combination of two parts are used for the dPL models, while the RMSE was calculated on the normalized predictions for the LSTM model since the transformation to represent low flow has been applied in the data preprocessing. Each training instance had two years' worth of meteorological forcings, but the first year was used as a warmup period so the loss was only calculated on the subsequent one year of simulation. We also used streamflow simulations from the multiscale parameter regionalization (MPR) scheme applied to the mHM hydrologic model (Rakovec et al., 2019) to represent a traditional regionalized hydrologic model, but only the temporal test is available for this model.



* Not all parameters and detailed processes of HBV sketched here for the sake of simplicity.

Figure 2. The flow diagram of δ models with HBV as the backbone. An LSTM unit estimates the parameters for the differentiable HBV, which has snow, evapotranspiration, surface runoff, shallow subsurface, and deep groundwater reservoirs. Outflows are released from different compartments with a linear formula with proportionality parameters (K's). g_A is the parameterization network with dynamic input x and static input attributes A. The buckets represent storage mass storage states (S's); θ refers to all HBV parameters. The model δ has static parameters while γ and β are two of the parameters. δ(β^t,γ^t) sets γ and β as time-dependent parameters, with a new value each day. Importantly, there are no intermediate target variables to supervise the neural networks -- the whole framework is trained on streamflow as the only loss, in an end-to-end fashion. For simplicity, we did not use the optional NN replacement in this study, but the high performance is retained. Abbreviations: P -- precipitation; T -- temperature; E_p -- potential evapotranspiration; Q_0 -- quick flow; Q_1 -- shallow subsurface flow; Q_2 -- baseflow; E_a -- actual evapotranspiration; S_p -- snowpack water storage; S_s -- soil water storage; S_{uz} -- upper subsurface zone water storage; S_{lz} -- lower subsurface zone water storage; L -- upper subsurface threshold for quick flow.

2.3. Data

We used the CAMELS dataset (Addor et al., 2017a, 2017b) which includes 671 basins across the contiguous United States (CONUS) to run the experiments. The Maurer et al. (2002) forcing was selected from the three forcings available in CAMELS. To train regionalized models for dPL and LSTM, we used 35 attributes as shown in Table A1 in the Appendix A. For the LSTM streamflow model, the attribute data were directly concatenated with the forcings and provided as inputs. With the δ models, the neural network g_A takes attributes and historical forcing data as inputs, and outputs parameters for the evolved HBV model. The LSTM model takes 5 forcing variables including precipitation, temperature, solar radiation, vapor pressure, and day length, while the HBV model only takes precipitation (P), temperature (T), and potential evapotranspiration (E_p). We used the temperature-based Hargreaves (1994) method to calculate E_p and the daily Maurer minimum and maximum temperature for CAMELS basins were acquired from Kratzert et al., (2019). The training target for all the models was streamflow observations. We trained all models on all 671 basins in CAMELS and reported the test performance on a widely

https://doi.org/10.5194/hess-2022-245
Preprint. Discussion started: 8 August 2022

used 531-basin subset, which excludes some basins due to unclear watershed boundaries. The results of some previous regionalized modeling efforts are also used to provide benchmark context (Kratzert, Klotz, Shalev, et al., 2019; Rakovec et al., 2019). For the comparison of evapotranspiration, we used a product derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite (Mu et al., 2013).

2.4. PUB and PUR experiments

As mentioned earlier, we designed two sets of experiments to benchmark the models: predictions in ungauged basins (PUB) and predictions in large ungauged regions (PUR) (illustrated in Figure 1). For PUB experiments, we randomly divided the whole CAMELS basins into 10 groups, trained the models on 9 groups, and tested it on the one group held out. By running this experiment for 10 rounds, we can get the out-of-sample PUB result for all basins. For the PUR experiment, we divided the whole CONUS into 7 continuous regions (as shown in Figure A1 in Appendix B), trained the model on 6 regions, and tested it on the holdout region. We ran the experiment 7 times so that each region could serve as the test region once. The study period was from October 1, 1989, to September 30, 1999. These spatial generalization tests were done in the same time period as the training samples (but for different basins).

From the daily hydrograph, we calculated the Nash-Sutcliffe (NSE) (Nash & Sutcliffe, 1970) and Kling-Gupta (KGE) (Gupta et al., 2009) model efficiency coefficients as performance metrics. NSE characterizes the variance in the observations explained by the simulation and KGE accounts for correlation, variability bias, and mean bias. We also reported the percent bias of the top 2% peak flow range (FHV) and the percent bias of the bottom 30% low flow range (FLV) (Yilmaz et al., 2008), which characterizes peak flows and baseflow, respectively.

We also evaluated the multi-year trend for streamflow values at different percentiles (Q_{98}, Q_{50}, Q_{10}) as well as the mean annual flow. Q_{98}, Q_{50}, and Q_{10} represent the peak flow, median flow, and low flow value, respectively. To this end, we calculated for each year, one data point corresponding to a flow percentile. Then, Sen's slope estimator (Sen, 1968) for the trend of that flow percentile was calculated for the 10 years in the test period and compared with the equivalent slope for the observations. Since streamflow records contain missing values, we only considered years with <61(about two months) daily missing values (not necessarily consecutive) for this purpose.

3. Results and Discussion

In this section, we first compared LSTM and the differentiable models (and, when available, the traditional regionalized model) for PUB and PUR, in terms of both the daily hydrograph metrics (NSE, KGE, FLV, and FHV) and decadal-scale trends. We then attempted to examine why δ models had robust performance and how well they could predict untrained variables (evapotranspiration). We use "δ models" to generically refer to the whole class of differentiable models with evolved HBV, while we use δ, δ(β^t) or δ(β^t,γ^t) to refer to particular models with static, one-parameter dynamic and two-parameter dynamic parameterization, respectively. The meanings of β and γ are described in the Appendix B.

3.1. The randomized PUB test

For the randomized PUB test, which represents a data-dense scenario (Figure 1a), the δ models approached the performance of the LSTM in terms of median NSE and KGE but with less spread in NSE and KGE, indicating robustness in the predictions. The δ(β^t,γ^t) model's PUB test produced median NSEs of 0.64 (Figure 3), only slightly below the LSTM median NSE (0.66) and considerably higher than the MPR+mHM NSE (0.53, in sample -- all sites included in training), similar to our previous temporal tests (Feng et al., 2022). For KGE, δ(β^t) and δ(β^t,γ^t) model had a median of 0.66 and 0.65, respectively, which were the same as LSTM, but also with a smaller spread. It is worthwhile to note, however, this performance is for a PUB test with more holdout data (lower k fold) and less computation, which degrades the performance compared to the higher metrics we reported earlier (Feng et al., 2021). LSTM had lower errors for FLV and FHV than the δ models, which is likely because LSTM is not subject to physical constraints and therefore possesses more flexibility in terms of base and peak flow generation than HBV. LSTM does not obey mass balances and may potentially learn biases in precipitation (Beck, Wood, et al., 2020) and other forcing terms and make internal corrections for them. Such biases could cause issues for models honoring mass balances, but the impact of precipitation bias is under debate (Frame et al., 2022). Overall, LSTM represents a high benchmark and the similar performance and smaller spread of the δ models are highly encouraging.

In terms of the projection of future trends, δ models again demonstrated high competitiveness, showing mixed comparisons to LSTM (Figure 4). Both LSTM and δ models accurately captured the trends in annual mean flow and high-flow bands (R^2>0.80), but both struggled somewhat with low flow Q_{10} (trend evaluated in the annual 10th-percentile flow, R^2<0.40). δ(β^t,γ^t) superseded LSTM in terms of annual mean flow and 98 percentile peak flow, while LSTM had a small advantage for Q_{50} (trend evaluated in the annual median flow) and Q_{10}. Overall, just as with LSTM, δ models seem appropriate for long-term trend predictions in the data-dense PUB scenario. They even have advantages over LSTM with respect to assessing future risks of flooding.

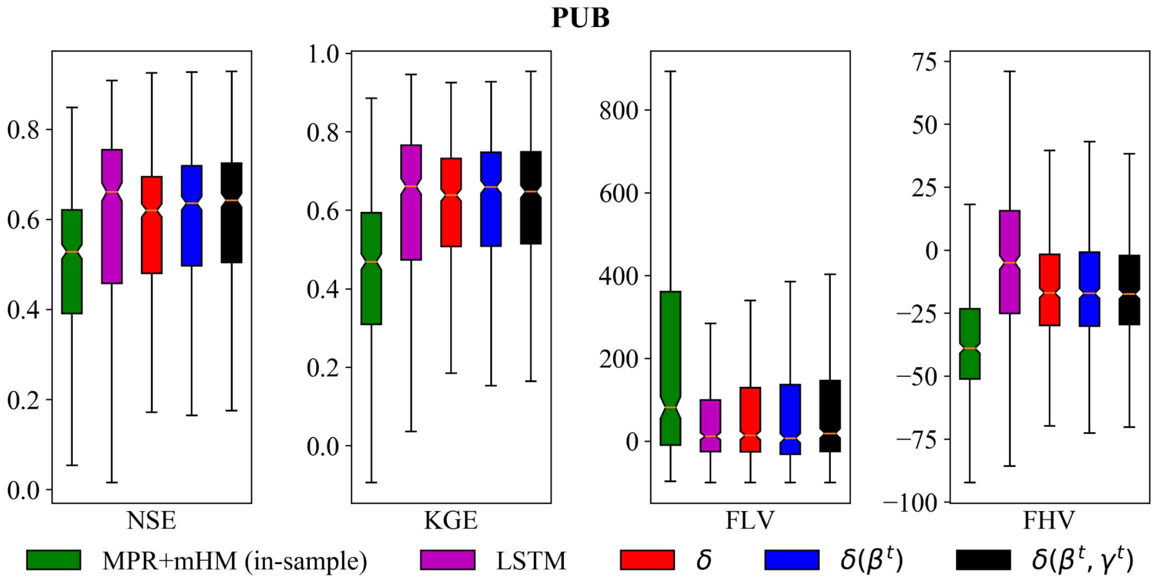


Figure 3. Performance of simulated daily hydrographs from the models for the randomized PUB experiment. Each box summarizes 531 values (one for each CAMELS basin) obtained in a cross-validation manner. All models except MPR+mHM (noted "in-sample", which means all sites are included in the training set) were evaluated out-of-sample spatially, i.e. they were trained on some basins and tested on other holdout basins. For MPR+mHM (Rakovec et al., 2019), all test basins were included in the training dataset. NSE is the Nash Sutcliffe model efficiency coefficient, KGE is the Kling Gupta efficiency, FLV is the low flow bias, and FHV is the high flow bias. δ and δ(β^t,γ^t) (or δ(β^t)) are respectively the differentiable, learnable HBV model without and with dynamical parameters. The horizontal line in each box represents the median and the bottom and top of the box represent the first and third quantiles, respectively, while the whiskers represent the minimum and maximum values, respectively. The PUB was run in a less computationally-expensive training experiment to be comparable to other models and also to reduce computational demand: we used only 10 years of training period, did not use an ensemble, and used a lower k-fold. When we ran the experiments using the same setting as Kratzert et al. (2019), our LSTM was able to match the PUB performance in their work (Feng et al., 2021).

The challenge with low flow projection for all models is probably attributable to multiple factors: (i) a lack of reliable information on subsurface hydraulic properties which hampers all models; (ii) the inherent challenge with baseflow trends --- the magnitude of the Q_{10} change trends is in the range of -0.5 to 1 *m^3/s/year* while that for the annual mean flow is -2 to 10 *m^3/s/year*. Even a small error in absolute terms can result in a large decrease in R^2; (iii) inadequacy of the low-flow modules --- the linear reservoir formulation in the present HBV groundwater modules may not capture the real-world dynamics, while even LSTM may not have the memory that is long enough to represent a gentle multi-year baseflow trend change; and (iv) the greater impact of human activities such as reservoir operations on low flow (Döll et al., 2009; Suen & Eheart, 2006). (v) the greater sensitivity of the training loss function to high flows than low flows due to the difference in their magnitudes. High flows are direct reflections of recent precipitation events in the basin while low flows are under large impacts of the geological system.

https://doi.org/10.5194/hess-2022-245
Preprint. Discussion started: 8 August 2022

For completeness, we also evaluated the trend for the temporal tests (trained and tested on the same basins but different time periods) (Figure 5). For the temporal test, the model δ's Q_{98} trends (0.88) are as accurate as those of LSTM for high flows (0.87), but LSTM outperformed δ models for the median and low flows (Q_{50} and Q_{10}). δ(β^t,γ^t) follow closely behind. This test, which excluded the impact of spatial generalization, suggests δ models' surface runoff routine has the ability to transform long-term forcing changes into the correct streamflow changes, but the current groundwater module may be suboptimal (or, stated in another way, it loses information). Also, compared to LSTM, δ models are more subject to trade-offs due to maintaining mass balances and thus could be trained to put more focus on the peaks of the hydrograph while sacrificing the low flow end.

Both LSTM and δ models surpassed MPR+mHm in the temporal test, by varying extents, for all flow percentiles, which demonstrated the potential from adaptive, learnable models. MPR+mHM's high flow (R^2=0.69) and median flow (R^2=0.63) trends lagged noticeably behind while the difference in the 10-th percentile flow was smaller. It was previously shown in Feng et al., (2022) (thus omitted here) that median NSEs of MPR+mHm, δ(β^t,γ^t) and LSTM were 0.53, 0.715, and 0.722, respectively. Compared to the learnable models, MPR+mHM tends to underestimate the wetting trend for the high flow and overestimate the wetting trend for the low flow. The fact that the annual mean flow trend is correct despite different flow percentiles getting lower metrics suggests in MPR+mHM some rainfall input was released from the wrong compartments. Note that the temporal test is the only comparison that we can carry out with existing process-based hydrologic models. Common benchmark problems certainly help the community understand the advantages and disadvantages of each model (Shen et al., 2018) and a PUB or PUR experiments from existing models would facilitate such comparisons.

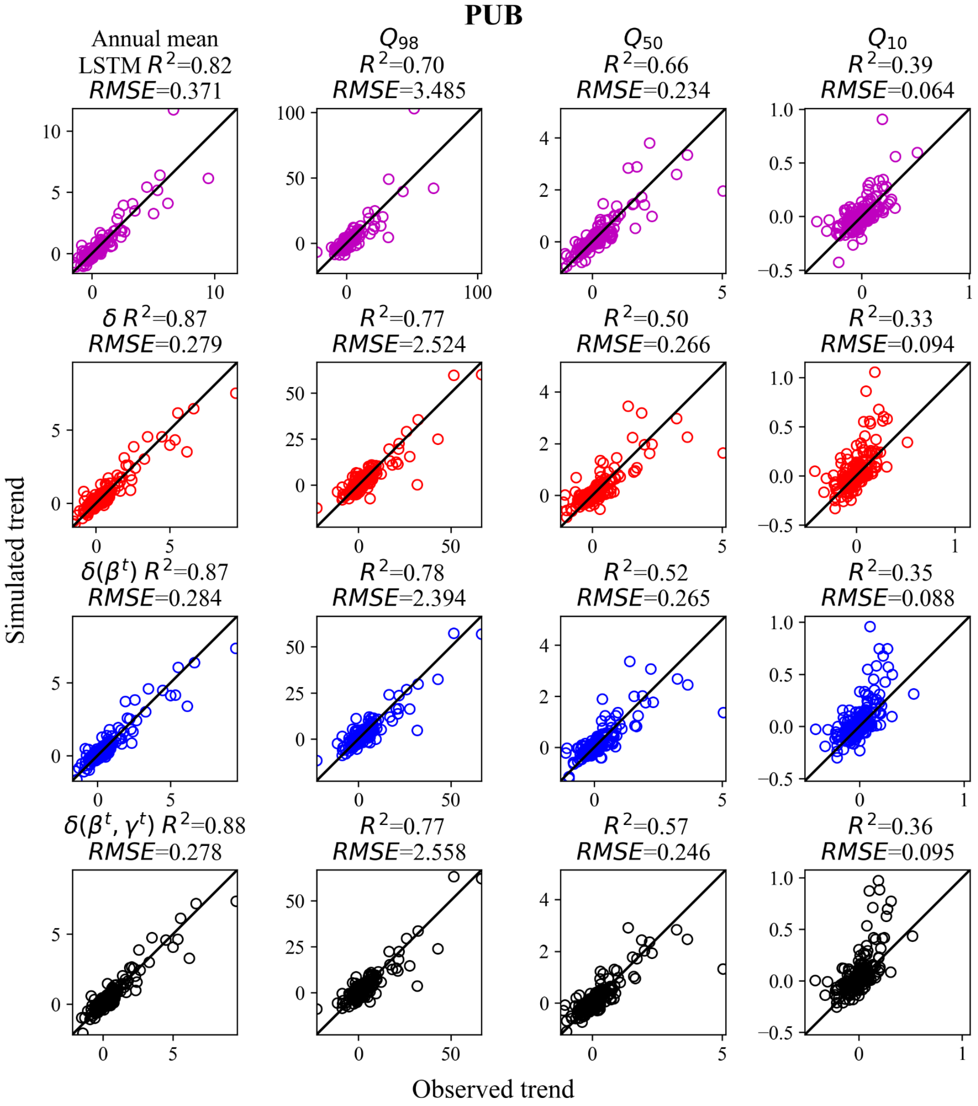
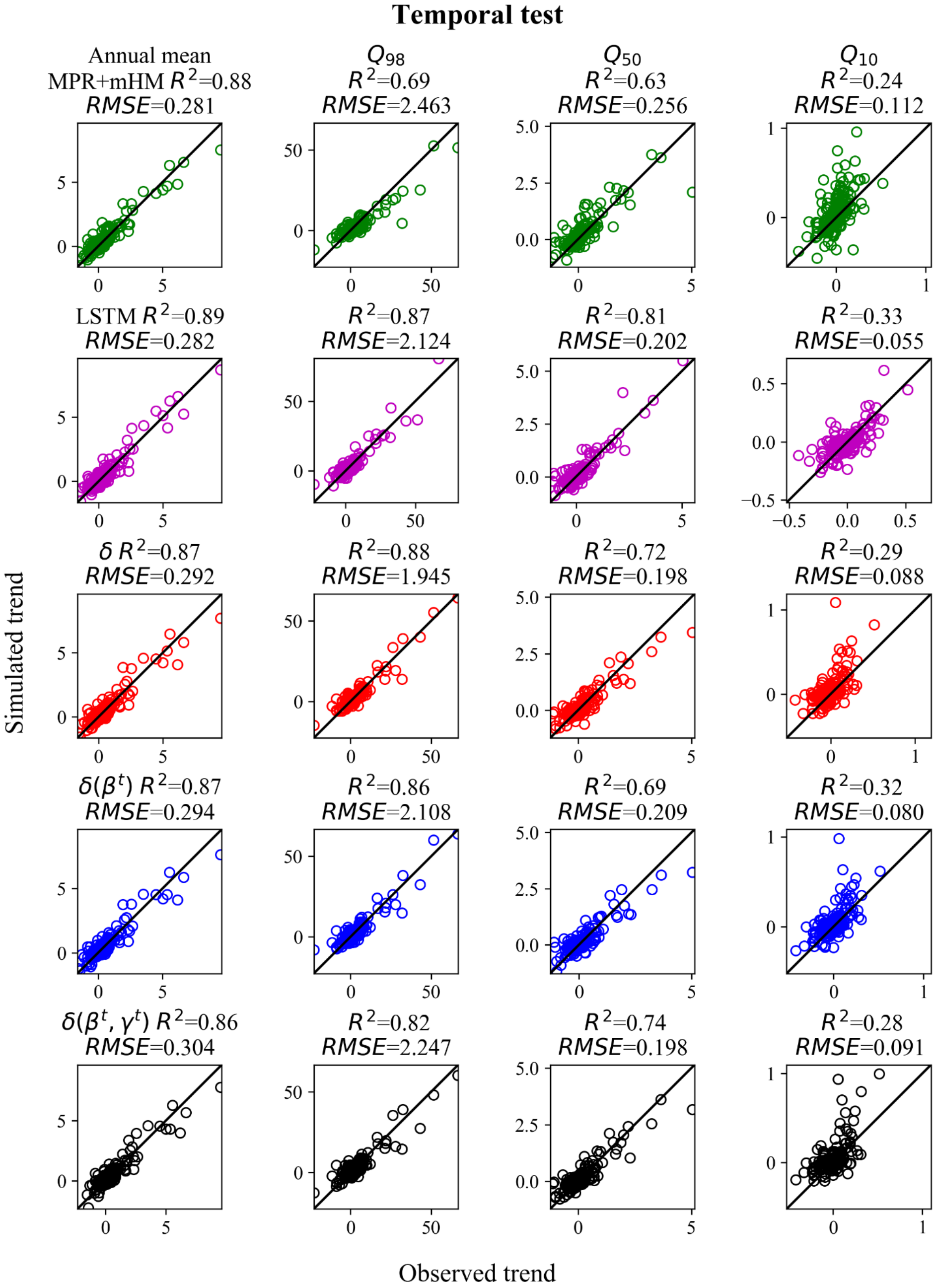


Figure 4. Decadal trends (m³/s/year) of flow for different flow percentiles for the randomized PUB cross-validation experiment, as compared to the observed trends. Q_{10}, Q_{50} and Q_{98} mean the trends were evaluated in the annual 10th-, 50th- and 98th- percentile flows, respectively. For each flow percentile, a corresponding value was extracted from each year's daily data and Sen's slope was estimated between hydrologic years 1989 and 1999.

https://doi.org/10.5194/hess-2022-245
Preprint. Discussion started: 8 August 2022



295 **Figure 5. Observed vs. simulated decadal trends (m³/s/year) of streamflow for the temporal test for 447 basins where MPR+mHM has predictions (all models trained from 1999 to 2008 and tested from 1989 to 1999 of hydrologic years on the same basins). We could only compare the trends with an existing process-based model with a parameter regionalization scheme on the temporal test because we did not have their systematic PUB results on the same dataset.**

https://doi.org/10.5194/hess-2022-245
Preprint. Discussion started: 8 August 2022

3.2. The region-based PUR test

For the regional holdout test (PUR), surprisingly, δ models moderately outperformed LSTM in terms of the daily hydrograph metrics (KGE, NSE, and FHV) and again had smaller spreads in these metrics (Figure 6). The LSTM performance dropped substantially from PUB to PUR, while the δ model performance dropped less. The median NSE values for LSTM, δ, and δ(β^t,γ^t) models were 0.56, 0.57 and 0.59, respectively, and the corresponding KGE values were 0.49, 0.58 and 0.61, respectively. We see that for the low flow dynamics, δ(β^t,γ^t) had a slightly smaller low flow bias (FLV). For high flow, δ models still had negative biases but they were smaller than those of LSTM.

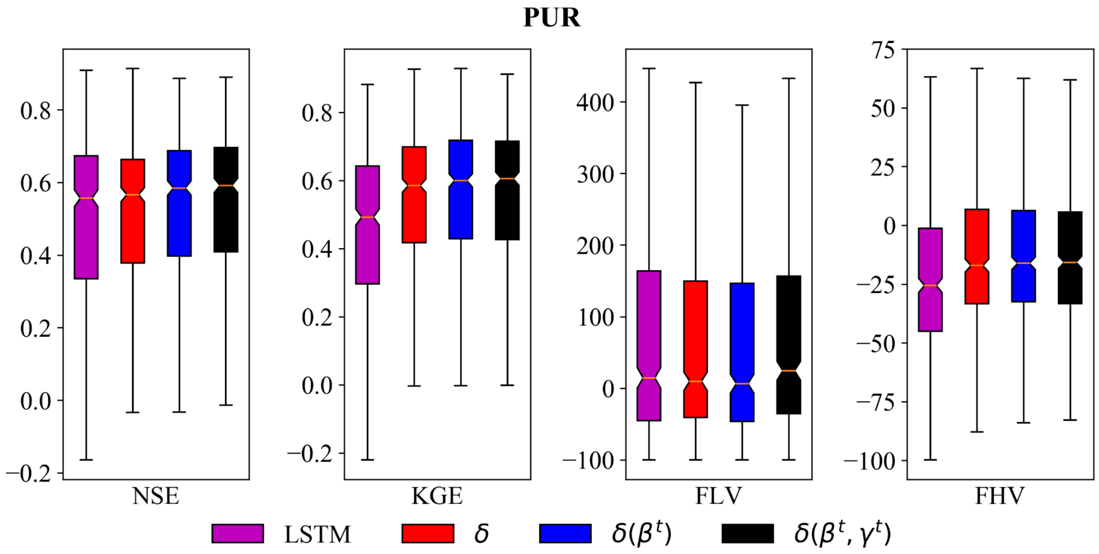


Figure 6. Same as Figure 3 but for the regional holdout (PUR) test. Each box summarizes the metrics of 531 basins obtained in a regional cross-validation fashion. We see clear outperformance of LSTM by the δ models for these daily hydrograph metrics (NSE, KGE, and FHV).

The decadal flow trends showed a stronger contrast -- while LSTM's trend metrics declined noticeably from PUB to PUR, the δ and δ^t models' trend accuracy barely budged. For the annual mean flow, the points for δ(β^t,γ^t) tightly surrounded the ideal 1-to-1 line and correctly captured the basins with strong wetting trends toward the higher end of the plot. In contrast, LSTM showed an underestimation bias and a tendency to plateau for the wetting basins. The same pattern is obvious for the high flow (Q_{98}). We previously also noticed such a flattening tendency in multi-year soil moisture trend projection (see Figure 9 in Fang et al., (2019)), although there the model was trained on satellite data which could also have played a role. LSTM's R^2 for annual mean discharge dropped from 0.82 for PUB to 0.72 for PUR, but R^2 remained at 0.88 for δ(β^t,γ^t). LSTM's R^2 for high flow (Q_{98}) trends dropped significantly, from 0.70 for PUB to 0.55 for PUR, whereas this metric remained around 0.77 for the δ models. The results highlight the δ models' robust ability to generalize in space, possibly due to the simple physics built into the model.

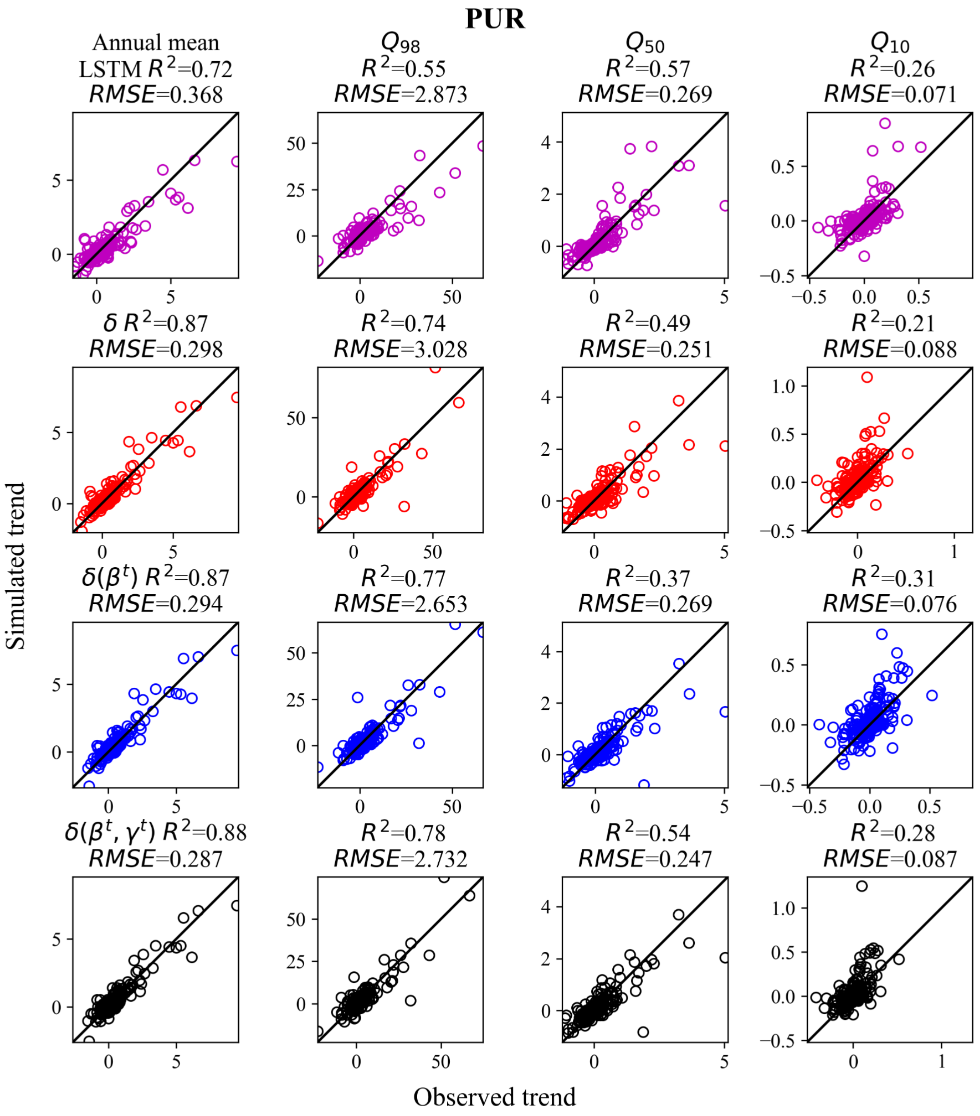


Figure 7. Same as Figure 4 but for the regional holdout (PUR) test. δ models outperformed LSTM for the trends (m^3/s/year) of mean annual flow and the high flow regime.

What makes δ models more robust than LSTM for PUR, especially in terms of high flow and mean annual flow? As indicated earlier, δ models can be considered as machine learning models that are restricted to a subspace allowable by the backbone structure. There are two structural constraints: (i) the static attributes can only influence the model via fixed interfaces (model parameters); and (ii) the whole system can only simulate flow as permitted by the backbone model, HBV. Hence, we can force the parameterization to learn a simpler and more generic mapping relationship, and when it succeeds, the relationship could be more transferable than that from LSTM, which mixes information from all variables in most steps.

The δ model-based parameter maps reveal that the in-sample, PUB, and PUR models all produced similar overall parameter patterns (Figure 8 -- for PUB and PUR, these parameters were generated when the basins were used as the test basins). Between in-sample and PUB, most of the points had similar colors, except for a few isolated basins (e.g., some basins in New Mexico). Between PUB and PUR, there were more regional differences (e.g., in the Dakotas, North Carolina, and Florida), but the overall CONUS-scale patterns were still similar. Recall that (i) these parameters were estimated by the parameter network g_A, which was trained on streamflow, and there are no ground-truth values for the parameters; and (ii) in the PUR experiments, a large region was held out. Despite these strong perturbations to the training data, such parameter stability under PUB is impressive. This stability is part of the reason for the mild performance drop under PUR. Had we used a basin-by-basin parameter calibration approach, the parameter values would have been much more stochastic and interspersed (similar to Figure 6b in Tsai et al. (2021)).

We note that δ models found advantages in the annual mean flow and high flow regimes rather than the low flow regime for the PUR test. As described above, we attribute the advantage in high flow to learning a more generalizable mapping between raw attributes and runoff parameters. For the low flow component, the δ models were close to the LSTM for performance in PUR and PUB but were outperformed by the LSTM for the temporal tests. We hypothesize that this was because the groundwater module inherited from the HBV model, which is based on a simple linear reservoir, cannot adequately represent long-term groundwater storage changes. This part of the model will require additional structural changes, e.g., by adopting nonlinearity (Seibert & Vis, 2012) or considering feedback between layers in the groundwater modules. Further, due to the guaranteed mass balance, the δ models face more tension (or trade-offs) between the low and high flow regimes during training. The peak flow part tends to receive more attention due to its larger values. Because pure LSTM models do not guarantee the conservation of mass, they are subject to fewer trade-offs and are more likely to capture both high and low flows. We believe future work can further improve the groundwater representation by considering better topographic distributions.

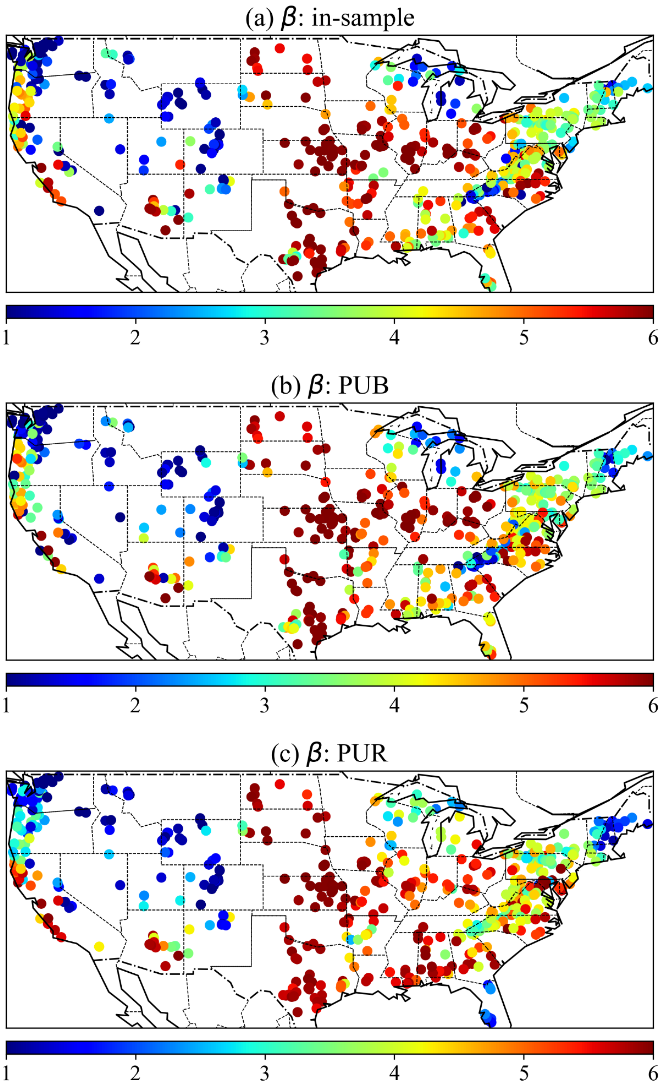


Figure 8. Parameter maps for the β parameter of the HBV model for (a) the in-sample temporal test; (b) PUB; and (c) PUR. For PUB and PUR, all the parameters were produced from cross-validation experiments when the sites were used as test sites and were not included in training. With other conditions being the same, higher β yields less runoff, but other parameters such as the maximum soil water storage also influence runoff. For simplicity, this parameter is generated from a δ model without dynamical parameterization and is the output of the parameterization network (g_A). Again, there is no ground truth parameter to supervise g_A.

3.3. The impacts of extrapolation on evapotranspiration

Spatial interpolation and extrapolation seemed to have a moderate impact on ET seasonality and muted impact on annual mean ET (Figure 9). For $\delta(\gamma^t, \beta^t)$, from temporal tests to PUB and then PUR, the median correlation and RMSE between simulated

ET and ET from the MODIS satellite product didn't vary much, around 0.84 and 6.4 mm/8day, respectively. The impact of extrapolation on ET was more muted compared to streamflow. Understandably, ET is controlled by the energy input and physics-based calculations, and thus the models cannot deviate too much from each other.

Moreover, the dynamic parameterization (DP) models, δ(γ^t, β^t) and δ(β^t), were better than static parameter models in all comparable cases (temporal test, PUB, or PUR). The decline due to spatial interpolation or extrapolation was minimal. Even for the most adverse case, i.e., PUR, δ(γ^t, β^t) provided a high-quality ET seasonality as compared to MODIS (median correlation of 0.84) and low RMSE. It appears that DP indeed captured missing dynamics in data, possibly attributable to long-term water storage and vegetation dynamics, and presented "*better models for the right reasons*".

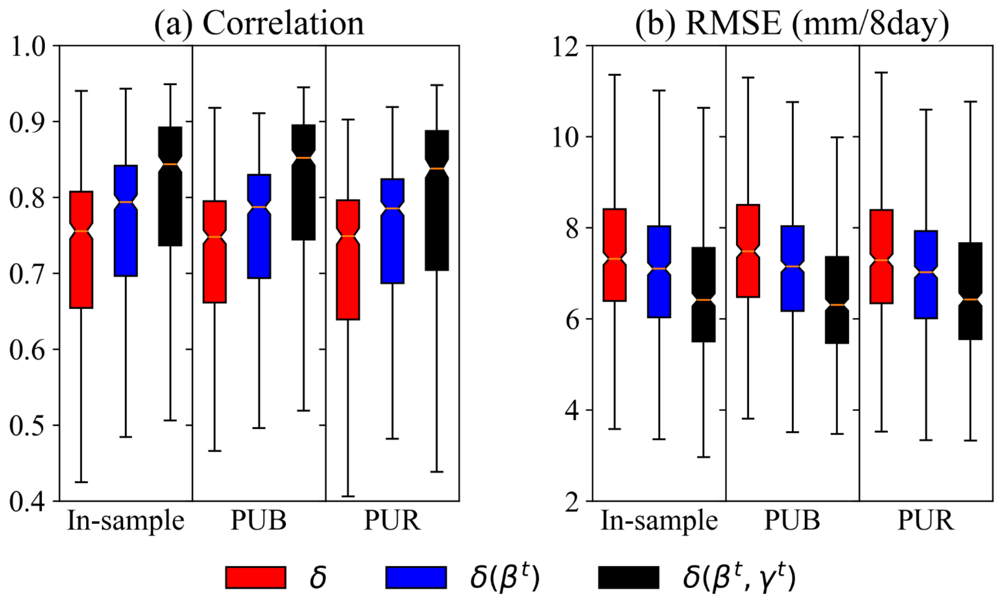


Figure 9. Comparison of the agreement of simulated ET and the MODIS satellite product for different models under the temporal test, PUB and PUR scenarios using two different metrics - (a) correlation and (b) root-mean-square error (RMSE). All models were trained only with streamflow as the target.

3.3. Further discussion

For all cases tested and for both streamflow and ET, the model with dynamical parameterization (DP), δ(γ^t, β^t), had better generalization than the δ models without DP. In theory, the model with DP has more flexibility, and, correspondingly, we had expected DP to be more overfitted in some cases. However, the results showed δ(γ^t, β^t) to be comparable or slightly better in most cases (either trends or NSE/KGE) than δ and δ(β^t), thus the expected overfitting did not occur. Although the LSTM-based parameterization unit g_A has a large amount of weights, it can only influence the computation through restricted interfaces (the parameters). In contrast, the full LSTM model we tested allows attributes to influence all steps of the calculations. The fact that δ(γ^t, β^t) was more generalizable also suggests that whether the model will overfit or not depends on the way the computation is regulated, rather than simply the number of weights. It seems DP may have enabled the learning of some true processes that are missing from HBV, possibly related to deep soil water storage and/or vegetation dynamics (Feng et al., 2022).

While not directly tested here, it is easy to imagine that in the future we can constrain the δ models using multiple sources of observations. So far, the simulation quality seems consistent between streamflow and ET, e.g., δ(γ^t, β^t) is better than δ in streamflow (NSE/KGE) and also ET. This was not always true traditionally due to equifinality (Beven, 2006), and it means a better conditioning of one of these variables could have positive impacts on other variables. Over the globe, while gauged basins are limited, there are many sources of information on soil moisture (ESA, 2022; NSIDC, 2022; Wanders et al., 2014), water storage (Eicker et al., 2014; Landerer et al., 2020), in-situ measurements of ET (LBNL, 2022; Velpuri et al., 2013), snow cover (Duethmann et al., 2014), and other measurements that provide additional opportunities for learning.

4. Conclusions

We demonstrated the high competitiveness of differentiable, learnable hydrologic models (δ models) for both spatial interpolation (PUB) and extrapolation (PUR). Evidence for such high competitiveness are provided in terms of daily hydrograph metrics including NSE and KGE and in terms of decadal-scale trends, which are of particular importance for climate change impact assessments. For the daily hydrograph metrics, the δ models closely approached the LSTM model in the PUB test (while showing less spread) and outperformed the LSTM model in the PUR test. For the decadal-scale trends, the δ models outperformed the LSTM model for the PUB test and more noticeably in the PUR tests, especially for the annual mean flow and high flows, although LSTM still fared better for the temporal (in-sample) test. In the temporal test, both LSTM and δ models surpassed an existing process-based model to varying extents for different flow percentiles, indicating better rainfall-runoff dynamics.

Out of the variants of differentiable models tested, δ(γ^t, β^t) stood out for having the best overall test performance, attesting to the strength of the structural constraints. Even though its structure is more complex, it was not more overfitted than other models. It also showed markedly better ET seasonality, which barely deteriorated in PUB or PUR scenarios, than δ or δ(β^t). As δ models simulate a wide variety of variables, they stand to benefit from assimilating multiple data sources. The need for additional memory units (in the LSTM that infers dynamical parameters) suggests that there is still significant room for structural improvement of the backbone model (HBV).

While LSTM models have achieved monumental advances, the δ models combine the fundamental strength of neural network learning with an interpretable, physics-based backbone to provide more constraints and better interpretability. The training of the δ models resulted in remarkably stable parameter fields despite large differences in training datasets (temporal test vs. PUB vs. PUR). δ models are not only reliable candidates for global climate change impact assessment but can also highlight potential deficiencies in current process-based model structures (in the case of HBV, in the representations of vegetation and deep subsurface water storage). δ models can thus be used as a guide to future improvements of the model mechanisms and what we learn from δ models can in fact be ported to traditional process-based models. Lastly, we clarify that this conclusion does not mean LSTM or existing models are not suitable for global applications. As one can see, LSTM remained a ferocious competitor for both PUB and PUR and existing models also presented decent trend metrics. We call for more benchmarking on large datasets for different scenarios such as PUB, PUR, and more variables.

Acknowledgments

This work is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under contract DESC0016605 and US National Science Foundation EAR 1832294 and OAC 1940190. The code for the LSTM is available at https://github.com/mhpi/hydroDL. The code for the δ models will be made available upon paper acceptance. Part of this work has been presented as an oral presentation in HydroML symposium, May 2022.

Appendix A.

Table A1 The attribute variables used in this study for regionalized models

Attribute variables	Description	Unit
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/yr
high_prec_dur	Average duration of high precipitation events	days
low_prec_freq	Frequency of dry days	days/yr
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	-
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	-

https://doi.org/10.5194/hess-2022-245
Preprint. Discussion started: 8 August 2022

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

Appendix B.

Here we describe the equations related to the parameters β and γ:

$$P_{eff} = \min\{(S_s/\theta_{FC})^{\beta}, 1\} \;*\; (P_r+I_{snow})$$

$$E_a = \min\{[S_s/(\theta_{FC}\theta_{LP})]^{\gamma}, 1\} * E_p$$

Here P_{eff} represents the effective rainfall to produce runoff, P_r represents the rainfall, I_{snow} represents the snowmelt infiltration to soil, S_s represents the surface soil water, E_p represents the potential evapotranspiration (ET), E_a represents the actual ET, parameters θ_{FC} and θ_{LP} (a fraction of θ_{FC}) represent the thresholds for maximum soil moisture storage and actual ET reaching to potential ET, respectively. β is the shape coefficient of the runoff relation, while γ is a newly added shape coefficient of the ET relation. For the dPL models with dynamic parameters in this study, we modify the static β and γ into dynamic parameters β^t and γ^t which change with time, based on the meteorological forcings.

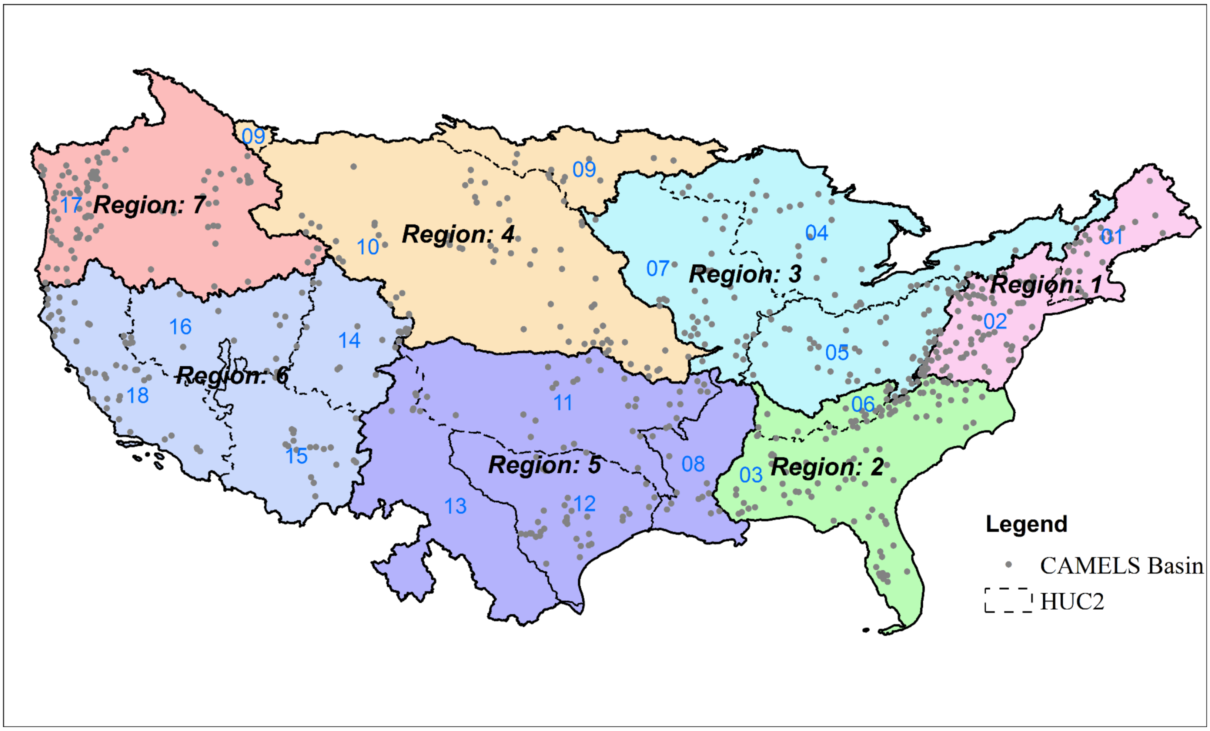


Figure A1. Division of the CAMELS dataset into 7 large regions for the PUR cross validation test: for every fold, the models were trained on 6 of the 7 regions and tested on the one held out. We ran the experiments for 7 rounds so that each region would be the test region once. The results for the test basins were then collected and the metrics were reported for this collection.

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017a). Catchment Attributes and MEteorology for Large-Sample studies (CAMELS) version 2.0 [Dataset]. In *UCAR/NCAR*. https://doi.org/10.5065/D6G73C3Q

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017b). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Aghakouchak, A., & Habib, E. (2010). Application of a Conceptual Hydrologic Model in Teaching Hydrologic Processes. *International Journal of Engineering Education*, *26*(4 (S1)). http://escholarship.org/uc/item/3sv066q5

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., Dijk, A. I. J. M. van, & 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

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*(5), 3599–3622. https://doi.org/10.1002/2015WR018247

Beck, H. E., Wood, E. F., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garreton, C., Baez-Villanueva, O. M., Sheffield, J., & Karger, D. N. (2020). Bias correction of global high-resolution precipitation climatologies using streamflow observations from 9372 catchments. *Journal of Climate*, *33*(4), 1299–1315. https://doi.org/10.1175/JCLI-D-19-0332.1

Bergström, S. (1976). *Development and application of a conceptual runoff model for Scandinavian catchments* [PhD Thesis, Swedish Meteorological and Hydrological Institute (SMHI)]. 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). 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/ccx2ks

Boretti, A., & Rosa, L. (2019). Reassessing the projections of the World Water Development Report. *Npj Clean Water*, *2*(1), 1–6. https://doi.org/10.1038/s41545-019-0039-9

Döll, P., Fiedler, K., & Zhang, J. (2009). Global-scale analysis of river flow alterations due to water withdrawals and reservoirs. *Hydrology and Earth System Sciences*, *13*(12), 2413. https://doi.org/10.5194/hess-13-2413-2009

Duethmann, D., Peters, J., Blume, T., Vorogushyn, S., & Güntner, A. (2014). The value of satellite-derived snow cover images for calibrating a hydrological model in snow-dominated catchments in Central Asia. *Water Resources Research*, *50*(3), 2002–2021. https://doi.org/10.1002/2013WR014382

Eicker, A., Schumacher, M., Kusche, J., Döll, P., & Schmied, H. M. (2014). Calibration/data assimilation approach for integrating GRACE data into the WaterGAP Global Hydrology Model (WGHM) using an ensemble Kalman filter: First results. *Surveys in Geophysics*, *35*(6), 1285–1309. https://doi.org/10.1007/s10712-014-9309-8

ESA. (2022). *About SMOS - Soil Moisture and Ocean Salinity mission*. European Space Agency (ESA). https://earth.esa.int/eogateway/missions/smos

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/gghp3v

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), 11,030-11,039. https://doi.org/10/gcr7mq

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

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). *Differentiable, learnable, regionalized process-based models with physical outputs can approach state-of-the-art hydrologic prediction accuracy*. https://doi.org/10.48550/arXiv.2203.14827

Frame, J. M., Kratzert, F., Raney II, A., 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

Frame, J. M., Ullrich, P., Nearing, G., Gupta, H., & Kratzert, F. (2022). *On Strictly Enforced Mass Conservation Constraints for Modeling the Rainfall-Runoff Process*. https://eartharxiv.org/repository/view/3028/

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 modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., & Tallaksen, L. M. (2011). Large-scale river flow archives: Importance, current status and future needs. *Hydrological Processes*, *25*(7), 1191–1200. https://doi.org/10.1002/hyp.7794

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)

Hattermann, F. F., Krysanova, V., Gosling, S. N., Dankers, R., Daggupati, P., Donnelly, C., Flörke, M., Huang, S., Motovilov, Y., Buda, S., Yang, T., Müller, C., Leng, G., Tang, Q., Portmann, F. T., Hagemann, S., Gerten, D., Wada, Y., Masaki, Y., … Samaniego, L. (2017). Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. *Climatic Change*, *141*(3), 561–576. https://doi.org/10.1007/s10584-016-1829-4

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., … Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10/gfsq5q

Innes, M., Edelman, A., Fischer, K., Rackauckas, C., Saba, E., Shah, V. B., & Tebbutt, W. (2019). *A Differentiable Programming System to Bridge Machine Learning and Scientific Computing* (arXiv:1907.07587). arXiv. http://arxiv.org/abs/1907.07587

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

Kim, Y. W., Kim, T., Shin, J., Go, B., Lee, M., Lee, J., Koo, J., Cho, K. H., & Cha, Y. (2021). Forecasting Abrupt Depletion of Dissolved Oxygen in Urban Streams Using Discontinuously Measured Hourly Time-Series Data. *Water Resources Research*, *57*(4), e2020WR029188. https://doi.org/10.1029/2020WR029188

Kratzert, F. (2019). *CAMELS Extended Maurer Forcing Data* [Data set]. 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/gg4ck8

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

Landerer, F. W., Flechtner, F. M., Save, H., Webb, F. H., Bandikova, T., Bertiger, W. I., Bettadpur, S. V., Byun, S. H., Dahle, C., Dobslaw, H., Fahnestock, E., Harvey, N., Kang, Z., Kruizinga, G. L. H., Loomis, B. D., McCullough, C., Murböck, M., Nagel, P., Paik, M., … Yuan, D.-N. (2020). Extending the global mass change data record: GRACE follow-on instrument and science data performance. *Geophysical Research Letters*, *47*(12), e2020GL088306. https://doi.org/10.1029/2020GL088306

LBNL. (2022, March 22). *Introducing the AmeriFlux FLUXNET data product*. Lawrence Berkeley National Laboratory (LBNL). https://ameriflux.lbl.gov/introducing-the-ameriflux-fluxnet-data-product/

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

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/dk5v56

Mu, Q., Zhao, M., & Running, S. W. (2013). *MODIS Global Terrestrial Evapotranspiration (ET) Product (NASA MOD16A2/A3) Algorithm Theoretical Basis Document, Collection 5* (No. 08-ATDB11-0001). National Aeronautics and Space Administration (NASA). https://lpdaac.usgs.gov/documents/93/MOD16_ATBD.pdf

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/fbg9tm

NSIDC. (2022). *SMAP Overview—Soil Moisture Active Passive*. National Snow & Ice Data Center (NSIDC). https://nsidc.org/data/smap

O, S., & Orth, R. (2021). Global soil moisture data derived through machine learning trained with in-situ measurements. *Scientific Data*, *8*(1), 170. https://doi.org/10.1038/s41597-021-00964-1

Qiu, R., Wang, Y., Rhoads, B., Wang, D., Qiu, W., Tao, Y., & Wu, J. (2021). River water temperature forecasting using a deep learning method. *Journal of Hydrology*, *595*, 126016. https://doi.org/10.1016/j.jhydrol.2021.126016

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

Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., Clark, M. P., & Samaniego, L. (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

Ray, K., Pandey, P., Pandey, C., Dimri, A. P., & Kishore, K. (2019). On the recent floods in India. *Current Science*, *117*(2), 204. https://doi.org/10.18520/cs/v117/i2/204-218

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/f22r5x

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389. https://doi.org/10.2307/2285891

Shen, C. (2018). 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/gd8cqb

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X., & Tsai, W.-P. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

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

Sivapalan, M. (2003). Prediction in ungauged basins: A grand challenge for theoretical hydrology. *Hydrological Processes*, *17*(15), 3163–3170. https://doi.org/10/cdc664

Suen, J.-P., & Eheart, J. W. (2006). Reservoir management to balance ecosystem and human needs: Incorporating the paradigm of the ecological flow regime. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004314

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., & Shen, C. (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

Wanders, N., Bierkens, M. F. P., de Jong, S. M., de Roo, A., & Karssenberg, D. (2014). The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models. *Water Resources Research*, *50*(8), 6874–6891. https://doi.org/10/f6j4b2

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9). https://doi.org/10/fpvsgb

Zaherpour, J., Gosling, S. N., Mount, N., Schmied, H. M., Veldkamp, T. I. E., Dankers, R., Eisner, S., Gerten, D., Gudmundsson, L., Haddeland, I., Hanasaki, N., Kim, H., Leng, G., Liu, J., Masaki, Y., Oki, T., Pokhrel, Y., Satoh, Y., Schewe, J., & Wada, Y. (2018). Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts. *Environmental Research Letters*, *13*(6), 065015. https://doi.org/10.1088/1748-9326/aac547

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