Water Resources Research

TECHNICAL REPORTS: METHODS

10.1029/2020WR028600

Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions

Kai Ma[1,2], Dapeng Feng[2], Kathryn Lawson[2,3], Wen-Ping Tsai[2], Chuan Liang[1], Xiaorong Huang[1], Ashutosh Sharma[2], and Chaopeng Shen[2,3]

[1]State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu, Sichuan, China, [2]Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA, [3]HydroSapient, Inc., State College, PA, USA

Key Points:

- Basins around the world can be better modeled by applying transfer learning (TL) to a deep network trained in the US, and tuning it locally
- The benefits of TL increased with the amount and diversity of the source data, and were larger than from pretraining with a hydrologic model
- This work greatly expands the reach of deep learning, adds to the value of existing big data, and calls for synergy of global data sets

Supporting Information:

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

Correspondence to:

C. Shen,
cshen@engr.psu.edu

Citation:

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*, e2020WR028600. https://doi.org/10.1029/2020WR028600

Received 21 AUG 2020
Accepted 12 MAR 2021

Abstract There is a drastic geographic imbalance in available global streamflow gauge and catchment property data, with additional large variations in data characteristics. As a result, models calibrated in one region cannot normally be migrated to another without significant modifications. Currently in these regions, non-transferable machine learning models are habitually trained over small local data sets. Here we show that transfer learning (TL), in the senses of weight initialization and weight freezing, allows long short-term memory (LSTM) streamflow models that were pretrained over the conterminous United States (CONUS, the source data set) to be transferred to catchments on other continents (the target regions), without the need for extensive catchment attributes available at the target location. We demonstrate this possibility for regions where data are dense (664 basins in Great Britain), moderately dense (49 basins in central Chile), and scarce with only remotely sensed attributes available (5 basins in China). In both China and Chile, the TL models showed significantly elevated performance compared to locally trained models using all basins. The benefits of TL increased with the amount of available data in the source data set, and seemed to be more pronounced with greater physiographic diversity. The benefits from TL were greater than from pretraining LSTM using the outputs from an uncalibrated hydrologic model. These results suggest hydrologic data around the world have commonalities which could be leveraged by deep learning, and synergies can be had with a simple modification of the current workflows, greatly expanding the reach of existing big data. Finally, this work diversified existing global streamflow benchmarks.

Plain Language Summary We introduced a method to utilize available big data to better start and warm up a machine learning streamflow model that is later fine-tuned for prediction in basins on other continents (Asia, South America and Europe). This procedure noticeably improved streamflow volume prediction for different scenarios with varying amounts of data in the target basins (in terms of time period, length of collected data, and number of basins having data). This allows thousands of basins across the world with only a few years' worth of streamflow observations to benefit from improved modeling and accuracy resulting from the use of deep learning.

1. Introduction

There is a great deal of geographic imbalance in global hydrologic data sets. Outside of the US and parts of Europe, there are many parts of the world that have only sparsely available streamflow gauge networks with only a few years' worth of data (Do et al., 2017; Fekete & Vörösmarty, 2007). Besides streamflow gauges, these regions also lack data on physiographic attributes such as geology and soil depth. Nevertheless, climate change is stressing these parts of the world, and accurate hydrologic simulations are needed for these regions just as much, or even more than for data-rich regions.

Catchments across the world are often perceived as being unique from each other, requiring customized model development for each basin (Teutschbein & Seibert, 2012). As a rule of thumb, when we create process-based hydrologic models, our development effort scales roughly linearly to the modeled area, computational effort scales linearly at best, and accuracy is unrelated to the number of basins modeled. It is typically difficult to apply knowledge gained from one basin to another, as parameters or experiences do not transfer easily. As a result, although there have been calls for hydrologic studies to transcend the uniqueness

of places (McDonnell et al., 2007), success at modeling some basins does not in general translate into equivalent success or reduced effort for other basins, especially in other continents.

Recently, data-driven hydrologic models, especially those based on the deep learning algorithm of long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997), have shown strong skills in learning streamflow dynamics for forward runs and forecasting (Feng et al., 2020; Kratzert, Klotz, Hochreiter, & Nearing, 2020; Li et al., 2020). Such performance has benefited from the availability of big data uniquely available over the conterminous United States (CONUS). In other parts of the world, however, we cannot apply these same techniques due to a shortage of streamflow gauge data. Moreover, these techniques require uniform input variables that not only have the same physical concepts but also roughly consistent characteristics, which makes it difficult to apply outside of a unified data set like the US-specific CAMELS data set (Addor et al., 2017). Across different continents, climate forcing data and static physiographic attributes are collected from different sources and can have systematic differences. Therefore, even if the compilation of a global database like the CAMELS series were possible, it is still uncertain if a uniform global model could be trained on such a database.

In data-scarce regions, there are often daily streamflow measurements that have been recorded for a few years (Alipour & Kibler, 2018; Bitew & Gebremichael, 2011), but not with the consistency and breadth of the CONUS data. The Global Runoff Data Center (GRDC, available at http://grdc.bafg.de), for example, shows that a large portion of basins around the world have fewer than 3 years' worth of daily streamflow observations. In these scenarios, machine learning models have still been employed but mostly in a local model setting, where a model is fitted to the data from one basin or a few neighboring basins (S. Zhu et al., 2020; Yaseen et al., 2015; Liang et al., 2018; Bowes et al., 2019; de la Fuente et al., 2019). Shen (2018) provided a summary and an entry point into a vast body of work in this realm, with many other papers also attesting to the huge demand for solutions (Beven, 2020; Guillon et al., 2020).

Transfer learning (TL) (Pan & Yang, 2010; Thrun & Pratt, 1998) is a method to migrate knowledge learned from one task to another. Because some different tasks have similar mathematical principles or require similar responses, their representations in a deep learning network are similar. Therefore, it should be possible to train a model with one task and data set and transfer it to another task, keeping part of the original model while retraining a different portion of the model. TL has become a highly popular technique in the artificial intelligence community (George et al., 2017; Shen, 2018) and is widely used in image classification (Yosinski et al., 2014; Y. Zhu et al., 2011). It is now a popular practice to transfer models from image-recognition data sets, for example, ImageNet, for remote sensing tasks such as land use classification (Marmanis et al., 2016; X. X. Zhu et al., 2017). TL has also been used to predict crop yields (Wang et al., 2018). While hydrologic parameter regionalization (Beck, van Dijk, et al., 2016; Wagener & Wheater, 2006) allows the transfer of parameters to nearby basins, transferring knowledge from one region to another - for example, across continents - had not been attempted before, and it was unclear whether such a transfer would be fruitful.

In this work, we applied TL to streamflow modeling to better understand if and when such knowledge transfer could be useful for hydrological time series modeling problems. We demonstrate how LSTM models trained over the data-rich CONUS can be transferred to data-scarce regions such as Asia and South America to mitigate the limitations of local observations and input attributes. For clarity, the CONUS data set is called the source data set, while the basins in the second, transferred location are referred to as the target. We reveal the benefits of TL by comparing models employing TL (hereafter called TL models) with those that are trained using only data from the target region (hereafter called local models) (Section 3.1). We also investigate the impacts of data quantity in the source and target data sets (Section 3.2) and how they compare to alternative initialization methods (Section 3.3). As a side note, because our approach does need local data for tuning, it does not solve the problem of prediction in ungauged basins (PUB) (Sivapalan, 2003).

2. Data and Methods

2.1. Data

To examine the effectiveness of TL in different data-density scenarios, we used data sets from four countries (Figure S1): the original Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) data set for the contiguous United States (Newman et al., 2014); CAMELS-GB, a dense data set for Great Britain

based on the CAMELS framework (Coxon, Addor, et al., 2020); CAMELS-CL, a moderately dense data set for Chile based on the CAMELS framework (Alvarez-Garreton et al., 2018); and CHINA-MR, hydrological and meteorological data for the upper Min River region of China (Ma et al., 2020).

The data sets all included daily streamflow volume, meteorological forcings (precipitation, temperature, etc.), and static basin attributes (slope, soil texture, etc), but specifics varied by data set (Text S1 and Table S1). The CAMELS data set contains 671 basins with minimal anthropogenic impacts spanning a wide range of geographies and climates in the conterminous US United States (CONUS). The CAMELS-GB basins are mainly of temperate oceanic climate, and are underlain by varied hydrogeological conditions. Chile's Southern Zone includes both humid Mediterranean and temperate rain-oceanic climates. The CHINA-MR data set contains five basins with large spatial and vertical variations in climate and land cover, elevations ranging from 653 to 6,054 m, and mostly urban agglomerations in the southern part.

We selected all the basins in CAMELS to train the source model (training period from October 1, 1985 to October 1, 2015). For context, if we trained the model using 10 years' worth of data (October 1, 1985 to October 1, 1995) and tested it in the next 10 years, the median NSE for the test period was 0.72, which was essentially identical to other models reported in the literature with NLDAS forcing data (Kratzert, Klotz, Hochreiter, & Nearing, 2020). For the TL model pretraining process, however, we wanted to maximize information learned, and hence provided models with the full source dataset containing all 30 years' worth of data. In Chile, we selected 49 basins by screening for basins in CAMELS-CL located in moderate central Chile, between latitudes 38°S and 42°S, with less than 20% of missing streamflow data from January 1, 2000 to January 1, 2010. In our preliminary tests, LSTM models gave poor results for the extremely dry deserts in the North and glacier-influenced cold regions in the south, a phenomenon worth future investigation. Given that the scope of this study was to improve LSTM-based modeling, we excluded these regions as the current LSTM models seem to be unsuitable there, which matches other observations of where CONUS-trained LSTM-based models have struggled (Feng et al., 2020). The 664 basins in the CAMELS-GB data set, representing a data-rich case in the target region, were selected with the same conditions for available streamflow data.

In data-scarce regions, streamflow records also tend to be short. To evaluate the impact of target-region data length, we ran a 1-year training scenario and a multi-year training scenario. For the multi-year training scenario, CAMELS-CL and CAMELS-GB were trained on data from 5 years (January 1, 2000 to January 1, 2005) and tested on data from 5 years (January 1, 2005 to January 1, 2010), while due to data set limitations, CHINA-MR models were trained on data from 4 years (January 1, 2009 to January 1, 2013), and tested on data from 3 years (January 1, 2013 to January 1, 2016). For the 1-year training scenario, CAMELS-CL and CAMELS-GB basins were trained in January 1, 2004 to January 1, 2005, and for CHINA-MR were trained in January 1, 2009 to January 1, 2010, with the same testing periods as for the multi-year training scenario.

2.2. Transfer Learning Model Based on LSTM

For this work, the TL models were based on an established LSTM architecture which was already successfully tested for predictions of streamflow (Feng et al., 2020), soil moisture (Fang, Pan, & Shen, 2019; Fang & Shen, 2020; Fang, Shen, et al., 2017) with uncertainty estimates (Fang, Kifer, et al., 2020), water temperature (Rahmani et al., 2020) and other water quality variables like dissolved oxygen (Zhi et al., 2021). Long short-term memory (LSTM) is a deep learning algorithm, a type of recurrent neural network that learns from sequential data. LSTM has "memory states" and "gates" which enable it to retain long memory and learn how long to retain state information, what to forget, and what to output. More detailed descriptions of LSTM are provided in Text S2 and can also be found in Feng et al. (2020). Deep networks such as LSTM are defined by a basic architecture with a number of weights and nonlinear activation functions across many layers. During the training process, information in the training data is stored as weights, with some parts of the network self-organizing to perform certain functionalities as dictated by the architecture (e.g., for a LSTM model trained for streamflow prediction, some of the cell states could be used to track accumulated snow storage).

Here we discuss TL in two senses: weight initialization (also known as pretraining) by a source data set, and weight freezing. To apply transfer learning (an overview is provided in Figure S2), we train the model after

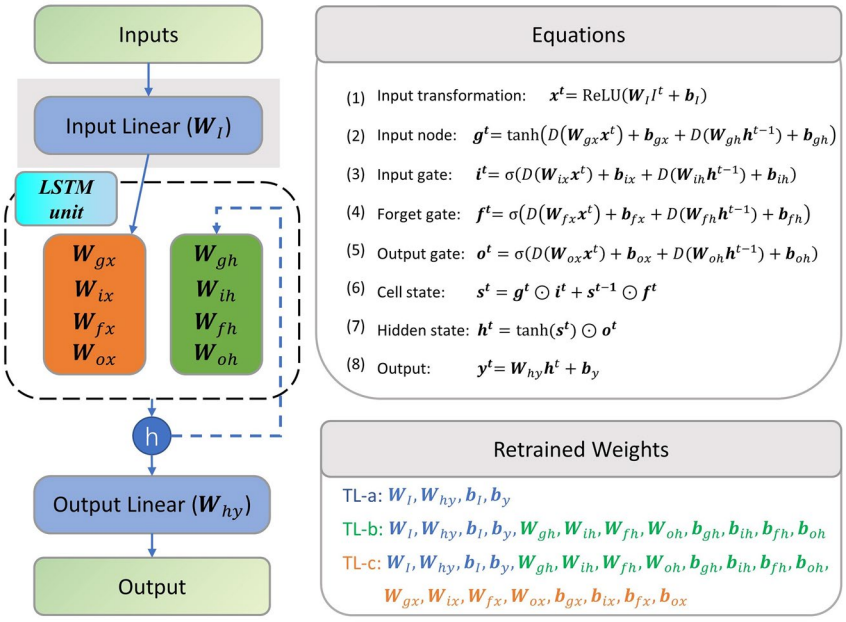

Figure 1. The architecture of long short-term memory with transfer learning (TL) options. TL-a, TL-b, and TL-c have progressively more weights which are allowed to be tuned during training with target data, which follows the pretraining process using source data. For equation details, please see Text S2.

the pretraining process, this time using data from the target location. Here, we have a choice: to allow some or all of the weights to further change during training on the target task (so the weights obtained during pretraining simply provide an initialization), or to prevent these weights from any further changes (weight freezing).

We tested three different combinations (TL-a, TL-b, and TL-c) of transfer learning (visualized in Figure 1). All options used weight initialization, and from TL-a to TL-c we progressively allowed more weights to be adjusted during training on the target data set. For option TL-a, the input and output linear transformation layers were the only weights allowed to be updated; the rest of the weights were frozen to prevent the values from changing. For the other two options (TL-b and TL-c), the weights of the LSTM gates themselves were also allowed to be updated. For TL-b, the weights of the recurrent connections from hidden states were allowed to change, but the weights from inputs were frozen. For TL-c, all weights in the network were allowed to be updated.

Compared to training a network from a blank (or cold) initial state (where the weights of the network are zero or randomly generated), the pretraining process initializes the weights so the network can roughly perform streamflow modeling according to the source data set. This allows the network to converge faster and requires fewer training data points from the target data. If we freeze some weights, the training forces the unfrozen weights to change and adapt around the frozen part. Typically, weight freezing reduces demand for training data and lowers the chance for overfitting compared to weight initialization, because it reduces the number of trainable parameters. However, there may also be some performance penalties due to reduced flexibility.

All of our TL models allowed updating of the input linear transformations (Text S2, Equation 1) along with the connection from the hidden states to the outputs. There are different forcing and attribute variables in data sets, and even for the same variable name, variable characteristics (such as biases) are often substantially different across data sets. Therefore, the input transformations were necessarily different, and allowing only these values to update (TL-a) essentially transforms target-region inputs into the roles of the source-region inputs. It is worth noting that the local models were trained with all the basins in the respective region, just as the CAMELS model was trained using all the basins in the CAMELS data set.

2.3. Experiments

2.3.1. The Effect of Source and Target Data Quantity on TL

We hypothesized that TL models would benefit from the quantity and diversity of the source data set. To test this hypothesis, we ran experiments where we varied the amount of source training data the TL models were given. In theory, we expected the benefits of TL to increase with more source data, with the benefits becoming incrementally smaller. We set the number of basins used in the source data to 10, 50, 100, 300, 500, and all 671, and randomly sampled the CONUS basins, defining the groups with fewer basins as being subsets of the larger ones. We also ran trials where we used CAMELS-GB as the source data set, and the CONUS CAMELS basins as the target.

Along with sparse gauges, target data-scarce regions often only have observations from a limited period of time. To validate the effects of TL when we have different lengths of training data in the target region, we ran a 1-year training scenario and a multi-year (3 or 4 years) training scenario for all regions. For comparison, their testing periods were the same (detailed descriptions are provided in Text S1).

2.3.2. Comparison of TL to Pretraining by a Process-Based Model

It has been suggested that pretraining a machine learning model (determining an improved initialization of the network weights) using outputs from a process-based model could improve the model (Jia et al., 2019; Read et al., 2019). The idea is that the process-based model outputs, even if imperfect or downright flawed, could teach the deep learning model basic hydrologic responses to inputs. Could TL essentially serve the same purpose as the physical equations encoded in a process-based model?

We set up a Soil & Water Assessment Tool (SWAT) model (Text S3) for the Min River in China. For the purpose of assessing the benefit of physics encoded in SWAT, the SWAT model could not be calibrated: calibrating this model would leak information from the observations, which would defeat the purpose of the test. The SWAT model was run from January 1, 2008 to January 1, 2013 with 1 year of warm up (Ma et al., 2020). The LSTM model was pretrained with the SWAT model, and then trained with the observed streamflow data from the Min River using all of the TL options (TL-a, TL-b, and TL-c), which were collectively referred to as SWAT-MR. We then compared the most effective option from SWAT-MR with the TL models from CAMELS.

Some research has shown that using the output of a process-based model as an input to LSTM could improve the robustness of the model (Jia et al., 2019). We concur, and have found minor evidence of these effects in previous work (Fang, Shen, et al., 2017). However, training a model like that would entail creating the process-based model for all the basins in the source data set, which would be a time- and effort-intensive option that hence we did not explore.

2.4. Evaluation Metrics and Hyperparameters

The first metric we used for model evaluation was the Nash–Sutcliffe efficiency coefficient (NSE), a widely used metric in hydrologic modeling (Nash & Sutcliffe, 1970). We ran all training experiments with five random seeds and used the average streamflow from these ensemble members for metric evaluation. We also evaluated the percent bias of the top 2% peak flow range (flow duration curve (FDC) high-segment volume, or FHV) and the percent bias of the bottom 2% low flow range (FDC low-segment volume, or FLV) to understand performance for more extreme values (Yilmaz et al., 2008).

Hyperparameters such as learning rate, hidden size, and dropout rate are configurations of LSTM model and therefore have some level of control on model performance (Goodfellow et al., 2016). We manually adjusted the hyperparameters by sensitivity analysis, and selected values such that each model had optimal performance, for more fair comparison. We tried many combinations, with Table S2 listing all the tested hyperparameters and the final values that were chosen, along with the meanings of these hyperparameters. We used a training-instance length of 365 days for all models. The batch size chosen for the TL model was the same as for the local model, and the hidden size used for TL was consistent with that for the corresponding source model.

Table 1
NSE Values of the 5-Member Ensemble Mean Discharge for Different Training Scenarios

		NSE_{mean}			NSE_{median}		
Model		CHINA-MR	CAMELS-CL	CAMELS-GB	CHINA-MR	CAMELS-CL	CAMELS-GB
(a)							
1-year training	Local	0.564	0.587	0.726	0.571	0.597	0.794
	TL-a	0.597	**0.705**	0.765	0.609	**0.725**	0.824
	TL-b	0.593	0.650	**0.770**	0.620	0.657	**0.827**
	TL-c	**0.603**	0.636	0.767	**0.624**	0.645	0.822
Multi-year training	Local	0.666	0.810	0.769	0.733	0.830	0.853
	TL-a	0.706	**0.845**	0.789	0.708	**0.868**	0.847
	TL-b	0.718	0.820	0.794	0.698	0.840	**0.861**
	TL-c	**0.734**	0.801	**0.796**	**0.749**	0.823	0.859

Model		NSE_{mean}	NSE_{median}
(b)			
1-year training	Local	0.564	0.571
	TL-c (SWAT-MR)	0.580	0.603
	TL-c (CONUS)	**0.603**	**0.624**
Multi-year training	Local	0.666	0.733
	TL-c (SWAT-MR)	0.693	0.748
	TL-c (CONUS)	**0.734**	**0.749**

Note. Local models were trained with all the basins from the target regions, for example, the 5 basins in CHINA-MR are trained together. TL options were TL-a, TL-b, and TL-c (Figure 1). Bold numbers indicate the best performing model for each category. (a) NSE values of the 5-member ensemble mean discharge for different training scenarios; (b) Comparison between the locally trained models for CHINA-MR, the TL model initialized with SWAT model outputs, and the best-performing TL model originally trained over the CONUS data (option TL-c).

3. Results

3.1. Performance of TL Models in Each Region

TL is an effective strategy that significantly improves streamflow predictions (Table 1; Figure 2). For the 1-year TL model, the improvement in NSE compared to the locally trained model ranged from 0.033 to 0.128. For each region, the optimal TL model had better metrics than the local model, and the advantages tended to be larger for smaller target data sets. The 1-year models of CAMELS-CL showed the highest benefits, with the optimal TL model respectively improving the mean and median NSE values by 0.118 and 0.128 as compared to the local model. For CHINA-MR, the mean NSE was elevated by 0.039 (ensemble member values ranged from 0.564–0.603) for 1-year training models. With the multi-year training scenario, CHINA-MR also showed the highest TL benefit, where option TL-a improved the mean NSE by 0.068 (0.666–0.734).

For both CAMELS-CL and CHINA-MR, the benefits from TL are likely substantial enough to be of interest to most modelers. The benefit was less pronounced for CAMELS-GB, but might still be considered non-trivial to those who want to have the best possible performance. These results agree with our intuition: re-training on the local target region fine-tuned the network weights and adapted them to local conditions, but when there was a small target data set, the model needed to heavily rely on knowledge obtained from the source data set. The more data were available for the target region; the more adjustments were applied to the network weights.

All the multi-year training models performed better than the corresponding 1-year training models, and the TL benefits tended to be smaller, although there were exceptions. Across all the regions, multi-year training of the local models improved the median NSE by an average of 0.151 (ranging between 0.059 and 0.233) as

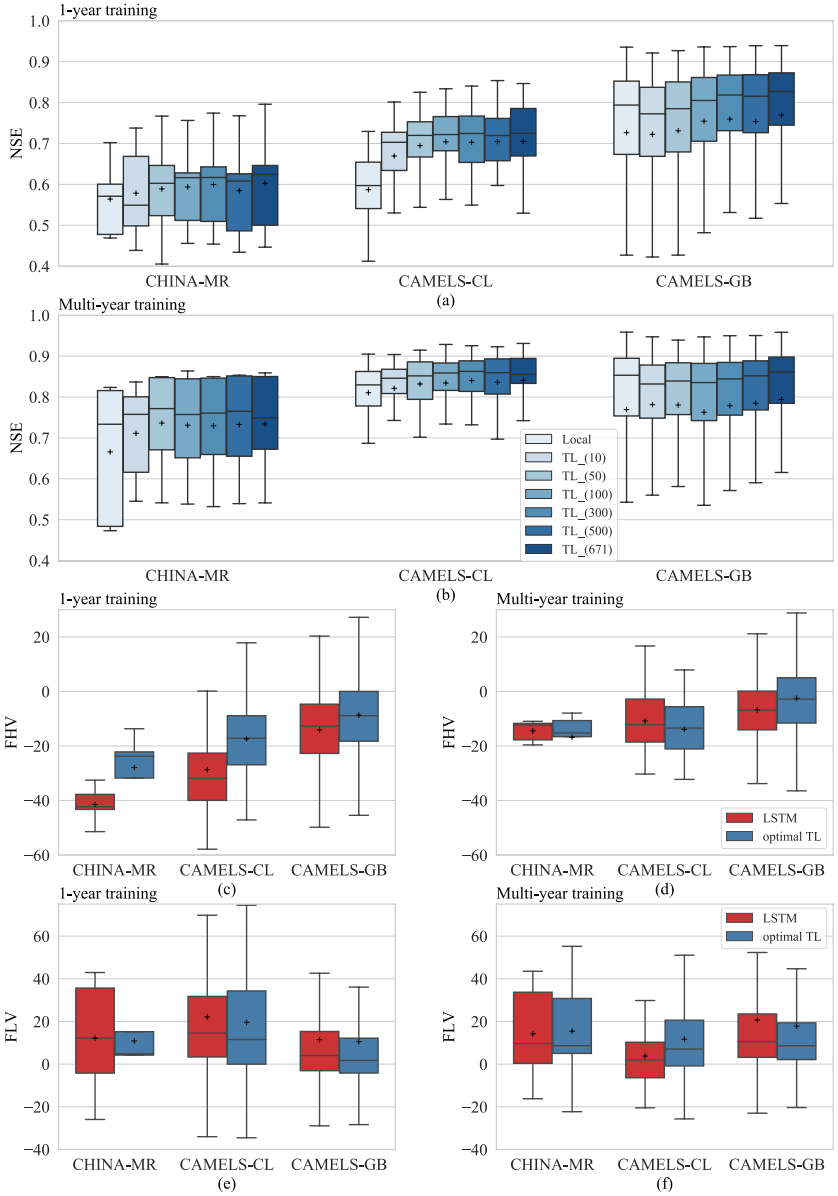

Figure 2. Performance of local and optimal TL models (selected based on Table 1) pretrained with different numbers of CONUS (source) basins for (a) 1-year training and (b) multi-year training. (c and d) The percent bias of the top 2% peak flow range (FHV) and (e and f) the percent bias of the bottom 2% low flow range (FLV) for (c and e) 1-year training and (d and f) multi-year training. All the metrics were calculated for the five-member ensemble mean discharge during the test period (ensemble standard deviation provided in Table S3). Plus symbols indicate mean values. For CHINA-MR, there were only five basins, so the two "whiskers", the two edges of the boxes, and the median line each represent performance for one basin.

compared to the 1-year training local models, while multi-year training of the TL models improved median NSE values by an average of 0.100 (ranging between 0.024 and 0.183) as compared to the 1-year training TL models.

For the more extreme parts of the hydrograph (represented by FHV and FLV), we noted a reliable improvement of TL with 1-year local training. TL alleviated the negative FHV bias significantly for all three target regions (Figure 2c) and also noticeably reduced the positive bias with FLV. To put things into context, the top 2-percentile events are generally considered difficult to capture due to various reasons such as inaccuracies of precipitation data. A similar −20% FHV was noted in other work with CAMELS as well (Feng

et al., 2020; Kratzert, Klotz, Shalev, et al., 2019). We found the TL model to better capture some of the peaks in CHINA-MR (Figure S3). For multi-year training scenarios, the benefits were less pronounced than for the 1-year training scenarios for CHINA-MR and CAMELS-GB, but were still observable. However, for CAMELS-CL, TL actually increased the bias of FHV and FLV. Since TL improved for median NSE by ~0.38 (Table 1a), TL seems to have improved the medium part of the hydrograph for CAMELS-CL for the multi-year training scenario to compensate for this worsened bias.

3.2. The Effect of Source Data Quantity on TL

In general, the performance of TL was enhanced as the number of basins in the source data set increased, though some inherent randomness did exist. We evaluated both 1-year training and multi-year training models (Figure 2a and 2b). For 1-year training, both the mean and median NSE of the TL model trained only on 10 CONUS basins were still higher than those of the local models for CAMELS-CL and CHINA-MR, suggesting that even a relatively small source data set was already beneficial. A main benefit of transfer learning arises from enabling the LSTM model to warm up and learn basic hydrologic dynamics - for example, that streamflow follows rainfall except when it is cold - and using the diverse CAMELS basins as the source data set provides a more robust initial model, as it has seen a rich combination of inputs and responses. For multi-year training TL models, all showed optimal performance in Chile and Great Britain when the number of basins was maximized, and the model performance progressively improved with the number of basins (Since CHINA-MR contained only five sites, the fluctuations were relatively larger). The basins shown in Figure 2 were selected randomly, but using different random seeds gave characteristically similar results (data not shown here). The temporal length of the source data set also affected the TL models. We also tried training using only 10 years' worth of source data, which gave lowered performance (Figure S4). This advantage may be seen when the quantity of the source data set is small, such as for the 1-year models tested in Chile and China using source data from 10 CONUS basins.

As expected, the gain resulting from increasing the source data set size became smaller and smaller as the data set became larger: the initial 50 basins showed the most notable benefit, raising median NSE from 0.629 to 0.720. As the source data set continued to increase, the benefit per added basin became smaller and smaller, albeit still non-zero if we discount some stochastic fluctuations. This observation was consistent with other results from big data machine learning: we typically see diminishing returns as the marginal benefit of a larger data set gradually decreases toward very large sample size, but the benefit may still be non-zero even for a very large data set (Sun et al., 2017).

3.3. Comparison With Model Initialization Using a Process-Based Hydrologic Model

Our experiments showed that the benefits of TL were larger than what would have been contributed by weight initialization using outputs from SWAT. The initialization by SWAT outputs raised the mean NSE from 0.564 to 0.580 in 1-year training and from 0.666 to 0.693 for multi-year training, suggesting this approach is useful (Table 1b). Nevertheless, the optimal TL model from CAMELS had much bigger benefit, with median NSE values of 0.603 for 1-year training and 0.734 for multi-year training. A possible explanation is that the source LSTM model is a more accurate hydrologic model than many process-based models (Feng et al., 2020; Kratzert, Klotz, Hochreiter, & Nearing, 2020; Shen, 2018).

Because initialization can only be done once, the two different approaches cannot accommodate each other. We must also consider the cost of process-based model initialization, which is very high in this case, because we would need to create process-based models for each target basin of interest. It was impossible for us to implement this method for CAMELS-CL and CAMELS-GB within the scope of this work.

4. Discussion

4.1. Selection of the Optimal TL for the Target Region

The optimal TL options differed for each data set, but they seemed to be the same for each region, regardless of training length. Table 1a shows that the optimal TL options were TL-c (all weights unfrozen) for CHINA-MR, TL-a (just input and output transformation layers unfrozen) for CAMELS-CL, and TL-b

(many, but not all weights unfrozen) for CAMELS-GB. The difference between different options was substantial for CHINA-MR and CAMELS-CL, but relatively minor for CAMELS-GB.

These results show that it will be difficult to find the best option *a priori*. TL-b was the best option for CAMELS-GB by a very small margin over TL-c, which was largely consistent with our intuition that more local data can be better exploited by models with larger complexity. TL-a was found to be better for CAMELS-CL, which suggests central Chile may be climatologically and hydrologically similar to some basins in CONUS and was sufficient to only perform linear transformations of inputs and outputs. However, TL-c, for which the pretraining only provided weight initialization, was found to be the best option for all CHINA-MR experiments, which countered our intuition that a smaller target data set would benefit from a partially frozen model. One potential explanation is that CHINA-MR contains larger basins than the CONUS source basins, which could have made the routing process comparably more important, and thus obtaining optimal results required the retraining of all the LSTM weights to sufficiently alter the model's memory dynamics.

4.2. The Impact of Source Data Diversity on TL

Using CAMELS-GB (30 years) as the source and CAMELS as the target produced an optimal TL model with a median NSE of 0.738, compared to our previous value of 0.732 (Feng et al., 2020). Transferring CAMELS-GB to CAMELS-CL resulted in a median NSE of 0.843, compared to the value of 0.868 obtained by using CAMELS as the source. CAMELS also has a wider range of values for topographic, climatic and hydrological attributes than CAMELS-GB, and covers more of the range of CAMELS-CL (Figure S5). Using CAMELS-GB as a source was still useful, but the benefits were less notable compared to using CAMELS as the source, suggesting it is preferable to use a more diverse source data set which likely helped reduce overfitting.

Not much work in the literature is directly comparable to the work reported here. With NLDAS forcing data, Kratzert, Klotz, Shalev, et al. (2019) reports a CAMELS-median NSE of ~0.74 (which we were able to reproduce) and 0.63 for LSTM and the SAC-SMA hydrologic model, respectively, while recent work with a 4,229-basin global data set produced a median Kling-Gupta efficiency (KGE, comparable to NSE) of 0.69 for locally calibrated conceptual hydrologic models in the validation period (Beck, Pan, et al., 2020). Our multi-year training with TL presented median NSE values of 0.75, 0.87, and 0.86 for CHINA-MR, CAMELS-CL, and CAMELS-GB, respectively, certainly representing state-of-the-art results. As a side note, the good performance on the global data set suggests that NLDAS, although more easily accessible, may have a lower quality than other climate data sets in the world.

4.3. Limitations

This study did not address streamflow measurement uncertainty (Coxon, Freer, et al., 2015; McMillan et al., 2017), which may be impactful for some stations but should be of limited importance for target data sets with observations in many basins if there is no systematic bias in the measurements. We also recognize that LSTM is not a silver bullet: for example, it did not work well in extremely arid or glacier-dominated regions in Chile, perhaps because the time scales of runoff generation in these basins (flash flood or glacier melt) are not easily handled by LSTM.

5. Conclusion

We introduced a transfer learning scheme to leverage information from data-rich regions to mitigate the limitations of small data sets and incomplete input attributes in data-scarce regions on different continents. Trained on the CAMELS data set over the CONUS, our LSTM model was transferred to data-scarce regions in Asia, Europe, and South America to enhance the accuracy of streamflow predictions as compared to locally trained models. There is tremendous value in this transfer learning procedure, as a huge number of basins around the world with only a few years' worth of local observations are now amenable to accurate modeling with deep learning.

These results suggest that hydrologic dynamics around the world, while often perceived as being unique for each location, have commonalities that could be leveraged by modelers across different continents. It

also means that enticing rewards in terms of model performance are "right at the fingertips" of the steadily rising amount of streamflow forecasters in data-scarce regions who employ LSTM on small data sets. Multiple transfer learning options are possible, and the choices need to be evaluated for each target region's use cases. This work suggests that modelers across the world can and should look beyond their watersheds or even their continents for useful data. Efforts such as the Global Runoff Data Center and the CAMELS data set series are highly meritorious, and could be leveraged for these efforts. A global synergy, which was not envisioned before, is now possible with deep learning frameworks.

Data Availability Statement

Data for CAMELS can be downloaded at https://ral.ucar.edu/solutions/products/camels. Data for CAMELS-CL can be downloaded at http://www.cr2.cl/camels-cl/. Data for CAMELS-GB can be downloaded at https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9. Atmospheric data for CHINA-MR can be downloaded at http://www.cmads.org/. The hydrologic deep learning code used in this work can be accessed at http://doi.org/10.5281/zenodo.3993880.

Acknowledgments
K. Ma was supported by the China Scholarship Council for 1 year study at the Pennsylvania State University. CS and KL were supported by National Science Foundation Award OAC #1940190. KL and CS have financial interests in HydroSapient, Inc., a company which could potentially benefit from the results of this research. This interest has been reviewed by the University in accordance with its Individual Conflict of Interest policy, for the purpose of maintaining the objectivity and the integrity of research at The Pennsylvania State University.

References

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

Alipour, M. H., & Kibler, K. M. (2018). A framework for streamflow prediction in the world's most severely data-limited regions: Test of applicability and performance in a poorly-gauged region of China. *Journal of Hydrology*, *557*, 41–54. https://doi.org/10.1016/j.jhydrol.2017.12.019

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., et al. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies – Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846. https://doi.org/10.5194/hess-22-5817-2018

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

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

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

Bitew, M. M., & Gebremichael, M. (2011). Assessment of satellite rainfall products for streamflow simulation in medium watersheds of the Ethiopian highlands. *Hydrology and Earth System Sciences*, *15*(4), 1147–1155. https://doi.org/10.5194/hess-15-1147-2011

Bowes, B. D., Sadler, J. M., Morsy, M. M., Behl, M., & Goodall, J. L. (2019). Forecasting groundwater table in a flood prone coastal city with long short-term memory and recurrent neural networks. *Water*, *11*(5), 1098. https://doi.org/10.3390/w11051098

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., et al. (2020). *Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB)*. UK Centre for Ecology & Hydrology. Retrieved from https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, *51*(7), 5531–5546. https://doi.org/10.1002/2014WR016532

de la Fuente, A., Meruane, V., & Meruane, C. (2019). Hydrological early warning system based on a deep learning runoff model coupled with a meteorological forecast. *Water*, *11*(9), 1808. https://doi.org/10.3390/w11091808

Do, H. X., Westra, S., & Leonard, M. (2017). A global-scale investigation of trends in annual maximum streamflow. *Journal of Hydrology*, *552*, 28–43. https://doi.org/10.1016/j.jhydrol.2017.06.015

Fang, K., Kifer, D., Lawson, K., & Shen, C. (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. *Water Resources Research*, *56*(12), e2020WR028095. https://doi.org/10.1029/2020wr028095

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

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

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

Fekete, B. M., & Vörösmarty, C. J. (2007). The current status of global river discharge monitoring and potential new technologies complementing traditional discharge measurements (Vol. 8). IAHS Publication.

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

George, D., Shen, H., & Huerta, E. A. (2017). *Deep transfer learning: A new deep learning glitch classification method for advanced LIGO*. Arxiv Preprint. Retrieved from http://arxiv.org/abs/1706.07446

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press. Retrieved from https://www.deeplearningbook.org/

Guillon, H., Byrne, C. F., Lane, B. A., Sandoval Solis, S., & Pasternack, G. B. (2020). Machine learning predicts reach-scale channel types from coarse-scale geospatial data in a large river basin. *Water Resources Research*, *56*(3). https://doi.org/10.1029/2019WR026691

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

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019). Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 558–566). Calgary, Alberta, Canada: Society for Industrial and Applied Mathematics. Retrieved from https://doi.org/10.1137/1.9781611975673.63

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

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. *Hydrology and Earth System Sciences Discussions*, 1–32. https://doi.org/10.5194/hess-2019-368

Liang, C., Li, H., Lei, M., & Du, Q. (2018). Dongting lake water level forecast and its relationship with the three gorges dam based on a long short-term memory network. *Water*, *10*(10), 1389. https://doi.org/10.3390/w10101389

Li, W., Kiaghadi, A., & Dawson, C. (2020). High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-020-05010-6

Ma, K., Huang, X., Liang, C., Zhao, H., Zhou, X., & Wei, X. (2020). Effect of land use/cover changes on runoff in the Min River watershed. *River Research and Applications*, *36*(5), 749–759. https://doi.org/10.1002/rra.3608

Marmanis, D., Datcu, M., Esch, T., & Stilla, U. (2016). Deep learning Earth observation classification using ImageNet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, *13*(1), 105–109. https://doi.org/10.1109/LGRS.2015.2499239

McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., et al. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

McMillan, H., Seibert, J., Petersen-Overleir, A., Lang, M., White, P., Snelder, T., et al. (2017). How uncertainty analysis of streamflow data can reduce costs and promote robust decisions in water management applications. *Water Resources Research*, *53*(7), 5220–5228. https://doi.org/10.1002/2016WR020328

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

Newman, A. J., Sampson, K., Clark, M. P., Bock, A., Viger, R. J., & Blodgett, D. (2014). *A large-sample watershed-scale hydrometeorological dataset for the contiguous USA*. Boulder, CO: UCAR/NCAR. Retrieved from https://doi.org/10.5065/D6MW2F4D

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2020). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*, *16*(2), 024025. https://doi.org/10.1088/1748-9326/abd501

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, *55*(11), 9173–9190. https://doi.org/10.1029/2019WR024922

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

Sivapalan, M. (2003). Prediction in ungauged basins: A grand challenge for theoretical hydrology. *Hydrological Processes*, *17*(15), 3163–3170. https://doi.org/10.1002/hyp.5155

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 843-852, IEEE. Retrieved from http://arxiv.org/abs/1707.02968

Teutschbein, C., & Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, *456–457*, 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052

Thrun, S., & Pratt, L. (1998). *Learning to learn*. Norwell, MA: Kluwer Academic Publishers. Retrieved from http://dl.acm.org/citation.cfm?id=296635

Wagener, T., & Wheater, H. S. (2006). Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *Journal of Hydrology*, *320*(1), 132–154. https://doi.org/10.1016/j.jhydrol.2005.07.015

Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018). Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) – COMPASS '18* (pp. 1–5). Menlo Park and San Jose, CA, USA: ACM Press. https://doi.org/10.1145/3209811.3212707

Yaseen, Z. M., El-Shafie, A., Jaafar, O., Afan, H. A., & Sayl, K. N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology*, *530*, 829–844. https://doi.org/10.1016/j.jhydrol.2015.10.038

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2007WR006716

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *NIPS 2014*. arXiv preprint. Retrieved from http://arxiv.org/abs/1411.1792

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., & Li, L. (2021). From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environmental Science & Technology*, *55*(4), 2357–2368. https://doi.org/10.1021/acs.est.0c06783

Zhu, S., Ptak, M., Yaseen, Z. M., Dai, J., & Sivakumar, B. (2020). Forecasting surface water temperature in lakes: A comparison of approaches. *Journal of Hydrology*, *585*, 124809. https://doi.org/10.1016/j.jhydrol.2020.124809

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Zhu, Y., Chen, Y., Lu, Z., Pan, S. J., Xue, G., Yu, Y., & Yang, Q. (2011). Heterogeneous transfer learning for image classification. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (pp. 1304–1309). San Francisco: AIAA Press. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewFile/3671/4073

References From the Supporting Information

Fischer, G., Nachtergaele, F., Prieler, S., van Velthuizen, H. T., Verelst, L., & Wiberg, D. (2008). *Global Agro-Ecological Zones Assessment for Agriculture (GAEZ 2008)*. Laxenburg, Austria, Rome, Italy: IIASA, FAO.

Meng, X., & Wang, H. (2018). *China meteorological assimilation driving datasets for the SWAT model version 1.1 (2008–2016)*. National Tibetan Plateau Data Center. https://doi.org/10.3972/westdc.002.2016.db