

Deep Dive into Hydrologic Simulations at Global Scale: Harnessing the Power of Deep Learning and Physics-informed Differentiable Models (δ HBV-globe1.0-hydroDL)

Dapeng Feng^{1,2,3}, Hylke Beck⁴, Jens de Bruijn^{3,5}, Reetik Kumar Sahu³, Yusuke Satoh⁶, Yoshihide Wada⁷, Jiangtao Liu¹, Ming Pan⁸, Kathryn Lawson¹, Chaopeng Shen^{1*}

¹ Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA

² Earth System Science, Stanford University, Stanford, CA, USA

³ Water Security Research Group, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

⁴ Climate and Livability Initiative, Physical Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

⁵ Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, Amsterdam, Netherlands

⁶ Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

⁷ Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

⁸ Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA

*Correspondence to: Dapeng Feng (dpfeng@stanford.edu) Chaopeng Shen (cshen@engr.psu.edu)

Abstract. Accurate hydrologic modeling is vital to characterizing how the terrestrial water cycle responds to climate change. Pure deep learning (DL) models have shown to outperform process-based ones while remaining difficult to interpret. More recently, differentiable, physics-informed machine learning models with a physical backbone can systematically integrate physical equations and DL, predicting untrained variables and processes with high performance. However, it was unclear if such models are competitive for global-scale applications with a simple backbone. Therefore, we use - for the first time at this scale - differentiable hydrologic models (full name δ HBV-globe1.0-hydroDL, shortened to δ HBV here) to simulate the rainfall-runoff processes for 3753 basins around the world. Moreover, we compare the δ HBV models to a purely data-driven long short-term memory (LSTM) model to examine their strengths and limitations. Both LSTM and the δ HBV models provide competent daily hydrologic simulation capabilities in global basins, with median Kling-Gupta efficiency values close to or higher than 0.7 (and 0.78 with LSTM for a subset of 1675 basins with long-term discharge records), significantly outperforming traditional models. Moreover, regionalized differentiable models demonstrated stronger spatial generalization ability (median KGE 0.64) than a traditional parameter regionalization approach (median KGE 0.46) and even LSTM for ungauged region tests across continents. Nevertheless, relative to LSTM, the differentiable model was hampered by structural deficiencies for

32 cold or polar regions, and highly arid regions, and basins with significant human impacts. This study also sets the benchmark
33 for hydrologic estimates around the world and builds foundations for improving global hydrologic simulations.

34
35 **Short Summary.** Accurate hydrologic modeling is vital to characterizing water cycle responses to climate change. For the
36 first time at this scale, we use differentiable physics-informed machine learning hydrologic models to simulate rainfall-runoff
37 processes for 3753 basins around the world and compare them with purely data-driven and traditional modeling approaches.
38 This sets a benchmark for hydrologic estimates around the world and builds foundations for improving global hydrologic
39 simulations.

40
41 **Key Words.** Physics-informed machine learning; Differentiable hydrologic models; Global hydrologic modeling; high
42 resolution evaluation; Parameter regionalization; Prediction in ungauged regions

43 1. Introduction

44 Hydrologic models are vital tools to model and elucidate the terrestrial water cycle, and have been widely used in flood
45 forecasting (Maidment, 2017), water resources management (Jayakrishnan et al., 2005), and assessing climate change impacts
46 (Hagemann et al., 2013). Recently, deep learning (DL) models have demonstrated superior performance compared to
47 traditional process-based hydrologic models in accurately predicting different components of the hydrologic cycle (Shen,
48 2018), such as soil moisture (Fang et al., 2017, 2019; Fang and Shen, 2020), streamflow (Feng et al., 2020; Konapala et al.,
49 2020; Kratzert et al., 2019b; Liu et al., 2024), snow water equivalent (Cui et al., 2023; Song et al., 2024b), groundwater
50 (Wunsch et al., 2021) and water quality (Hansen et al., 2022; Rahmani et al., 2021; Saha et al., 2023; Song et al., 2024a; Zhi
51 et al., 2021). Long short-term memory (LSTM) networks, which are a type of recurrent neural network (Hochreiter and
52 Schmidhuber, 1997), and Transformers (Vaswani et al., 2017) are currently popular DL algorithms for handling time series
53 dynamics in hydrology, while other architectures can also be employed. LSTM models have established state-of-the-art
54 accuracy for streamflow prediction at continental and smaller scales (Feng et al., 2020, 2021; Kratzert et al., 2019a, b; Lees et
55 al., 2021; Mai et al., 2022).

56
57 Although DL models have shown great prediction accuracy compared to traditional models, they usually do not possess clear
58 physical constraints inside the model and are often considered to be “black boxes”, despite recent efforts shed by some
59 interpretive efforts (Lees et al., 2022). Thus, purely data-driven models are limited in that they cannot predict unobservable or
60 untrained physical variables, which impedes the investigation of the physical relations of different hydrologic variables behind
61 the change in the target variable. They may also become overfitted and acquire incorrect sensitivities to inputs (Reichert et al.,
62 2024). In contrast, traditional process-based hydrologic models following physical laws like mass balances can provide a full
63 set of diagnostic outputs for hydrologic variables like soil water storage, groundwater recharge, evapotranspiration and snow

64 water equivalent, even though they are usually only calibrated on discharge observations (Burek et al., 2020; Müller Schmied
65 et al., 2014). The multivariate output nature of these models provides an opportunity for calibration on one or more observable
66 variables to better predict other, perhaps unobservable, variables (in reality, whether this is the case or not depends on if the
67 issue of parameter non-uniqueness is addressed). However, it seems quite difficult for the traditional physical model to
68 approach the performance level of the DL models in daily hydrograph metrics (Feng et al., 2020; Kratzert et al., 2019b) or to
69 improve in generalization with increasing training data (Tsai et al., 2021). In addition, traditional calibration is typically done
70 site-by-site and can be time- and labor-intensive. Therefore, it logically follows that integrating DL and process-based models
71 might enable harnessing their respective strengths while circumventing their weaknesses (Shen et al., 2023).

72

73 By combining a physical model with a DL model, differentiable modeling (Feng et al., 2022; Shen et al., 2023) provides a
74 systematic solution to leveraging the strengths of both model types while circumventing their limitations. In differentiable
75 models, we use process-based models as a backbone and insert neural networks to either provide parameters (Tsai et al., 2021)
76 or process substitutes for physical models (Aboelyazeed et al., 2023; Feng et al., 2022, 2023; Höge et al., 2022; Jiang et al.,
77 2020), or they could use limited physical constraints (Kraft et al., 2022). They are collectively called “differentiable models”
78 in the sense that they can rapidly compute gradients of outputs with respect to inputs or parameters using automatic
79 differentiation (or any other means). The differentiability enables the training of neural network components placed anywhere
80 in the model via backpropagation. Inserting neural networks into process-based models can be perceived as posing questions
81 regarding some uncertain relationships given some known ones (priors) and we want to get answers for these questions by
82 automatically learning from big data.

83

84 Some of our recent work has applied differentiable modeling to the conceptual hydrologic model named Hydrologiska Byråns
85 Vattenbalansavdelning (HBV) (Bergström, 1976, 1992; Seibert and Vis, 2012), and built a physics-informed hybrid model for
86 basins in the contiguous United States (CONUS) (Feng et al., 2022, 2023). The model is “regionalized” in the sense that the
87 embedded neural network components are trained simultaneously on all basins in the study region in order to provide physical
88 HBV parameters which are learned from raw information of basin attributes, resulting in improved generalizability and reduced
89 overfitting to local noise. With the help of differentiable modeling to flexibly evolve the original structure of HBV, the
90 differentiable hybrid models can approach the performance level of the LSTM model, whilst being constrained to physical
91 laws and keeping process clarity to predict untrained diagnostic variables with decent accuracy (Feng et al., 2022). Since the
92 framework is regionalized, this differentiable model can be used to predict in ungauged regions and even extrapolates better
93 spatially than LSTM in data-sparse regions when tested across the CONUS (Feng et al., 2023).

94

95 Owing to the complexity of calibration, current global hydrologic models are largely either uncalibrated (Hattermann et al.,
96 2017; Zaherpour et al., 2018) or only calibrated on mean annual water budgets or in limited regions (Burek et al., 2020; Müller
97 Schmied et al., 2014). Only very limited studies attempt to calibrate global models on monthly discharge variations (Werth

98 and Güntner, 2010). We desire efficient regionalized models that maximally leverage available information and provide
99 accurate predictions to diverse basins across different climate groups and geographic characteristics in the world. We also want
100 the models to perform decently even in data-sparse regions, showing competitive extrapolation ability, given that many large
101 regions such as in Africa and Asia lack publicly available streamflow data. DL and differentiable models seem plausible
102 candidates for such simulations. Nevertheless, previous studies on DL and physics-informed differentiable models mainly
103 focus on continental or smaller scales, with a relatively homogeneous forcing dataset --- it is unclear if their observed strengths,
104 e.g., high performance and strong generalization ability, can carry over to global scales, where the climate is much more diverse
105 and datasets differ widely in their biases and uncertainty characteristics. In particular, we want to thoroughly examine how
106 well these models can leverage information learned in data-rich continents to characterize the hydrologic processes in
107 ungauged regions across the world. Meanwhile, DL models also show favorable scaling relationships (or data synergy) where
108 more data leads to more robust models (Fang et al., 2022). Thus, training on a larger dataset may provide additional benefits.
109
110 In this study, we test physics-informed differentiable models (with the full version name δ HBV-globe1.0-hydroDL, where “ δ ”
111 represents “differentiable”, globe1.0 is the version, and “hydroDL” refers to our particular code implementation. δ HBV is
112 used as the abbreviation in this paper) to simulate hydrologic processes for global basins and compare results to purely data
113 driven methods and traditional modeling approach. We focus on regionalized modeling and emphasize the importance of
114 spatial generalization in data-sparse scenarios, since observed streamflow data in many parts of the world are scarce. This
115 means one framework with parameter regionalization from geographic attributes will be used to model all the global basins
116 rather than calibrating a separate model in each individual basin (Beck et al., 2020b; Feng et al., 2022; Mizukami et al., 2017).
117 We first investigate what prediction accuracy can be achieved by different models at global scale by learning from a large and
118 diverse dataset. We then relate the global spatial patterns of model performance to geographic characteristics and hydrologic
119 processes to identify model structural deficiencies and gain hydrologic insights. Finally, we provide evidence indicating which
120 type of model may be more appropriate for next-generation global modeling by rigorously examining their generalizability to
121 ungauged regions across the world.

122 2. Data and methods

123 2.1 Global datasets

124 We use a global database compiled in a previous study (Beck et al., 2020b) which contains a total of 4229 headwater
125 catchments. The dataset includes basin mean meteorological forcings, catchment characteristics such as the climate,
126 topography, land cover, soil composition, and geology information to support parameter regionalization, along with streamflow
127 gauge discharge observations. Meteorological forcings are the driving inputs of hydrologic models. This global dataset
128 includes daily precipitation from Multi-Source Weighted-Ensemble Precipitation (MSWEP), a product that merges gauge,
129 satellite, and reanalysis precipitation data (Beck et al., 2017c, 2019), and maximum and minimum temperature from Multi-

Source Weather (MSWX), a product that bias-corrects and harmonizes meteorological data from atmospheric reanalyses and weather forecast models (Beck et al., 2022). Potential evapotranspiration was estimated using the method from Hargreaves (1994). The discharge observations at the outlet gauges were used as prediction targets to train the hydrologic models. We excluded some basins with potential erroneous discharge records such as showing unreasonable magnitude way larger than precipitation or dramatic differences between two time intervals, by manually performing visual screening, and also excluded those with severe amounts of missing data (less than 5 years' worth of data points in the study period from 2000 to 2016). Thus, 3753 basins were finally used to evaluate different models. These basins had been classified into five Köppen-Geiger climate classes in Beck et al., (2020b), including tropical (489 basins), arid (109 basins), temperate (1423 basins), cold (1593 basins), and polar (139 basins), as shown in Figure 1. To evaluate the simulations of untrained variables like evapotranspiration (ET), the MOD16A2GF (Running et al., 2021), a gap-filled 8-day composite ET product estimated from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data and meteorological reanalysis data, were used as independent observations to compare against the simulated ET from differentiable hydrologic models.

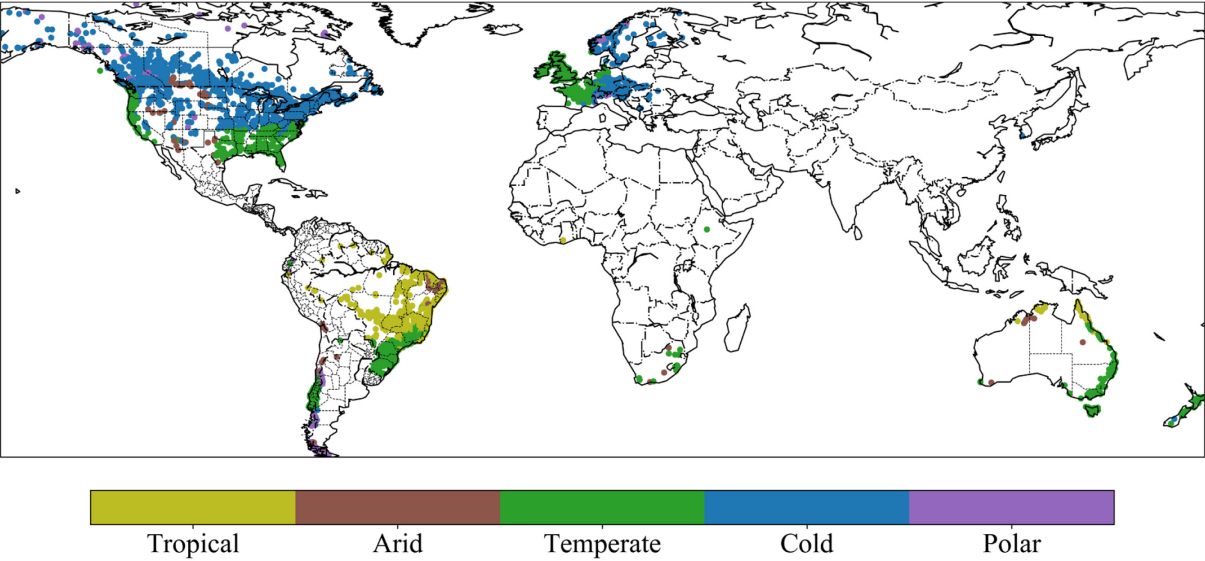
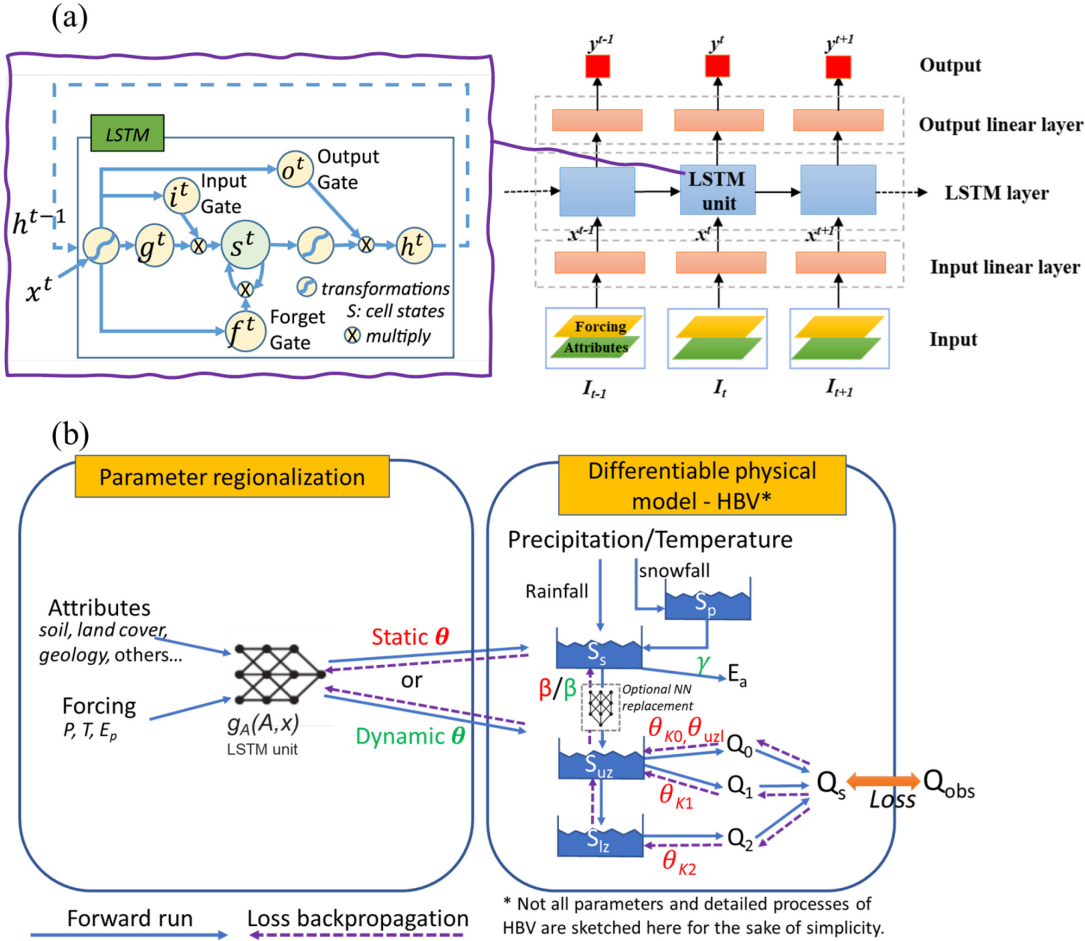


Figure 1. Locations and climate groups of the 3753 global basins used in this study, which were originally compiled by Beck et al., 2020b. Plotted in Python using Matplotlib Basemap Toolkit.

2.2 The long short-term memory (LSTM) streamflow model for comparison

Here the LSTM model is used as a benchmark for purely data-driven DL. The LSTM has “cell states” and “gates” to maintain and filter information, as shown in Figure 2a. The input, forget, and output gates control the flow of information, respectively controlling what to let in, what to forget, and what to output from the system. In this study we use the LSTM streamflow model demonstrated in Feng et al. (2020) which has been successfully applied to simulate streamflow in hundreds of basins across the CONUS. The framework takes meteorological forcings and basin attributes as inputs and generates daily streamflow

151 predictions for each basin at each time step (Figure 2a). We used mini-batches to train the LSTM model, where each minibatch
 152 was composed of two-year sequences from 256 randomly-selected basins. The first-year sequences are only used for
 153 initializing the cell states, so we calculate the batch loss function only on the second-year sequences. The training sequences
 154 were also randomly selected from the whole training period, and one epoch was finished when the model had seen all the
 155 training data. Note that this sequence length is a subset of, and different concept from, the length of training period. Sequence
 156 length specifically refers to the length of the training instance that comprises a minibatch, whereas training period refers to the
 157 whole period when observations are available for training, from which the minibatch sequence length is randomly selected.
 158 The model was forwarded on each minibatch iteratively and its weights were updated using gradient descent after each
 159 forwarding. One epoch was considered to have occurred when the model is iterated over all the training data. We trained the
 160 LSTM model for 300 epochs to achieve convergence.



161

162 **Figure 2. Illustrations of two different types of regionalized hydrologic models. (a) Framework of the purely data-driven LSTM**
 163 **streamflow model (adapted from Figure 2 in Feng et al., 2020), and (b) framework of the differentiable HBV model (δ HBV-globe1.0-**
 164 **hydroDL) with parameter regionalization developed in Feng et al. (2022) (adapted from Figure 1 in Feng et al. (2022)). The neural**

165 *network g_A here is a LSTM unit which is trained by the observed streamflow to produce the static or dynamic physical HBV parameters*
166 *(θ, β, γ) from basin characteristics.*

167 **2.3 The hybrid differentiable hydrologic models**

168 We used the hybrid differentiable models (δ HBV-globe1.0-hydroDL) developed in Feng et al., (2022) for regionalized
169 modeling in global basins. The HBV model used here as the physical backbone is a conceptual hydrologic model with
170 representations of snowpack, soil, and groundwater storages, and can simulate flux variables such as snow melting,
171 evapotranspiration, and quick and slow outflows (Beck et al., 2020b; Bergström, 1976, 1992; Seibert and Vis, 2012). The
172 differentiable parameter learning (dPL) framework (Tsai et al., 2021) is used to provide parameter regionalization for HBV,
173 as shown by the g_A neural network in Figure 2b. The g_A network, which is a LSTM unit here, takes basin attributes and
174 meteorological forcings as inputs, and outputs static or dynamic physical HBV parameters. The differentiable HBV model
175 then takes these parameters as well as the meteorological forcings to simulate the hydrologic process and predict daily
176 streamflow discharge along with other key flux variables. The whole framework including HBV itself was implemented in a
177 DL platform (PyTorch 1.0.1 was used for the original development and the model has also shown good compatibility with
178 more recent PyTorch versions, (Paszke et al., 2017)) supporting automatic differentiation and trained with gradient descent to
179 minimize the difference between the simulated and observed streamflow (the loss function). As in Feng et al., (2022), we
180 employed the loss function based on root-mean-square error (RMSE) with two weighted parts. The first part calculates RMSE
181 directly on the simulated and observed discharge, while the second part calculates RMSE on the transformed discharge records
182 to improve low flow representations. Note that we do not directly train the HBV parameters; rather, we focus on training the
183 weights of the g_A neural network to map the relationship between basin-averaged characteristics and HBV parameters.
184 Differentiable models are also trained in mini-batches that are formed in the same way as for training the LSTM streamflow
185 model. Within one epoch, differentiable models are forwarded and optimized over the randomly formed mini-batches until the
186 iterations have used all the training data points. We train the differentiable models for 50 epochs in total.

187

188 As described in Feng et al. (2022), the differentiable modeling framework enables optional modification of the structures of
189 the original HBV model to enable better performance and we use two versions of evolved HBV models in this study. We used
190 16 parallel subbasin-scale response units, each with a separate set of parameters to describe a fraction of the basin with different
191 hydrologic responses. These components implicitly represent subbasin-scale spatial heterogeneity. The simulated fluxes (e.g.,
192 streamflow) are the average of all the response units. The parameters of the multiple components are different and all are
193 produced simultaneously by the same g_A network. The first version of our model (referred to as “dPL + evolved HBV”) only
194 has static parameters which are kept constant during the hydrologic simulation. The second version (referred to as “dPL +
195 evolved HBV with DP) further allows some formerly static parameters of the multi-component model to vary daily with the
196 meteorological forcings. These dynamic parameters (DP) were also produced by the g_A LSTM unit. If we were to apply the
197 dynamic parameterization to all parameters, the model could become overly flexible, potentially leading to overfitting to the
198 training data (which would lead to issues with extrapolation beyond the training data). To reduce the risk of overfitting, we

restricted the dynamism to only two empirical parameters: the shape coefficient β in the equation that describes the relationships between soil storage and potential runoff, and a newly added shape parameter (γ) which is involved in the calculation of evapotranspiration. For more details regarding these differentiable HBV models, please refer to our previous studies (Feng et al., 2022, 2023).

2.4 Experiments and evaluation metrics

We ran one temporal and two spatial generalization experiments to evaluate the performance of different regionalized models. For the temporal generalization experiment, the models were trained for the period of 2000 to 2016 on all global basins, and tested for the period of 1980 to 1997. Basins without discharge records or with less than 5 years' worth of data points in the testing period were excluded from the evaluation. Without spatially holding out any basin during training, this experiment aimed at evaluating the model's generalizability in the time dimension by testing prediction ability on the same basins but in a different time period from the training data. The other two spatial generalization experiments served as the true litmus tests for evaluating the effectiveness of regionalization schemes, i.e., how well the model can be applied to basins that have never been seen during training. The first spatial generalization experiment was a traditional "prediction in ungauged basins" (PUB) problem, where we randomly divided the whole global basin set into 10 folds (groups) and performed cross-validation across these folds to obtain spatial out-of-sample predictions for all basins (training on 9 of the folds with the 10th fold held out and testing on the 10th, then rotating such that each fold is used for testing once). The second spatial generalization experiment, which we refer to as cross-continent "prediction in ungauged regions" (PUR), was more challenging. In this experiment, we assumed that all the basins in certain continents are ungauged and excluded from the training dataset, trained a regionalized model in other data-rich continents, and then tested the trained model to make predictions in the ungauged continents. With random hold-out, an ungauged test basin in the first spatial generalization experiment always has training gauges surrounding it. Therefore, the first PUB experiment can be interpreted as spatial interpolation. The second spatial experiment (cross-continent PUR) holds out all the basins in one continent as testing targets, and thus is the much harder test of spatial extrapolation.

To evaluate the overall performance of the hydrologic models, we used the Kling-Gupta Efficiency (KGE) (Gupta et al., 2009; Kling et al., 2012) as compared in Beck et al., (2020b) and Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970). KGE has three components that account for correlation, mean bias (the ratio of simulated and observed means), and variability bias (the ratio of simulated and observed coefficients of variation), while NSE mainly represents the variance explained by the simulations. Both metrics indicate better performance when their values are closer to the maximum value of 1. We also examined the percent bias of the top 2% peak flow range (FHV) and bottom 30% low flow range (FLV) of streamflow predictions to evaluate the model's ability to simulate extreme events (Yilmaz et al., 2008). All the reported performance metrics in this study are from model evaluation on the testing dataset, which is not seen by the model during the training process.

232 **3. Results and discussions**

233 **3.1 General patterns over global basins**

234 From the standpoint of daily hydrograph metrics (KGE and NSE), LSTM and the two differentiable models all achieved highly
235 competitive performance for the global basins in the temporal test (trained and tested on the same basins, but in different time
236 periods) (Figure 3). For the global dataset, all three models obtained median KGE values close to or higher than 0.7, but the
237 LSTM model performed the best of the three models here, achieving a median NSE (KGE) value of 0.70 (0.74) for all the
238 evaluated basins. For a subset of 1675 basins with long-term records (at least 15 years' worth of streamflow data available in
239 the training period and 5 years' worth of data available in the testing period, though not necessarily continuous), LSTM even
240 reached a median KGE of 0.78 (see Figure A1). Both versions of the differentiable models approached the performance level
241 of the LSTM, in agreement with our previous assessment for the CONUS (Feng et al., 2022). The model with dynamic
242 parameters achieved a median NSE (KGE) of 0.67 (0.69), followed by the model with static parameters, which obtained a
243 median NSE (KGE) of 0.65 (0.68).

244
245 The LSTM exhibited advantages for the low flow predictions compared with the differentiable models, as shown by the FLV
246 metric (Figure 3). However, for the peak flow predictions, the LSTM and differentiable models were quite similar, and they
247 all underestimated the observed peaks (FHV in Figure 3). The underestimation for peak flows is consistent with what was
248 found in previous studies. For example, all the physical and deep learning models have significant negative peak flow bias
249 when benchmarked in the CONUS dataset (Feng et al., 2020; Kratzert et al., 2019b). We hypothesize that the systematic
250 underestimation of peaks may be partially related to bias in precipitation forcings. MSWEP is based on the ERA5 reanalysis,
251 which is known to underestimate precipitation peaks (Beck et al., 2019). Furthermore, the use of basin-averaged, daily-
252 averaged precipitation may further suppress the peaks (Chen et al., 2017). In addition, the errors with peak flow could also be
253 partly due to some numerical and structural issues with the differentiable models, e.g., numerical errors introduced by the
254 explicit and sequential solution scheme of HBV with excessive use of threshold functions that lead to different results when
255 the sequence changes, and structure limitations, e.g., deeper groundwater storage cannot feed back to the upper layers. Given
256 the commonality of this issue, we call for community efforts and collaboration to address this issue.

257
258 Both the LSTM model and the differentiable models performed well over diverse landscapes, including North America
259 (especially along the Rocky and Appalachian Mountain ranges and the Southeastern Coastal Plains), Western Europe, Asia
260 (mostly Japan), the southern part of Brazil, and the northeast coast of Australia (Figure 4a and b). There are other regions
261 where none of the three models performed well, such as the longitudinally-central part of North America (Great Plains and
262 Interior Lowlands), the southern edge of Chile (with many glaciers), the Tasmania state of Australia, and the few basins in
263 Africa. These regions, for example, the Northern Great Plains and the state of Texas in the CONUS, have always been difficult
264 for all kinds of models, likely due to incorrect basin boundary, highly localized precipitation, the dry conditions with small

runoff amounts and flash flooding mechanisms (Berghuijs et al., 2014; Driscoll et al., 2002; Feng et al., 2020; Martinez and Gupta, 2010; Newman et al., 2017), to be explored below. Despite some challenges, however, these values represent currently the best metrics reported at the global scale compared to earlier studies, e.g., (Alfieri et al., 2020; Beck et al., 2017a, 2020b; Hou et al., 2023), attesting to these models’ great potential as global modeling tools.

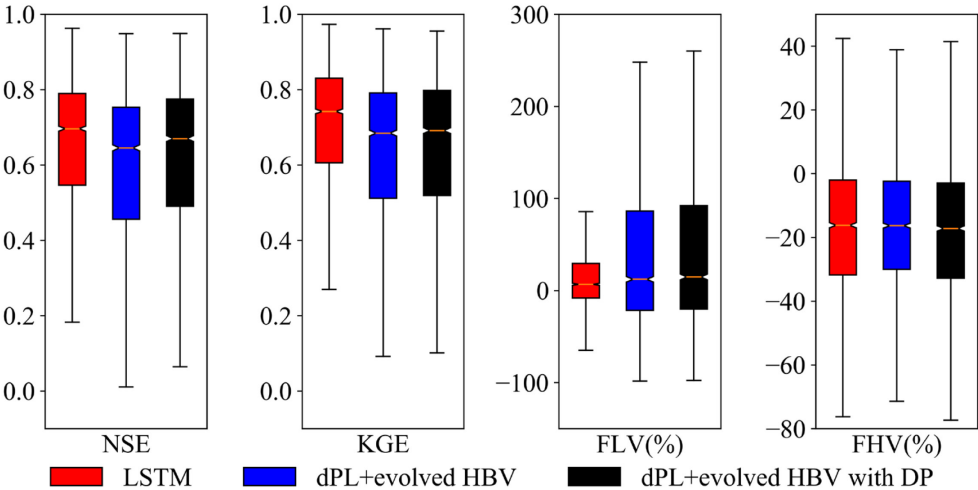
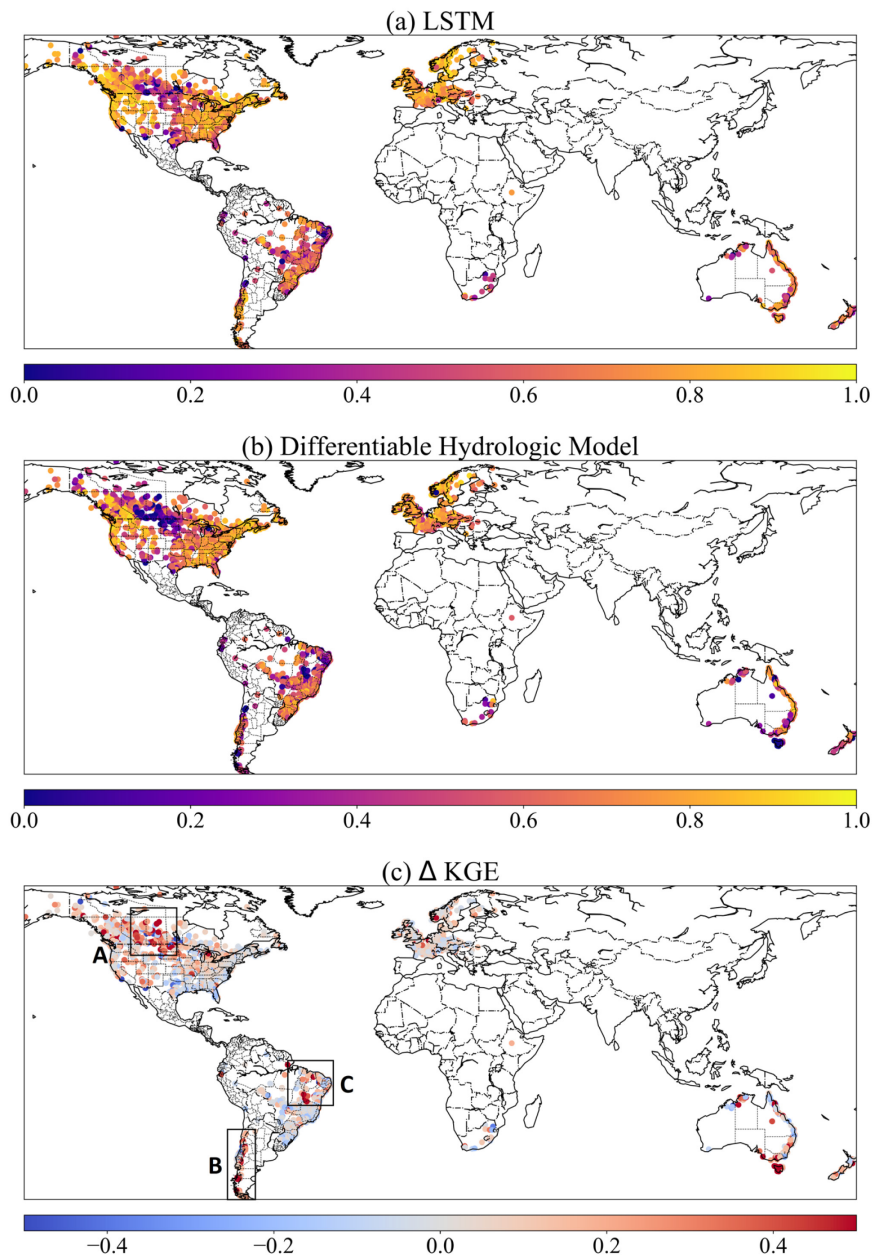


Figure 3. Performance comparison between the LSTM and differentiable models on global basins. dPL refers to the differentiable parameter learning framework, while “evolved HBV” refers to some modifications to improve the standard HBV model, and “with DP” indicates that some parameters were allowed to be dynamic rather than static. Here, the horizontal line inside the colored box represents the median, while the top and bottom of the colored box indicate the first and third quartiles. The bars extending from the colored boxes indicate 1.5 times the interquartile range from the first and third quartiles. NSE is Nash-Sutcliffe Efficiency, KGE is Kling-Gupta Efficiency, FLV indicates the model’s percent bias on the bottom 30% low flow range of streamflow, and FHV indicates percent bias on the top 2% peak flow range of streamflow.



279

280

281

282

Figure 4. The spatial patterns of different model performance and their differences shown by KGE metric. (a) the LSTM model; (b) the differentiable model with dynamic parameters (dPL + evolved HBV with DP); and (c) the KGE difference between two models (KGE of LSTM – KGE of dPL + evolved HBV with DP). Plotted in Python using Matplotlib Basemap Toolkit.

283 3.2 Model behaviors and limitations across climate groups and regions

284 All three models' performances vary significantly across different climate groups of the global basins (Figure 5), revealing
285 their strengths and limitations. The LSTM model behaved the best in the polar, cold, and temperate groups, while the
286 performance deteriorated in the tropical and arid basins. Similar to LSTM, differentiable models showed strong performance
287 in temperate and cold groups and worse performance in tropical ones, with the worst performance in arid basins. These clusters
288 of challenging basins can also be identified on the map (Figure 4a and b). The differentiable model with dynamic parameters
289 performed better than the model with static parameters in all climate groups except the most challenging arid group. Dynamic
290 parameterization with more structural flexibility generally provides stronger modeling ability, while also showing a higher risk
291 of overfitting and degraded generalizability in basins which are very difficult to simulate. As we examine how LSTM and
292 differentiable models behave differently, we find that such differences can be attributed to processes missing from the simple
293 backbone process-based model (HBV here) as explained below. Here we use LSTM as an indicator of upper bound, that is, it
294 shows the ideal performance of a model, given the available information from forcing and input data. Thus the distance from
295 LSTM indicates either systematic and predictable forcing errors (which can be remediated by LSTM) or structural issues with
296 the differentiable model.

297

298 For example, the polar group stands out as a climate type favoring LSTM, while the cold group shows a similar but less
299 pronounced contrast, both of which may be related to HBV's physical deficiencies and forcing issues with snow undercatch.
300 For the polar (cold) groups, LSTM surprisingly had a median KGE of 0.81 (0.78) while the differentiable model only reached
301 0.62 (0.71). The polar regions include, for example, Southern Chile (in region B in Figure 4c). As glaciers can store water for
302 extended periods of time and are driven mostly by temperature rather than rainfall, it is possible for LSTM to capture the
303 temperature-driven dynamics (Lees et al., 2022) while the original HBV itself does not have a glacial module. HBV does not
304 have the ability to simulate frozen soil, sublimation or snow cover fractions. Furthermore, as snow gauges in high altitude are
305 known to suffer systematic bias due to undercatch problems (Beck et al., 2020a), LSTM can learn to address such systematic
306 bias while physical differentiable models cannot due to mass balance. For the cold regions, e.g., high-latitude regions of the
307 North American Great Plains (Region A in Figure 4c --- this also includes the Prairie Pothole Region, or PPR), HBV may
308 suffer from not having descriptions for frozen ground conditions (soil ice) which can influence infiltration, and rainfall
309 underestimation due to undercatch, ice blockage, and other potential reasons (Beck et al., 2020a). In addition, another reason
310 why LSTM and differentiable HBV may have trouble with PPR (but HBV performed especially poorly) is the countless
311 wetlands that store water until full and become connected after snowmelt and large rainfall. HBV does not have modules that
312 can describe such large-scale fill-connect-spill processes (Shaw et al., 2013; Vanderhoof et al., 2017).

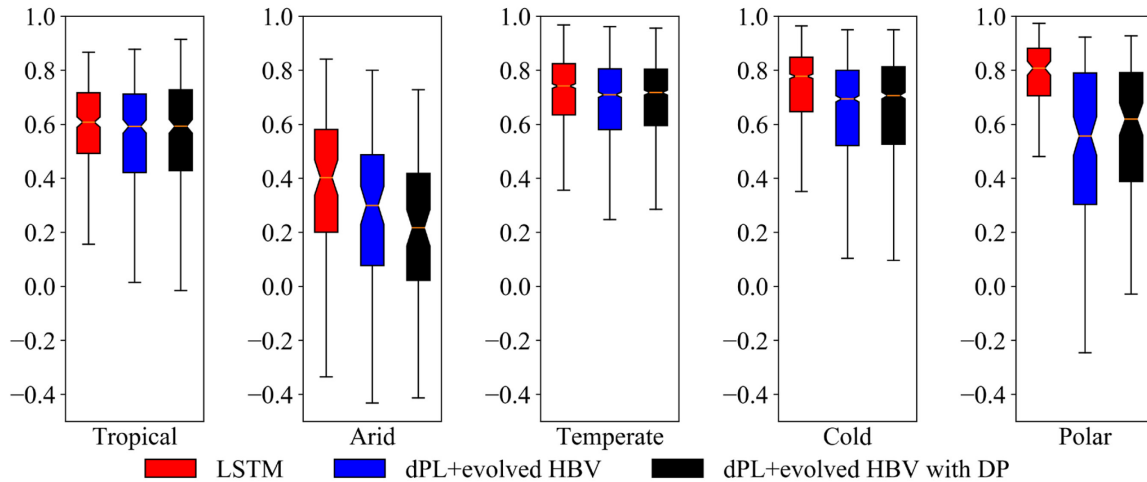
313

314 A more prominent challenge is the arid regions (middle CONUS, north Chile and east Brazil in Figure 1 and Figure 4). This
315 challenge can be attributed to the long duration of low flows which requires long-term memory, and flash floods which result

316 from intense short-duration storms not well represented at the daily scale. Even the LSTM model cannot retain year-long
 317 memory and cannot perform well for the baseflow (Feng et al., 2020). Because HBV has a linear reservoir for its slow-flow
 318 (lowest) bucket, it cannot generate zero base flows. Neither can it well simulate the impact of intense hourly-scale rainfall.
 319 These process improvements need to be considered in the future. Another reason for the challenge in arid regions is the lack
 320 of reservoir management modules. Arid regions tend to have water management infrastructure that significantly influences
 321 streamflow (Veldkamp et al., 2018). Since the HBV model doesn't have any module representing human impacts on the natural
 322 water cycle, the poor performance in middle Brazil in region C may have come from the missing representation of human
 323 interferences. There are large population and intensive agricultural activities in this region which could induce significant
 324 impacts on the hydrologic process. Parameter compensations apparently cannot make up for all the missing mechanisms.

325
 326 The sensitivity of model performance to missing processes in the differentiable models is both good and bad news. It's good
 327 news because this means we can identify suitable or insufficient process representations by learning from data. On the other
 328 hand, this means more challenges as we need to increase the process complexity of this model before it can perform well for
 329 these basins, unlike the purely data-driven LSTM which is not explicitly concerned with physical processes.

330



331

332 **Figure 5. The performance comparison (KGE, Kling-Gupta Efficiency) of different models for five climate groups. dPL refers to the**
 333 **overall differentiable parameter learning framework, while “evolved HBV” refers to some modifications to improve the standard HBV**
 334 **model, and “with DP” indicates that some parameters were allowed to be dynamic rather than static. Here, the horizontal line inside the**
 335 **colored box represents the median, while the top and bottom of the colored box indicate the first and third quartiles. The bars extending**
 336 **from the colored boxes indicate 1.5 times the interquartile range from the first and third quartiles.**

337 3.3 Spatial generalization for prediction in ungauged regions

338 While LSTM maintains mild advantages over differentiable models in data-dense settings, it was outperformed by
 339 differentiable models in a highly data-scarce scenario. As mentioned above, the data-dense setting was tested in the randomized

340 holdout test called prediction in ungauged basins (PUB), while the data-scarce scenario was tested in the regional holdout test,
341 or prediction in ungauged regions (PUR). In the global PUB test, LSTM has a small edge (median KGE=0.67) over
342 differentiable models (median KGE=0.64). Both were noticeably higher than the traditional regionalization method using
343 linear transfer functions reported by Beck et al. (2020b) (Beck20, median KGE=0.46), which already represents the previous
344 state-of-the-art performance of global parameter regionalization. Differentiable modeling does not rely on strong assumptions
345 of the functional form for the parameter transfer function. It leverages the powerful ability of neural networks to represent
346 complicated functions, and automatically learns robust and generalizable relationships between geographic attributes and
347 physical model parameters from large data. Therefore, we can expect significant performance advantages from differentiable
348 modeling compared to traditional methods relying on linear transfer functions. In the PUR scenario where European basins
349 were held out for testing, differentiable models (median KGE=0.58) performed significantly better (p-value less than 0.01
350 using the one-sided Wilcoxon signed-rank test) than LSTM (median KGE=0.52). In the South American PUR experiment,
351 lower performance was seen for all models which can be expected considering the prediction difficulties in this region even
352 for the in-sample scenario (Region B and C in Figure 4). The median KGE of LSTM is 0.28 while the differentiable model
353 with static parameters achieves a higher median KGE of 0.31 for the PUR scenario. It seemed that the differentiable model
354 with dynamic parameterization was somewhat overfitted in this case, resulting in a median KGE that was lower than the static-
355 parameter differentiable model. We do not have PUR results from traditional models available to compare against, since this
356 is a very challenging issue for traditional regionalization methods to make predictions across continents.

357

358 With these results, we show that differentiable models have demonstrated a high simulation capability that cannot be obtained
359 with traditional parameter regionalization approaches, and also provide a robust extrapolation capability in large data-sparse
360 regions that is stronger than purely data-driven models like LSTM. This conclusion was not only verified in the USA, but now
361 has also been confirmed in global catchments with generalization tests including prediction in neighboring ungauged basins
362 and cross-continent predictions, each of which have different conditions with respect to data availability and density.

363

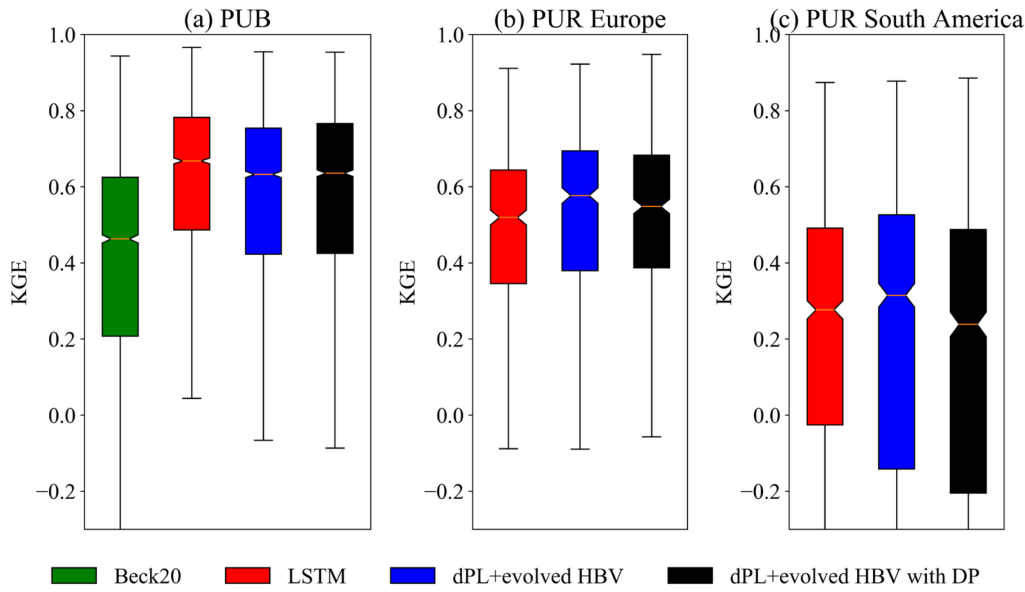


Figure 6. The performance comparison (KGE, Kling-Gupta Efficiency) of different models for spatial generalization tests. (a) Random hold-out test for prediction in ungauged basins (PUB), (b) and (c) holding out all the basins in Europe or South America, respectively, for cross-continent predictions in ungauged regions (PUR). Beck20 refers to a traditional regionalization method using linear transfer functions (Beck et al., 2020b), LSTM is the purely data-driven long short-term memory network, dPL refers to the differentiable parameter learning framework, while “evolved HBV” refers to some modifications to improve the standard HBV model, and “with DP” indicates that some parameters were allowed to be dynamic rather than static. Here, the horizontal line inside the colored box represents the median, while the top and bottom of the colored box indicate the first and third quartiles. The bars extending from the colored boxes indicate 1.5 times the interquartile range from the first and third quartiles.

3.4 Predicting untrained variables

The evapotranspiration (ET) simulations from differentiable models are consistent with independent MODIS satellite estimates of ET in both temporal dynamics and spatial patterns. We did not use any ET observations as training targets to supervise the differentiable models. At the global scale, the mean annual ET comparison shows overall consistency with MODIS, with most basins lying close to the 1:1 line and a correlation of 0.75 for all the basins (Figure 7a). Spatially, the model was able to represent energy limitations in the cold regions, e.g., high-latitude North America and Europe, and water limitations, e.g., southwestern US and arid basins of Australia (Figure 7a and b). The model also represented high ET in basins adjacent to the Amazon forest, those along the US southeastern and Australian coast. Temporally, the median correlation of ET time series between simulations and MODIS products achieves 0.82 and 0.89 for two differentiable models in 3753 basins, respectively (Figure 7c).

The ET simulations show high correlation with MODIS in most North American and European basins (Figure 7d), in line with the good performance of streamflow modeling in these regions. However, the correlation is relatively lower in South America but the coefficient of variation of ET residuals (CoV, the ratio of standard deviation of ET residuals to the annual mean) is also

387 small (Figure 7e), in part because the ET here is large and less driven by the seasonal energy cycle (Niu et al., 2017). MODIS
388 ET itself is not the ground truth and always has large uncertainties in Amazonia regions due to the cloud coverage and
389 difficulties for observation (Hilker et al., 2015; Xu et al., 2019). Furthermore, the simulations could be negatively influenced
390 by the data quality issues with streamflow records in these regions. Upon examining the records, some stations in South
391 America show unrealistic hydrographs that may indicate data processing errors. To address such issues in the future, more in-
392 depth data screening and correction or constraining the model using datasets other than streamflow, e.g., eddy covariance flux
393 data, should be considered. The CoV is less than 0.3 for most of the world, showing that ET errors are mostly small relative to
394 its annual averages (Figure 7e). Noticeable exceptions are US southwest, where ET varies strongly from year to year and is
395 highly dependent on the precipitation, and Chile, where glaciers and deserts are both present, posing challenges to the model.
396 As the present study is basin-focused, we will leave the evaluation of global gridded ET to future work.
397

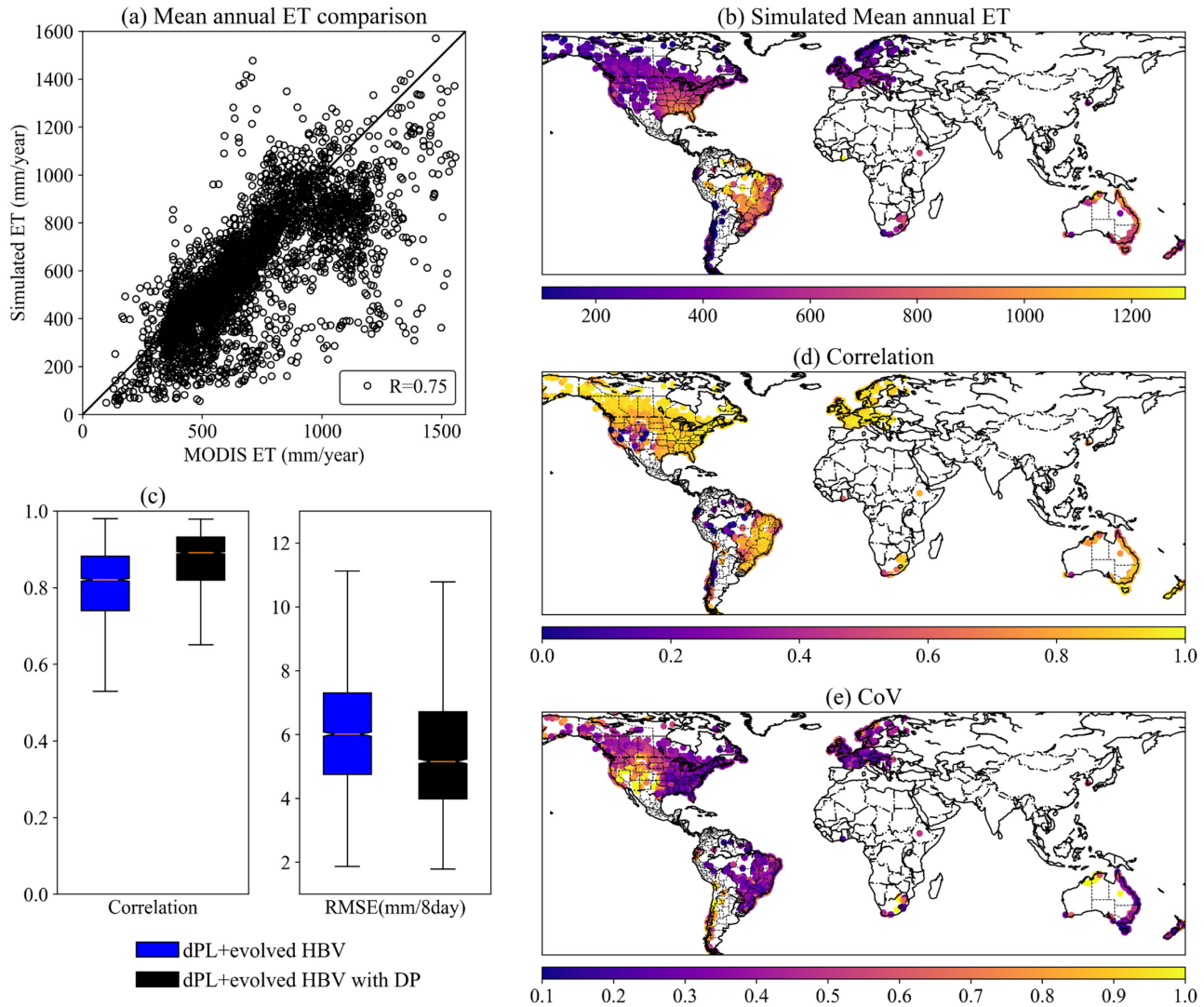


Figure 7. The comparison between simulated ET from the differentiable hydrologic models and independent MODIS ET product. (a) mean annual ET comparison, (b) simulated mean annual ET for global basins, (c) boxplots for the temporal dynamic evaluation by correlation and RMSE, (d) correlation and (e) coefficient of variation for ET comparison in global basins. Maps plotted in Python using Matplotlib Basemap Toolkit.

3.5 Further discussion

Compared to the LSTM model which only outputs discharge simulations, differentiable models offer a suite of interpretable variables including ET, soil water, recharge, baseflow, etc., thus providing a comprehensive description for the hydrologic cycle and far better interpretability. To create a new differentiable model or turn an existing model into a differentiable one, we need to implement the model on a differentiable platform like PyTorch, Tensorflow, or JAX, while better enabling model parallelism in order to maximally leverage the computing power of modern graphical processing units (GPUs). If a model

contains mostly explicit calculations, automatic differentiation (AD) offered by the above platforms can effortlessly provide gradient calculations, requiring only a syntax-level translation which can nowadays be done easily. Sometimes, a limited amount of adjustments are needed to turn non-differentiable operations into equivalent differentiable ones. However, when a model contains iterative solutions to nonlinear systems, large matrix solvers or constrained optimizations, we can employ the adjoint method (Song et al., 2023). The adjoint method explicitly defines the gradient-calculation method and alters the order of calculations so iteration is avoided during gradient calculations, which can dramatically reduce memory demand and improve efficiency. Another important consideration is the effective use of parallelism and the modern computing infrastructure for AI (i.e., GPUs). In our context, the regionalized parameterization (in this case, training one neural network on a large amount of basins), which is crucial to ensuring the generalizability of the model, requires going through large data in high-throughput parallelism. Embracing parallelism may necessitate some coding adjustments. At this point, several versions of differentiable hydrologic models have been proposed with varying complexities and different handling of parameterization, post-processing (which we didn't use in this study, as it can interfere with interpretability of the internal variables, mass balances, and the sensitivity to inputs encoded by the process-based components), and dynamical parameters. Across geoscientific domains, differentiable ecosystem (Aboelyazeed et al., 2023; Zhao et al., 2019), flow and routing (Bindas et al., 2024), water quality (Rahmani et al., 2023), and ice sheet (Bolibar et al., 2023) models have already been demonstrated.

The challenges facing the differentiable models in this study include not only missing processes like reservoir management, ground ice, and glaciers, but also large errors in meteorological forcings and streamflow target data. Substantial bias could exist in precipitation, e.g., due to snow-gauge undercatch (Hou et al., 2023), or in discharge, e.g., streamflow are measured using different approaches which exhibit large variability; for another example, gridded climate forcing data often consistently underestimate the magnitudes of heavy storms (Beck et al., 2017b). While LSTM can easily adapt to systematic bias, such forcing errors put the differentiable models under stress because they cannot reconcile streamflow observations with such forcings given the constraint of mass balances. If our objective is to learn core physics and parameterizations that are reliable despite forcing discrepancies, we can set up forcing data correction layers that can, to some extent, shield the core processes from being influenced by such errors. This will be an important aspect of future work to ensure reliable prediction of future water resources.

The backbone of a differentiable process-based model thus serves as a double-edged sword: when such backbones are essentially correct, they serve as a stabilizing element of the model that mitigates overfitting and improves generalization; when they lack critical processes or when observations have large, unexplained bias, they can drag down model performance and cause compensation between processes. However, the limitations are tractable: future work can gradually incorporate critical processes and include more observations to constrain the learning process, making sure each addition is valuable and accretive. The research community collectively has already substantial experience in evolving earth system models to include many processes. We expect some processes to be invited back in the differentiable modeling framework. Nevertheless, with

differentiable modeling, we now have a new tool that was not previously available: highly flexible deep neural networks that can be placed anywhere in the model, which provide a systematic way of managing model complexity. With their help, such model evolution may take much less time than previously required. However, we still expect the development cycle to take longer than that for purely data-driven models like LSTM, requiring us to view differentiable models as evolving rather than static entities, which need a bit of patience while maturing.

This study builds a benchmark and a basis for model selection and diagnosis for the next-generation global hydrologic modeling, which previously did not learn from such large observations. With rigorous tests at the global scale, this study proves that differentiable models are strong candidates as global water models. With powerful spatial generalization ability, they can be applied to characterizing the hydrologic processes in ungauged regions by leveraging learned information in data-rich continents. Differentiable models in this study have already learned the generalizable and robust relationships between geographic attributes and physical model parameters from thousands of global catchments. Therefore, these models can be easily applied towards providing seamless global hydrologic modeling with parameters directly generated from worldwide geographic attributes. Future work can use such models and continuously improved observational datasets to produce global hydrologic fluxes while enhancing some process representations in extremely arid, glaciated, or heavily human-influenced basins.

4. Conclusions

In this work, we used both purely data-driven models and, for the first time, physics-informed, differentiable models to simulate rainfall-runoff processes in 3753 global basins. Both types of models achieved overall highly competitive performance for global basins with diverse climate conditions, yielding median KGE values close to or higher than 0.7 which is state-of-the-art at this large scale. The LSTM still achieved the best performance for the temporal generalization test, but the differentiable HBV models with evolved structure (δ HBV-globe1.0-hydroDL) approach the LSTM's performance level. Furthermore, the spatial generalization experiments highlighted the stronger regionalization and extrapolation ability of differentiable models than the traditional modeling approach and LSTM, demonstrating its promise to be applied to data-scarce regions in the world. River routing is not included in this work and will be investigated in the future, possibly also with differentiable approaches (Bindas et al., 2022).

Different models appear to have generally consistent spatial performance patterns, though obvious distinctions stand out in several local regions. All models achieve good performance in the temperate and cold climate groups, while they all behave unsatisfactorily in the arid group. For the polar group, the differentiable model performed significantly worse than the LSTM. Without any physical constraints, LSTM shows strong power in simulating storage (snow and glacier) dominated processes, while differentiable models are limited by the structure of their physical backbone model, which in this case does not simulate

multiyear ice buildup and melt. Another limitation could be soil sealing processes in extremely arid regions. These regional performance comparisons thus reveal some deficiencies of the physical backbone in δ HBV that cannot be mitigated even by advanced neural network-based parameterization. These insights provide directions for future improvements. Different from purely data-driven models only trained by the target variable, differentiable models constrained by the physical backbone can give accurate simulations for a full set of hydrologic variables in the water cycle including evapotranspiration, snow water equivalent, water storage, infiltration and baseflow. As some process limitations are addressed in the future, we believe differentiable models will be strong candidates for the next generation global water models to characterize and predict the hydrologic processes in ungauged regions across the world.

Appendix

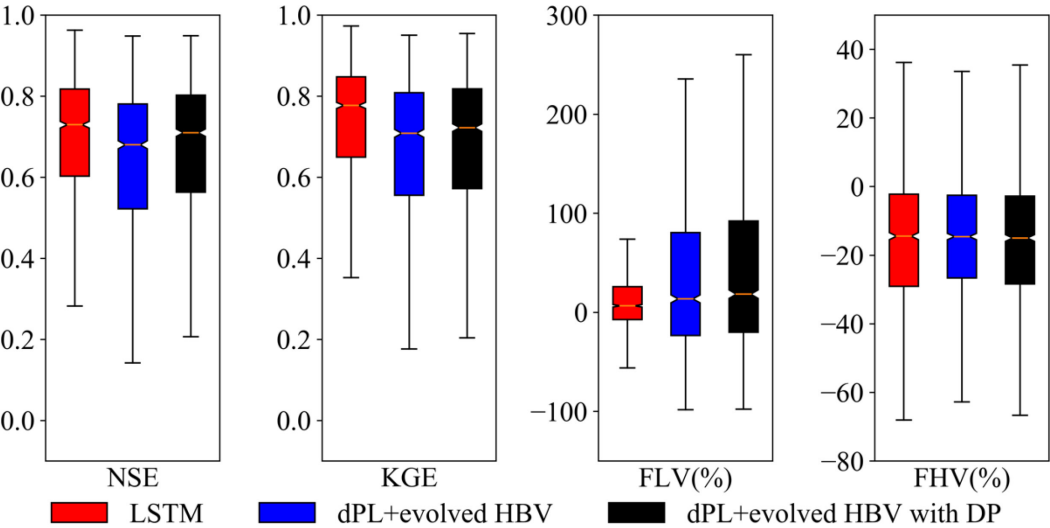


Figure A1. Performance comparison on the 1675 subset basins with long-term streamflow records (at least 15 years' worth of streamflow data available in the training period and 5 years' worth of data available in the testing period, not necessarily continuous). Other items are the same as in Figure 3.

Author contributions

DF and CS conceived this study. DF set up the hydrologic models and ran all the experiments. DF and CS performed the major analysis, with HB, JdB, RKS, YS, YW and MP contributing substantially to the discussions on the methodology and results. HB provided the benchmark results of the traditional parameter regionalization scheme. JL prepared the ET product for comparison. DF wrote the initial draft and CS revised the manuscript. HB, JdB, RKS, YS, YW, and KL substantially edited the manuscript.

495 **Financial support**

496 DF was supported by the National Science Foundation Award EAR-2221880. This work was also partially supported and
497 inspired by the Young Scientists Summer Program (YSSP) of International Institute for Applied Systems Analysis (IIASA).
498 JL was supported by Google.org’s AI Impacts Challenge Grant 1904-57775. CS and KL were supported by Cooperative
499 Institute for Research to Operations in Hydrology (CIROH), award number A22-0307-S003. Computation was partially
500 supported by the National Science Foundation Major Research Instrumentation Award PHY-2018280.

501 **Competing interests**

502 Kathryn Lawson and Chaopeng Shen have financial interests in HydroSapient, Inc., a company which could potentially benefit
503 from the results of this research. This interest has been reviewed by The Pennsylvania State University in accordance with its
504 individual conflict of interest policy, for the purpose of maintaining the objectivity and the integrity of research at The
505 Pennsylvania State University.

506 **Code and Data Availability**

507 The source codes for the differentiable hydrologic models can be accessed at <https://doi.org/10.5281/zenodo.7091334>, and this
508 study evaluates these models at global scale. The MOD16A2GF ET product can be downloaded at
509 <https://lpdaac.usgs.gov/products/mod16a2gfv061/>. Meteorological forcing datasets MSWEP and MSWX can be downloaded
510 at <https://www.gloh2o.org/mswep/> and <https://www.gloh2o.org/mswx/>, respectively. The streamflow observations used in this
511 study were initially compiled by Beck et al., (2020b) and can be accessed from the original data sources including the United
512 States Geological Survey (USGS) National Water Information System (NWIS; <https://waterdata.usgs.gov/nwis>), the Global
513 Runoff Data Centre (GRDC; <https://grdc.bafg.de>), the HidroWeb portal of the Brazilian Agência Nacional de Águas
514 (<https://www.snirh.gov.br/hidroweb>), the European Water Archive (EWA) of EURO-FRIEND-Water
515 (https://www.bafg.de/GRDC/EN/04_spcldtbss/42_EWA/ewa.html) and the CCM2-JRC CCM River and Catchment Database
516 (<https://data.jrc.ec.europa.eu/collection/ccm>), Water Survey of Canada (WSC) National Water Data Archive (HYDAT;
517 <https://wateroffice.ec.gc.ca/>), the Australian Bureau of Meteorology (BoM; <http://www.bom.gov.au/waterdata/>), and the
518 Chilean Center for Climate and Resilience Research (CR2) website (<https://www.cr2.cl/datos-de-caudales/>).

519 **References**

520 Aboelyazeed, D., Xu, C., Hoffman, F. M., Liu, J., Jones, A. W., Rackauckas, C., Lawson, K., and Shen, C.: A differentiable,
521 physics-informed ecosystem modeling and learning framework for large-scale inverse problems: demonstration with
522 photosynthesis simulations, *Biogeosciences*, 20, 2671–2692, <https://doi.org/10.5194/bg-20-2671-2023>, 2023.

523 Alfieri, L., Lorini, V., Hirpa, F. A., Harrigan, S., Zsoter, E., Prudhomme, C., and Salamon, P.: A global streamflow reanalysis
524 for 1980–2018, *Journal of Hydrology X*, 6, 100049, <https://doi.org/10.1016/j.hydroa.2019.100049>, 2020.

525 Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff
526 from 10 state-of-the-art hydrological models, *Hydrology and Earth System Sciences*, 21, 2881–2903,
527 <https://doi.org/10.5194/hess-21-2881-2017>, 2017a.

528 Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F.,
529 Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and
530 hydrological modeling, *Hydrology and Earth System Sciences*, 21, 6201–6217, <https://doi.org/10.5194/hess-21-6201-2017>,
531 2017b.

532 Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-
533 hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, *Hydrology and Earth
534 System Sciences*, 21, 589–615, <https://doi.org/10.5194/hess-21-589-2017>, 2017c.

535 Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Dijk, A. I. J. M. van, McVicar, T. R., and Adler, R. F.:
536 MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, *Bulletin of the American
537 Meteorological Society*, 100, 473–500, <https://doi.org/10.1175/BAMS-D-17-0138.1>, 2019.

538 Beck, H. E., Wood, E. F., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garretón, C., Baez-Villanueva, O. M., Sheffield,
539 J., and Karger, D. N.: Bias correction of global high-resolution precipitation climatologies using streamflow observations from
540 9372 catchments, *Journal of Climate*, 33, 1299–1315, <https://doi.org/10.1175/JCLI-D-19-0332.1>, 2020a.

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

544 Beck, H. E., van Dijk, A. I. J. M., Larraondo, P. R., McVicar, T. R., Pan, M., Dutra, E., and Miralles, D. G.: MSWX: Global
545 3-hourly 0.1° bias-corrected meteorological data including near real-time updates and forecast ensembles, *Bulletin of the
546 American Meteorological Society*, 103, E710–E732, <https://doi.org/10.1175/BAMS-D-21-0145.1>, 2022.

547 Berghuijs, W. R., Sivapalan, M., Woods, R. A., and Savenije, H. H. G.: Patterns of similarity of seasonal water balances: A
548 window into streamflow variability over a range of time scales, *Water Resources Research*, 50, 5638–5661,
549 <https://doi.org/10.1002/2014WR015692>, 2014.

550 Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, PhD Thesis, Swedish
551 Meteorological and Hydrological Institute (SMHI), Norköping, Sweden, 1976.

552 Bergström, S.: The HBV model - its structure and applications, Swedish Meteorological and Hydrological Institute (SMHI),
553 Norrköping, Sweden, 1992.

554 Bindas, T., Tsai, W.-P., Liu, J., Rahmani, F., Feng, D., Bian, Y., Lawson, K., and Shen, C.: Improving river routing using a
555 differentiable Muskingum-Cunge model and physics-informed machine learning, *Water Resources Research*, 60,
556 e2023WR035337, <https://doi.org/10.1029/2023WR035337>, 2024.

557 Bolibar, J., Sapienza, F., Maussion, F., Lguensat, R., Wouters, B., and Pérez, F.: Universal differential equations for glacier
558 ice flow modelling, *Geoscientific Model Development*, 16, 6671–6687, <https://doi.org/10.5194/gmd-16-6671-2023>, 2023.

559 Burek, P., Satoh, Y., Kahil, T., Tang, T., Greve, P., Smilovic, M., Guillaumot, L., Zhao, F., and Wada, Y.: Development of
560 the Community Water Model (CWatM v1.04) – a high-resolution hydrological model for global and regional assessment of
561 integrated water resources management, *Geoscientific Model Development*, 13, 3267–3298, [https://doi.org/10.5194/gmd-13-](https://doi.org/10.5194/gmd-13-3267-2020)
562 3267-2020, 2020.

563 Chen, B., Krajewski, W. F., Liu, F., Fang, W., and Xu, Z.: Estimating instantaneous peak flow from mean daily flow,
564 *Hydrology Research*, 48, 1474–1488, <https://doi.org/10.2166/nh.2017.200>, 2017.

565 Cui, G., Anderson, M., and Bales, R.: Mapping of snow water equivalent by a deep-learning model assimilating snow
566 observations, *Journal of Hydrology*, 616, 128835, <https://doi.org/10.1016/j.jhydrol.2022.128835>, 2023.

567 Driscoll, D. G., Carter, J. M., Williamson, J. E., and Putnam, L. D.: *Hydrology of the Black Hills Area*, South Dakota, 2002.

568 Fang, K. and Shen, C.: Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive
569 data integration kernel, *J. Hydrometeor.*, 21, 399–413, <https://doi.org/10.1175/jhm-d-19-0169.1>, 2020.

570 Fang, K., Shen, C., Kifer, D., and Yang, X.: Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S.
571 using a deep learning neural network, *Geophys. Res. Lett.*, 44, 11,030–11,039, <https://doi.org/10.1002/2017gl075619>, 2017.

572 Fang, K., Pan, M., and Shen, C.: The value of SMAP for long-term soil moisture estimation with the help of deep learning,
573 *IEEE Trans. Geosci. Remote Sensing*, 57, 2221–2233, <https://doi.org/10.1109/TGRS.2018.2872131>, 2019.

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

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

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

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

585 Feng, D., Beck, H., Lawson, K., and Shen, C.: The suitability of differentiable, physics-informed machine learning hydrologic
586 models for ungauged regions and climate change impact assessment, *Hydrology and Earth System Sciences*, 27, 2357–2373,
587 <https://doi.org/10.5194/hess-27-2357-2023>, 2023.

588 Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance
589 criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, 377, 80–91,
590 <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.

591 Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., Hanasaki, N., Heinke, J., Ludwig, F., Voss,
592 F., and Wiltshire, A. J.: Climate change impact on available water resources obtained using multiple global climate and
593 hydrology models, *Earth System Dynamics*, 4, 129–144, <https://doi.org/10.5194/esd-4-129-2013>, 2013.

594 Hansen, L. D., Stokholm-Bjerregaard, M., and Durdevic, P.: Modeling phosphorous dynamics in a wastewater treatment
595 process using Bayesian optimized LSTM, *Computers & Chemical Engineering*, 160, 107738,
596 <https://doi.org/10.1016/j.compchemeng.2022.107738>, 2022.

597 Hargreaves, G. H.: Defining and using reference evapotranspiration, *Journal of Irrigation and Drainage Engineering*, 120,
598 1132–1139, [https://doi.org/10.1061/\(ASCE\)0733-9437\(1994\)120:6\(1132\)](https://doi.org/10.1061/(ASCE)0733-9437(1994)120:6(1132)), 1994.

599 Hattermann, F. F., Krysanova, V., Gosling, S. N., Dankers, R., Daggupati, P., Donnelly, C., Flörke, M., Huang, S., Motovilov,
600 Y., Buda, S., Yang, T., Müller, C., Leng, G., Tang, Q., Portmann, F. T., Hagemann, S., Gerten, D., Wada, Y., Masaki, Y.,
601 Alemayehu, T., Satoh, Y., and Samaniego, L.: Cross-scale intercomparison of climate change impacts simulated by regional
602 and global hydrological models in eleven large river basins, *Climatic Change*, 141, 561–576, [https://doi.org/10.1007/s10584-](https://doi.org/10.1007/s10584-016-1829-4)
603 016-1829-4, 2017.

604 Hilker, T., Lyapustin, A. I., Hall, F. G., Myneni, R., Knyazikhin, Y., Wang, Y., Tucker, C. J., and Sellers, P. J.: On the
605 measurability of change in Amazon vegetation from MODIS, *Remote Sensing of Environment*, 166, 233–242,
606 <https://doi.org/10.1016/j.rse.2015.05.020>, 2015.

607 Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, 9, 1735–1780,
608 <https://doi.org/10.1162/neco.1997.9.8.1735>, 1997.

609 Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., and Fenicia, F.: Improving hydrologic models for predictions and
610 process understanding using neural ODEs, *Hydrology and Earth System Sciences*, 26, 5085–5102,
611 <https://doi.org/10.5194/hess-26-5085-2022>, 2022.

612 Hou, Y., Guo, H., Yang, Y., and Liu, W.: Global Evaluation of Runoff Simulation From Climate, Hydrological and Land
613 Surface Models, *Water Resources Research*, 59, e2021WR031817, <https://doi.org/10.1029/2021WR031817>, 2023.

614 Jayakrishnan, R., Srinivasan, R., Santhi, C., and Arnold, J. G.: Advances in the application of the SWAT model for water
615 resources management, *Hydrological Processes*, 19, 749–762, <https://doi.org/10.1002/hyp.5624>, 2005.

616 Jiang, S., Zheng, Y., and Solomatine, D.: Improving AI system awareness of geoscience knowledge: Symbiotic integration of
617 physical approaches and deep learning, *Geophysical Research Letters*, 47, e2020GL088229,
618 <https://doi.org/10.1029/2020GL088229>, 2020.

619 Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change
620 scenarios, *Journal of Hydrology*, 424–425, 264–277, <https://doi.org/10.1016/j.jhydrol.2012.01.011>, 2012.

621 Konapala, G., Kao, S.-C., Painter, S. L., and Lu, D.: Machine learning assisted hybrid models can improve streamflow
622 simulation in diverse catchments across the conterminous US, *Environ. Res. Lett.*, 15, 104022, [https://doi.org/10.1088/1748-](https://doi.org/10.1088/1748-9326/aba927)
623 9326/aba927, 2020.

624 Kraft, B., Jung, M., Körner, M., Koirala, S., and Reichstein, M.: Towards hybrid modeling of the global hydrological cycle,
625 *Hydrology and Earth System Sciences*, 26, 1579–1614, <https://doi.org/10.5194/hess-26-1579-2022>, 2022.

626 Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in
627 ungauged basins: Exploiting the power of machine learning, *Water Resources Research*, 55, 11344–11354,
628 <https://doi.org/10/gg4ck8>, 2019a.

629 Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and
630 local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrology and Earth System Sciences*,
631 23, 5089–5110, <https://doi.org/10.5194/hess-23-5089-2019>, 2019b.

632 Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–
633 runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual
634 models, *Hydrology and Earth System Sciences*, 25, 5517–5534, <https://doi.org/10.5194/hess-25-5517-2021>, 2021.

635 Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.:
636 Hydrological concept formation inside long short-term memory (LSTM) networks, *Hydrology and Earth System Sciences*, 26,
637 3079–3101, <https://doi.org/10.5194/hess-26-3079-2022>, 2022.

638 Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert,
639 F., O’Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and
640 Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), *Hydrology and Earth
641 System Sciences*, 26, 3537–3572, <https://doi.org/10.5194/hess-26-3537-2022>, 2022.

642 Maidment, D. R.: Conceptual Framework for the National Flood Interoperability Experiment, *JAWRA Journal of the
643 American Water Resources Association*, 53, 245–257, <https://doi.org/10/f97pz3>, 2017.

644 Martinez, G. F. and Gupta, H. V.: Toward improved identification of hydrological models: A diagnostic evaluation of the
645 “abcd” monthly water balance model for the conterminous United States, *Water Resources Research*, 46,
646 <https://doi.org/10.1029/2009WR008294>, 2010.

647 Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O., and Samaniego, L.:
648 Towards seamless large-domain parameter estimation for hydrologic models, *Water Resources Research*, 53, 8020–8040,
649 <https://doi.org/10/gcg2dm>, 2017.

650 Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., and Döll, P.: Sensitivity of simulated
651 global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration,
652 *Hydrology and Earth System Sciences*, 18, 3511–3538, <https://doi.org/10.5194/hess-18-3511-2014>, 2014.

653 Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, *Journal
654 of Hydrology*, 10, 282–290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6), 1970.

655 Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., Nearing, G., Newman, A. J., Mizukami, N., Clark, M.
656 P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a Physically Based Hydrologic Model, *Journal of
657 Hydrometeorology*, 18, 2215–2225, <https://doi.org/10/gbwr9s>, 2017.

658 Niu, J., Shen, C., Chambers, J., Melack, J. M., and Riley, W. J.: Interannual variation in hydrologic budgets in an Amazonian
659 watershed with a coupled subsurface - land surface process model, *Journal of Hydrometeorology*, 18, 2597–2617,
660 <https://doi.org/10.1175/JHM-D-17-0108.1>, 2017.

661 Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A.:
662 Automatic differentiation in PyTorch, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long
663 Beach, CA, 2017.

664 Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C.: Exploring the exceptional performance of a deep
665 learning stream temperature model and the value of streamflow data, *Environ. Res. Lett.*, 16, 024025,
666 <https://doi.org/10.1088/1748-9326/abd501>, 2021.

667 Rahmani, F., Appling, A., Feng, D., Lawson, K., and Shen, C.: Identifying structural priors in a hybrid differentiable model
668 for stream water temperature modeling, *Water Resources Research*, 59, e2023WR034420,
669 <https://doi.org/10.1029/2023WR034420>, 2023.

670 Reichert, P., Ma, K., Höge, M., Fenicia, F., Baity-Jesi, M., Feng, D., and Shen, C.: Metamorphic testing of machine learning
671 and conceptual hydrologic models, *Hydrology and Earth System Sciences*, 28, 2505–2529, [https://doi.org/10.5194/hess-28-](https://doi.org/10.5194/hess-28-2505-2024)
672 2505-2024, 2024.

673 Running, S., Mu, Q., Zhao, M., and Moreno, A.: MODIS/Terra Net Evapotranspiration Gap-Filled 8-Day L4 Global 500m
674 SIN Grid V061, <https://doi.org/10.5067/MODIS/MOD16A2GF.061>, 2021.

675 Saha, G., Rahmani, F., Shen, C., Li, L., and Cibin, R.: A deep learning-based novel approach to generate continuous daily
676 stream nitrate concentration for nitrate data-sparse watersheds, *Science of The Total Environment*, 878, 162930,
677 <https://doi.org/10.1016/j.scitotenv.2023.162930>, 2023.

678 Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package,
679 *Hydrology and Earth System Sciences*, 16, 3315–3325, <https://doi.org/10/f22r5x>, 2012.

680 Shaw, D. A., Pietroniro, A., and Martz, L. w.: Topographic analysis for the prairie pothole region of Western Canada,
681 *Hydrological Processes*, 27, 3105–3114, <https://doi.org/10.1002/hyp.9409>, 2013.

682 Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources scientists, *Water Resources*
683 *Research*, 54, 8558–8593, <https://doi.org/10.1029/2018wr022643>, 2018.

684 Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu,
685 X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y.,
686 Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.:
687 Differentiable modelling to unify machine learning and physical models for geosciences, *Nat Rev Earth Environ*, 4, 552–567,
688 <https://doi.org/10.1038/s43017-023-00450-9>, 2023.

689 Song, Y., Knoben, W. J. M., Clark, M. P., Feng, D., Lawson, K. E., and Shen, C.: When ancient numerical demons meet
690 physics-informed machine learning: adjoint-based gradients for implicit differentiable modeling, *Hydrology and Earth System*
691 *Sciences Discussions*, 1–35, <https://doi.org/10.5194/hess-2023-258>, 2023.

692 Song, Y., Chaemchuen, P., Rahmani, F., Zhi, W., Li, L., Liu, X., Boyer, E., Bindas, T., Lawson, K., and Shen, C.: Deep
693 learning insights into suspended sediment concentrations across the conterminous United States: Strengths and limitations,
694 *Journal of Hydrology*, 639, 131573, <https://doi.org/10.1016/j.jhydrol.2024.131573>, 2024a.

695 Song, Y., Tsai, W.-P., Gluck, J., Rhoades, A., Zarzycki, C., McCrary, R., Lawson, K., and Shen, C.: LSTM-based data
696 integration to improve snow water equivalent prediction and diagnose error sources, *Journal of Hydrometeorology*, 25, 223–
697 237, <https://doi.org/10.1175/JHM-D-22-0220.1>, 2024b.

698 Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C.: From calibration to parameter learning:
699 Harnessing the scaling effects of big data in geoscientific modeling, *Nat Commun*, 12, 5988, [https://doi.org/10.1038/s41467-](https://doi.org/10.1038/s41467-021-26107-z)
700 021-26107-z, 2021.

701 Vanderhoof, M. K., Christensen, J. R., and Alexander, L. C.: Patterns and drivers for wetland connections in the Prairie Pothole
702 Region, United States, *Wetlands Ecol Manage*, 25, 275–297, <https://doi.org/10.1007/s11273-016-9516-9>, 2017.

703 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is All
704 you Need, in: *Advances in Neural Information Processing Systems*, 2017.

705 Veldkamp, T. I. E., Zhao, F., Ward, P. J., Moel, H. de, Aerts, J. C. J. H., Schmied, H. M., Portmann, F. T., Masaki, Y., Pokhrel,
706 Y., Liu, X., Satoh, Y., Gerten, D., Gosling, S. N., Zaherpour, J., and Wada, Y.: Human impact parameterizations in global
707 hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study,
708 *Environ. Res. Lett.*, 13, 055008, <https://doi.org/10.1088/1748-9326/aab96f>, 2018.

709 Werth, S. and Güntner, A.: Calibration analysis for water storage variability of the global hydrological model WGHM,
710 *Hydrology and Earth System Sciences*, 14, 59–78, <https://doi.org/10.5194/hess-14-59-2010>, 2010.

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

714 Xu, D., Agee, E., Wang, J., and Ivanov, V. Y.: Estimation of Evapotranspiration of Amazon Rainforest Using the Maximum
715 Entropy Production Method, *Geophysical Research Letters*, 46, 1402–1412, <https://doi.org/10.1029/2018GL080907>, 2019.

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

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

722 Zhao, W. L., Gentile, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., and Qiu, G. Y.: Physics-constrained
723 machine learning of evapotranspiration, *Geophysical Research Letters*, 46, 14496–14507,
724 <https://doi.org/10.1029/2019gl085291>, 2019.

725 Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., and Li, L.: From hydrometeorology to river water quality:
726 Can a deep learning model predict dissolved oxygen at the continental scale?, *Environ. Sci. Technol.*, 55, 2357–2368,
727 <https://doi.org/10.1021/acs.est.0c06783>, 2021.

728