

Water Resources Research®



RESEARCH ARTICLE

10.1029/2024WR037736

A Multi-Resolution Deep-Learning Surrogate Framework for Global Hydrological Models

B. Droppers¹ , M. F. P. Bierkens^{1,2} , and N. Wanders¹ 

¹Department of Physical Geography, Utrecht University, Utrecht, The Netherlands, ²Unit Subsurface and Groundwater Systems, Deltares, Utrecht, The Netherlands

Key Points:

- We introduce a framework for the development of multi-resolution deep-learning global hydrological surrogates
- To test our framework, we develop and evaluate a surrogate for the PCRaster Global Water Balance (PCR-GLOBWB) model
- Our framework allows the global hydrological community to develop multi-resolution deep-learning surrogates for their own models

Supporting Information:

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

Correspondence to:

B. Droppers,
b.droppers@gmail.com

Citation:

Droppers, B., Bierkens, M. F. P., & Wanders, N. (2025). A multi-resolution deep-learning surrogate framework for global hydrological models. *Water Resources Research*, 61, e2024WR037736. <https://doi.org/10.1029/2024WR037736>

Received 16 APR 2024

Accepted 6 MAR 2025

Author Contributions:

Conceptualization: B. Droppers,

M. F. P. Bierkens, N. Wanders

Data curation: B. Droppers

Formal analysis: B. Droppers

Investigation: B. Droppers

Methodology: B. Droppers,

M. F. P. Bierkens, N. Wanders

Supervision: M. F. P. Bierkens,
N. Wanders

Visualization: B. Droppers

Writing – original draft: B. Droppers

Writing – review & editing: B. Droppers,
M. F. P. Bierkens, N. Wanders

Abstract Global hydrological models are important decision support tools for policy making in today's water-scarce world as their process-based nature allows for worldwide water resources assessments under various climate-change and socio-economic scenarios. Although efforts are continuously being made to improve water resource assessments, global hydrological model computational demands have dramatically increased and calibrating them has proven difficult. To address these issues, deep-learning approaches have gained prominence in the hydrological community, in particular the development of deep-learning surrogates. Nevertheless, the development of deep-learning global hydrological model surrogates remains limited, as most surrogate frameworks only focus on natural water states and fluxes at a single spatial resolution. Therefore, we introduce a global hydrological model surrogate framework that integrates spatially distributed runoff routing, including lake outflow and reservoir operation, includes human activities, such as water abstractions, and can scale across spatial resolutions. To test our framework, we develop a deep-learning surrogate for the PCRaster Global Water Balance (PCR-GLOBWB) global hydrological model. Our surrogate performed well when compared to the model outputs, with a median Kling-Gupta Efficiency of 0.50, while predictions were at least an order of magnitude faster. Moreover, the multi-resolution surrogate performed similarly to several single-resolution surrogates, indicating limited trade-offs between the surrogate's broad spatial applicability and its performance. Model surrogates are a promising tool for the global hydrological modeling community, given their potential benefits in reducing computational demands and enhancing calibration. Accordingly, our framework provides an excellent foundation for the community to create their own multi-scale deep-learning global hydrological model surrogates.

1. Introduction

Global hydrological models are important decision support tools for policy-making in today's water-scarce world. Emerging around the turn of the 21st century, global hydrological models initially stemmed from land-surface models (M. F. Bierkens, 2015). While land-surface models primarily focus on estimating vertical water and energy fluxes at the Earth's surface to support general circulation models, global hydrological models distinguish themselves by decoupling from the general circulation models (Haddeland et al., 2011; Sood & Smakhtin, 2015). Instead, they focus on better representation of lateral river streamflow and the impacts of human activities on the water resources. These human activities encompass a range of factors, such as water abstractions for irrigation, domestic, industrial, energy, and livestock use, as well as water redistribution through reservoir storage and release (Alcamo et al., 2003; Döll & Siebert, 2002; Hanasaki et al., 2006; Wada et al., 2011). As a result, global hydrological models allow us to assess worldwide water management and scarcity. Moreover, owing to their process-based nature, these models can project worldwide water resources under various climate-change and socio-economic scenarios (Haddeland et al., 2014; Schewe et al., 2014; Vorosmarty et al., 2000). These scenario analyses provide essential information to underpin future sustainable development and water management.

However, efforts to improve worldwide water assessments have dramatically increased the computational demands of global hydrological models. For example, global hydrological models have increasingly turned to scenario projections and ensemble forecasting analyses to better capture the inherent uncertainties in water-resource projections (Hut et al., 2021; Warszawski et al., 2014). These analyses can encompass a large number of simulations, including several climate-change and socio-economic scenarios, a variety of general circulation models and a multitude of initial states. Furthermore, the spatial resolution of global hydrological models has transitioned from approximately 60 arc-minute (about 10,000 km²) to approximately 30 arc-second (about 1 km²) to better incorporate landscape heterogeneity and provide more detailed local information (M. F. Bierkens et al., 2015;

© 2025. The Author(s).

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Wada et al., 2017; Wood et al., 2011). Consequently, global hydrological models are now pushing the boundaries of what can be feasibly achieved with current computational resources.

In addition, global hydrological model calibration efforts to better align model simulations with observations have proven to be difficult. Not only are iterative calibration techniques computationally demanding, but hydrological observations (e.g., discharge) are sparse and unequally distributed across the globe. Therefore, model calibration often results in a discontinuous patchwork of model parameter combinations (Yang et al., 2019), biased toward areas where observations are available. Furthermore, parameters are often scale-dependent and cannot directly be applied at other spatial resolutions (Samaniego et al., 2017; Wood, 1997). Although transfer functions have been used to derive scale-specific calibration parameters (Mizukami et al., 2017; Samaniego et al., 2010), their performance is limited by the function and parameter selection, as determined by experts. Therefore, improving worldwide water assessments remains challenging.

To address these issues, deep-learning approaches have recently gained prominence in the hydrological community (Pal & Sharma, 2021; Shen, 2018; Sit et al., 2020). In particular, deep-learning approaches using Long-Short Term Memory (LSTM) networks (Hu et al., 2018; Kratzert et al., 2018) to predict observed hydrological fluxes such as discharge based on meteorological forcings and catchment characteristics. Studies have demonstrated that these approaches can match and even surpass the performance of conceptual and process-based hydrological models in predicting observed discharge, even in ungauged basins (Kratzert et al., 2019a, 2019b). Moreover, while training these deep-learning networks can be time-consuming, their prediction speeds far outpace those of process-based models.

Nevertheless, there are two major drawbacks to deep-learning approaches that are based on (semi-)observed data (i.e., without integrating physics). First, predictions from such pure data-driven approaches are limited to available observable hydrological states and fluxes that can be included during training. Therefore, these approaches provide only a limited view of the (spatially distributed) water balance. For example, LSTMs trained to predict discharge at the catchment outlet cannot simultaneously provide (spatially distributed) evapotranspiration and groundwater storage, as observations for these components do not exist. Second, predictions from such pure data-driven approaches are limited to the historical context in which they are trained. Deep-learning approaches tend to exhibit erratic behavior outside their training conditions. Therefore, these approaches prove inadequate for water-resource projections, especially under non-analogous future climatic conditions due to climate change and human water management shifts due to socio-economic developments.

To avoid these drawbacks, deep-learning and process-based approaches are increasingly being integrated (Feng et al., 2022; Höge et al., 2022; L. Sun et al., 2020; Shen et al., 2023; Wang et al., 2024). One of these integration methods is the development of deep-learning surrogates that emulate process-based model simulations. From a deep-learning standpoint, these surrogates mitigate the downsides of deep-learning approaches that are based on observations because they provide a complete view of the water balance and, if trained on non-historical simulations, can extrapolate beyond the historical context. From a process-based standpoint, these surrogates can reduce the computational demands as their simulation speed far outpaces that of process-based models. However, more importantly, these surrogates enable novel calibration and assimilation approaches that incorporate observations (Gong et al., 2015; Shen et al., 2023; Tang et al., 2020). Therefore, although the deep-learning surrogate inherits the shortcomings of the process-based model, it enables improved hydrological assessments through faster (i.e., more detailed and numerous) simulations and better calibration.

Although deep-learning surrogates hold promise for both deep-learning and process-based communities, their application to global hydrological models remains limited. Two critical components are required in a deep-learning surrogate framework in order to be effective for global hydrological models. First, whereas many surrogates predominantly focus on estimating natural vertical water states and fluxes (e.g., soil moisture and evapotranspiration), global hydrological model surrogates should explicitly include lateral water fluxes (e.g., runoff routing) and their interaction with human activities (e.g., abstractions). Second, whereas many surrogates are only trained at a single spatial resolution, limiting their deployment to that spatial resolution, global hydrological model surrogates should be deployable at the diverse range of spatial resolutions at which global hydrological models are applied.

The objective of our study is to develop a framework for multi-resolution deep-learning global hydrological surrogates. Our framework is characterized by two features. First, the architecture of our surrogate explicitly

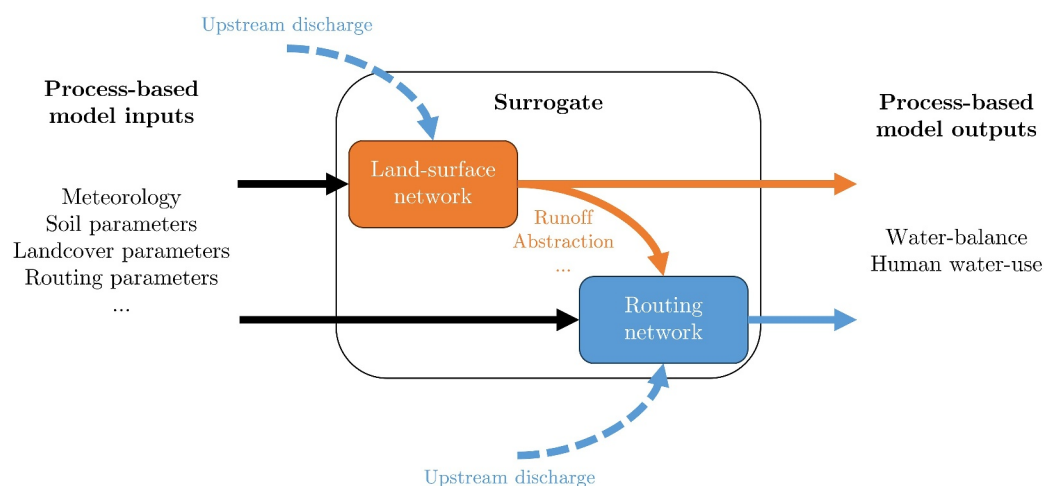


Figure 1. Overview of the surrogate framework, representing a single cell from a process-based global hydrological model. The surrogate is split into a land-surface and a routing component that each consist of a neural network. During prediction, land-surface and routing networks are processed sequentially where the required land-surface hydrological fluxes are used as inputs for the routing network.

integrates spatially distributed runoff routing, including lake outflow and reservoir operation. This runoff-routing integration, allows us to estimate surface water availability and, in turn, human water abstractions throughout the river basin. Second, the training of our surrogate is adapted to include multiple spatial resolutions, spanning from 30 arc-minute to 30 arc-second. This training strategy allows the surrogate to learn to scale across different resolutions, a characteristic that is useful during application and calibration.

To assess the performance of this multi-resolution deep-learning global hydrological surrogate framework, we apply it to the process-based PCRaster Global Water Balance (PCR-GLOBWB) global hydrological model (Sutanudjaja et al., 2018; van Beek & Bierkens, 2009; van Beek et al., 2011). PCR-GLOBWB simulates the natural water cycle, including processes like evapotranspiration, runoff, and groundwater recharge, as well as human water abstractions for various purposes, such as domestic, irrigation, and industrial use. Performance is determined by comparing the deep-learning surrogate outputs to the process-based model outputs.

2. Methods

Our surrogate's primary objective is to reproduce the same outputs (e.g., hydrological states and fluxes) as any cell in the process-based model when provided with the same inputs (e.g., meteorology and parameters) as the process-based model. To incorporate integrated and spatially distributed runoff routing, our surrogate framework comprises of a land-surface and a routing component (Section 2.1). Each component predicts a daily time series of its respective hydrological states and fluxes using a neural network, primarily consisting of a LSTM layer (Section 2.2). Each network is trained and evaluated independently (Section 2.3) using data from three distinct PCR-GLOBWB simulations at different spatial resolutions (Section 2.4).

2.1. Surrogate Framework

Our surrogate represents a single cell from a process-based global hydrological model (i.e., a single surrogate for all cells). The surrogate is split into a land-surface and a routing component that each consist of a neural network (Figure 1). The land-surface network estimates vertical (within-cell) hydrological states and fluxes, such as evapotranspiration, runoff, abstraction and subsurface water storage components. The routing networks estimate lateral (between-cell) hydrological states and fluxes, such as discharge and surface water storage components.

A single land-surface network and three routing networks for rivers, lakes and reservoirs were trained. These three routing networks were necessary as river, lake and reservoir routing possess distinct and challenging characteristics that are not easily captured in a single network. Particularly, reservoir operations were difficult to capture, as these operations are determined based on complex relationships with the historical reservoir inflow. Thus, depending on the cell, the routing network can be exchanged to represent a river, lake or reservoir routing.

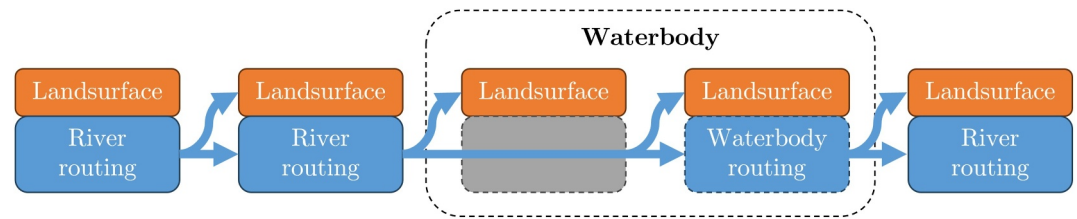


Figure 2. Overview of the surrogate processing order. During prediction, surrogate cells are processed in upstream-to-downstream order following the basin flow direction. Depending on the cell, the routing network can be exchanged to represent a river, lake or reservoir routing. Lake and reservoir waterbodies are processed when the routing network reaches their outflow cell.

Although the land-surface and routing networks could be combined, this approach has two notable disadvantages. First, the routing network performance would trade-off with the land-surface network performance. This trade-off is important as routing errors will accumulate throughout the river basin. Second, maintaining separate land-surface and routing networks allows for more flexibility in application, where poorly performing networks (e.g., the reservoir network) could be replaced by other models (e.g., conceptual models).

The surrogate's land-surface and routing networks are interdependent, as routing requires land-surface hydrological fluxes such as runoff and abstraction (Figure 1). Therefore, the land-surface and routing networks are processed sequentially within our surrogate. The land-surface network is processed first, after which the required hydrological fluxes are used as inputs for the routing network. Note that the land-surface and routing networks are trained separately (i.e., not sequentially) using hydrological fluxes from the process-based model as inputs (Section 2.3).

In addition, the surrogate cells are interdependent, as routing requires discharge fluxes from upstream cells (Figure 2). Therefore, cells are processed in upstream-to-downstream order according to the basin flow direction. Discharge fluxes from the upstream surrogates are used as inputs for the downstream surrogate. As lakes and reservoirs can cover multiple cells, not all land-surface hydrological fluxes are immediately available. Therefore, lakes and reservoirs are processed when the basin flow direction reaches their outflow cell. Note that the cells are trained separately (i.e., not upstream-to-downstream) using discharge fluxes from the process-based model as inputs (Section 2.3).

2.2. Network Architecture

The surrogate land-surface and routing networks consist of a series of neural layers with, at their core, a LSTM layer (Figure 3). LSTMs, introduced by Hochreiter and Schmidhuber (1997) and further refined by Gers et al. (2000), are specialized in handling sequential information. Like other Recurrent Neural Networks (RNNs) (Salehinejad et al., 2017), LSTMs maintain an internal (or hidden) state (or memory) that updates throughout the sequence to incorporate prior information. This state and the current sequence inputs then predict the current sequence output. Unlike other RNNs, LSTMs are better at retaining their internal state over long sequences and

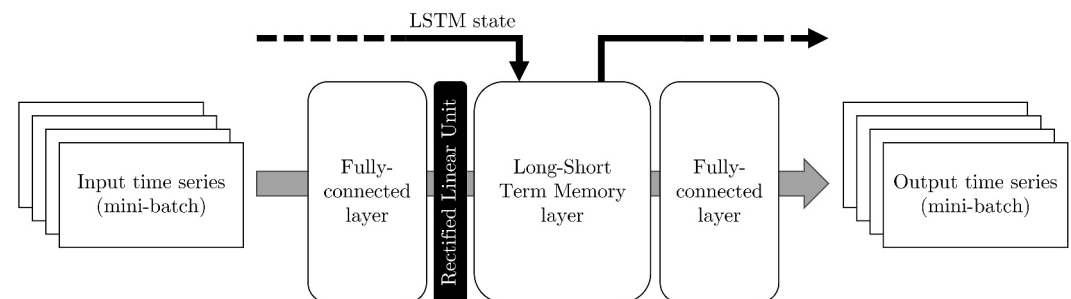


Figure 3. Overview for the land-surface and routing networks. Each network, primarily consists of an Long-Short Term Memory (LSTM) layer, preceded by a Rectified Linear Unite activation layer and enclosed by fully-connected layers. The networks are provided with time series at a daily timestep. Inputs comprise mini-batches that are provided to the model sequentially in time. Between batches, the LSTM state is transferred to maintain continuity.

find extensive use in time-series prediction tasks with long-range dependencies (Hochreiter & Schmidhuber, 1997). Therefore, the LSTM architecture suits hydrological applications well.

As our networks mainly use an LSTM, our deep-learning surrogate, like its process-based counterpart, is essentially a time-series predictor similar for each cell in the domain. Note that lake and reservoir routing networks also predict time series but for the lake and reservoir entity as a whole, even though the waterbody area can cover multiple cells (see Section 2.1). The surrogate is provided with input time series at a daily timestep and predicts daily output time series. Static, monthly and yearly inputs are repeated for each relevant timestep in the time series. The land-surface and routing networks thus learn the delays (and processes) associated with each hydrological state and flux.

However, global hydrological models contain many highly non-linear interactions, especially those related to human activities. Therefore, the LSTM layer is preceded by a Rectified Linear Unit (ReLU) activation layer and enclosed by fully-connected (linear) layers in our network. These additional layers are essential for learning complex mappings between inputs and outputs and better capture non-linear interactions in the simulation. Additionally, these layers help translate between the LSTM value space (ranging from -1 to 1) and the output value space (ranging from $-\infty$ to ∞). The various network hyperparameters, such as the number and size of the layers, are optimized as outlined in Appendix A.

2.3. Training and Prediction

Each surrogate network is trained independently from the others, meaning all hydrological fluxes transferred between networks during prediction (e.g., upstream discharge, runoff and abstraction) are taken directly from the process-based model outputs. Training is done on a small subset of the available process-based model outputs (see Section 2.4). The network's weights and biases are updated using the Adam algorithm (Kingma & Ba, 2014) to minimize the Mean Squared Error (MSE) difference between the network's predictions and the model outputs in the training subset. However, the best-performing network is selected based on the MSE difference between the network's predictions and the model outputs in a separate validation subset, to avoid overfitting to the training subset.

Training is done in mini-batches comprising several cells over a certain period (Figure 3). These mini-batches are presented to the surrogate sequentially in time. The initial mini-batch initializes (or spins up) the LSTM's internal state. Therefore, the network's predictions for this initial mini-batch are not included in the MSE calculation. For subsequent mini-batches, the LSTM's internal state is transferred between batches to maintain continuity and help the LSTM learn long-range (i.e., across batch) dependencies in the data. The various training hyperparameters, such as the mini-batch size and the learning rates, are optimized as outlined in Appendix A.

To test the surrogate's ability to scale across spatial resolutions, four different surrogates are trained using data at different spatial resolutions (see Section 2.4). The main surrogate is a multi-resolution surrogate trained on a third of the available data for each spatial resolution. Besides the multi-resolution surrogate, three single-resolution surrogate variants are trained on all the data of a single spatial resolution. These surrogate variants are trained to allow for a performance comparison between the multi-resolution surrogate and specialized single-resolution surrogates at each resolution.

During prediction, the network ceases to update its weights and biases, significantly reducing computation times. First, an initialization run over the whole period is performed, at the end of which the internal state of the LSTM is stored. The initialized internal state is subsequently used for the LSTM during the actual prediction. This long initialization period is especially important for storage components with long residence times such as ground-water and reservoir stores.

Note that, during prediction, discharge is calculated based on the network's predicted surface water storage changes. This approach was taken as the highly skewed discharge data required a log transformation before training. Although this transformation greatly improves the network's performance, minor prediction errors at the high end of the log-transformed discharge will result in major prediction errors in the untransformed discharge. These errors resulted in cases where discharge became inconsistent and violated the water balance (i.e., discharge would decrease and increase substantially along the river flow path) near the river mouth.

Table 1
Available Samples, Total and for Each Subset, for Each Network and Resolution

Network	30 arc-minute samples (#)	5 arc-minute samples (#)	30 arc-second samples (#)
Land-surface	8,407 (12.50%)	16,602 (0.78%)	32,689 (0.20%)
training (50.0% of total)	4,204	8,301	16,345
validation (16.7% of total)	1,401	2,767	5,448
test (33.3% of total)	2,802	5,534	10,896
River routing	7,145 (12.50%)	16,005 (0.78%)	31,950 (0.20%)
training (50.0% of total)	3,576	8,005	15,978
validation (16.7% of total)	1,189	2,666	5,324
test (33.3% of total)	2,380	5,334	10,648
Lake routing	560 (25.00%)	680 (25.00%)	45 (25.00%)
training (50.0% of total)	283	343	24
validation (16.7% of total)	92	112	7
test (33.3% of total)	185	225	14
Reservoir routing	708 (25.00%)	1,580 (25.00%)	292 (25.00%)
training (50.0% of total)	356	793	149
validation (16.7% of total)	117	262	47
test (33.3% of total)	235	525	96

Note. The multi-resolution surrogate is trained using only a third of the available samples for each resolution.

2.4. Data

To train and evaluate our surrogate, input and corresponding output data from three distinct PCR-GLOBWB simulations between 2000 and 2010 are used. A full list of all PCR-GLOBWB inputs and outputs can be found in Text S1 in Supporting Information S1. These simulations were derived from previous studies and cover various spatial domains and resolutions: a global 30 arc-minute simulation (van Beek et al., 2011), a global 5 arc-minute simulation (Sutanudjaja et al., 2018), and a European 30 arc-second simulation (Hoch et al., 2023). Together these simulations cover the wide range of spatial resolutions at which global hydrological models are typically applied.

Besides their spatial domain and resolution, the simulation inputs are also not standardized and differ substantially. For instance, the 30 arc-minute simulation contains two natural land-cover types, uses the Food and Agriculture Organization of the United Nations (FAO) Digital Soil Map of the World (DSMW) data set for its soil parameters (FAO, 1998) and is forced by the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis–Interim (ERA-Interim) data set (Dee et al., 2011) whereas the 30 arc-second simulation contains just one natural land-cover type, uses the SoilGrids250 data set for its soil parameters (Hengl et al., 2017) and is forced by the WATCH Forcing Data methodology applied to ERA5 data (W5E5) data set (Cocchi et al., 2020). Nevertheless, the deep-learning surrogate should be able to accommodate these differences effectively as the processes in the simulations are consistent.

Only a subset of the PCR-GLOBWB model simulation data is used (Table 1). For the land-surface network, 1 out of 8 (12.50%), 128 (0.78%), and 512 (0.20%) cells are randomly selected for the 30 arc-minute, 5 arc-minute, and 30 arc-second resolutions, respectively. At each spatial resolution increment, the selection fraction decreases. This selection approach is necessary to prevent an excessive number of cells for the higher-resolution simulations, thus avoiding an overly strong training bias toward these simulations. However, the subset size still roughly doubles for each increment, to incorporate the greater input and output variability in the higher-resolution simulations. The influence of the subset size on the model's performance is assessed in Appendix C.

For the routing networks, simulation outputs are highly skewed, especially at higher resolutions, as rivers take up a small proportion of the total cells in the simulation. Therefore, routing cells are systematically selected for a

more uniform discharge distribution. Cells are divided into 5 equal-magnitude discharge bins for each resolution. Subsequently, the same number of cells were selected for each bin, if possible. The river routing network uses the same fractions as the land-surface network. For the lake and reservoir routing networks, 1 out of 4 cells (25%) is selected for all resolutions as these cells represent individual lakes and reservoirs, whose count does not substantially increase at higher resolutions.

The subset is further subdivided into a training set, a validation set, and a test set (without cross-validation). Jointly, the training and validation sets encompass two-thirds of the subset, with three-quarters allocated to the training set and one-quarter designated for the validation set. Additionally, these sets encompass only the first two-thirds of the temporal sequence, approximately 7 years. The training and validation sets are employed in updating the networks' weights and biases and selecting the best performing networks (see Section 2.3). The test set constitutes the remaining one-third of the subset, encompassing all time steps. This independent set is exclusively deployed post-training to gauge and compare the model's performance across both space (i.e., for samples beyond those used during training) and time (i.e., for time steps not included in the training process).

3. Results

To evaluate our surrogate framework, we use it to develop a multi-resolution deep-learning surrogate for the process-based global hydrological model PCR-GLOBWB. The surrogate's performance and generalization are assessed by comparing the deep-learning surrogate outputs to the process-based model outputs on a test data set that contains samples and periods not used during training (Section 3.1). In addition, the surrogate's ability to scale across resolutions is assessed by comparing the multi-resolution surrogate to several single-resolution surrogate variants (Section 3.2). Lastly, the deep-learning surrogate is used to predict over the whole domain and its spatial and computational performance is compared with that of the process-based model for each spatial resolution (Section 3.3 and 3.4).

3.1. Performance and Generalization

The deep-learning surrogate performed well in capturing the process-based model output variables (Figure 4, Figure S1 and Table S1 in Supporting Information S1). Over all samples and variables of the test data set, the surrogate has a median Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) of 0.50 (with a correlation of 0.83, variability ratio of 0.95 and bias ratio of 0.98) when compared with the process-based model. Moreover, time series of sample-average discharge, runoff, evapotranspiration and abstraction generally show good temporal agreement between the deep-learning surrogate and the process-based model.

However, substantial performance differences can be discerned between the output variables. Over all samples and variables, the surrogate has a KGE interquartile range of 1.18 (with a range for correlation of 0.56, variability ratio of 0.57 and bias ratio of 0.39). More specifically, median KGEs range from >0.8 , for upper soil storage, upper transpiration, bare soil evaporation and interception evaporation variables to <-1 , for desalination abstraction, fossil groundwater storage and (fossil) groundwater abstraction variables. These differences are likely determined by the complexity of the processes the variable represents and the variable data distribution (see Discussion section).

Nevertheless, the surrogate developed in this study has a similar performance compared to surrogates from other studies. For example, Tsai et al. (2021) use an LSTM to predict simulated surface soil moisture and evapotranspiration with a correlation of 0.91 and 0.92, respectively, whereas our surrogate shows a correlation of 0.93 and 0.97, respectively.

3.2. Scaling Across Spatial Resolutions

Additionally, the deep-learning surrogate performs well in capturing the process-based model outputs at various spatial resolutions (Figure 5). The surrogate performs best on the high-resolution 30 arc-second test data set (including all output variables) with a median KGE of 0.61, followed by the medium-resolution 5 arc-minute and the low-resolution 30 arc-minute test data sets with median KGEs of 0.38 and 0.24, respectively.

Notably, the multi-resolution surrogate performance, trained on a third of the training and validate data sets of each resolution, is similar to that of the single-resolution surrogate variant, trained on the full training and validate data sets of a single resolution, for each resolution. Although the multi-resolution surrogate performs worse,

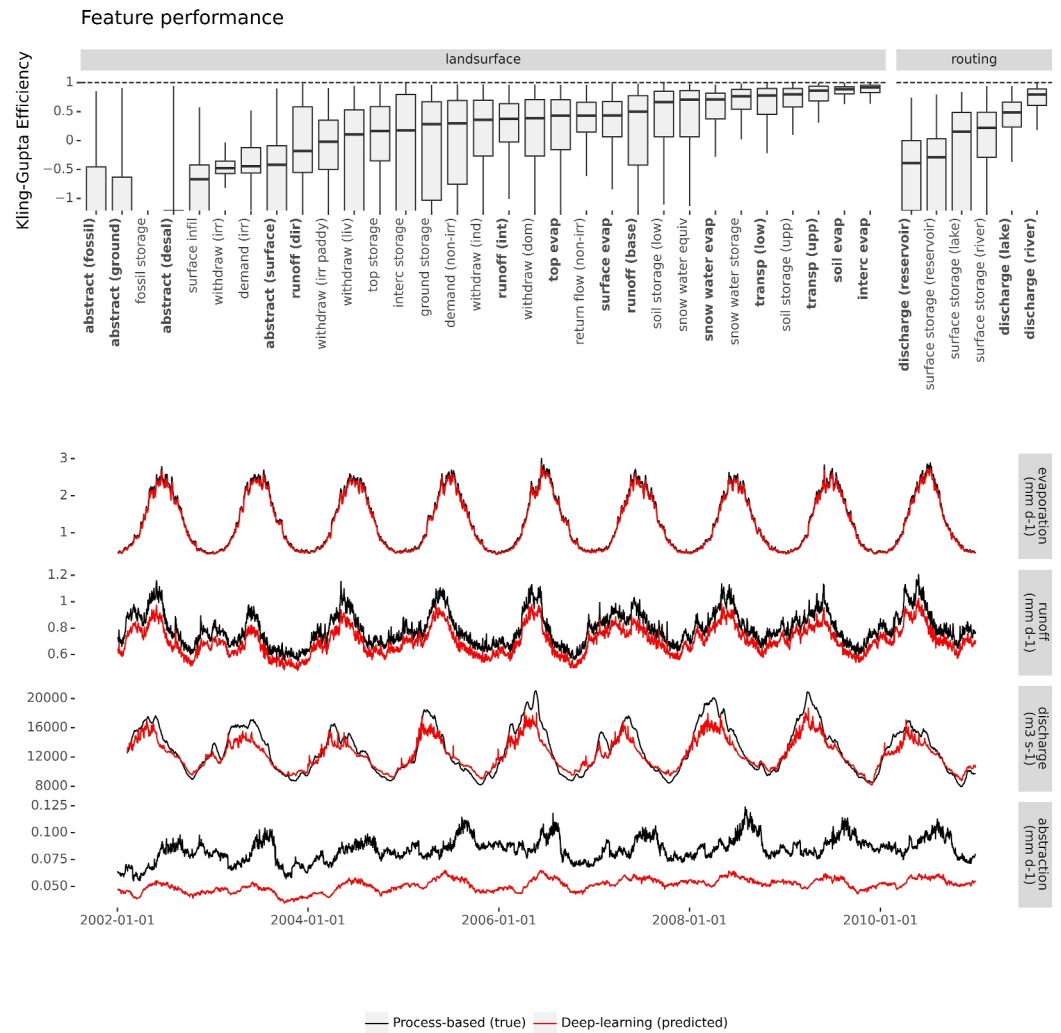


Figure 4. Deep-learning surrogate performance measured on the test data set. The top figure shows Kling-Gupta Efficiency (KGE) (–) distributions for all land-surface and routing surrogate output variables. The bottom figure shows process-based model and deep-learning surrogate time-series of sample-average discharge ($\text{m}^3 \text{s}^{-1}$), runoff (mm d^{-1}), evapotranspiration (mm d^{-1}) and abstraction (mm d^{-1}). Note that the bottom figure time-series are a combination of multiple variables that have been indicated in bold in the top figure. Nash-Sutcliffe Efficiency (NSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) performances can be found in Figure S1 and Table S1 in Supporting Information S1.

differences are small with single-resolution surrogate variant median KGEs of 0.69, 0.49 and 0.35 for the 30 arc-second, 5 arc-minute and 30 arc-minute test data sets, respectively. Moreover, where the multi-resolution performs well on all spatial resolutions, the single-resolution surrogate variants perform poorly on test data sets from other spatial resolutions than the resolution they are trained on.

3.3. Integrated Predictions

When predicting over the full simulation domain (including the training, validation and test samples), the deep-learning surrogate shows good spatial agreement with the process-based model outputs (Figures 6 and 7 and Figures S1 to S15 in Supporting Information S1). For the average discharge, the surrogate has a spatial efficiency (SPEAF) (Koch et al., 2018) of 0.89, 0.91 and 0.97 (Pearson correlation of 1.00, 1.00 and 0.97, variability ratio of 1.11, 1.10, 1.01 and histogram intersection of 0.99, 0.99 and 0.98) when compared with the process-based model outputs at 30 arc-minute, 5 arc-minute and 30 arc-second, respectively.

As routing is done sequentially, errors in routing inputs (i.e., upstream routing outputs and land-surface inputs) and inaccuracies in the routing network accumulate throughout the basin. Therefore, absolute discharge

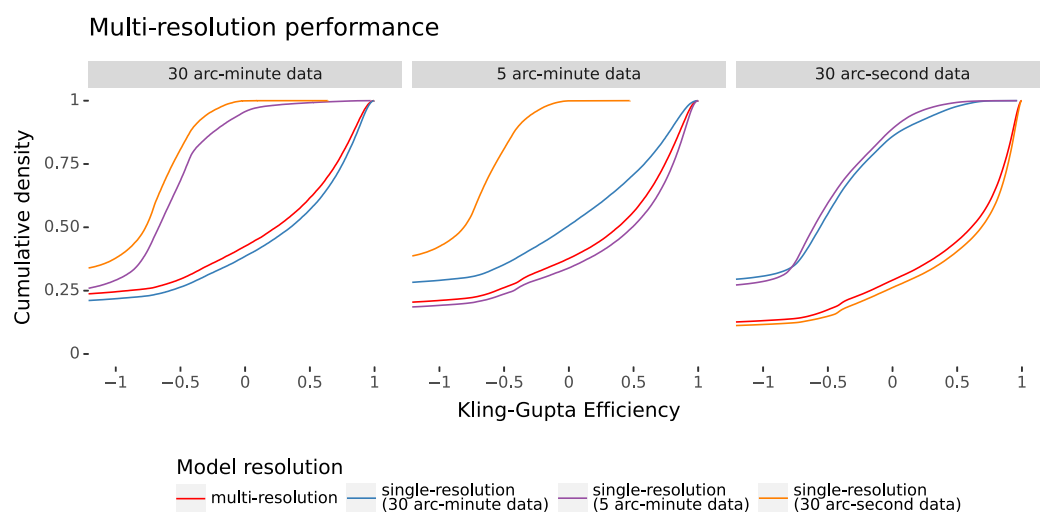


Figure 5. Deep-learning surrogate resolution performance measured on the test data set (including all outputs). The figure shows KGE (–) cumulative density for the multi-resolution surrogate and various single-resolution surrogate variants on the 30 arc-minute, 5 arc-minute and 30 arc-second test data sets.

differences at the river mouth are generally larger than at the river origin. Nevertheless, the median Nash-Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970) at the mouth of major rivers (i.e., basin area >50 000 km²) is 0.99, 0.96 and 0.73 when compared with the process-based model outputs at 30 arc-minute, 5 arc-minute and 30 arc-second, respectively.

This routing performance is in line with routing surrogates from other studies. For example, Gu et al. (2020) use an LSTM with a K-means clustering algorithm to predict simulated discharge at the basin outflow point based on simulated upstream runoff with an NSE of 0.97, similar to our surrogate. Moreover, due to the routing approach in our surrogate framework, discharge routing is consistent and spatially distributed throughout the river basin and explicitly includes lake outflow and reservoir operation.

3.4. Computational Demand

In terms of runtime, the deep-learning surrogate is at least an order of magnitude faster than the process-based model (Table 2). Runtimes improved from approximately 50 $\mu\text{sec cell}^{-1} \text{ year}^{-1}$ for the process-based model to 3 $\mu\text{sec cell}^{-1} \text{ year}^{-1}$ for the surrogate. Note, however, that this comparison is relatively crude, as the deep-learning surrogate and process-based model run on different hardware (Central Processing Unit vs. Graphics Processing Unit) and their implementation is different. When considering the actual number of operations being done, the number needed to process a single land-surface cell in the deep-learning surrogate exceeds that of the process-based model by several orders of magnitude, with approximately 2 million and 10 thousand point operations, respectively. Despite this difference, the computational optimizations provided by the GPU still ensure a faster runtime for the surrogate.

4. Discussion

Using our newly developed framework, we show that deep learning can effectively be harnessed to develop multi-resolution global hydrological model surrogates. The surrogate framework explicitly integrates hydrological states and fluxes, including the human impacts on water resources, such as water abstractions and reservoir operations, and spatially distributed runoff routing. In addition, we show that, by training the surrogate at multiple spatial resolutions, the surrogate can effectively scale its predictions across spatial resolutions as well. Therefore, our framework constitutes the first deep-learning surrogate framework that can be used to emulate global hydrological models.

From our results, two main implications can be derived. First, although our surrogate can capture the spatial and temporal distributions of the process-based model output variables, performance strongly depends on the process

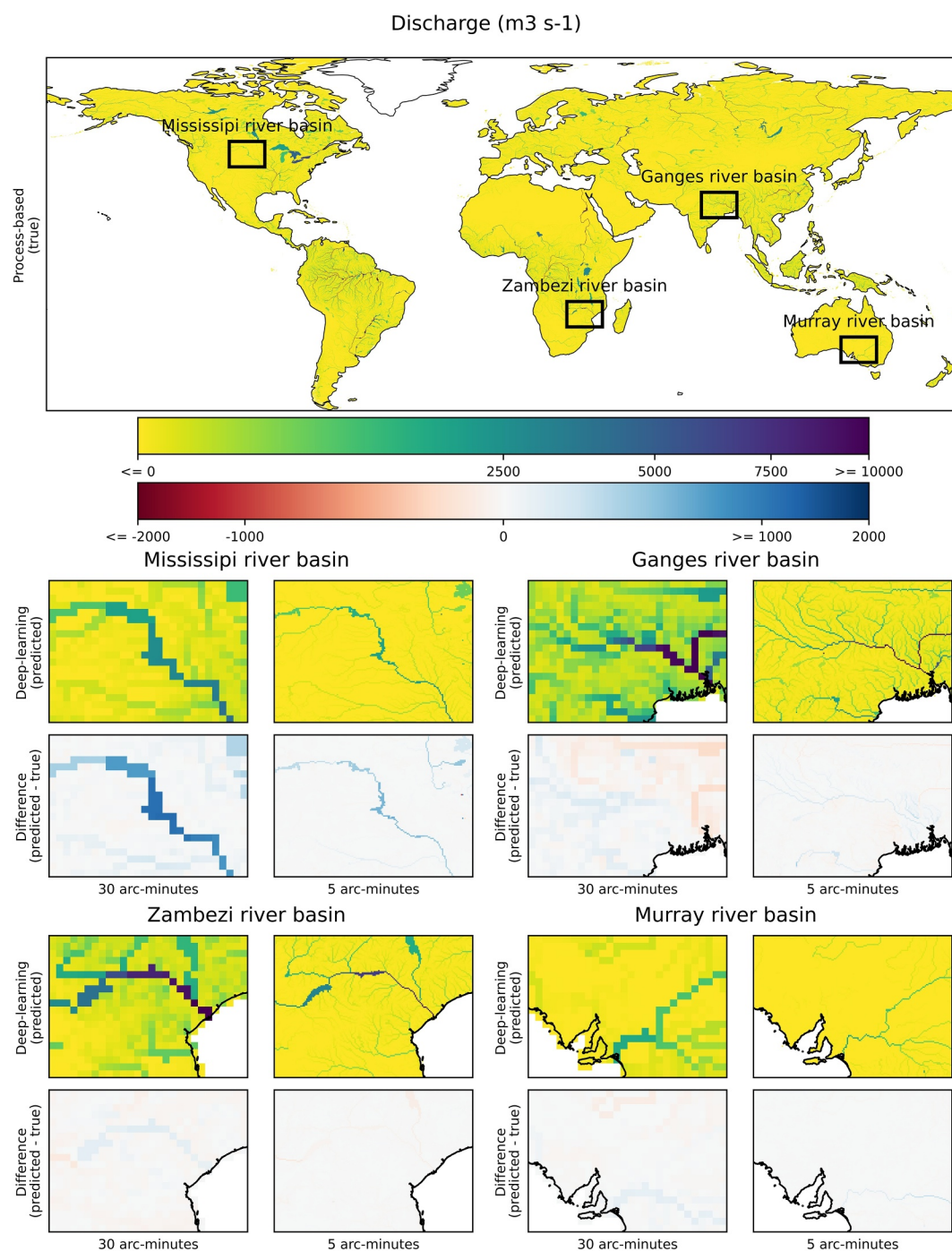


Figure 6. Comparison of average discharge ($\text{m}^3 \text{s}^{-1}$) between the process-based model and the deep-learning surrogate for the 30 arc-minute and 5 arc-minute data resolutions. The top figure shows the worldwide 5 arc-minute process-based model results and the bottom figures show deep-learning surrogate results, and their differences, for several regions at multiple spatial resolutions. Note that discharge values are squared to better display the wide range of discharges. Worldwide comparisons of discharge (including distributed NSE values), runoff, evapotranspiration and abstraction can be found in Figures S2 to S15 in Supporting Information S1.

complexity. Generally, variables representing simple processes are better predicted than those representing complex processes. For example, the best-performing variable, interception evaporation, is strongly independent and highly correlated to inputs such as precipitation. Conversely, the worst-performing variable, fossil

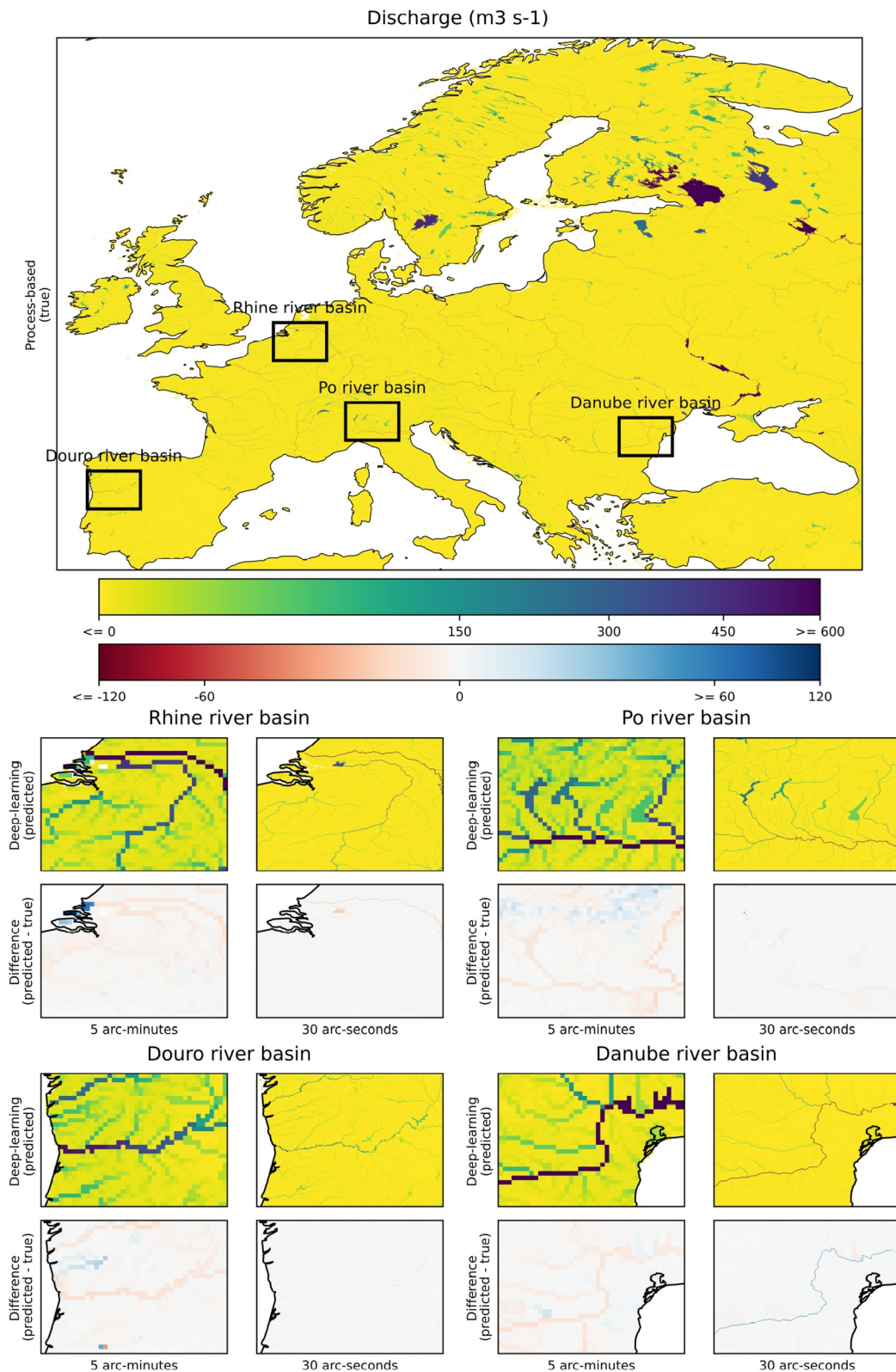


Figure 7. Comparison of average discharge ($\text{m}^3 \text{s}^{-1}$) between the process-based model and the deep-learning surrogate for the 5 arc-minute and 30 arc-second data resolutions. The top figure shows the European 30 arc-second process-based model results and the bottom figures show deep-learning surrogate results, and their differences, for several regions at multiple spatial resolutions. Note that average discharge values are squared to better display the wide range of discharges. European comparisons of discharge (including distributed NSE values), runoff, evapotranspiration and abstraction can be found in Figures S2 to S15 in Supporting Information S1.

Table 2
Runtimes of the Deep-Learning Surrogate and the Process-Based Model

	Deep-learning surrogate runtime ($\mu\text{sec cell}^{-1} \text{ year}^{-1}$)	Process-based model runtime ($\mu\text{sec cell}^{-1} \text{ year}^{-1}$)
Land-surface only	1.90 (± 0.49)	
Routing only	0.78 (± 0.15)	
Fully integrated	2.81 (± 0.49)	52.47 (± 64.69)

Note. Values are based on simulations from a single 30 arc-minute domain and multiple (53) 5 arc-minute domains. Simulations were run on a single Central Processing Unit (CPU) for the process-based model and a single Graphics Processing Unit (GPU) for the deep-learning surrogate.

groundwater abstraction, depends on a complex chain of other processes. Besides process complexity, data distribution plays an important role in variable performance. Highly skewed data distributions are generally predicted worse as outliers are poorly represented during training. For example, variables such as desalination abstraction or fossil groundwater storage which are near zero in most locations, perform relatively poorly. Nevertheless, not all output variables contribute to the total water balance with the same magnitude. Therefore, surrogate application should depend on the performance of the water balance components of interest or the training objective should be weighted to improve performance for these components.

Second, multi-resolution training is essential for deep-learning surrogates to capture the process-based model's underlying processes. Although these processes are theoretically resolution-independent, the surrogate variants

trained at only a single resolution show poor performance when applied to other resolutions. These results indicate that the underlying processes are often obfuscated in the process-based output at a single spatial resolution. As a result, single-resolution surrogate variants are learning input-output relationships that are based on correlations rather than actual processes (i.e., overfitting). Although the multi-resolution surrogate is trained on less data, the surrogate can use the information learned from a specific resolution for the other spatial resolutions as well with minimal losses in performance. Therefore, there is only a limited trade-off between the broad spatial applicability of the multi-resolution surrogate and its performance.

Although our framework incorporates essential components for global hydrological model surrogates, our approach has important caveats. Most importantly, our approach does not allow for neighborhood operations. As cells are processed upstream to downstream the surrogate can, at most, only process information from the current and upstream cells. Although neighborhood operations exist in the PCR-GLOBWB model, most notably in the water allocation scheme, the surrogate can only approximate such operations. To incorporate neighborhood operations, there are possibilities to use a Convolutional Neural Network (CNN) in combination with the LSTM network, as shown in several other studies (Li et al., 2023; R. Sun et al., 2023; Wang et al., 2024). However, as most global hydrological model operations occur on a cell-by-cell basis and the CNN resolution is fixed, the performance, accuracy and multi-resolution capabilities of CNN networks should be further explored.

Another caveat of our approach is the substantial storage requirements of (high-resolution) predictions. Although our surrogate reduces computational resources, the surrogate requires inputs transformed to a different format, the Tensor format, and structure, upstream to downstream. Therefore, these inputs need to be restored. There are several ways to reduce storage requirements (and runtimes) that were neglected in our study to maintain the best resemblance between the process-based model and its surrogate. Feature importance analysis can determine important input features for surrogate predictions (see Appendix B). Based on this analysis, unimportant input features can be omitted, reducing storage requirements for similar model performance. Furthermore, experiments can be done to see whether temporally aggregated (e.g., weekly instead of daily) inputs are sufficient for approximating coarser temporal-resolution (e.g., monthly) outputs. If sufficient, coarse temporal-resolution predictions would require significantly less storage and runtime.

Nevertheless, deep-learning global hydrological model surrogates can enable substantial water resource assessment improvements. Note that these surrogates will likely not (yet) replace process-based models. Although deep-learning surrogates perform well in capturing the process-based model outputs, they remain an approximation. Moreover, surrogates cannot be deliberately developed. If any new processes need to be introduced, they should first be incorporated into the process-based model and the deep-learning surrogate will subsequently need to be (re-)trained. Rather than replacing process-based models, deep-learning surrogates can help improve process-based models and their applications.

In particular, we see three major use cases to improve worldwide water resource assessments. First, deep-learning surrogates can help capture the uncertainties in water resource assessments by providing hydrological approximations where process-based models are too computationally intensive. An example would be to use surrogates for ensemble forecasting approximations (Kuehnert et al., 2022). As ensemble forecasting requires a high number

of (recurring) simulations, deep-learning surrogates could replace process-based models to enable a larger number of simulations. These ensemble approximations could inform the forecasting uncertainty range or identify important scenarios or periods for which a more computationally demanding process-based model simulation should be done.

Second, deep-learning surrogates can help improve process-based model performance by enabling hybrid simulation approaches. Within PCR-GLOBWB, the kinematic wave routing implementation substantially increases runtimes (Sutanudjaja et al., 2018). However, since the surrogate's routing networks act as stand-alone parts of the surrogate, these networks could be integrated into the process-based model to replace the current computationally expensive routing implementations. In addition, deep-learning reservoir operation surrogates (Zhang et al., 2018) that better capture observations could be used to replace the default reservoir operation scheme in the process-based model.

Third, deep-learning surrogates can help improve simulation accuracy by incorporating observations through calibration and assimilation. Deep-learning surrogates are differentiable (Shen et al., 2023), meaning they allow for calculating gradients between output errors (i.e., differences between simulations and observations) and surrogate weights, biases and inputs. Based on these gradients, surrogate weights, biases and inputs can be adjusted to better match the observations. For example, parameter learning (Tsai et al., 2021) uses these gradients to train another neural network to map from better-known physical inputs to poorly-known calibration parameters without the need for a priori definition of transfer function form. These calibrated parameters can subsequently also be used for the process-based model.

Given these use cases, we believe deep-learning multi-resolution surrogates are a promising tool for the global hydrological community. As our framework is broadly applicable and flexible, it is suitable for a wide range of other global hydrological models and thus provides an excellent foundation for the community to create their own multi-scale deep-learning model surrogates.

5. Conclusions

Using our newly developed framework, we show that deep learning can effectively be harnessed to develop multi-resolution global hydrological model surrogates. Our framework integrates spatially distributed runoff routing, including lake outflow and reservoir operation, includes human activities, such as water abstractions, and can scale across spatial resolutions. These components make our framework especially useful for global hydrological models.

When applied to the PCR-GLOBWB global hydrological model, our deep-learning surrogate framework performs well. Our surrogate has a median KGE of 0.45 when compared to all process-based model outputs, although performance varies substantially between output variables. Moreover, our multi-resolution surrogate performed similarly to several single-resolution surrogate variants, indicating limited trade-offs between the surrogate's broad spatial applicability and its performance. Compared to the global hydrological model, surrogate predictions show good temporal and spatial agreement with the process-based model and are at least an order of magnitude faster.

Model surrogates are a promising tool for the global hydrological modeling community, given their potential benefits in reducing computational demands and enhancing calibration. Accordingly, our framework provides an excellent foundation for the community to create their own multi-scale deep-learning model surrogates.

Appendix A: Hyper-Parameters and Optimization

The deep-learning surrogate necessitates specifying various network-related parameters, referred to as hyper-parameters. 10 hyper-parameters are optimized: the mini-batch sample size and date size, the transformation function applied to the input and output data, the number and size of the pre-process linear layers before the LSTM layer, the number of LSTM layers, the number and size of the post-process linear layers after the LSTM layer, the learning rate and the dropout rate (Table A1). Note that the transformation function is only applied to features with non-normally distributed data, as measured with the Fisher-Pearson coefficient of skewness (Zwillinger & Kokoska, 2000), and that all input and output data was standardized after transformation. In addition, during training, a cyclic learning rate schedule is used (Smith, 2017), whereby the learning rate cycles

Table A1
Hyperparameters Used For the Land-Surface, River, Lake and Reservoir Networks

Hyperparameter	Land-surface network	River network	Lake network	Reservoir network
Batch sample size	32	32	32	32
Batch date size	512	768	768	768
Transformation function	sqrt	log	log +1	log
Pre-process linear layer number	1	1	1	1
Pre-process linear layer size	512	256	256	256
LSTM layers number	1	1	1	1
Post-process linear layers number	1	1	1	1
Post-process linear layers size	512	256	256	256
Learning rate	$1e^{-4}$	$5e^{-5}$	$5e^{-5}$	$1e^{-5}$
Dropout rate	0.15	0.15	0.2	0.7

Note. Log +1 indicates a value of one is added to the values before log transformation to decrease the logarithmic value range at smaller values.

three times between an order of magnitude larger and an order of magnitude smaller than the specified learning rate.

Determining the optimal combinations of model parameters is achieved through the Optuna hyper-parameter optimization framework (Akiba et al., 2019). Initially, 128 randomly generated parameter combinations are explored. Subsequently, the Tree-structured Parzen Estimator algorithm (Ozaki et al., 2020) is used to sample an additional 320 parameter combinations, aiming to identify the optimal parameter combinations. Optimal parameters were identified based on the network performance after training. Only a subset of the training, validation and test data sets were used to reduce optimization times: 1 out of 32 samples for the land-surface and river routing data sets and 1 out of 8 samples of the lake and reservoir routing data sets.

Performance was measured by the MSE and the KGE for the land-surface network and the MSE and the KGE variability ratio (KGE-alpha) for the routing networks. The KGE-alpha metric is used for the routing networks as the surface water storage variability is the most important for calculating the discharge during prediction. Taking into account the best-performing parameter combinations and the estimated parameter importance (see Appendix B) a unique combination of hyper-parameters is determined for each network.

Appendix B: Feature Importance

To assess the importance of each input feature, a multitude of surrogate predictions are made with a permutation to that input feature. After each prediction, the input feature importance is measured by comparing the surrogate prediction with the unperturbed (original) surrogate prediction. Four permutations are used: the minimum value over all samples, the maximum value over all samples, the mean value over all samples, time-shuffled values per sample and substituted values from another sample. Importance is measured as the (min-max) Normalized Mean Squared Error (NMSE) between the unperturbed and permuted surrogate output.

For the land-surface network, the meteorological inputs (i.e., precipitation, reference evapotranspiration and temperature) are by far the most important (Figure B1). After the meteorological input features, the upstream discharge land-cover fractions and land-cover evapotranspiration coefficients play a role. For the routing network, the input feature importance is more distributed. Note that, many input features are provided both as a flux and as the mean of the waterbody area. Generally, the runoff, upstream discharge, surface water abstraction and surface water evaporation are roughly equally important for the routing network.

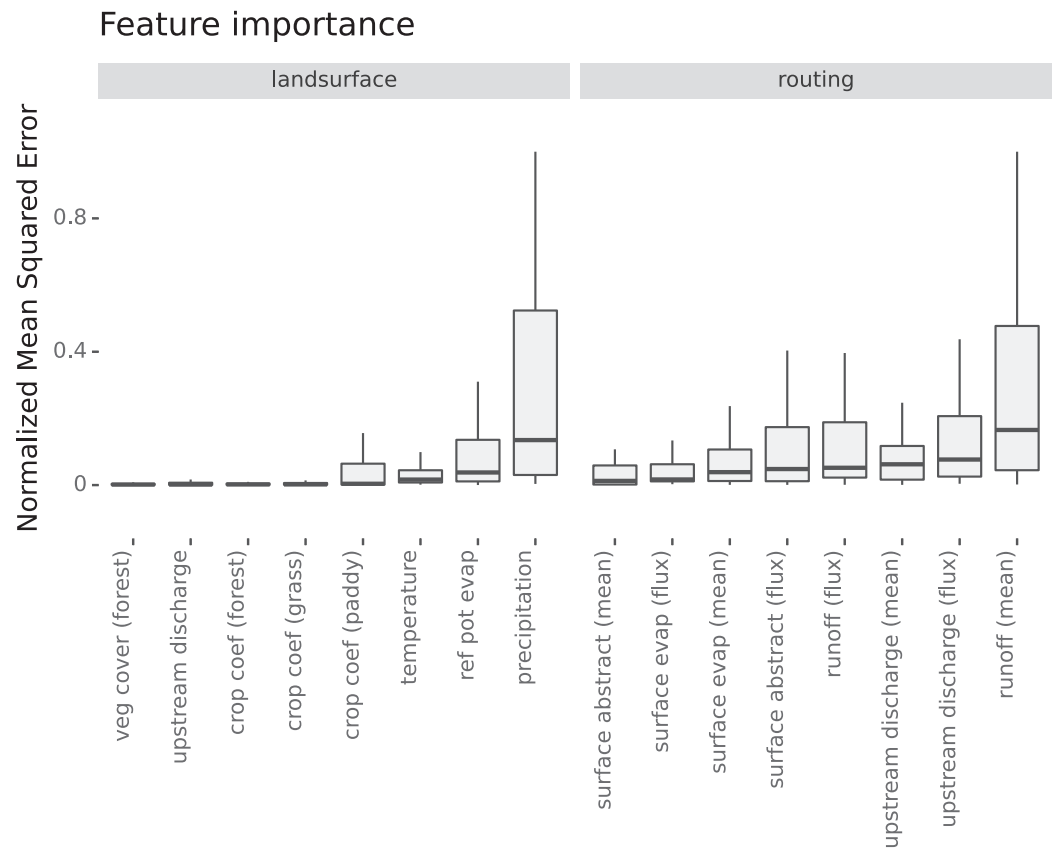


Figure B1. Deep-learning surrogate feature importance measured on the test data set. Feature importance is based on an input feature permutation analysis. Figure shows permutation Normalized Mean Squared Error (–) distributions for the most sensitive land-surface and routing features.

Appendix C: Sample Size Sensitivity

To assess the deep-learning surrogate sensitivity to the number of training samples, five additional surrogates are trained with various subsets of the original training and validation data sets. These subsets consist of a random selection of 1 out of 2, 4, 8, 16 and 32 of all samples. After training, the surrogate's performance is measured on the full test data set.

As the subset size increases, the deep-learning surrogate performance also increases (Figure C1). In the land-surface network, the median KGE increases from 0.00 to 0.50 whereas the routing network KGE increases from 0.09 to 0.52 between the 1/32 subset and the 1/1 subset. In addition, performance variability of the land-surface network decreases from a KGE interquartile range of 1.92 to 1.25 between the 1/32 subset and the 1/1 subset. As the performance increases level off closer to the 1/1 subset, including more training and validation samples will likely not result in substantial performance increases.

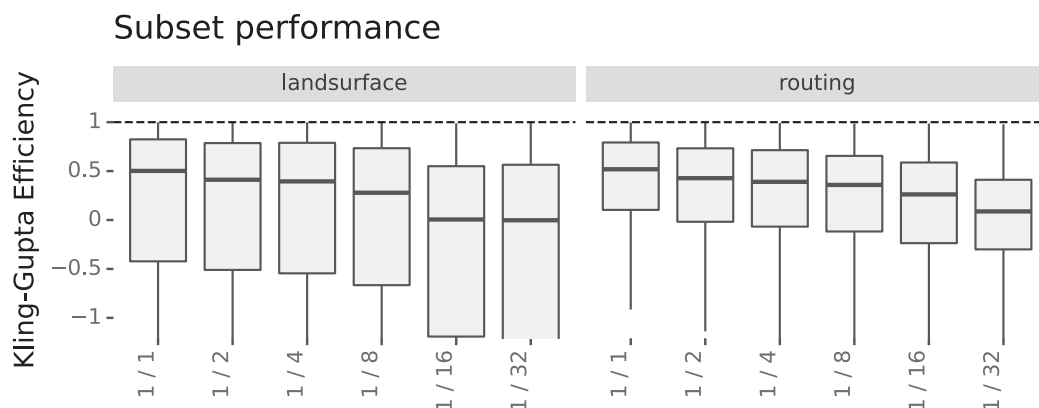


Figure C1. Deep-learning surrogate sample-size performance measured on the test data set. The figure shows KGE (–) distributions for the land-surface and routing networks trained on various sample subsets (i.e., fractions of the original data sets).

Data Availability Statement

All code and data related to our deep-learning multi-resolution surrogate framework as well as the PCRaster Global Water Balance (PCR-GLOBWB) global hydrological model surrogate are available on GitHub (Droppers, 2024a) and Zenodo (Droppers, 2024b).

Acknowledgments

Our research was supported by the National Geographic Society, as part of the World Water Map and Freshwater Initiative (M. Bierkens et al., 2023).

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery and data mining* (pp. 2623–2631). <https://doi.org/10.1145/3292500.3330701>
- Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., & Siebert, S. (2003). Development and testing of the watgap 2 global model of water use and availability. *Hydrological Sciences Journal*, 48(3), 317–337. <https://doi.org/10.1623/hysj.48.3.317.45290>
- Bierkens, M., Wanders, N., Droppers, B., Leijnse, M., Tait, A., Gemache, M., et al. (2023). World water map. Retrieved from <https://worldwatermap.nationalgeographic.org/>
- Bierkens, M. F. (2015). Global hydrology 2015: State, trends, and directions. *Water Resources Research*, 51(7), 4923–4947. <https://doi.org/10.1002/2015WR017173>
- Bierkens, M. F., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., et al. (2015). Hyper-resolution global hydrological modelling: What is next? “everywhere and locally relevant”. *Hydrological Processes*, 29(2), 310–320. <https://doi.org/10.1002/hyp.10391>
- Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., et al. (2020). Wfde5: Bias-adjusted era5 reanalysis data for impact studies. *Earth System Science Data*, 12(3), 2097–2120. <https://doi.org/10.5194/essd-12-2097-2020>
- Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656), 553–597. <https://doi.org/10.1002/qj.828>
- Döll, P., & Siebert, S. (2002). Global modeling of irrigation water requirements. *Water resources research*, 38(4), 8–1–8–10. <https://doi.org/10.1029/2001WR000355>
- Droppers, B. (2024a). PCR-GLOBWB surrogate [software]. *GitHub*. https://github.com/BramDr/PCR-GLOBWB_surrogate/releases/tag/v1.1.1
- Droppers, B. (2024b). PCR-GLOBWB surrogate [software]. *Zenodo*. <https://doi.org/10.5281/zenodo.13933252>
- FAO. (1998). Digital soil map of the world (Tech. Rep.). *Food and Agriculture Organization of the United Nations*. Retrieved from <https://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1026564/>
- 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>
- Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10), 2451–2471. <https://doi.org/10.1162/089976600300015015>
- Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Dai, Y., et al. (2015). Multi-objective parameter optimization of common land model using adaptive surrogate modeling. *Hydrology and Earth System Sciences*, 19(5), 2409–2425. <https://doi.org/10.5194/hess-19-2409-2015>
- Gu, H., Xu, Y.-P., Ma, D., Xie, J., Liu, L., & Bai, Z. (2020). A surrogate model for the variable infiltration capacity model using deep learning artificial neural network. *Journal of Hydrology*, 588, 125019. <https://doi.org/10.1016/j.jhydrol.2020.125019>
- 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>
- Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., et al. (2011). Multimodel estimate of the global terrestrial water balance: Setup and first results. *Journal of Hydrometeorology*, 12(5), 869–884. <https://doi.org/10.1175/2011JHM1324.1>
- Haddeland, I., Heinke, J., Biemans, H., Eisner, S., Flörke, M., Hanasaki, N., et al. (2014). Global water resources affected by human interventions and climate change. *Proceedings of the National Academy of Sciences*, 111(9), 3251–3256. <https://doi.org/10.1073/pnas.1222475110>
- Hanasaki, N., Kanae, S., & Oki, T. (2006). A reservoir operation scheme for global river routing models. *Journal of Hydrology*, 327(1–2), 22–41. <https://doi.org/10.1016/j.jhydrol.2005.11.011>

- Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al. (2017). Soilgrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748. <https://doi.org/10.1371/journal.pone.0169748>
- Hoch, J. M., Sutanudjaja, E. H., Wanders, N., Van Beek, R. L., & Bierkens, M. F. (2023). Hyper-resolution pcr-globwb: Opportunities and challenges from refining model spatial resolution to 1 km over the european continent. *Hydrology and Earth System Sciences*, *27*(6), 1383–1401. <https://doi.org/10.5194/hess-27-1383-2023>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., & Fenicia, 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>
- Hu, C., Wu, Q., Li, H., Jian, S., Li, N., & Lou, Z. (2018). Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water*, *10*(11), 1543. <https://doi.org/10.3390/w10111543>
- Hut, R., Drost, N., van de Giesen, N., van Werkhoven, B., Abdollahi, B., Aerts, J., et al. (2021). The ewatercycle platform for open and fair hydrological collaboration. *Geoscientific Model Development Discussions*, *2021*(13), 1–31. <https://doi.org/10.5194/gmd-15-5371-2022>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. <https://doi.org/10.48550/arXiv.1412.6980>
- Koch, J., Demirel, M. C., & Stisen, S. (2018). The spatial efficiency metric (spaef): Multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, *11*(5), 1873–1886. <https://doi.org/10.5194/gmd-11-1873-2018>
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using long short-term memory (Lstm) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>
- Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. <https://doi.org/10.1029/2019WR026065>
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- Kuehnert, J., McGlynn, D., Remy, S. L., Walcott-Bryant, A., & Jones, A. (2022). Surrogate ensemble forecasting for dynamic climate impact models. *arXiv preprint arXiv:2204.05795*. <https://doi.org/10.48550/arXiv.2204.05795>
- Li, B., Li, R., Sun, T., Gong, A., Tian, F., Khan, M. Y. A., & Ni, G. (2023). Improving Lstm hydrological modeling with spatiotemporal deep learning and multi-task learning: A case study of three mountainous areas on the Tibetan plateau. *Journal of Hydrology*, *620*, 129401. <https://doi.org/10.1016/j.jhydrol.2023.129401>
- Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. <https://doi.org/10.1002/2017WR020401>
- 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)
- Ozaki, Y., Tanigaki, Y., Watanabe, S., & Onishi, M. (2020). Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. In *Proceedings of the 2020 genetic and evolutionary computation conference* (pp. 533–541). <https://doi.org/10.1145/3377930.3389817>
- Pal, S., & Sharma, P. (2021). A review of machine learning applications in land surface modeling. *Earth*, *2*(1), 174–190. <https://doi.org/10.3390/earth2010011>
- Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*. <https://doi.org/10.48550/arXiv.1801.01078>
- Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5), W05523. <https://doi.org/10.1029/2008WR007327>
- Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., et al. (2017). Toward seamless hydrologic predictions across spatial scales. *Hydrology and Earth System Sciences*, *21*(9), 4323–4346. <https://doi.org/10.5194/hess-21-4323-2017>
- Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., et al. (2014). Multimodel assessment of water scarcity under climate change. *Proceedings of the National Academy of Sciences*, *111*(9), 3245–3250. <https://doi.org/10.1073/pnas.1222460110>
- 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 and Environment*, *4*(8), 552–567. <https://doi.org/10.1038/s43017-023-00450-9>
- Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, *82*(12), 2635–2670. <https://doi.org/10.2166/wst.2020.369>
- Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 464–472). <https://doi.org/10.48550/arXiv.1506.01186>
- Sood, A., & Smakhtin, V. (2015). Global hydrological models: A review. *Hydrological Sciences Journal*, *60*(4), 549–565. <https://doi.org/10.1080/02626667.2014.950580>
- Sun, L., Gao, H., Pan, S., & Wang, J.-X. (2020). Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data. *Computer Methods in Applied Mechanics and Engineering*, *361*, 112732. <https://doi.org/10.1016/j.cma.2019.112732>
- Sun, R., Pan, B., & Duan, Q. (2023). A surrogate modeling method for distributed land surface hydrological models based on deep learning. *Journal of Hydrology*, *624*, 129944. <https://doi.org/10.1016/j.jhydrol.2023.129944>
- Sutanudjaja, E. H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H., Drost, N., et al. (2018). Pcr-globwb 2: A 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453. <https://doi.org/10.5194/gmd-11-2429-2018>
- Tang, M., Liu, Y., & Durlafsky, L. J. (2020). A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems. *Journal of Computational Physics*, *413*, 109456. <https://doi.org/10.1016/j.jcp.2020.109456>
- 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>
- van Beek, L., & Bierkens, M. F. (2009). The global hydrological model pcr-globwb: Conceptualization, parameterization and verification (Tech. Rep.). In *Department of Physical Geography, Faculty of Earth Sciences*. Utrecht University. Retrieved from <http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf>
- van Beek, L., Wada, Y., & Bierkens, M. F. (2011). Global monthly water stress: 1. water balance and water availability. *Water Resources Research*, *47*(7). <https://doi.org/10.1029/2010WR009791>

- Vorosmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global water resources: Vulnerability from climate change and population growth. *Science*, 289(5477), 284–288. <https://doi.org/10.1126/science.289.5477.284>
- Wada, Y., Bierkens, M. F., De Roo, A., Dirmeyer, P. A., Famiglietti, J. S., Hanasaki, N., et al. (2017). Human–water interface in hydrological modelling: Current status and future directions. *Hydrology and Earth System Sciences*, 21(8), 4169–4193. <https://doi.org/10.5194/hess-21-4169-2017>
- Wada, Y., Van Beek, L., Viviroli, D., Dürr, H. H., Weingartner, R., & Bierkens, M. F. (2011). Global monthly water stress: 2. Water demand and severity of water stress. *Water Resources Research*, 47(7), W07518. <https://doi.org/10.1029/2010WR009792>
- 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(4), e2023WR036170. <https://doi.org/10.1029/2023WR036170>
- Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The inter-sectoral impact model intercomparison project (isi–mip): Project framework. *Proceedings of the National Academy of Sciences*, 111(9), 3228–3232. <https://doi.org/10.1073/pnas.1312330110>
- Wood, E. F. (1997). Effects of soil moisture aggregation on surface evaporative fluxes. *Journal of Hydrology*, 190(3–4), 397–412. [https://doi.org/10.1016/S0022-1694\(96\)03135-6](https://doi.org/10.1016/S0022-1694(96)03135-6)
- Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L., Bierkens, M. F., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring earth’s terrestrial water. *Water Resources Research*, 47(5), W05301. <https://doi.org/10.1029/2010WR010090>
- Yang, Y., Pan, M., Beck, H. E., Fisher, C. K., Beighley, R. E., Kao, S.-C., et al. (2019). In quest of calibration density and consistency in hydrologic modeling: Distributed parameter calibration against streamflow characteristics. *Water Resources Research*, 55(9), 7784–7803. <https://doi.org/10.1029/2018WR024178>
- Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., et al. (2018). Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. *Journal of Hydrology*, 565, 720–736. <https://doi.org/10.1016/j.jhydrol.2018.08.050>
- Zwillinger, D., & Kokoska, S. (2000). *Crc standard probability and statistics tables and formulae*. Crc Press.