

Water Resources Research®

RESEARCH ARTICLE







10.1029/2023WR036170

Distributed Hydrological Modeling With Physics-Encoded Deep Learning: A General Framework and Its Application in the Amazon



Key Points:

- A fully differentiable framework that seamlessly integrates physics and deep learning was developed for distributed hydrological modeling
- The framework flexibly fuses multi-source observations and improves the efficiency and accuracy of large-scale hydrological modeling
- The hybrid model for the Amazon Basin exhibits excellent fidelity and physical plausibility and provides insights into the ET process

Chao Wang¹ , Shijie Jiang^{2,3,4} , Yi Zheng^{1,5,6} , Feng Han^{1,5} , Rohini Kumar⁴ , Oldrich Rakovec⁴ , and Siqi Li¹

¹School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China, ²Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany, ³ELLIS Unit Jena, Jena, Germany, ⁴Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany, ⁵Shenzhen Municipal Engineering Lab of Environmental IoT Technologies, Southern University of Science and Technology, Shenzhen, China, ⁶State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, Southern University of Science and Technology, Shenzhen, China

Supporting Information:

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

Correspondence to:

Y. Zheng,
zhengy@sustech.edu.cn

Citation:

Wang, C., Jiang, S., Zheng, Y., Han, F., Kumar, R., Rakovec, O., & Li, S. (2024). Distributed hydrological modeling with physics-encoded deep learning: A general framework and its application in the Amazon. *Water Resources Research*, 60, e2023WR036170. <https://doi.org/10.1029/2023WR036170>

Received 31 AUG 2023
Accepted 1 APR 2024

Author Contributions:

Conceptualization: Shijie Jiang, Yi Zheng

Data curation: Chao Wang

Formal analysis: Chao Wang, Shijie Jiang, Yi Zheng, Rohini Kumar, Oldrich Rakovec

Funding acquisition: Yi Zheng

Investigation: Chao Wang

Methodology: Chao Wang, Shijie Jiang, Yi Zheng, Feng Han, Siqi Li

Project administration: Yi Zheng

Resources: Yi Zheng

Supervision: Yi Zheng

Validation: Chao Wang, Feng Han

© 2024. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Abstract While deep learning (DL) models exhibit superior simulation accuracy over traditional distributed hydrological models (DHMs), their main limitations lie in opacity and the absence of underlying physical mechanisms. The pursuit of synergies between DL and DHMs is an engaging research domain, yet a definitive roadmap remains elusive. In this study, a novel framework that seamlessly integrates a process-based hydrological model encoded as a neural network (NN), an additional NN for mapping spatially distributed and physically meaningful parameters from watershed attributes, and NN-based replacement models representing inadequately understood processes is developed. Multi-source observations are used as training data, and the framework is fully differentiable, enabling fast parameter tuning by backpropagation. A hybrid DL model of the Amazon Basin ($\sim 6 \times 10^6$ km²) was established based on the framework, and HydroPy, a global-scale DHM, was encoded as its physical backbone. Trained simultaneously with streamflow observations and Gravity Recovery and Climate Experiment satellite data, the hybrid model yielded median Nash-Sutcliffe efficiencies of 0.83 and 0.77 for dynamic and distributed simulations of streamflow and total water storage, respectively, 41% and 35% higher than those of the original HydroPy model. Replacing the original Penman–Monteith formulation in HydroPy with a replacement NN produces more plausible potential evapotranspiration (PET) estimates, and unravels the spatial pattern of PET in this giant basin. The NN used for parameterization was interpreted to identify the factors controlling the spatial variability in key parameters. Overall, this study lays out a feasible technical roadmap for distributed hydrological modeling in the big data era.

1. Introduction

The process-based distributed modeling approach has played an important role in promoting hydrology as a cornerstone discipline of Earth science (Fatichi et al., 2016). The detailed characterization of complex and heterogeneous internal conditions within a catchment renders distributed models more suitable than lumped models for facilitating fundamental and theoretical discoveries regarding hydrologic processes and supporting effective water resource management (Simmons et al., 2020). Despite the past success of this approach (Paniconi & Putti, 2015), the main challenge that process-based distributed models still encounter is the selection of suitable process equations and parameterization schemes (Clark et al., 2016; Semenova & Beven, 2015). In contrast, machine learning (ML) approaches construct direct input-to-output mappings using efficient training algorithms such as backpropagation (BP), bypassing the explicit representation of hydrological processes and the cumbersome parameterization procedure. In recent years, deep learning (DL), a state-of-the-art ML approach, has demonstrated remarkable success in hydrological modeling (Jiang, Zheng, et al., 2022; Sadler et al., 2022; Wunsch et al., 2022). Existing DL applications have mainly focused on streamflow prediction at the catchment outlet, disregarding hydrological fluxes and states within the catchment, despite some recent studies utilizing spatially distributed inputs (C. Chen et al., 2022; Xu et al., 2022). DL-based hydrological models have been criticized for their poor interpretability (Shen, 2018; Xu & Liang, 2021), and the absence of spatial details exacerbates this issue. Post hoc interpretation methods (Jiang, Bevacqua, & Zscheischler, 2022; Samek et al., 2019) may enhance model interpretability, but they do not address the issue of absence of spatial details. Recently, the potential of applying DL in distributed hydrological modeling to enhance the understanding of hydrological

Writing – original draft: Chao Wang
Writing – review & editing: Shijie Jiang,
Yi Zheng, Rohini Kumar, Oldrich Rakovec

processes has been discussed (Beven, 2020; Nearing et al., 2021; Shen et al., 2021). Convolutional Long Short-Term Memory (ConvLSTM) is a popular DL architecture for modeling streamflow with spatially distributed input data (C. Chen et al., 2022; Zhu et al., 2023). For example, Xu et al. (2022) used a physically based, spatially distributed snow model to simulate snowmelt and used the results to drive a ConvLSTM streamflow model. While this one-way loosely coupled approach improved modeling accuracy, it only simulates streamflow at the watershed outlet, keeping the flow routing process within the watershed a black box. Novel approaches to merge distributed hydrological modeling and DL need further exploration (Nearing et al., 2021; Shen et al., 2023).

Furthermore, DL does not guarantee physical consistency, which can lead to spurious and inaccurate predictions, especially outside the training range (Konapala et al., 2020; Nearing et al., 2021). Physics-informed ML paradigms provide an avenue for DL-based hydrological modeling to address above issues (Herath et al., 2021; Karniadakis et al., 2021; Xie et al., 2021). For example, Xie et al. (2021) adopted a monotonic relationship between rainfall and runoff as an additional regularization term in the loss function for network training to ensure physical consistency. Bhasme et al. (2022) used water storages simulated by a physics-based model as the inputs for a DL streamflow model, improving the DL model's physical consistency. Physical knowledge can also be directly encoded into a DL model (M. Chen et al., 2023; Höge et al., 2022; Kraft et al., 2022; Reichstein et al., 2019). For example, Jiang et al. (2020) wrapped the physical processes depicted by a conceptual hydrological model (i.e., EXP-HYDRO) into a recurrent neural network (RNN). This symbiotic integration between DL and physical knowledge led to enhanced runoff simulation accuracy and intelligence for inferring unobserved processes (e.g., snow accumulation). Similarly, Feng et al. (2022) implemented another conceptual hydrological model called Hydrologiska Byråns Vattenbalansavdelning in a DL architecture and demonstrated improved performance for runoff simulation and reasonable inference of variables for which training was not directly performed, such as evapotranspiration (ET) and baseflow. However, the above hybrid models are still lumped in space without considering internal heterogeneity and river routing. One major difficulty is encoding the river routing process into a NN. Graph neural networks (GNNs) have been substituted for routing schemes (Sun et al., 2022), but the trained GNN does not guarantee the water balance in a river network. Bindas et al. (2024) used a physics-based NN to model river routing, but the process of runoff generation in sub-basins was separately modeled using long short-term memory (LSTM) models. A unified DL architecture for large-scale distributed hydrological modeling is still lacking.

Developing a distributed hybrid model with explicit representation of internal heterogeneity and river routing not only ensures the physical consistency of DL but also allows us to take full advantage of spatial observation data sets to better constrain model simulations. In many cases, hydrological observations are available at multiple locations within a watershed. Novel data sets may also be available for model training. For example, the Gravity Recovery and Climate Experiment (GRACE) satellite data provide an unprecedented opportunity to quantify spatiotemporal variations in Earth's surface mass, mainly reflecting changes in total water storage (TWS) (Landerer & Swenson, 2012; Tapley et al., 2004). GRACE data provide global-scale information related to the total water storage anomaly (TWSA) over more than 10 years (2003–2016), with a spatiotemporal resolution of a few hundred kilometers and a monthly temporal resolution; thus, these data can be used to constrain and improve hydrological modeling (Güntner, 2008; Soltani et al., 2021). Incorporating GRACE data can help distributed hydrological models (DHMs) better represent underlying hydrological processes (e.g., runoff, ET, and TWS) (Dembélé et al., 2020; Rakovec et al., 2016). However, integrating multi-source hydrological observations, including novel GRACE data, has rarely been attempted for DL-based hydrological models.

To address the above challenges, a distributed hybrid modeling framework based on physics-encoded deep NNs is developed in this study. The framework seamlessly integrates process-based runoff and river routing models, a module for spatially distributed parameterization, and NN-based replacement models for unknown or vaguely known processes into a unified DL architecture. The framework is fully differentiable, conforming to the emerging concept of differentiable modeling (Shen et al., 2023), and accommodates multi-source hydrological observations as training data. We selected the Amazon Basin, the largest river basin on Earth and a highly complex hydrological system, as the testbed for the new framework. We encoded a process-based DHM, named HydroPy (Stacke & Hagemann, 2021a), into the NN architecture as the physical backbone of the hybrid DL model. The study is aimed to demonstrate the feasibility and strengths of distributed hybrid modeling in spatiotemporal learning of hydrological processes at the continental scale. Overall, this study lays out a feasible technical roadmap for distributed hydrological modeling in the big data era.

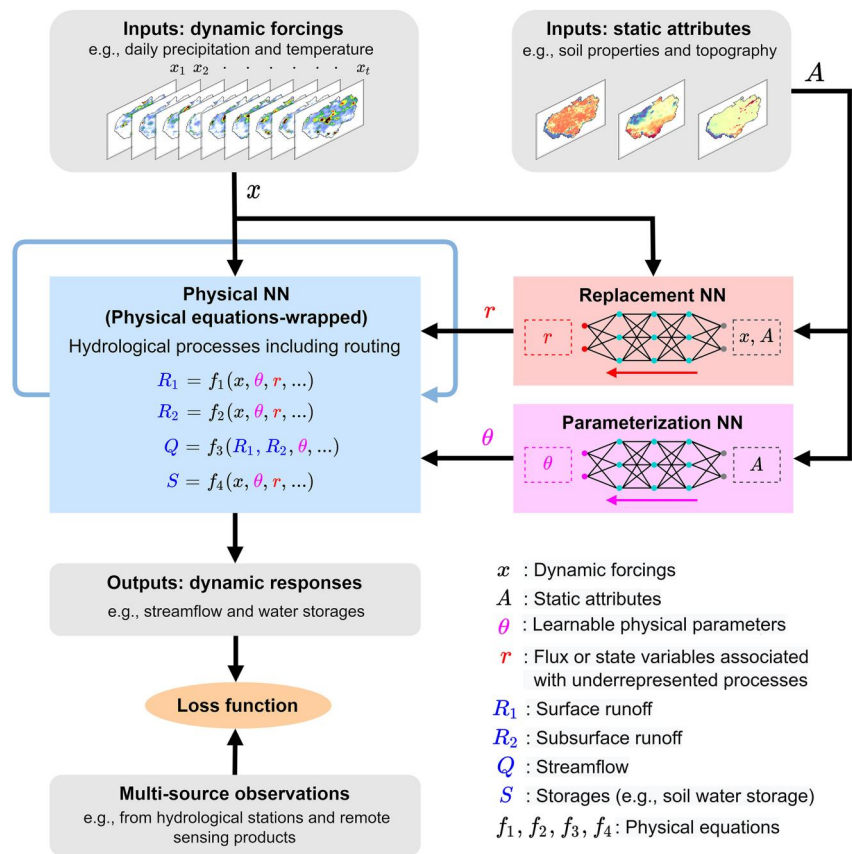


Figure 1. Schematic representation of the distributed hybrid modeling framework. The blue, magenta, and red boxes denote the physical neural network (NN), parameterization NN, and replacement NN, respectively. f_3 in the blue box stands for the river routing module. The black arrows indicate data flows, and the blue arrow indicates a recurrent NN architecture.

2. The New Modeling Framework

The distributed hybrid modeling framework (Figure 1) developed in this study fuses three types of NNs with different functions: the physical NN, the parameterization NN, and the replacement NN. The physical NN encodes physical equations including those for surface runoff generation (f_1), subsurface runoff generation (f_2), river flow routing (f_3), and water storages (f_4). This NN therefore ensures the physical consistency and interpretability of the DL model. For the subprocesses underrepresented by the process-based model, the replacement NN can be used to model flux and/or state variables associated with the subprocesses in a data-driven way. Dynamic forcings (x) and region-dependent static attributes (A) can be used as the inputs of the replacement NN. Additionally, the parameterization NN maps static attributes (A) to the distributed physical parameters (θ) of the process-based model (i.e., the physical NN). All NNs are seamlessly unified in a single DL architecture, allowing for global backpropagation and joint optimization across both physical and data-driven components (Jiang et al., 2020; Tsai et al., 2021). The parameters of the entire NN can be calibrated against multi-source observations. Designed to strictly adhere to the law of mass conservation, the framework can generate physically meaningful variables for intermediate processes in addition to the model outputs.

3. Data and Methods

3.1. Implementation of the Framework in the Amazon Basin

As the largest river basin ($\sim 6 \times 10^6 \text{ km}^2$) on Earth, the Amazon Basin (Figure 2) is a complex hydrological system characterized by large variations in topography (1–6,251 m above sea level), precipitation (500–6,000 mm per year), and land cover (Figure 2b). A distributed model that can capture these spatial heterogeneities is essential for

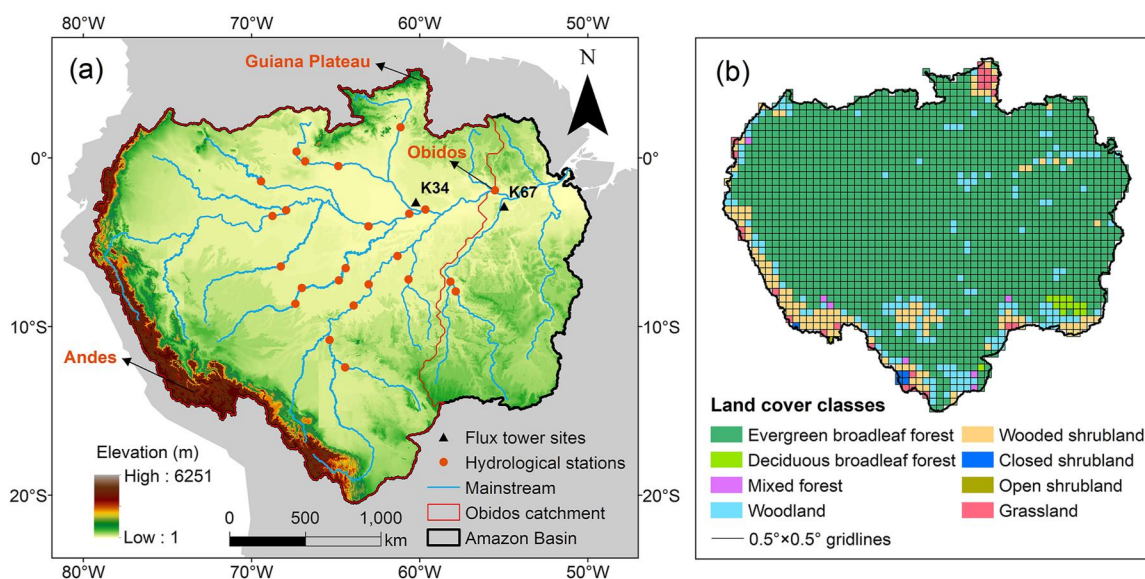


Figure 2. The Amazon Basin. (a) Topography, river network, flux tower sites and hydrological stations. The red line delineates the drainage area above the Obidos station. (b) Land cover and modeling grid ($0.5^\circ \times 0.5^\circ$ cells).

gaining an accurate understanding of water and energy fluxes in the basin (de Paiva et al., 2013). The region is at risk from climate and anthropogenic changes, and alterations in Amazon hydrology could have major global impacts (Chagas et al., 2022). Over the past few decades, the basin has experienced several intense climatic events, such as extreme droughts and floods (Marengo & Espinoza, 2016). Characterizing and understanding the dynamics of the Amazon water cycle is therefore crucial for water resource management. While many studies have addressed the water balance in this giant basin, uncertainty remains significant, particularly for ET (Fassoni-Andrade et al., 2021). By applying the new distributed hybrid modeling framework in the Amazon Basin, we aim to improve the understanding of the water balance and gain new insights into the complex interactions among climate, vegetation, and hydrology in this critical region. Hybrid hydrological modeling was performed with a grid with cells of $0.5^\circ \times 0.5^\circ$ (Figure 2b) for the period of 2001–2016. This large basin, with its diverse landscapes and climates, uniquely enables us to test the model's spatial generalizability.

We encoded a DHM, HydroPy (Figure 3a), into the physical NN in the framework depicted in Figure 1. HydroPy originated from the Max Planck Institute for Meteorology's Hydrology Model (Hagemann & Dümenil, 1997; Stacke & Hagemann, 2012) and has been enhanced with new processes (Stacke & Hagemann, 2021a). The model is for large-scale (regional to global scale) simulation by design, typically operating at daily time steps and a spatial resolution of $0.5^\circ \times 0.5^\circ$. The model simulates five main types of hydrological processes: snow processes, skin and canopy processes, soil and surface processes, groundwater processes, and river routing processes. The first four processes are modeled in individual grid cells, and the river routing process involves interactions between adjacent grid cells. Each grid cell in HydroPy has multiple conceptualized water storage “buckets” (Figure 3a): snow, skin, canopy, root zone soil, surface water, shallow groundwater, and river. The vertical water balance accounts for water transport through the snow, skin, canopy, and soil buckets, partitioning precipitation into ET and runoff. The mathematical equations describing these processes can be found in Stacke and Hagemann (2021a). The river routing process is realized based on the equations of the Hydrological Discharge Model (Hagemann & Dümenil, 1997), where the streamflow is routed through river buckets according to a predefined river routing network. The predefined river routing network uses a D8 flow direction scheme, which assigns one of the eight directions to each grid cell. Specifically, the river bucket in each grid cell receives the streamflow from upstream cells and releases outflow. The streamflow to the downstream grid cell is calculated by aggregating the runoff (including groundwater runoff and overland flow) generated in the local grid cell and the river bucket outflow. The routing scheme and corresponding equations are provided in Texts S1 and S2 in Supporting Information S1. This routing method assumes linear reservoir schemes and spatially varying but temporally constant flow velocities. The lag time (i.e., retention time) parameters are utilized in the river routing equations to characterize flow velocities. In general, HydroPy adequately simulates daily streamflow on a large scale.

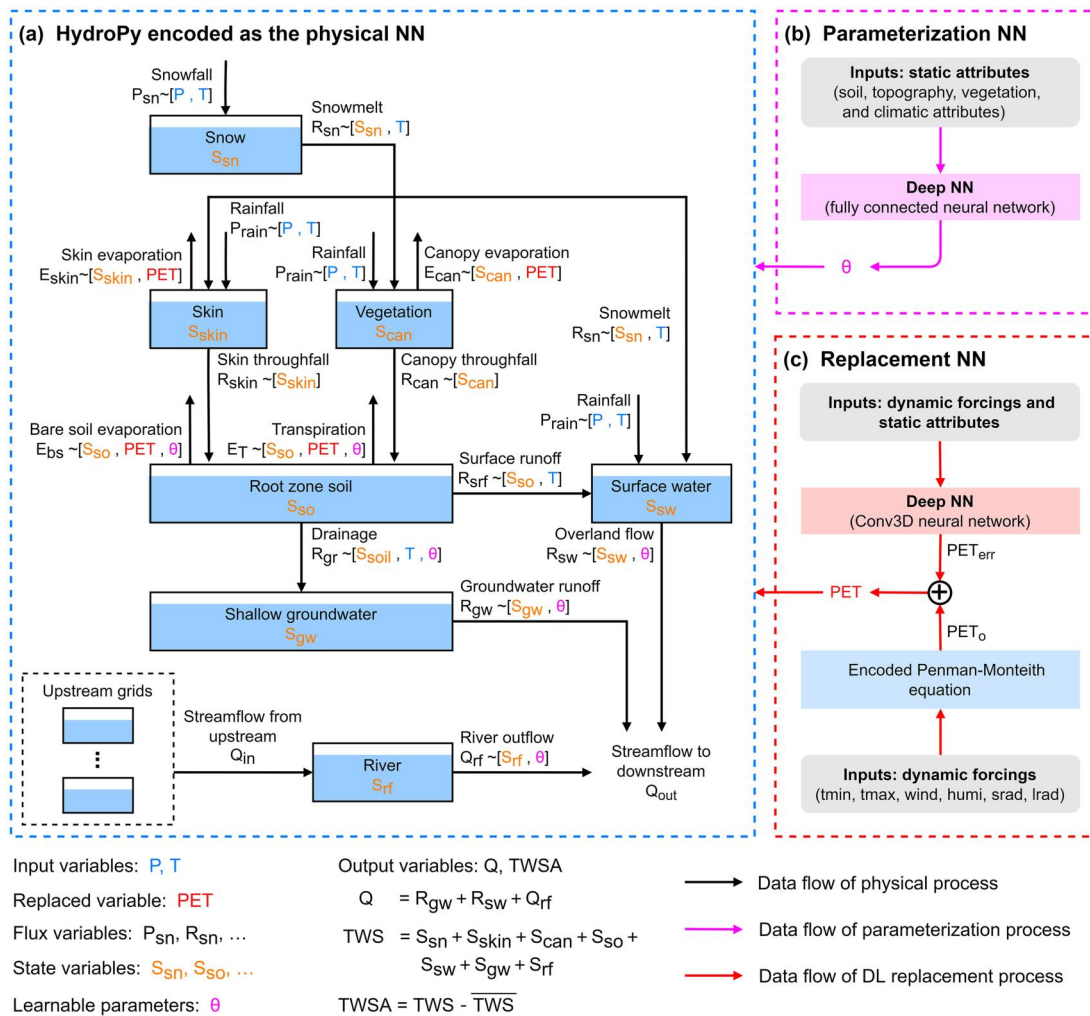


Figure 3. The hybrid deep learning model for distributed hydrological modeling in the Amazon Basin. (a) HydroPy encoded as the physical neural network (NN), (b) the parameterization NN, which provides grid-specific parameters for the encoded HydroPy, and (c) the replacement NN substituted for the plausible potential evapotranspiration sub-model in the original HydroPy.

However, it may exaggerate peak flow for the Amazon and Congo rivers and underpredict it for Arctic rivers (Stacke & Hagemann, 2021a).

We encoded HydroPy into a physical RNN architecture, which is a modified version of ordinary RNN, following the approach of Jiang et al. (2020). Within a physical RNN unit, the connections among neurons (inputs, states, and outputs) are defined by the numerical solutions of hydrological process equations instead of activation functions, and the weights and biases are consequently physically meaningful parameters (i.e., the parameters of the wrapped equations). Text S1 and Figure S1 in Supporting Information S1 provide more details of the encoding approach. However, the original approach can only be used to encode a spatially lumped model. To enable distributed modeling, we further represent the river routing process in the RNN architecture, which is a significant improvement over the original approach. As Figure 4a illustrates, the river routing equations of the Hydrological Discharge Model (Hagemann & Dümenil, 1997) are encoded into the step function of the physical RNN, which outputs the matrix of streamflow to the downstream grid cell (Q_{out}). The encoded routing equations can be found in Text S2 in Supporting Information S1. To program the routing scheme in a DL framework (e.g., TensorFlow) and ensure effective backpropagation and efficient computation, a flow transfer algorithm was also proposed, which turns the data matrix of the outflow at the current time step (Q'_{out}) into the data matrix of the inflow at the next time step (Q'^{t+1}_{in}). Figure 4b provides a visual explanation of the algorithm's workflow. Texts S1 and S2 in Supporting Information S1 provide more details of the routing-enabled encoding approach.

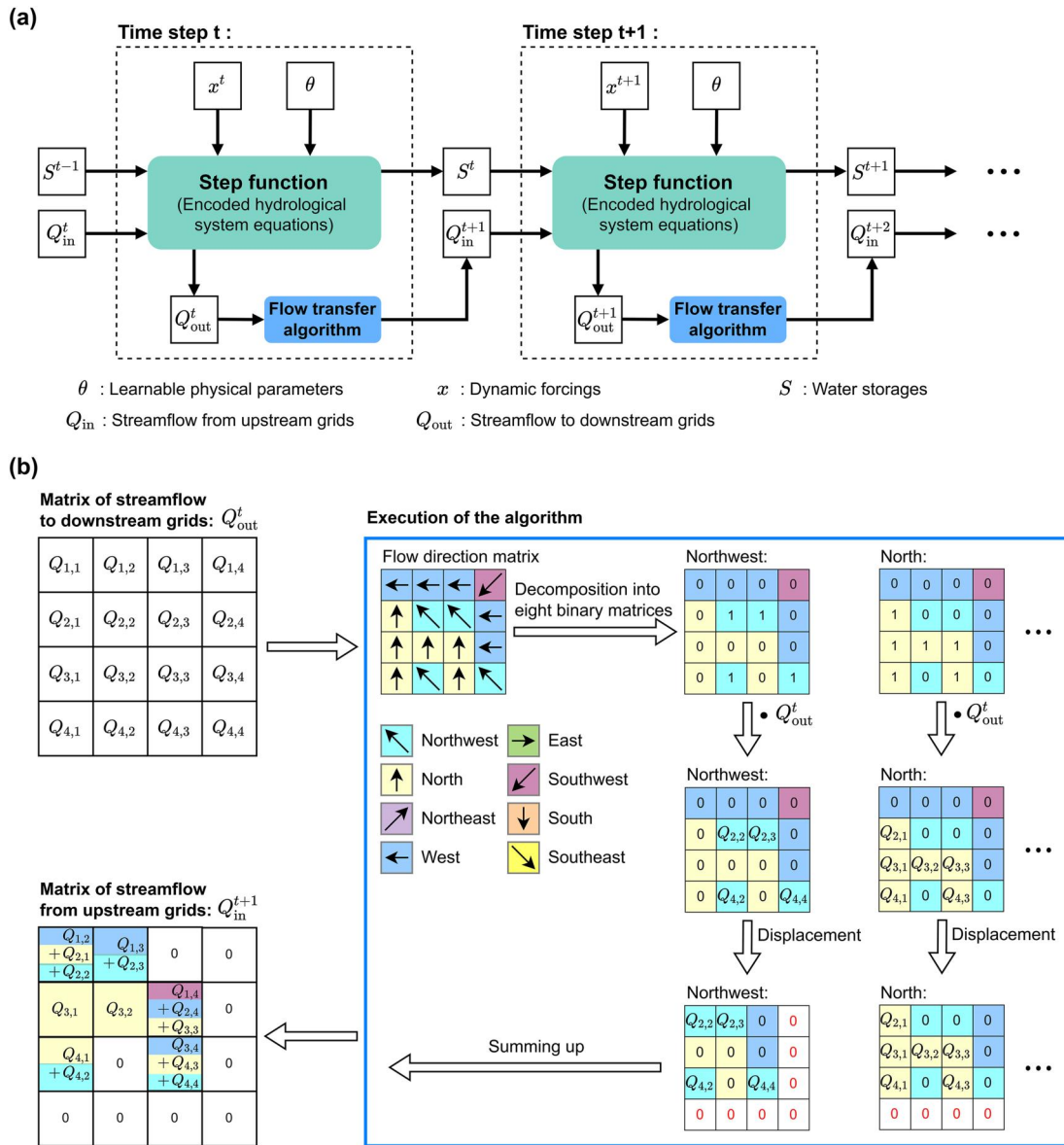


Figure 4. The hydrology-encoded recurrent neural network (RNN). (a) The unfolded RNN unit. The step function represents the explicit discrete form of hydrological system equations including the river routing equations. At each time step, the step function is evaluated over the entire modeling domain. The flow transfer algorithm turns the data matrix of the outflow at the current time step (Q_{out}^t) into the data matrix of the inflow at the next time step (Q_{in}^{t+1}). (b) A visual explanation of the flow transfer algorithm using a 4×4 grid domain as an illustrative case. The flow direction matrix is disaggregated into eight binary matrices, whose elements represent flow presence (1) or absence (0) in a specific direction. Each binary matrix is element-wise multiplied with the data matrix Q_{out}^t . Elements within the resultant matrices are shifted to their corresponding downstream positions, aligning with the flow directions. The eight matrices, following displacement, are summed to generate the final data matrix Q_{in}^{t+1} . In the Amazon case, the modeling domain has a 60×72 grid size, and the entire Amazon is run in each batch to ensure the integrity of the river routing process.

The physically meaningful parameters (θ) of the physical NN (Figure 3a) are mapped from static attributes via the parameterization NN (Figure 3b). In this Amazon case, there are nine parameters in total (Table 1) with grid-specific values. Similar to previous studies (Feng et al., 2022; Jiang et al., 2020), we selected 21 attributes (see Table S1 in Supporting Information S1) as the inputs of the parameterization NN. The parameterization NN is a fully connected neural network (FCNN) consisting of five layers. The numbers of neurons in the five layers are 21, 256, 64, 16, and 9, respectively. The network configuration was determined after preliminary trials.

In the original HydroPy model, plausible potential evapotranspiration (PET) is estimated using the Penman-Monteith reference ET approach (Allen et al., 1998). This classic approach tends to underestimate PET in the

Table 1
The Nine Grid-Specific Physical Parameters Included in the Physical Neural Network

Parameter	Default	Range	Unit	Description	Reference for the range
d_{\min}	0.024	0.001–0.1	mm/d	Slow drainage velocity	Troy et al. (2008)
d_{\max}	2.40	0.1–10	mm/d	Fast drainage velocity	
$f_{d\min}$	0.05	0–0.20	/	Fraction of soil moisture content when slow drainage occurs	
$f_{d\max}$	0.90	0.2–1	/	Fraction of soil moisture content when fast drainage occurs	
f_{crit}	0.75	0.55–0.95	/	Fraction of critical root zone soil moisture	Stacke and Hagemann (2021b)
f_{evap}	0.05	0–0.25	/	Fraction of minimum soil moisture for soil evaporation	
lag_{sw}	10–50	1–200	d	Lag time of overland flow	Hagemann and Dümenil (1997)
lag_{gw}	300–800	30–1,500	d	Lag time of groundwater runoff	
$\text{lag}_{\text{river}}$	0.05–0.5	0–5	d	Lag time of river outflow	

tropics (Trambauer et al., 2014; Weiland et al., 2015) and has been specifically identified as a major source of error in the HydroPy model (Stacke & Hagemann, 2021a). Reliable PET representation schemes in the tropics are still lacking (Trambauer et al., 2014). In the Amazon Basin, ET is primarily energy limited, and therefore, the calculation of PET significantly influences the calculation of other hydrological fluxes and states. In this study, we constructed a replacement NN to replace the Penman–Monteith approach and output PET estimates to the physical NN. In the replacement NN, the Penman–Monteith equation was encoded to output the reference ET (PET_o), and a deep NN was used to derive an error term (PET_{err}) added to PET_o (Figure 3c). This error correction strategy is a common way to improve the accuracy of PET estimation (Gebremedhin et al., 2022). Potential data biases in the meteorological input variables may also contribute to the error term in PET calculations (Weedon et al., 2014; Zuluaga et al., 2021), and such biases at a specific grid cell can be corrected based on meteorological conditions and land surface attributes of neighboring cells (L. Han et al., 2021; Xia et al., 2023). In this study, a 3D convolutional (Conv3D) NN was adopted in the error correction to account for the influence of meteorological and land surface conditions from neighboring cells on a specific grid cell. Preliminary experiments showed that the Conv3D NN outperforms a FCNN in modeling the error term (see Table S2 in Supporting Information S1). Both the six meteorological variables (see Table S3 in Supporting Information S1) required by the Penman–Monteith equation and the 21 static attributes for the parameterization NN are used as the inputs of the Conv3D NN. This specific network structure was determined after preliminary trials, as explained in Text S3 and Table S2 in Supporting Information S1. Similar to the parameterization NN, the Conv3D NN includes an input layer connected to the input variables, three hidden layers with 32, 16, and 8 convolutional filters, and an output layer with one convolutional filter. Following common recommendations for kernel size selection, a $3 \times 3 \times 3$ convolution kernel was chosen for all convolutional layers, providing a good balance between modeling performance and computational efficiency (Tran et al., 2015). In the case of Amazon, the replacement NN is designed to model the error term (Figure 3). However, within the general framework (Figure 1), replacement NNs can be directly established for process variables. Note that this study uses PET as an example of an underrepresented process, and different processes in other modeling cases can be addressed similarly.

3.2. Data for Modeling

The daily precipitation and temperature time series used to drive model simulations were from the WATCH-Forcing-Data-ERA-Interim (WFDEI) data set with a spatial resolution of $0.5^\circ \times 0.5^\circ$ (Weedon et al., 2014). Among the 21 static attributes (see Table S1 in Supporting Information S1), the soil, topography, and vegetation attributes were derived from Stacke and Hagemann (2021b), and the climate attributes were derived from the WFDEI data set. Furthermore, the time series of the six meteorological input variables for the Penman–Monteith equation were obtained from the WFDEI data set (see Table S3 in Supporting Information S1). Because HydroPy requires a D8 flow direction scheme, and the original river routing network derived from Hagemann and Dümenil (1997) has the same scheme, we used it as a starting point. However, we found that the drainage areas of the 24 hydrological stations reported by the Global Runoff Data Centre (GRDC) differed by 10%–20% from those calculated using the original network. While the Major River Network data set by GRDC provides latitude and longitude information for global rivers, it does not use the D8 scheme and cannot be directly used in HydroPy.

Therefore, we modified the original network using the GRDC data set as a reference to ensure consistency in data. Manual modification was performed at a resolution of $0.5^\circ \times 0.5^\circ$, adhering to HydroPy's general specification for the downstream grid cell, which dictates that the downstream grid cell of each grid cell is lateral and uniquely fixed. The modified river routing network is shown in Figure S2 in Supporting Information S1. As the figure shows, the differences are quite noticeable, highlighting the necessity of the modifications.

We used both streamflow and TWSA data to calibrate and evaluate the hybrid model. We obtained daily streamflow observations from the Global Runoff Data Centre (GRDC; <https://www.bafg.de/GRDC/>), which provides daily streamflow records for over 200 stations in the Amazon Basin. We selected 24 key stations (Figure 2a and Table S4 in Supporting Information S1) based on several criteria, such as the corresponding catchment size (larger than $1.0 \times 10^5 \text{ km}^2$), temporal completeness (less than 10% missing data), and station locations (covering major tributaries). For TWSA, we used the monthly Mascon products from the GRACE satellite mission (Tapley et al., 2004), which have been widely used in hydrology. We considered the average values of three products prepared by different institutes, such as the Center for Space Research at the University of Texas (CSR) (Save et al., 2016), NASA's Jet Propulsion Laboratory (JPL) (Wiese et al., 2016), and NASA's Goddard Space Flight Center (GSFC) (Loomis et al., 2019), to account for data uncertainty. Constrained by the inherent frequency band limitation problem of GRACE (Save et al., 2016), the fundamental spatial resolution of the GRACE TWSA data is approximately 300 km. Following Schumacher et al. (2016), we spatially aggregated these products with an original resolution of $0.25^\circ \times 0.25^\circ$ or $0.5^\circ \times 0.5^\circ$ to a $3^\circ \times 3^\circ$ resolution to ensure data reliability. To match the aggregated GRACE TWSA data, we processed the model output of TWS, which is the sum of all water storage buckets, with the following steps. First, the modeled monthly TWSA was derived as the modeled TWS minus the average TWS from 2004 to 2009 (i.e., the GRACE TWSA baseline period). Second, the modeled TWSA was rescaled to the same $3^\circ \times 3^\circ$ grid cell as the GRACE TWSA data.

Direct observations of ET in the Amazon Basin are very limited, making it difficult to evaluate the model's performance in ET prediction over the entire modeling domain (Baker et al., 2021; Fassoni-Andrade et al., 2021). Remote sensing-derived ET data provide complete spatial coverage over the study area. However, dense vegetation canopies and frequent cloud cover in the Amazon significantly impair the accuracy of remote sensing-based ET estimates (Fassoni-Andrade et al., 2021; Swann & Koven, 2017). Previous studies in the Amazon Basin calculated catchment-wise ET based on the water balance approach as a benchmark, which is considered to provide the closest approximation to a direct ET "measurement" at large spatial scales (Baker et al., 2021; Maeda et al., 2017; Swann & Koven, 2017). In this study, the catchment-wise water balance, $ET = P - Q - \Delta S$, was evaluated, where P is the annual precipitation derived from the WFDEI data set; Q is the annual runoff derived from GRDC; and ΔS is the catchment-wise change in TWS in 1 year derived from GRACE TWSA data. In addition, we collected site-scale ET observations from two eddy flux towers (see Figure 2) established in the Large-Scale Biosphere–Atmosphere Experiment in Amazonia (LBA) research program (Restrepo-Coupe et al., 2021), which were used as independent references.

3.3. Numerical Experiments

In this study, five NN-based modeling schemes, namely, HydroPyNN, HydroPyNN(Q,S), Hybrid(Q), Hybrid(Q,S), and Hybrid(Q,S)+, were explored. In HydroPyNN and HydroPyNN(Q,S), the original HydroPy was implemented as the stand-alone physical NN in Figure 3a. Using the default parameter values provided by Stacke and Hagemann (2021b) (Table 1). HydroPyNN and the original HydroPy were compared to determine whether HydroPy was successfully embedded in the RNN architecture. HydroPyNN(Q,S) is trained using both streamflow (Q) and GRACE TWSA (S) data. Hybrid(Q), Hybrid(Q,S), and Hybrid(Q,S)+ are all hybrid DL models. In Hybrid(Q) and Hybrid(Q,S), the parameterization NN is activated, and the models are trained using streamflow data alone and both streamflow and GRACE TWSA data, respectively. In HydroPyNN, HydroPyNN(Q,S), Hybrid(Q), and Hybrid(Q,S), PET was estimated using the Penman-Monteith equation. Hybrid(Q,S)+ further activates the replacement NN for PET estimation and was trained using both streamflow and GRACE TWSA data.

The entire modeling time period was from 2001 to 2016, with the first two years for spin-up. We considered three data splitting strategies, temporal, spatial, and spatiotemporal splitting, to train and validate hybrid DL models. Only temporal splitting was applied to HydroPyNN(Q,S) as it has no parameterization NN. In temporal splitting,

the periods of 2003–2012 and 2013–2016 were selected as the training and validation periods, respectively. All 24 stations and/or 56 grid cells ($3^\circ \times 3^\circ$) were included in both periods. In spatial splitting, the 24 hydrological stations were randomly divided into four sets, and the entire period of 2003–2016 was considered in both the training and validation stages. Spatial splitting was applied for Hybrid(Q,S) and Hybrid(Q,S)+, and four-fold cross-validation was performed. Preliminary tests showed that the prediction accuracy of hybrid models is not sensitive to the fold number. In each fold, the streamflow data for three groups of stations plus the TWSA data for all 56 grid cells were used for training, and the streamflow data for the fourth group of stations was used for validation. The spatial splitting of streamflow data was performed to investigate the transfer learning capacity (i.e., predictive ability in ungauged basins) of the hybrid models. In the third strategy, spatiotemporal splitting, the data was divided into four parts: training stations in 2003–2012 (training period), training stations in 2013–2016 (validation period), validation stations in 2003–2012, and validation stations in 2013–2016. Hybrid(Q,S)+ was trained using the first part of the data, and its temporal, spatial, and spatiotemporal learning capabilities were evaluated based on the second, third, and fourth parts, respectively. Four-fold cross-validation was also performed, similar to the spatial splitting case.

The Nash-Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970) was considered when formulating the loss function for network training. As the most commonly used evaluation metric in hydrological modeling, the NSE ranges from $(-\infty, 1]$, with a large value indicating good performance. The loss function used in the study is generally expressed as shown in Equation 1:

$$\text{loss} = w_1 \cdot \frac{1}{N} \sum_{n=1}^N \text{NSE}_Q^n + w_2 \cdot \frac{1}{M} \sum_{m=1}^M \text{NSE}_{\text{TWSA}}^m \quad (1)$$

where N and M represent the numbers of streamflow stations and grid cells included in the training stage, respectively, and w_1 and w_2 represent the weights of streamflow and TWSA in the loss function, respectively. For HydroPyNN(Q,S) in temporal splitting, N is equal to 24 and M is equal to 56. For Hybrid(Q), Hybrid(Q,S) and Hybrid(Q,S)+, N is equal to 24 in the temporal splitting scheme and 18 in the spatial splitting scheme. For Hybrid(Q,S) and Hybrid(Q,S)+, M is equal to 56 in both temporal and spatial splitting schemes. For Hybrid(Q,S)+ in spatiotemporal splitting, N is equal to 18 and M is equal to 56. For Hybrid(Q), w_1 and w_2 were set to 1 and 0, respectively. For HydroPyNN(Q,S), Hybrid(Q,S) and Hybrid(Q,S)+, both w_1 and w_2 were set to 0.5, assuming that streamflow and TWSA are equally important. We also investigated the influence of unequally weighting schemes on the performance of Hybrid(Q,S)+, varying w_1 from 0 to 1 (w_2 from 1 to 0, simultaneously) with an interval of 0.1. The results are presented in Figure S3 in Supporting Information S1, which confirms that the equal weighting scheme is an appropriate choice for modeling streamflow, TWSA and the intermediate variable ET.

We implemented all the NN-based models based on the DL platform TensorFlow (Abadi et al., 2016). The Adam optimization algorithm (Kingma & Ba, 2014) with a learning rate of 0.0005 was used to train the models. The NN-based models were deployed and trained using two 32-core Xeon Gold 6338 CPUs. Each training iteration takes approximately 2 min and requires 200 GB of memory. Within the training period, the maximum number of iterations (i.e., the epoch number) was set to 300 to balance the computational accuracy and efficiency. The hyperparameters for network training were determined through preliminary trials. More details of the hyperparameters are provided in Table S5 in Supporting Information S1. For HydroPyNN(Q,S), Hybrid(Q), Hybrid(Q,S), and Hybrid(Q,S)+, the training process involved 10 random repeat runs with different initial weights and biases for the NNs to account for modeling uncertainty. Unless otherwise mentioned, the results presented in this paper are the mean values of the repeated runs.

3.4. Interpretation of the Neural Network

In this study, the expected gradient (EG, Erion et al., 2021) method was used to interpret the behavior of the parameterization NN in the hybrid model. The EG method uses the gradient of the model output with respect to the input features to trace the specific contributions of inputs (Erion et al., 2021). This method is based on the game theory approach, and an importance score is assigned to each feature of a given input. The score of an input feature indicates its local contribution to the prediction: a large positive value indicates that the feature (i.e., the static attribute in the case) substantially increases the network output (i.e., the physical parameter at the individual grid scale), and a large negative value indicates the opposite. Given a NN Θ , the EG value for feature i is defined as:

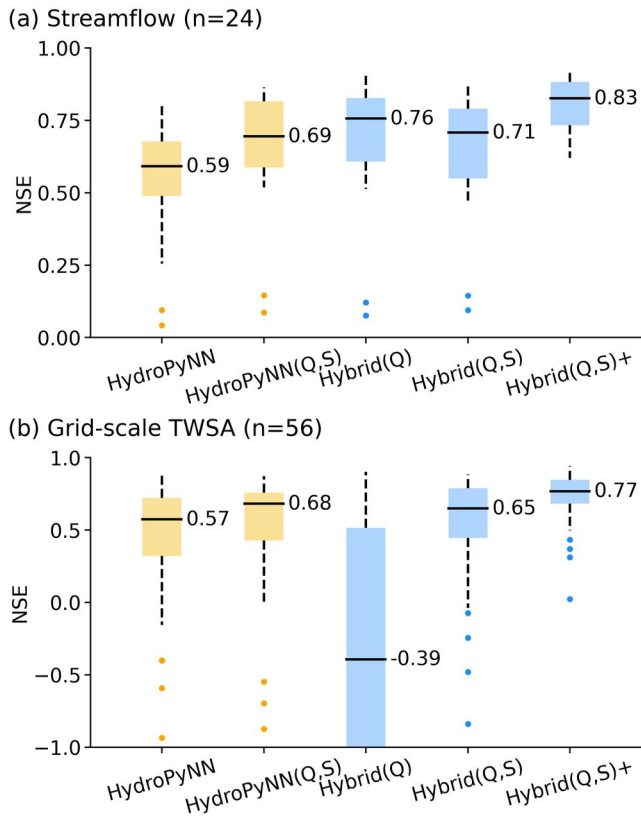


Figure 5. Performance in predicting dynamic streamflow and total water storage anomaly with different neural network (NN)-based models during the validation period. (a) Nash-Sutcliffe efficiency (NSE) values for the 24 selected streamflow stations and (b) NSE values for the 56 grid cells ($3^\circ \times 3^\circ$ in size). The box plots show the median (the values are showed), 25th percentile, 75th percentile, and 1.5 times interquartile range of NSE values. The dots denote the outliers outside the 1.5 times interquartile range. The yellow boxes represent models without the parameterization NN, and the blue boxes represent hybrid models with the parameterization NN.

drastically higher than training HydroPyNN. HydroPyNN(Q,S) has similar performance as Hybrid(Q,S). However, without the NN mapping geographical attributes to physical parameters, HydroPyNN(Q,S) lacks a physically grounded basis for its parameter values, significantly constraining its ability to accurately predict in unseen geographical locations. It is noted from Figure 5a that adding the GRACE TWSA in training actually degrades the performance of the model in streamflow prediction (Hybrid(Q,S) versus Hybrid(Q)), as has been reported in previous studies (Dembélé et al., 2020; Rakovec et al., 2016). The important implication is that improvements in a single objective does not necessarily suggest that the overall model is improved, as the improvement may be at the cost of degradations for other objectives. For TWSA (Figure 5b), Hybrid(Q) performs unacceptably, even worse than the uncalibrated HydroPyNN. The poor performance is due to model equifinality (Beven, 2006; Clark et al., 2017), as the streamflow observations at the 24 stations are insufficient to constrain the water storage simulation by HydroPyNN. The problem was resolved in Hybrid(Q,S) after the inclusion of GRACE TWSA data. Furthermore, Hybrid(Q,S)+ achieves the best predictive performance for both streamflow and TWSA, indicating the benefit of using the replacement NN for modeling PET.

Our new framework has significant advantages over the original form of HydroPy. Parameter calibration in HydroPy for such a complex case (the modeling domain of 5.96×10^6 km² is delineated into 1,951 grid cells, and multiple calibration objectives are considered) is practically infeasible using traditional gradient-based or heuristic search methods, which is the possible reason why the global HydroPy model was not calibrated (Hagemann & Dümenil, 1997; Stacke & Hagemann, 2021a). However, with the model transformed into a differentiable NN, the BP algorithm can efficiently infer the network parameters. The framework can also easily fuse multi-source

$$\phi_i(\Theta, x) = E_{x' \sim D, \alpha \sim U(0,1)} \left[(x_i - x'_i) \times \frac{\partial f(x' + \alpha(x - x'))}{\partial x'_i} \right] \quad (2)$$

where x is the target input; x' is the baseline input; D is the underlying distribution of the baseline input for the background data set (e.g., the training data set), $x' \in D$; $U(0,1)$ is the uniform distribution between 0 and 1, $\alpha \in U$; and $\partial f(x' + \alpha(x - x')) / \partial x'_i$ is the local gradient of the NN Θ at a point interpolated between the baseline input and target input. In this study, the procedure used to calculate the EG value (ϕ_i) was based on the SHapley Additive exPlanations package (Lundberg & Lee, 2017), which can provide various post hoc analyses for different neural networks. The mean absolute ϕ_i value of a specific attribute to a physical parameter reflects its global contribution over all grid cells. The contributions of different attributes are additive, and the sum of all attribute contributions and the average predicted value equals the final predicted value.

4. Results

4.1. Temporal Learning for Streamflow and TWSA

It was first confirmed that HydroPyNN (the RNN version) precisely emulates the original HydroPy (see Table S6 in Supporting Information S1). Figure 5 compares the performance of five NN models using NSE. Similar comparison results, based on root mean square error and the Kling-Gupta efficiency (KGE, Gupta et al., 2009), are displayed in Figure S4 in Supporting Information S1. In general, network training with observational data significantly improves the model prediction accuracy compared to the non-training case (i.e., HydroPyNN), except for Hybrid(Q) for modeling TWSA. It is worth emphasizing that training of HydroPyNN is much more computationally efficient than calibration of the original form of HydroPy. HydroPyNN is fully differentiable such that the training process based on backpropagation can be finished in several hours (on two 32-core Xeon Gold 6338 CPUs in our case). But calibration of HydroPy using heuristic algorithms inevitably encounters the curse of dimensionality, and the computational cost would be

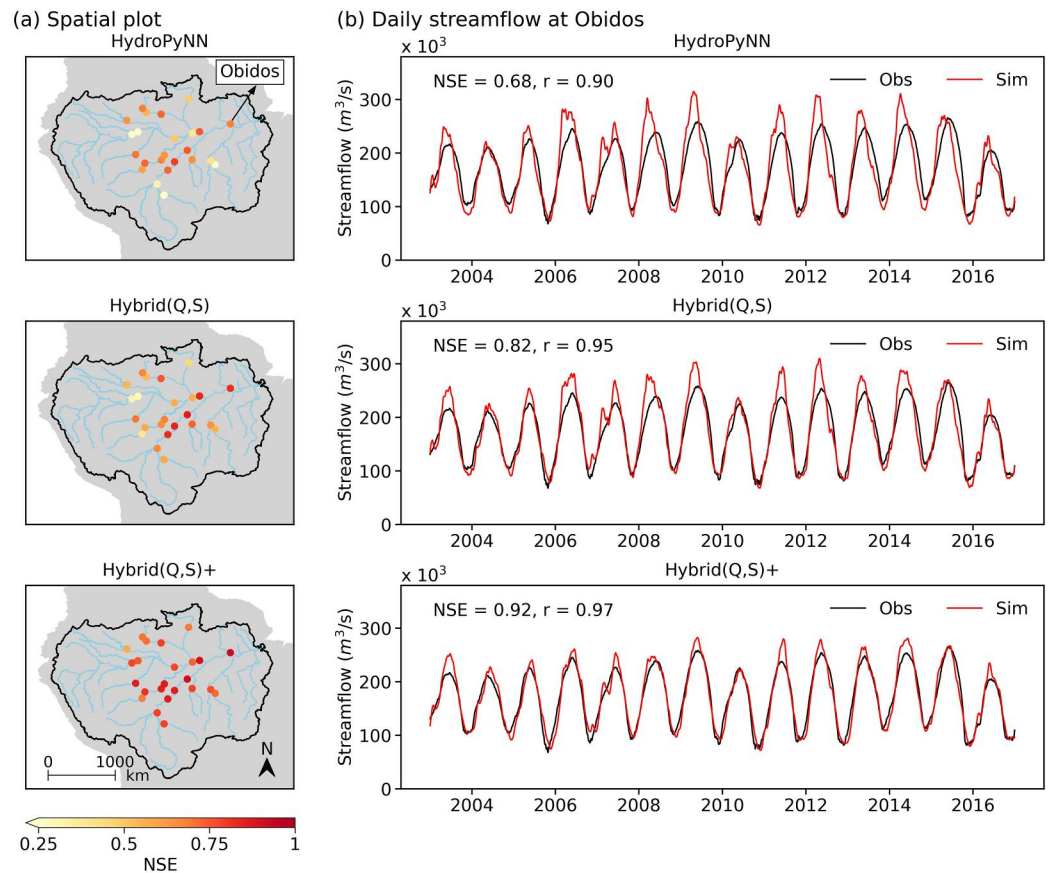


Figure 6. Performance of the neural network models in spatial transfer learning. (a) The Nash-Sutcliffe efficiency values based on four-fold cross-validation at the 24 stations as validation stations. (b) Observed and simulated hydrographs at the most downstream station (Obidos).

observations for the spatiotemporal learning of hydrological processes, which helps to plausibly constrain model behavior and to achieve potentially balanced simulation performance among multiple target variables. Furthermore, the end-to-end learning system allows co-training of the physical NN and the replacement NN to represent unknown or uncertain process(es), such that processes without direct observations (e.g., PET in this study) can be effectively constrained by other indirect observations (e.g., streamflow and TWSA in this study).

4.2. Spatial and Spatiotemporal Learning for Streamflow Simulation

The spatial cross-validation with spatial splitting reveals that both Hybrid(Q,S) and Hybrid(Q,S)+ perform obviously better than the untrained HydroPyNN model in the spatial transfer learning of streamflow (Figure 6a). Hybrid(Q,S)+ performs the best, with only three stations displaying an NSE below 0.70. The median NSE of Hybrid(Q,S)+ is 0.80, much higher than the values of HydroPyNN (0.63) and Hybrid(Q,S) (0.66), and the difference in model performance is statistically significant ($p < 0.01$ in paired t -test). The time series of simulated (by Hybrid(Q,S)+) and observed streamflow at the 24 hydrological stations are illustrated in Figure S5 in Supporting Information S1. Even in the challenging realm of spatiotemporal learning by spatiotemporal splitting, the median NSE of Hybrid(Q,S)+ reaches 0.76 (see Figure S6 in Supporting Information S1). The excellent transfer learning capability of Hybrid(Q,S)+ reflects the value of integrating the parameterization and replacement NNs with the physical NN (Figure 3).

Figure 6b illustrates the detailed simulation results for Obidos, the most downstream station in the Amazon Basin; notably, the drainage area of this station covers approximately 77% of the basin (Figure 1). Consistent with the findings of Stacke and Hagemann (2021a), the untrained HydroPyNN predicts earlier and higher flow peaks than the observed values, which may be associated with the modeled low ET in March to May. The two trained models

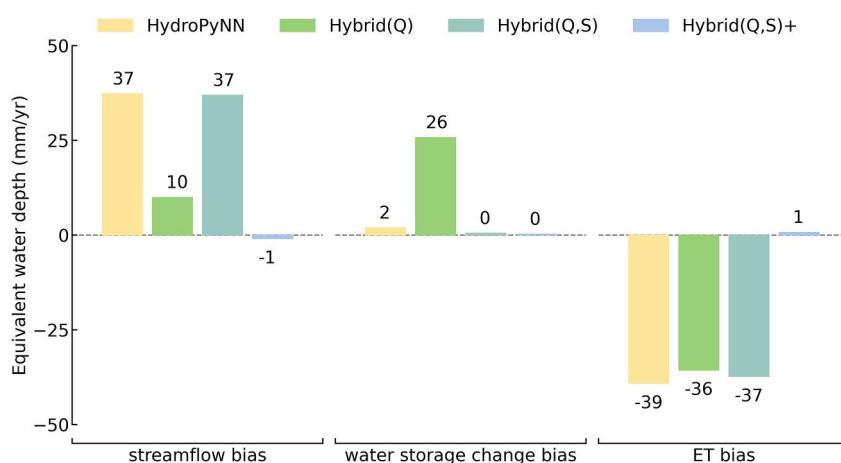


Figure 7. Comparison of the water budget terms in the Obidos catchment during the study period (2003–2016). The biases reflect the difference between the modeled and reference water budgets. A positive (or negative) bias indicates an overestimation (or underestimation) by the model.

overcome the timing issue to different extents. The overestimation of the peak flow is less notable for Hybrid(Q, S)+ than for Hybrid(Q,S). Hybrid(Q,S)+ almost perfectly reproduces the observed hydrograph, with an NSE of 0.92. In the tropical Amazon Basin, the humid climate leads to high variability in performance among PET calculation methods, and therefore, the choice of method can have a large impact on streamflow simulation (Sperna Weiland et al., 2012; Trambauer et al., 2014). The overestimation by Hybrid(Q,S) in the Amazon Basin may be due to the underestimation of overall PET estimated by the Penman–Monteith equation (Trambauer et al., 2014; Weiland et al., 2015). In Section 5.1, the PET issue is further discussed.

4.3. Water Budget Analysis

Data-driven hydrological models do not necessarily guarantee a water balance. In our hybrid modeling framework, incorporating a physical NN as the backbone ensures adherence to the water balance principle, potentially resulting in superior generalization performance compared to conventional DL methods (Hoedt et al., 2021). Here, we derived the simulated annual streamflow (Q), water storage change (ΔS), and ET for the drainage area of the most downstream station (Obidos) based on the models trained with the temporal splitting strategy. ΔS and ET were averaged over the entire drainage area. The simulated values were compared with the respective values derived based on observations, where the observed Q and ΔS were derived from GRDC and GRACE, respectively, and the observation-based ET was derived from the water balance calculation. Figure 7 illustrates the difference between the modeled and observation-based water budgets. When only streamflow is used for training (i.e., in Hybrid(Q)), a positive ΔS bias of 26 mm and a positive Q bias of 10 mm lead to a negative ET bias of -36 mm. Adding the TWSA data for model training reduces the ΔS bias to zero, but the Q and ET biases are 37 mm and -37 mm, respectively, suggesting that the good performance of Hybrid(Q,S) in simulating streamflow and TWSA (Figure 5) comes at the cost of a notable underestimation in ET. In contrast, Hybrid(Q,S)+ adequately reproduces the water budget in all three dimensions, indicating that replacing the original formulation of PET with that based on the replacement NN enhances the physical plausibility of the modeling results.

We further compared the ET modeled by Hybrid(Q,S) and Hybrid(Q,S)+ against the ET based on water balance calculations at the sub-basin scale (Figure 8a). Table S7 in Supporting Information S1 provides more information on the nine sub-basins for comparison. Water balance-based ET shows a generally increasing pattern from southwest to northeast across the basin (Figure 8b), consistent with previous findings (Baker et al., 2021). Hybrid(Q,S) does not render the above pattern (Figure 8c), while Hybrid(Q,S)+ does (Figure 8d). Hybrid(Q,S) yields a notable underestimation of ET in sub-basins 1, 3, and 5, and Hybrid(Q,S)+ ameliorates this deficiency (Figure 8e). In particular, Hybrid(Q,S)+ predicts low ET in sub-basin 4, which may be associated with lower solar radiation and partly associated with local characteristics (e.g., higher elevations and sparse vegetation) (Maeda et al., 2017). Accurately predicting low ET is crucial for simulating runoff in this region. As Figure 9

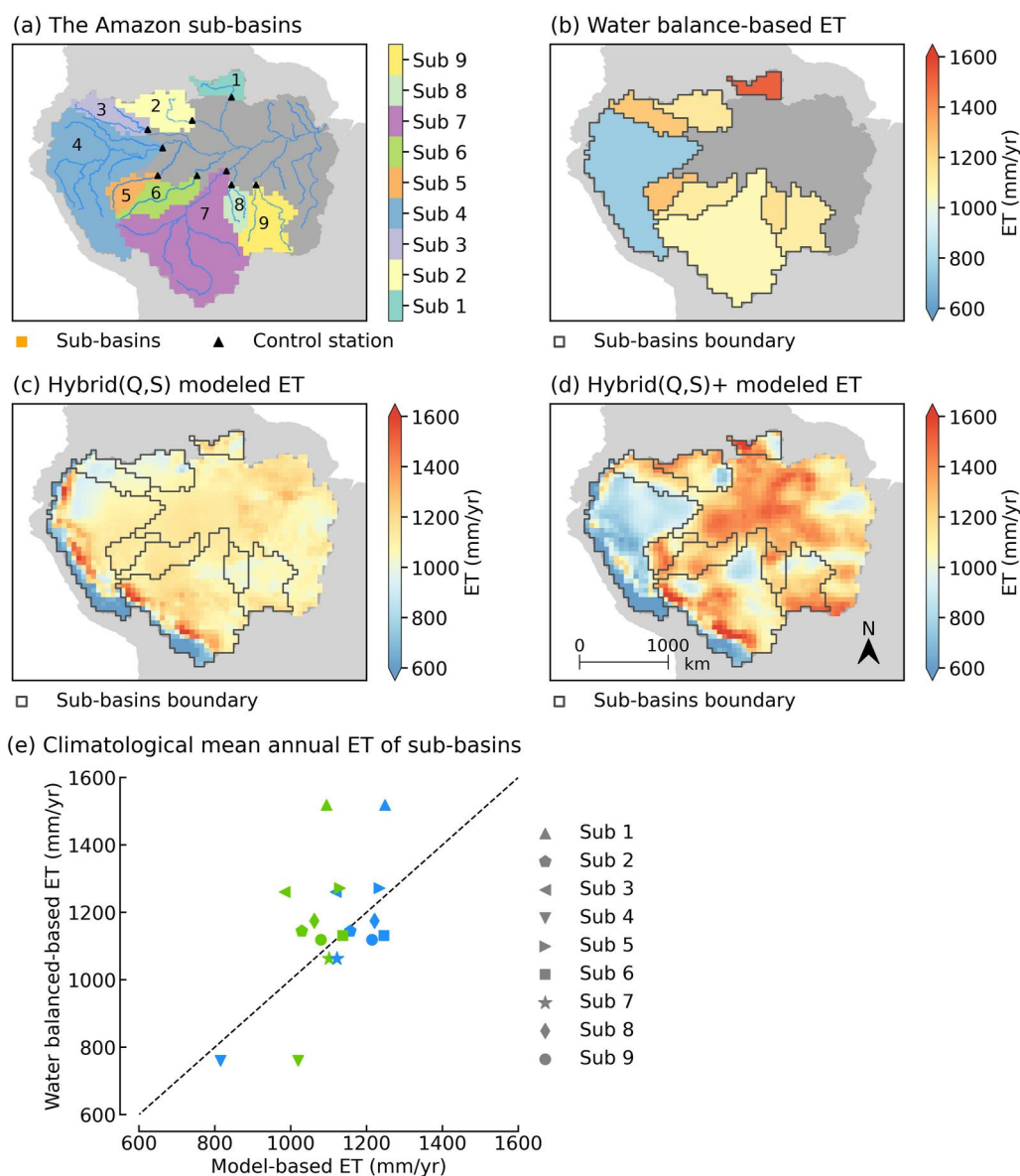


Figure 8. Annual evapotranspiration (ET) in nine sub-basins of the Amazon Basin from 2003 to 2016. (a) The locations of the sub-basins and their respective outlets; (b) the sub-basin ET based on the water balance calculation; (c) the sub-basin ET modeled with Hybrid(Q,S); (d) the sub-basin ET modeled with Hybrid(Q,S)+; and (e) comparisons between the water balance-based ET (Y axis) and the modeled ET (X axis), with the green and blue markers representing Hybrid(Q,S) and Hybrid(Q,S)+, respectively.

shows, the other three NN models all led to a pronounced underestimation of the runoff in this region. The above results further confirm the importance of replacing the Penman–Monteith equation with the NN in the hybrid model.

We compared the measurements from two flux towers, K34 and K67 (see Figure 2), with the modeled ET at the respective grid cells (Figure 10). These measurements exhibit a statistically significant ($p < 0.01$) correlation and are within the same range. Some discrepancies can be partially explained by the following two aspects. The representation scales of ET derived from eddy flux towers and the modeled ET are vastly different. Measurements by eddy flux tower represent a range of only 1–2 km (Paca et al., 2022; Restrepo-Coupe et al., 2021), whereas the grid resolution of the hybrid models exceeds 50 km. In addition, tropical forest flux towers suffer from energy balance closure problems, and an underestimation of total energy fluxes by 20%–30% was reported (Baker

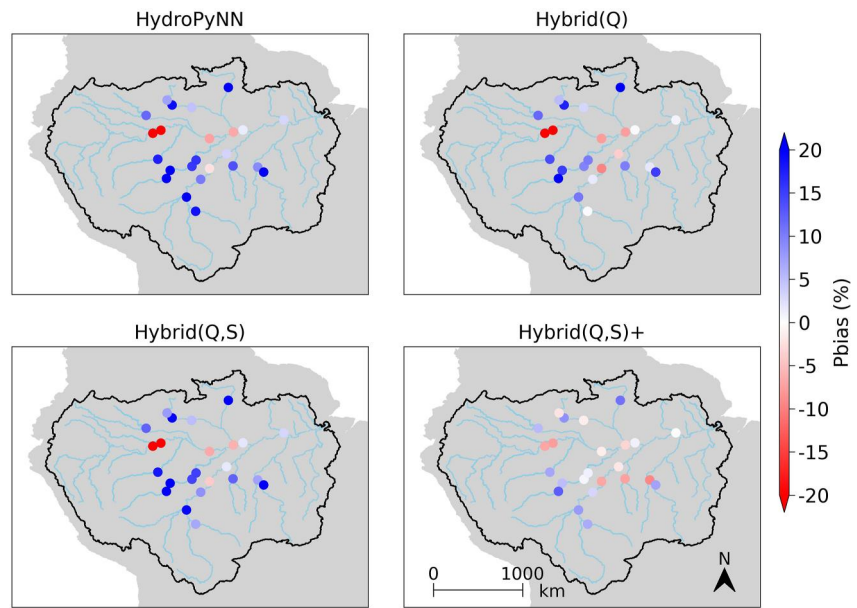


Figure 9. The percentage bias (Pbias) in the streamflow predictions of the neural network-based models at 24 stations from 2003 to 2016.

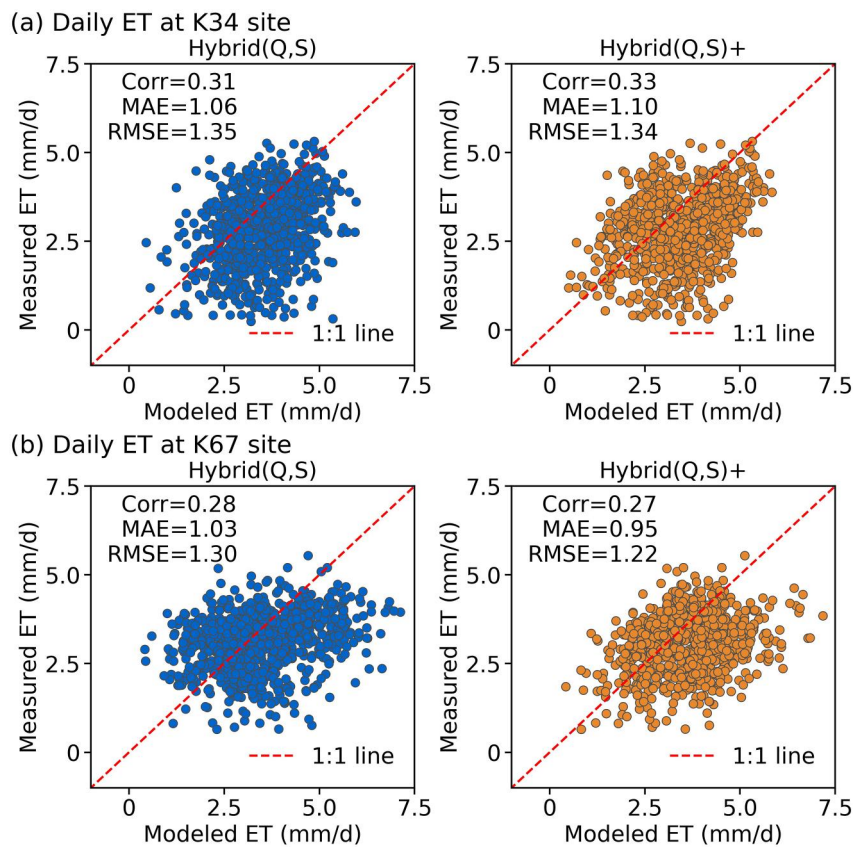


Figure 10. Comparison of daily ET between flux tower data and grid-scale modeling results. (a) Measured ET from the K34 flux tower versus modeled ET from Hybrid(Q,S) and Hybrid(Q,S)+. (b) Measured ET from the K67 flux tower versus modeled ET from Hybrid(Q,S) and Hybrid(Q,S)+. The indicators, Corr, MAE, and root mean square error (RMSE), represent the correlation coefficient, mean absolute error, and RMSE, respectively.

et al., 2021; Fisher et al., 2009). Despite these discrepancies, the results illustrated by Figure 10 strongly suggest that the modeled ET from hybrid models makes physical sense.

5. Discussion

5.1. PET in the Amazon Basin

Estimating PET in tropical regions presents a significant challenge, as a variety of approaches often yield inconsistent results (Sperna Weiland et al., 2012; Trambauer et al., 2014). For the Amazon Basin, we compared the PET derived with Hybrid(Q,S)+ with that from three classic estimation approaches: the Penman–Monteith equation (Allen et al., 1998), the temperature-based Hargreaves–Samani approach (Hargreaves & Samani, 1985) and the radiation-based Priestley–Taylor approach (Priestley & Taylor, 1972). The four approaches are consistent in terms of temporal pattern (Figure 11a), with high PET values during periods of high temperature and high radiation (see Figure S7 in Supporting Information S1). In terms of magnitude, the PET derived with Hybrid(Q, S)+ falls between the estimates of the Penman–Monteith and Priestley–Taylor approaches, consistent with the findings of Trambauer et al. (2014); that is, the Penman–Monteith approach is likely to underestimate PET in the tropics. The PET obtained with Hybrid(Q,S)+ closely aligns with that based on the Penman–Monteith estimation on the 90th to 270th days of the year, but it becomes obviously higher than the PET obtained with the Penman–Monteith method and closer to the PET obtained with the Priestley–Taylor method on the remaining days. Given that the PET derived from Hybrid(Q,S)+ is consistent with previous findings (Trambauer et al., 2014) and that this model performs best in various aspects of the evaluation (e.g., runoff, water storage, and ET), we argue that the PET derived with Hybrid(Q,S)+ may be the most plausible estimate for the Amazon Basin.

The spatial patterns of the four PET estimation approaches (Figure 11b) consistently exhibit low PET values in the southwestern Andes (Figure 2a) but differ in other regions. Both the Penman–Monteith and Hybrid(Q,S)+ methods render a low PET trend in the northwestern region, which is absent in the Priestley–Taylor and Hargreaves–Samani results. This inconsistency in spatial pattern may be due to the differences in input variables. The inputs of the Hargreaves–Samani and Priestley–Taylor approaches are limited to temperature and solar radiation, but the Penman–Monteith method and Hybrid(Q,S)+ consider additional factors, such as humidity and wind speed. In particular, humidity and wind speed were reported to have an important effect on PET calculations in the tropics (Jhajharia et al., 2012). The high PET values in the south revealed by Hybrid(Q,S)+ may be associated with the high local wind speed (Figure S8 in Supporting Information S1). Conversely, the low PET values in the western region revealed by Hybrid(Q,S)+ may be explained by the high humidity and low solar radiation (Figure S8 in Supporting Information S1).

Overall, the distributed hybrid modeling not only produces more plausible PET estimates but also reveals fine-scale details regarding the spatial pattern of PET, providing opportunities for identifying knowledge gaps in PET estimation. It eliminates the need for arbitrarily selecting a PET estimation approach from a variety of choices (Trambauer et al., 2014; Weiland et al., 2015). Note that our framework (Figure 1) is general and not limited to PET modeling; it can be used to address other processes that are not clearly understood or lack direct observations.

5.2. Insights From Model Parameterization

The model parameterization through NN training not only leads to computational efficiency but also provides hydrological insights (Feng et al., 2022; Tsai et al., 2021). In the Amazon Basin, vegetation transpiration largely influences the spatial pattern of basin ET, and therefore, interpreting the parameterization results for transpiration modeling may enhance our understanding of the ET process. Here, the critical root zone moisture fraction (f_{crit}) in HydroPy is selected as an example. HydroPy distinguishes two states of vegetation transpiration: the water-limited regime, in which transpiration is mainly controlled by soil moisture availability, and the energy-limited regime, in which transpiration is mostly governed by the energy supply. f_{crit} defines the threshold for the state transition of vegetation transpiration. When the soil moisture fraction (f_{soil}) is higher than f_{crit} , vegetation transpiration reaches a maximum; otherwise, vegetation transpiration is constrained by water stress. Therefore, this critical parameter controls soil-vegetation-atmosphere interaction behaviors (Denissen et al., 2020; Novák & Havrila, 2006). The specific functional relationship among transpiration, f_{soil} and f_{crit} is described in Text S4 in Supporting Information S1. Figure 12a illustrates the f_{crit} values derived from Hybrid(Q,S)+. Apparently, f_{crit} is low in the central Amazon plains and high in the Andes at the southwestern boundary of the basin and at the corner of the Guiana Plateau (Figure 2a) in the upper-middle basin, with a clear and spatially continuous pattern.

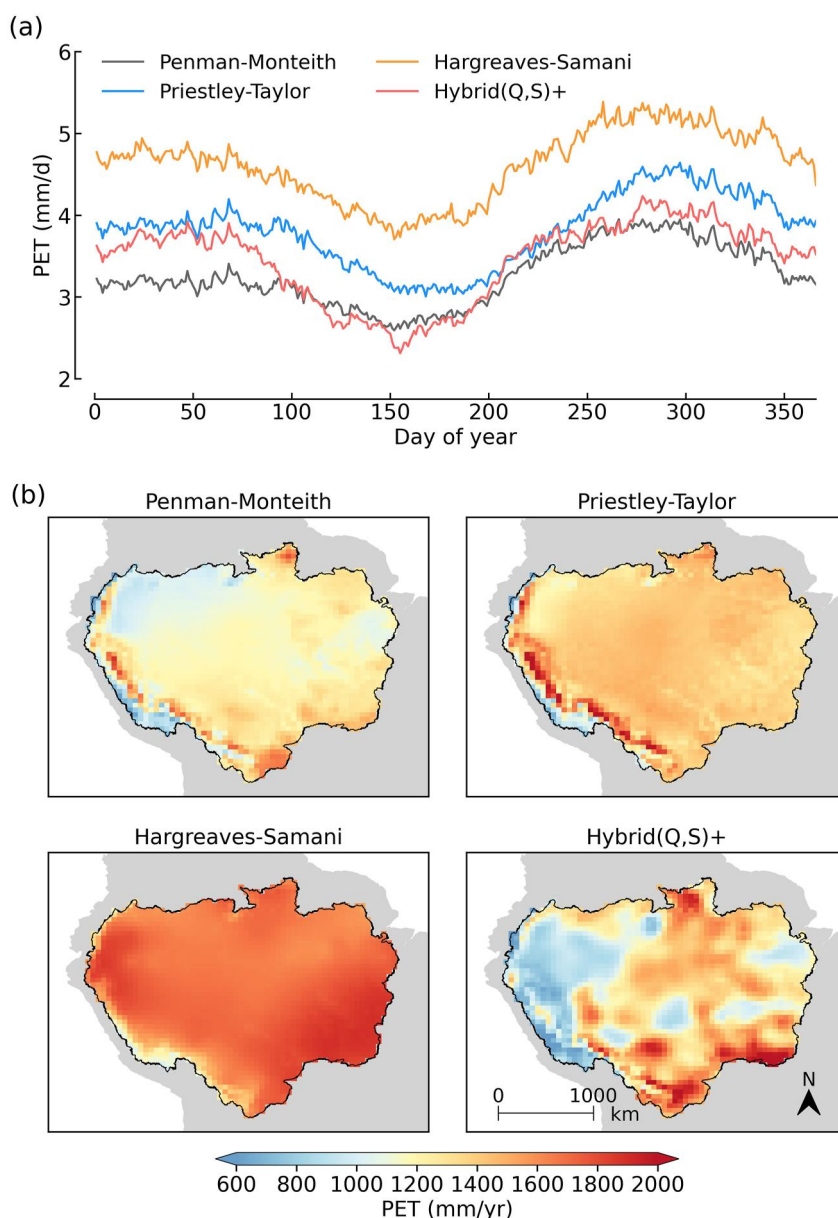


Figure 11. Spatiotemporal comparison of plausible potential evapotranspiration (PET) derived from classic empirical formulas (Penman–Monteith, Priestley–Taylor, and Hargreaves–Samani) and Hybrid(Q,S)+. (a) Daily basin-scale climatology for PET among the four approaches. (b) Spatial maps of annual PET at the $0.5^\circ \times 0.5^\circ$ grid scale for the four approaches. The results above are calculated for the period from 2003 to 2016.

To further investigate the possible causes of the distribution illustrated in Figure 12a, we used the EG method to analyze the contributions of static attributes to f_{crit} in the parameterization NN. Figure 12b summarizes the mean absolute ϕ_i for each individual attribute across all 1,951 grid cells, and their global importance to the parameterization of f_{crit} is assessed. A high value indicates a high contribution. As revealed by Figure 12b, the four most important attributes are the difference between the maximum and minimum monthly leaf area indices (LAI_{diff}), the average daily temperature ($Temp_{avg}$), the average monthly vegetation fraction ($Fveg_{avg}$) and the maximum monthly vegetation fraction ($Fveg_{max}$), highlighting the dominant impacts of vegetation type and climate. The spatial consistency between the distribution of vegetation types (Figure 12c) and the distribution of the total contribution of LAI_{diff} , $Fveg_{avg}$ and $Fveg_{max}$ to the parameterization of f_{crit} (Figure 12d) suggests that shrubland and grassland tend to lead to high f_{crit} , with the opposite trend observed for forest. This may be due to the high

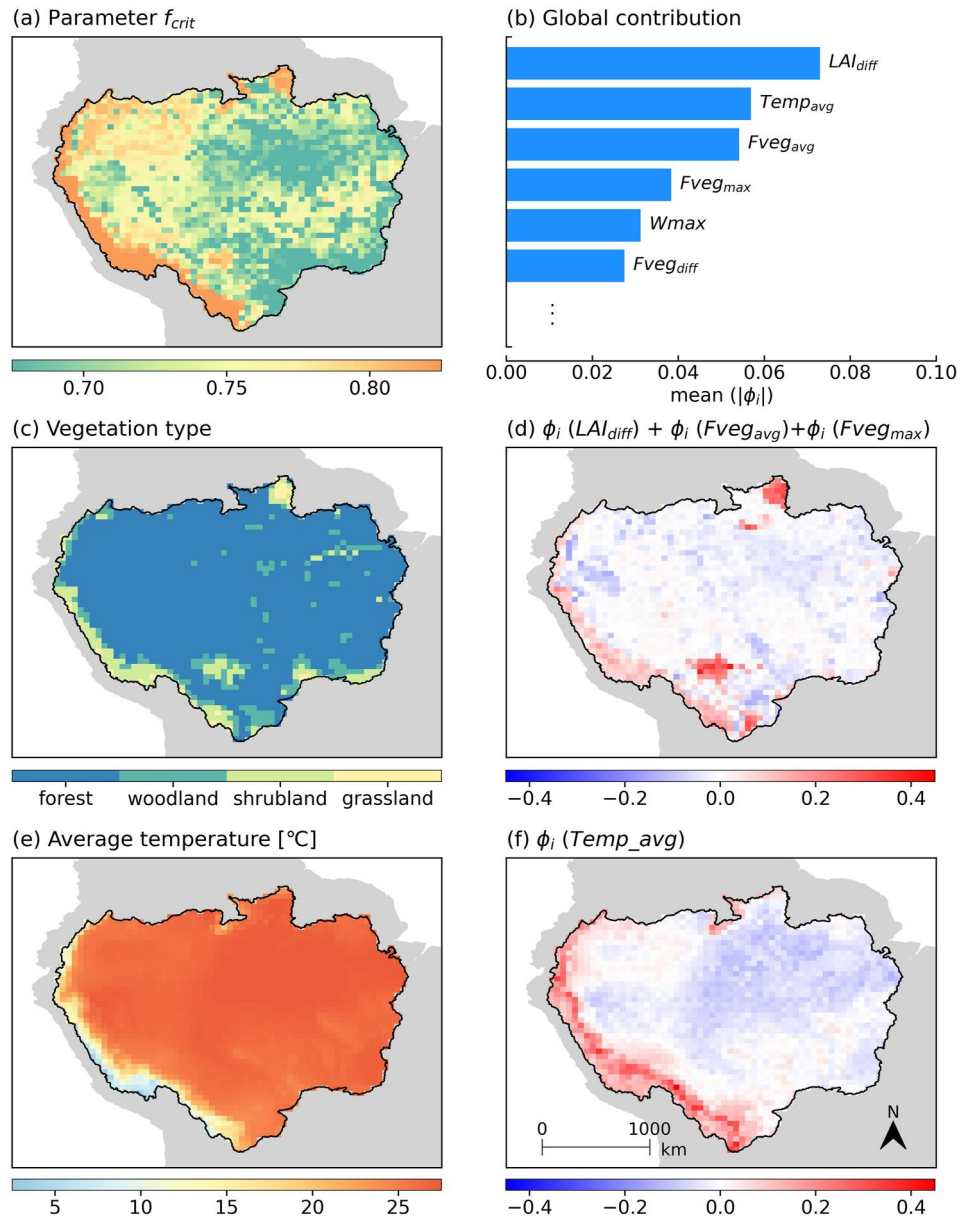


Figure 12. Parameterization of Hybrid(Q,S)+ and its backward interpretation. (a) Spatial distribution of the inferred values of f_{crit} . (b) Mean absolute ϕ_i for the six most important static attributes. The six important attributes are the difference between the maximum and minimum leaf area indices (LAI_{diff}), the average temperature ($Temp_{avg}$), the average monthly vegetation fraction ($Fveg_{avg}$), the maximum monthly vegetation fraction ($Fveg_{max}$), the maximum water holding capacity ($Wmax$), and the difference between the maximum and minimum vegetation fractions ($Fveg_{diff}$). (c) Spatial distribution of vegetation types based on the University of Maryland Global Land Cover Classifications. (d) Spatial distribution of the sum of ϕ_i for the three vegetation-related attributes. (e) Spatial distribution of average daily temperature from 2003 to 2016; (f) Spatial distribution of ϕ_i for average daily temperature. The spatial resolution of (a), (c), (d), (e), and (f) is $0.5^\circ \times 0.5^\circ$.

susceptibility of low vegetation (e.g., shrubland and grassland) with shallow root systems to water limitation issues (Denissen et al., 2020). In contrast, the root zones of trees in forested areas are deep and have access to deep water reservoirs to avoid such water limitations. Additionally, based on model interpretation, the possible contribution of temperature to f_{crit} can be assessed (Figures 12e and 12f). In the southwestern Andes region, low temperatures ($<15^\circ\text{C}$) corresponded to higher positive ϕ_i values, and in other regions, this contribution was weakly negative (or there was no contribution). The combination of low temperature and low vegetation likely leads to the high f_{crit} in this region. Overall, the above discussion reinforces the point that the distributed hybrid

modeling framework proposed in this study not only improves the efficiency and accuracy of large-scale hydrological modeling but also ensures the physical plausibility and consistency of the modeling.

5.3. Limitations and Future Work

The hybrid models developed in this study have not considered the impact of human activities, such as building dams and reservoirs. The Amazon has a growing number of hydropower dams (Latrubesse et al., 2017), and therefore the lack of reservoir module may limit the simulation accuracy and generalizability of the hybrid models. Nevertheless, our framework can be easily extended to include reservoir modules, either within the physical NN or as a replacement NN (Figure 1). Process-based reservoir modules (Dang et al., 2020) can be encoded into the physical NN in the same way as HydroPy's equations. DL models simulating reservoir operation in a black-box way (He et al., 2022) can also be adopted as the replacement NN which can be further attached to the physical backbone.

Uncertainty is a common issue in hydrological modeling. This study addressed random initial weights and biases, which represent a major source of uncertainty for NN-based models (Lakshminarayanan et al., 2017), by using 10 random repeat runs. Figures S9 and S10 in Supporting Information S1 show that this uncertainty is small in this study case. Figure S9 in Supporting Information S1 shows this uncertainty in terms of variations in prediction accuracy, while Figure S10 in Supporting Information S1 shows it in terms of variations in inferred parameter and PET values. However, DL-based hydrological models are also susceptible to uncertainty associated with model inputs and observational data for model calibration/training, similar to classic DHMs. For example, both the GRACE-based TWSA products and gauge-observed streamflow have errors (Di Baldassarre & Montanari, 2009; Loomis et al., 2019). In traditional hydrological modeling, these errors can be addressed by statistical approaches such as Bayesian methods (F. Han & Zheng, 2018). How to integrate classic uncertainty analysis approaches with deep NNs, particularly in a fully differentiable way, is an interesting topic for future research.

6. Conclusions

In this study, a fully differentiable framework was developed to perform distributed hydrological modeling with physics-encoded DL. The framework seamlessly integrates process-based runoff and river routing models encoded as NNs, an NN to map spatially distributed and physically meaningful parameters from watershed attributes, and NN-based replacement models for unknown or vaguely known processes. The framework accommodates multi-source hydrological observations as training data. The framework is applied to the Amazon Basin, with HydroPy, a global-scale DHM, encoded as the physical backbone of the hybrid DL model. The major study findings are summarized below.

The novel framework enables automatic parameter tuning for DHMs in large-scale (i.e., continental to global-scale) applications. In the Amazon case, tuning the parameters of HydroPy (the NN version, HydroPyNN) by backpropagation led to a significant increase in the accuracy of streamflow prediction (the median NSE value was enhanced from 0.59 to 0.76, see Figure 5a). The addition of GRACE data in the parameter tuning process (i.e., Hybrid(Q,S)) improved the model performance in TWSA prediction but slightly degraded that in streamflow prediction. Further activation of the replacement NN for PET calculation (i.e., Hybrid(Q,S)+) resolved the tradeoff between the two modeling objectives, and the median NSE values for both streamflow and TWSA reached their maxima in the temporal learning scheme (0.83 and 0.77, respectively, see Figure 5). Even in the more challenging scheme of spatial learning, the median NSE of Hybrid(Q,S)+ was 0.80 for streamflow, much higher than that of HydroPyNN (0.63). Replacing the original Penman–Monteith formulation for PET with the replacement NN not only produced more plausible PET estimates for the Amazon Basin but also aided in identifying fine-scale details in the spatial pattern of PET, thus providing opportunities for identifying knowledge gaps in PET estimation. In addition, the model parameterization through NN training provided important hydrological insights. Interpretation of the parameterization NN revealed that f_{crit} , a critical parameter controlling soil-vegetation-atmosphere interaction behaviors, in HydroPy is dominantly impacted by vegetation type and climate. The combination of low temperature and low vegetation is likely the cause of the high f_{crit} in this region.

Overall, this study lays out a feasible technical roadmap for distributed hydrological modeling in the big data era. The framework is general and extendable. Future research may investigate the potential and develop methods for

wrapping hydrological system equations into other NN architectures such as Gated Recurrent Unit (Cho et al., 2014) and LSTM. Other types of observational data could be considered. For example, satellite-based soil moisture data can be used to train hybrid models that simulate multi-layer soil moisture dynamics. In addition, investigating the effects of different spatial resolutions on the performance of hybrid models would be an interesting avenue for future research.

Data Availability Statement

The WFDEI meteorological data set used in this study was obtained from the ISIMIP portal (<https://www.isimip.org/>). The static attribute data set was obtained from Stacke and Hagemann (2021b, <https://zenodo.org/record/4541239>). The major river networks and streamflow observations for the Amazon Basin were obtained from the Global Runoff Data Centre (GRDC, 2022, <https://www.bafg.de/GRDC/>). The GRACE Mascon products were obtained from three agencies: the Center for Space Research at the University of Texas (CSR, <https://www2.csr.utexas.edu/>), NASA's Jet Propulsion Laboratory (JPL, <https://grace.jpl.nasa.gov/>), and NASA's Goddard Space Flight Center (GSFC, <https://earth.gsfc.nasa.gov/>). The source code of HydroPy was obtained from Stacke and Hagemann (2021a, <https://zenodo.org/record/4730160>). In addition, the code for the hybrid model in this study is available at <https://zenodo.org/record/8251987>.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (42325702 and 92047302). This work was technically supported by the Center for Computational Science and Engineering at Southern University of Science and Technology. Additional support was provided by the Carl Zeiss Foundation (Junior Research Group “Knowledge integration for spatio-temporal environmental modeling”).

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Paper presented at the OSDI*.
- Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56.
- Baker, J. C. A., Garcia-Carreras, L., Gloor, M., Marsham, J. H., Buermann, W., da Rocha, H. R., et al. (2021). Evapotranspiration in the Amazon: Spatial patterns, seasonality, and recent trends in observations, reanalysis, and climate models. *Hydrology and Earth System Sciences*, 25(4), 2279–2300. <https://doi.org/10.5194/hess-25-2279-2021>
- Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1–2), 18–36. <https://doi.org/10.1016/j.jhydrol.2005.07.007>
- Beven, K. (2020). Deep learning, hydrological processes and the uniqueness of place. *Hydrological Processes*, 34(16), 3608–3613. <https://doi.org/10.1002/hyp.13805>
- Bhasme, P., Vagadiya, J., & Bhatia, U. (2022). Enhancing predictive skills in physically-consistent way: Physics informed machine learning for hydrological processes. *Journal of Hydrology*, 615, 128618. <https://doi.org/10.1016/j.jhydrol.2022.128618>
- Bindas, T., Tsai, W. P., Liu, J., Rahmani, F., Feng, D., Bian, Y., et al. (2024). Improving river routing using a differentiable Muskingum-Cunge model and physics-informed machine learning. *Water Resources Research*, 60(1), e2023WR035337. <https://doi.org/10.1029/2023wr035337>
- Chagas, V. B. P., Chaffe, P. L. B., & Blöschl, G. (2022). Climate and land management accelerate the Brazilian water cycle. *Nature Communications*, 13(1), 5136. <https://doi.org/10.1038/s41467-022-32580-x>
- Chen, C., Jiang, J., Liao, Z., Zhou, Y., Wang, H., & Pei, Q. (2022). A short-term flood prediction based on spatial deep learning network: A case study for Xi County, China. *Journal of Hydrology*, 607, 127535. <https://doi.org/10.1016/j.jhydrol.2022.127535>
- Chen, M., Qian, Z., Boers, N., Jakeman, A. J., Kettner, A. J., Brandt, M., et al. (2023). Iterative integration of deep learning in hybrid Earth surface system modelling. *Nature Reviews Earth & Environment*, 4(8), 568–581. <https://doi.org/10.1038/s43017-023-00452-7>
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. <https://doi.org/10.48550/arXiv.1406.1078>
- Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., et al. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, 21(7), 3427–3440. <https://doi.org/10.5194/hess-21-3427-2017>
- Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, 52(3), 2350–2365. <https://doi.org/10.1002/2015wr017910>
- Dang, T. D., Chowdhury, A. F. M. K., & Galelli, S. (2020). On the representation of water reservoir storage and operations in large-scale hydrological models: Implications on model parameterization and climate change impact assessments. *Hydrology and Earth System Sciences*, 24, 397–416. <https://doi.org/10.5194/hess-2019-334>
- Dembéle, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. *Water Resources Research*, 56(1), e2019WR026085. <https://doi.org/10.1029/2019wr026085>
- Denissen, J. M. C., Teuling, A. J., Reichstein, M., & Orth, R. (2020). Critical soil moisture derived from satellite observations over Europe. *Journal of Geophysical Research: Atmospheres*, 125(6), e2019JD031672. <https://doi.org/10.1029/2019jd031672>
- de Paiva, R. C. D., Buarque, D. C., Collischonn, W., Bonnet, M.-P., Frappart, F., Calmant, S., & Bulhões Mendes, C. A. (2013). Large-scale hydrologic and hydrodynamic modeling of the Amazon River basin. *Water Resources Research*, 49(3), 1226–1243. <https://doi.org/10.1002/wrcr.20067>
- Di Baldassarre, G., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. *Hydrology and Earth System Sciences*, 13(6), 913–921. <https://doi.org/10.5194/hess-13-913-2009>
- Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., & Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7), 620–631. <https://doi.org/10.1038/s42256-021-00343-w>
- Fassoni-Andrade, A. C., Fleischmann, A. S., Papa, F., Paiva, R. C., Wongchuig, S., Melack, J. M., et al. (2021). Amazon hydrology from space: Scientific advances and future challenges. *Reviews of Geophysics*, 59(4), e2020RG000728. <https://doi.org/10.1029/2020rg000728>

- Faticchi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., et al. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology*, *537*, 45–60. <https://doi.org/10.1016/j.jhydrol.2016.03.026>
- Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research*, *58*(10), e2022WR032404. <https://doi.org/10.1029/2022wr032404>
- Fisher, J. B., Malhi, Y., Bonal, D., Da Rocha, H. R., De Araújo, A. C., Gamo, M., et al. (2009). The land–atmosphere water flux in the tropics. *Global Change Biology*, *15*(11), 2694–2714. <https://doi.org/10.1111/j.1365-2486.2008.01813.x>
- Gebremedhin, M. A., Lubczynski, M. W., Maathuis, B. H. P., & Tekla, D. (2022). Deriving potential evapotranspiration from satellite-based reference evapotranspiration, Upper Tekeze Basin, Northern Ethiopia. *Journal of Hydrology: Regional Studies*, *41*, 101059. <https://doi.org/10.1016/j.ejrh.2022.101059>
- GRDC. (2022). Retrieved from <https://www.bafg.de/GRDC>
- Güntner, A. (2008). Improvement of global hydrological models using GRACE data. *Surveys in Geophysics*, *29*(4–5), 375–397. <https://doi.org/10.1007/s10712-008-9038-y>
- 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*(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Hagemann, S., & Dümenil, L. (1997). A parametrization of the lateral waterflow for the global scale. *Climate Dynamics*, *14*(1), 17–31. <https://doi.org/10.1007/s003820050205>
- Han, F., & Zheng, Y. (2018). Joint analysis of input and parametric uncertainties in watershed water quality modeling: A formal Bayesian approach. *Advances in Water Resources*, *116*, 77–94. <https://doi.org/10.1016/j.advwatres.2018.04.006>
- Han, L., Chen, M., Chen, K., Chen, H., Zhang, Y., Lu, B., et al. (2021). A deep learning method for bias correction of ECMWF 24–240 h forecasts. *Advances in Atmospheric Sciences*, *38*(9), 1444–1459. <https://doi.org/10.1007/s00376-021-0215-y>
- Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99. <https://doi.org/10.13031/2013.26773>
- He, S., Guo, S., Zhang, J., Liu, Z., Cui, Z., Zhang, Y., & Zheng, Y. (2022). Multi-objective operation of cascade reservoirs based on short-term ensemble streamflow prediction. *Journal of Hydrology*, *610*, 127936. <https://doi.org/10.1016/j.jhydrol.2022.127936>
- Herath, H. M. V. V., Chadalawada, J., & Babovic, V. (2021). Hydrologically informed machine learning for rainfall–runoff modelling: Towards distributed modelling. *Hydrology and Earth System Sciences*, *25*(8), 4373–4401. <https://doi.org/10.5194/hess-25-4373-2021>
- Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., et al. (2021). MC-LSTM: Mass-conserving lstm. *CoRR*. <https://doi.org/10.48550/arXiv.2101.05186>
- Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., & Fencica, F. (2022). Improving hydrologic models for predictions and process understanding using neural ODEs. *Hydrology and Earth System Sciences*, *26*(19), 5085–5102. <https://doi.org/10.5194/hess-26-5085-2022>
- Jhajharia, D., Dinpashoh, Y., Kahya, E., Singh, V. P., & Fakheri-Fard, A. (2012). Trends in reference evapotranspiration in the humid region of northeast India. *Hydrological Processes*, *26*(3), 421–435. <https://doi.org/10.1002/hyp.8140>
- Jiang, S., Bevacqua, E., & Zscheischler, J. (2022). River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrology and Earth System Sciences*, *26*(24), 6339–6359. <https://doi.org/10.5194/hess-26-6339-2022>
- 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>
- Jiang, S., Zheng, Y., Wang, C., & Babovic, V. (2022). Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. *Water Resources Research*, *58*(1), e2021WR030185. <https://doi.org/10.1029/2021wr030185>
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440. <https://doi.org/10.1038/s42254-021-00314-5>
- Kingma, D. P., & Ba, J. J. (2014). Adam: A method for stochastic optimization.
- Konapala, G., Kao, S.-C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environmental Research Letters*, *15*(10), 104022. <https://doi.org/10.1088/1748-9326/aba927>
- Kraft, B., Jung, M., Körner, M., Koiraal, S., & Reichstein, M. (2022). Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences*, *26*(6), 1579–1614. <https://doi.org/10.5194/hess-26-1579-2022>
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, *30*. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html
- Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*(4), W04531. <https://doi.org/10.1029/2011wr011453>
- Latrubesse, E. M., Arima, E. Y., Dunne, T., Park, E., Baker, V. R., d’Horta, F. M., et al. (2017). Damming the rivers of the Amazon basin. *Nature*, *546*(7658), 363–369. <https://doi.org/10.1038/nature22333>
- Loomis, B. D., Luthcke, S. B., & Sabaka, T. J. (2019). Regularization and error characterization of GRACE mascons. *Journal of Geodynamics*, *93*(9), 1381–1398. <https://doi.org/10.1007/s00190-019-01252-y>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.
- Maeda, E. E., Ma, X., Wagner, F. H., Kim, H., Oki, T., Eamus, D., & Huete, A. (2017). Evapotranspiration seasonality across the Amazon Basin. *Earth System Dynamics*, *8*(2), 439–454. <https://doi.org/10.5194/esd-8-439-2017>
- Marengo, J. A., & Espinoza, J. C. (2016). Extreme seasonal droughts and floods in Amazonia: Causes, trends and impacts. *International Journal of Climatology*, *36*(3), 1033–1050. <https://doi.org/10.1002/joc.4420>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3), e2020WR028091. <https://doi.org/10.1029/2020wr028091>
- Novák, V., & Havrila, J. (2006). Method to estimate the critical soil water content of limited availability for plants. *Biologia*, *61*(S19), S289–S293. <https://doi.org/10.2478/s11756-006-0175-9>
- Paca, V. H., Espinoza-Dávalos, G. E., da Silva, R., Tapajós, R., & dos Santos Gaspar, A. B. (2022). Remote sensing products validated by flux tower data in Amazon rain forest. *Remote Sensing*, *14*(5), 1259. <https://doi.org/10.3390/rs14051259>
- Paniconi, C., & Putti, M. (2015). Physically based modeling in catchment hydrology at 50: Survey and outlook. *Water Resources Research*, *51*(9), 7090–7129. <https://doi.org/10.1002/2015wr017780>

- Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, 100(2), 81–92. [https://doi.org/10.1175/1520-0493\(1972\)100<0081:OTAOSH>2.3.CO;2](https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2)
- Rakovec, O., Kumar, R., Attinger, S., & Samaniego, L. (2016). Improving the realism of hydrologic model functioning through multivariate parameter estimation. *Water Resources Research*, 52(10), 7779–7792. <https://doi.org/10.1002/2016wr019430>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Restrepo-Coupe, N., da Rocha, H. R., Hutyrá, L. R., de Araujo, A. C., Borma, L. S., Christoffersen, B., et al. (2021). *LBA-ECO CD-32 flux tower network data compilation, Brazilian Amazon: 1999-2006*, V2. ORNL Distributed Active Archive Center.
- Sadler, J. M., Appling, A. P., Read, J. S., Oliver, S. K., Jia, X., Zwart, J. A., & Kumar, V. (2022). Multi-task deep learning of daily streamflow and water temperature. *Water Resources Research*, 58(4), e2021WR030138. <https://doi.org/10.1029/2021wr030138>
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Lecture Notes in Computer Science (Vol. 11700). Springer Nature. <https://doi.org/10.1007/978-3-030-28954-6>
- Save, H., Bettadpur, S., & Tapley, B. D. (2016). High-resolution CSR GRACE RL05 mascons. *Journal of Geophysical Research: Solid Earth*, 121(10), 7547–7569. <https://doi.org/10.1002/2016jb013007>
- Schumacher, M., Kusche, J., & Döll, P. (2016). A systematic impact assessment of GRACE error correlation on data assimilation in hydrological models. *Journal of Geodesy*, 90(6), 537–559. <https://doi.org/10.1007/s00190-016-0892-y>
- Semenova, O., & Beven, K. (2015). Barriers to progress in distributed hydrological modelling. *Hydrological Processes*, 29(8), 2074–2078. <https://doi.org/10.1002/hyp.10434>
- 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/10.1029/2018wr022643>
- Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, 4(8), 552–567. <https://doi.org/10.1038/s43017-023-00450-9>
- Shen, C., Chen, X., & Laloy, E. (2021). Editorial: Broadening the use of machine learning in hydrology. *Frontiers in Water*, 3, 681023. <https://doi.org/10.3389/frwa.2021.681023>
- Simmons, C. T., Brunner, P., Therrien, R., & Sudicky, E. A. (2020). Commemorating the 50th anniversary of the Freeze and Harlan (1969) blueprint for a physically-based digitally-simulated hydrologic response model. *Journal of Hydrology*, 584, 124309. <https://doi.org/10.1016/j.jhydrol.2019.124309>
- Soltani, S. S., Ataie-Ashtiani, B., & Simmons, C. T. (2021). Review of assimilating GRACE terrestrial water storage data into hydrological models: Advances, challenges and opportunities. *Earth-Science Reviews*, 213, 103487. <https://doi.org/10.1016/j.earscirev.2020.103487>
- Sperna Weiland, F. C., Tisseuil, C., Dürr, H. H., Vrac, M., & van Beek, L. P. H. (2012). Selecting the optimal method to calculate daily global reference potential evaporation from CFSR reanalysis data for application in a hydrological model study. *Hydrology and Earth System Sciences*, 16(3), 983–1000. <https://doi.org/10.5194/hess-16-983-2012>
- Stacke, T., & Hagemann, S. (2012). Development and evaluation of a global dynamical wetlands extent scheme. *Hydrology and Earth System Sciences*, 16(8), 2915–2933. <https://doi.org/10.5194/hess-16-2915-2012>
- Stacke, T., & Hagemann, S. (2021a). HydroPy (v1.0): A new global hydrology model written in Python. *Geoscientific Model Development*, 14(12), 7795–7816. <https://doi.org/10.5194/gmd-14-7795-2021>
- Stacke, T., & Hagemann, S. (2021b). Land surface parameter fields at 0.5deg resolution for use with the HydroPy model. <https://doi.org/10.5281/zenodo.4541239>
- Sun, A. Y., Jiang, P., Yang, Z.-L., Xie, Y., & Chen, X. (2022). A graph neural network (GNN) approach to basin-scale river network learning: The role of physics-based connectivity and data fusion. *Hydrology and Earth System Sciences*, 26(19), 5163–5184. <https://doi.org/10.5194/hess-26-5163-2022>
- Swann, A. L. S., & Koven, C. D. (2017). A direct estimate of the seasonal cycle of evapotranspiration over the Amazon Basin. *Journal of Hydrometeorology*, 18(8), 2173–2185. <https://doi.org/10.1175/jhm-d-17-0004.1>
- Tapley, B. D., Bettadpur, S., Watkins, M., & Reigber, C. (2004). The gravity recovery and climate experiment: Mission overview and early results. *Geophysical Research Letters*, 31(9), L09607. <https://doi.org/10.1029/2004gl019920>
- Trambauer, P., Dutra, E., Maskey, S., Werner, M., Pappenberger, F., van Beek, L. P. H., & Uhlénbrook, S. (2014). Comparison of different evaporation estimates over the African continent. *Hydrology and Earth System Sciences*, 18(1), 193–212. <https://doi.org/10.5194/hess-18-193-2014>
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Paper presented at the 2015 IEEE international conference on computer vision (ICCV)*. <https://doi.org/10.1109/iccv.2015.510>
- Troy, T. J., Wood, E. F., & Sheffield, J. (2008). An efficient calibration method for continental-scale land surface modeling. *Water Resources Research*, 44(9), W09411. <https://doi.org/10.1029/2007wr006513>
- Tsai, W. P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 12(1), 5988. <https://doi.org/10.1038/s41467-021-26107-z>
- Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH forcing data methodology applied to ERA-interim reanalysis data. *Water Resources Research*, 50(9), 7505–7514. <https://doi.org/10.1002/2014wr015638>
- Weiland, F. C. S., Lopez, P., Dijk, A. I. J. M., & Schellekens, J. (2015). Global high-resolution reference potential evaporation. In *21st international congress on modelling and simulation*.
- Wiese, D. N., Landerer, F. W., & Watkins, M. M. (2016). Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. *Water Resources Research*, 52(9), 7490–7502. <https://doi.org/10.1002/2016wr019344>
- Wunsch, A., Liesch, T., & Broda, S. (2022). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nature Communications*, 13(1), 1221. <https://doi.org/10.1038/s41467-022-28770-2>
- Xia, J., Liu, Q., & Tan, L. (2023). A deep learning method integrating multisource data for ECMWF forecasting products correction. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5. <https://doi.org/10.1109/lgrs.2023.3307717>
- Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., & Shen, C. (2021). Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. *Journal of Hydrology*, 603, 127043. <https://doi.org/10.1016/j.jhydrol.2021.127043>
- Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. *WIREs Water*, 8(5), e1533. <https://doi.org/10.1002/wat2.1533>
- Xu, T., Longyang, Q., Tyson, C., Zeng, R., & Neilson, B. T. (2022). Hybrid physically based and deep learning modeling of a snow dominated, mountainous, Karst watershed. *Water Resources Research*, 58(3), e2021WR030993. <https://doi.org/10.1029/2021wr030993>

- Zhu, S., Wei, J., Zhang, H., Xu, Y., & Qin, H. (2023). Spatiotemporal deep learning rainfall-runoff forecasting combined with remote sensing precipitation products in large scale basins. *Journal of Hydrology*, *616*, 128727. <https://doi.org/10.1016/j.jhydrol.2022.128727>
- Zuluaga, C. F., Avila-Diaz, A., Justino, F. B., & Wilson, A. B. (2021). Climatology and trends of downward shortwave radiation over Brazil. *Atmospheric Research*, *250*, 105347. <https://doi.org/10.1016/j.atmosres.2020.105347>

References From the Supporting Information

- Iliev, A., Kyurkchiev, N., & Markov, S. (2017). On the approximation of the step function by some sigmoid functions. *Mathematics and Computers in Simulation*, *133*, 223–234. <https://doi.org/10.1016/j.matcom.2015.11.005>