The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology

Kuai Fang1,2, Daniel Kifer3, Kathryn Lawson2, Dapeng Feng3, and Chaopeng Shen2

1Department of Earth System Science, Stanford University, Stanford, CA, USA, 2Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA, 3Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA

Abstract When fitting statistical models to variables in geoscientific disciplines such as hydrology, it is a customary practice to stratify a large domain into multiple regions (or regimes) and study each region separately. Traditional wisdom suggests that models built for each region separately will have higher performance because of homogeneity within each region. However, each stratified model has access to fewer and less diverse data points. Here, through two hydrologic examples (soil moisture and streamflow), we show that conventional wisdom may no longer hold in the era of big data and deep learning (DL). We systematically examined an effect we call data synergy, where the results of the DL models improved when data were pooled together from characteristically different regions. The performance of the DL models benefited from modest diversity in the training data compared to a homogeneous training set, even with similar data quantity. Moreover, allowing heterogeneous training data makes eligible much larger training datasets, which is an inherent advantage of DL. A large, diverse data set is advantageous in terms of representing extreme events and future scenarios, which has strong implications for climate change impact assessment. The results here suggest the research community should place greater emphasis on data sharing.

1. Introduction

As in many other geoscientific fields, there has been a long and pervasive history in hydrology of stratifying data points into different regions or regimes, for which one separately creates statistical models for the variables of interest. This has been done, for example, with hydraulic geometry curves (the relationship between discharge and channel geometries like width and depth): many studies have divided the United States into multiple regions, each of which was fitted with a separate hydraulic geometry curve (Bieger et al., 2015; Castro & Jackson, 2001). Regional regression formulas were prevalent since the early days for estimating annual streamflow (Vogel et al., 1999) and evapotranspiration (Fennessey & Vogel, 1996), as well as for flood frequency analysis (Archfield et al., 2013; Burn et al., 1997). Apart from regionalization schemes (discussed more below) which aimed at prediction in ungauged basins, rainfall-runoff models were mostly calibrated for each basin separately or for a small batch of basins in a region, for example, see relatively recent work (Li et al., 2018; Rajib et al., 2018). For a broader geoscientific example, the US was divided into many different fire regimes for modeling wildfires (Barrett et al., 2010). The assumed benefits from stratification may have also partially given popularity to many stratification and classification schemes such as ecoregions and hydrologic landscape regions (Wolock, 2003).

Related to stratifying by contiguous regions, hydrologists are also familiar with the concepts of hydrologic classification and similarity (Wagener et al., 2007). Many classification schemes exist in the attribute space, for
example, using hydrologic signatures (Sawicz et al., 2014), flood generating mechanisms (Berghuijs et al., 2016), hydrologic disturbance (McManamay et al., 2014), or storage-streamflow response regimes (Fang & Shen, 2017). The basic principle is that basins clustered in each class are, in certain metrics, similar, and thus the variability within each class is limited (McDonnell & Woods, 2004). These concepts provide the framework to guide our understanding and facilitate transfer of information (Sawicz et al., 2011; Wagener et al., 2007). Regardless of the scheme, however, the implicit assumption of classification is that grouping similar basins can better guide us to model the systems and project future changes.

In parallel, there are several classes of methods under the banner of hydrologic regionalization that seek to transfer calibrated hydrologic parameters to ungauged basins, as summarized by Brunner et al. (2018), Guo et al. (2021), and Razavi and Coulibaly (2013). Normally, information sharing is facilitated between catchments that are deemed similar, and discouraged between those deemed dissimilar. Some other classes of hydrologic regionalization approaches attempt to build whole-domain transfer functions (or regression relationships) between model parameters and catchment attributes (Beck et al., 2020; Kumar, Livneh, & Samaniego, 2013). Various modeling studies established the expectation that regionalization schemes would sacrifice some local performance for generality and transferability (Beck et al., 2020; Hogue et al., 2005; Kumar, Samaniego, & Attinger, 2013; Rosero et al., 2010). However, this experience has not been verified against recently popular deep learning models, to be discussed below.

The well-known learning theory of bias-variance tradeoff (Shalev-Shwartz & Ben-David, 2014) is at the core of this need for stratification. For a model class (loosely, the set of functions that can be obtained by varying the parameters of a given basic model architecture), bias measures the error of the model that best approximates the underlying true relationship (i.e., the error with the best possible choice of model parameters). Meanwhile, variance measures sensitivity to sampling variability and other noise in the training data (stated another way, model variance measures how much the model parameters can be constrained given the training data at hand). Large variance indicates the model is overfitting to the noise in the data, rather than to the general data trends. Both bias and variance contribute to the overall model error. The bias-variance tradeoff states that if a model class is too simple, it could have a small variance but a larger bias. On the other hand, if the model class is too complex, it will have a low bias but a large variance, often because there is not enough data to properly constrain the model. In the framework of the bias-variance tradeoff, the goal of stratification is to separate out regions with relatively homogeneous conditions so that each region may be characterized by a simple underlying relationship. A small hypothesis class can thus be fitted with acceptable bias. In addition, there are always latent variables which cannot be observed or provided as inputs, such as geologic characteristics. Assuming that the important latent variables are relatively homogeneous within each region, their effects can then be conveniently lumped into the constants and coefficients of the region-specific model. However, if one increases the number of region divisions allowable, the average number of data points per region decreases, thus increasing the variance of each region-specific model. Therefore, one must hope to wisely choose a stratification scheme such that the benefit of simplification due to stratification outweighs the drop in data quantity.

Recently, deep learning (DL) approaches have proven to be a promising tool in modeling hydrologic dynamics (Shen, 2018; Shen & Lawson, 2021; Sit et al., 2020). Among these, long short-term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) present excellent performance in modeling soil moisture (Fang et al., 2017, 2019), streamflow (Feng et al., 2020; Frame et al., 2021; Gauch, Kratzert, et al., 2021; Ha et al., 2021; Kratzert et al., 2019; Nearing, Klotz, et al., 2021; Xiang & Demir, 2020), water table depth (Zhang et al., 2018), water quality variables such as water temperature (Rahmani et al., 2020, 2021) and dissolved oxygen (Zhi et al., 2021), and reservoir modulation (Ouyang et al., 2021). DL can be adapted for tasks like uncertainty quantification (Fang et al., 2020; Li et al., 2021), data assimilation (Fang & Shen, 2020; Feng et al., 2020), and multiscale modeling (Liu et al., 2022). In many of these models, spatial attributes were included as static input attributes, allowing the models to differentiate between different basins, grid cells, or sites. This setup permitted simultaneous training and simulation over thousands of sites or more. However, in many other machine learning studies, following the conventional wisdom of stratification, geoscientists still tend to train separate models using data from each site (Duan et al., 2020; Herath et al., 2021; Petty & Dhingra, 2018), or each region composed of sites with similar environmental conditions (Abdalla et al., 2021; Sahoo et al., 2017).

Several research groups have presented scattered evidence that DL model performance improves as we include more sites (or basins), yet this effect has not been formalized, rigorously studied, or systematically summarized.
For example, Nearing, Kratzert, et al. (2021) showed models trained using all data from the conterminous United States (CONUS) were stronger than those trained on one basin alone, but the difference could have simply been attributed to the very limited data from one basin. For another example, Gauch, Mai, and Lin (2021) studied the impact of increasing training data size based on random sampling of the CAMELS data set, but this experiment was conducted by random sub-samples and focused on model performance over all basins. Their test scheme did not address whether one should include more training data if they were only interested in their own basins (which requires testing on the same basins they started with). It was also not clear whether samples inside a homogeneous region contained sufficient information to capture the hydrologic dynamics, or if including more samples from multiple regions would confuse the model. Moreover, none of these studies examined the impacts of geographic similarity or diversity, which require geographically clustered sampling. Due to the lack of a systematic study, there is generally an under appreciation of the value of more hydrologic training data from characteristically different regions.

In this study, we systematically study the interesting phenomenon with DL models where a large training set leads to a unified model that tends to be statistically stronger than a collection of stratified, locally trained models (i.e., the whole is greater than the sum of its parts). We call this effect data synergy, borrowed from Higginson et al. (2018). We hypothesize that deep learning networks use their internal representations to automatically form multilevel models that learn inter-regional homogeneities and heterogeneities (commonalities and differences between regions). This hypothesis has a range of implications. For instance, suppose one is interested in making predictions about region X. One could amass a large homogeneous data set purely from region X, as well as an equally sized heterogeneous data set that contains data not only from X but from other regions as well. According to the theory of data synergy, a model trained on the second data set should be able to model the commonalities better and should be less prone to overfitting than a model trained on the first data set. As a result, the data synergy effect would mean that the model trained on this second, heterogeneous data set would achieve higher predictive performance for region X. Given the current era of big data, such a phenomenon would suggest that researchers could increasingly benefit from sharing and pooling datasets together, even if the data were to come from outside of an individual researcher’s region of interest.

We demonstrate the effect of data synergy with time-series DL models in hydrology for (a) satellite-observed soil moisture and (b) streamflow measured at basin outlets. In these experiments, predictions from local models (trained using data only from inside the respective region), and predictions from global models (trained using more heterogeneous data that included sites both in the study region and from more distant regions), were evaluated in various regions of interest. The experiments were designed to address the following questions: (a) For these applications, are global models better than local models? and (b) Do the models benefit from the diversity of this training data, or simply the increased quantity of training data, or both? Answering these questions may guide us to better understand how DL networks work to improve model performance.

2. Methods and Data

In this section, we first present the datasets leveraged in this study (§2.1), followed by DL model structure (§2.2) and specific experimental designs (§2.3).

2.1. Input and Target Datasets

We investigated the phenomenon of data synergy as applied to two different types of hydrological predictions: soil moisture and streamflow.

2.1.1. Soil Moisture Data

In the soil moisture experiments, the Soil Moisture Active Passive (SMAP) satellite mission’s Level 3 radiometer product (L3SMP, version 6) was used as the training target. SMAP measures global surface soil moisture (<5 cm) on a 36 km Equal-Area Scalable Earth Grid (EASE-Grid) based on L-band passive brightness temperature, with a revisit time of about every 2–3 days starting on 2015/04/01. Our inputs contained dynamical forcings (meteorological conditions) and static geophysical attributes. Climate forcing data included precipitation, temperature, long-wave and short-wave radiation, specific humidity, and wind speed, which were extracted from the North American Land Data Assimilation System phase II (NLDAS-2) data set. Static physiographic data included land
cover classes, surface roughness, and vegetation density extracted from SMAP flags; soil properties like sand, silt, and clay percentages, bulk density, and soil water capacity obtained from the World Soil Information (ISRIC-WISE) database; and normalized difference vegetation index (NDVI) values obtained from the Global Inventory Monitoring and Modeling System (GIMMS). All the input data are aggregated into SMAP’s 36 km EASE-Grid using area weighting.

2.1.2. Streamflow Data

For streamflow experiments, we collected streamflow observations from the U.S. Geological Survey’s (USGS) National Water Information System (NWIS) database. Here our goal was to predict daily basin runoff (mm), which we calculated by dividing daily USGS streamflow observations recorded at the basin outlet by the area of the basin. The training period was 1979/01/01 to 2009/12/31, and the testing period was 2010/01/01 to 2019/12/31. We selected 2,773 USGS basins which had observations available for more than 90% of the days in both training and testing periods. Among those basins, 576 of them were categorized as reference basins, which are considered to have low human impacts and high data quality. We re-assembled this data set, instead of relying on existing datasets such as Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS; Newman et al., 2015), so that our experiments could use more basins than the 671 basins in CAMELS.

As with the soil moisture data set, we extracted basin-averaged climate forcings and geophysical attributes as input predictors. For streamflow, however, the daily climate forcing data were extracted from the gridMET (Abatzoglou, 2013) product, which contains precipitation, temperature, humidity, radiation, and reference evapotranspiration, with a spatial resolution of 1/24°. For each targeted USGS site, we integrated the gridMET data set with the drainage basin boundary from the Geospatial Attributes of Gages for Evaluating Streamflow II (GAGES-II) data set (Falcone, 2011). Geographic attributes were also extracted from GAGES-II, and we selected 17 fields likely to impact the rainfall-runoff process, including drainage area, basin compactness ratio, snow percent of precipitation, stream density, percentage of first-order streams, base flow index, subsurface flow contact time, dam density, permeability, water table depth, rock depth, slope, dominant ecoregion, nutrient region, geology region, hydrologic landscape, and land cover.

2.2. Model Architecture

Long short-term memory (LSTM) networks are general-purpose models for sequential data and have proven to be effective in hydrology applications. In this study, we used LSTM to predict two dynamical hydrologic variables (soil moisture and streamflow) using the inputs described in §2.1. LSTM models were trained on pixels for soil moisture and on basins for streamflow. In both cases, we used a similar network architecture, which consisted of a linear layer of 256 nodes with rectified linear unit (ReLU) activation, followed by a LSTM layer with 256 nodes, and then a linear output layer. The loss function, or metric which was the models’ primary goal to minimize, was root-mean-square error (RMSE) between observed and predicted values. The network was trained to minimize RMSE using the AdaDelta optimizer (Zeiler, 2012), which dynamically tunes learning rate through training iterations. For soil moisture models, the length of time step was set to be 30 (days) and batch size was 100; for streamflow models, time step length was 365 (days) and batch size was 500. All models were trained for 500 epochs, with a hidden size of 256 and dropout rate of 0.5.

Our model settings, like hidden size and dropout rate described above, followed our earlier work reported in Fang et al. (2019, 2020). As this work focuses on examining the data synergy effect, we did not further tune the hyperparameters. However, we further trained models with smaller hidden sizes (16, 32, 64, and 128) and different training epochs (100, 200, 300, and 400). The performance of those models (some of which are presented in Figures S1–S5 in the Supporting Information S1) suggests that these settings have little influence on our conclusions.

2.3. Experimental Design

Stratification of the data was guided by the United States Environmental Protection Agency (EPA) ecoregions (Omernik & Griffith, 2014), as these groupings were devised to provide similarity in terms of surface hydrologic responses. The CONUS was divided into ecoregions based on the compositions of geology, landforms, soils, vegetation, climate, land use, wildlife, and hydrology. Three hierarchical levels (denoted as I, II, and III) divide
the CONUS into 11, 25, and 105 regions, respectively. For example, ecoregion 8.3.5 (Southeastern Plains) is a level III ecoregion nested within ecoregion 8.3 (Southeastern USA Plains), which is a level II ecoregion nested inside level I ecoregion 8 (Eastern Temperate Forests). Figure 1 shows a map of level II EPA ecoregions and the boundaries of ecoregions from level I to III.

The first set of experiments, which we refer to as “global versus local” experiments, compares stratification (dividing the data and separately building models on each stratum) to unification (training a single model on the entire data). However, if in these experiments the global model was found to perform better than each of the local models, it could be argued that this was simply because the global model had more data to work with. Thus, the second set of experiments, which we refer to as “similar versus dissimilar” experiments, were designed to

---

Figure 1. (a) Map of U.S. Environmental Protection Agency (EPA) ecoregions colored based on level II regions. (b) Map of the 18 sub-regions used for the global versus local experiments, based on EPA ecoregions.
test whether the quantity of data fully explained the differences between the models, or if the diversity of the data set was also important. For both sets of experiments, the resulting models were evaluated inside various regions of interest (ROI) using temporal generalization tests, where the testing data came from a different time period than the training data (see §2.1).

2.3.1. Global Versus Local Experiments

These experiments were devised to directly compare unification and stratification. Considering the data quality and computational cost, the streamflow experiment here only includes 576 reference basins. In order to divide basins and SMAP pixels having similar environmental conditions into individual regions, we generally considered the level II ecoregions. However, as some level II ecoregions did not contain enough data, we merged them with their closest neighbors, merging 5.2 with 5.3, 9.5 with 9.6, and 13.1 with 13.2. In addition, we merged ecoregions 14.3 and 15.4 since both were tropical forests and ecoregion 15.4 was too small to stand alone. The resulting 18 “sub-regions,” referred to using letters A-R, had more similar areas between $1 \times 10^5$ km$^2$ and $1 \times 10^6$ km$^2$, with an average area of $5 \times 10^5$ km$^2$ (Figure 1, Table 1). Regions L, N, P, and R, were excluded from the streamflow analyses, as there were almost no reference basins present in those regions.

We then compared two scenarios: (a) a single LSTM model trained with data from all 18 sub-regions (1 global model), and (b) individual models for each sub-region trained only with data from that sub-region (18 local models). In the testing phase, for each sub-region we compared the predictions from the global model and from that sub-region’s corresponding local model. More specifically, the global model was tested on the same pixels (for soil moisture) or gages (for streamflow) inside each sub-region as the corresponding local model. The same comparison was also conducted using the 2-digit Hydrologic Unit Code (HUC2) divisions to ensure our conclusions are robust.

2.3.2. Similar Versus Dissimilar Experiments

The second set of experiments was designed to study the effect of training data diversity on model accuracy. Put more simply, if we are interested in creating a prediction model for a ROI, should we gather additional data from nearby/similar regions, or should we instead obtain a more diverse data set? We used the hierarchical nature of the EPA ecoregions as a proxy for (dis)similarity: two level III ecoregions were defined as being close neighbors if they belonged to the same level II ecoregion, far neighbors if they belonged to the same level I ecoregion (but different level II ecoregions), or dissimilar if they belonged to different level I ecoregions.

For soil moisture, the location of a gridcell centroid determined its ecoregion membership. For streamflow, we determined ecoregion membership based on which ecoregion covered the majority of the basin. Obviously, the amount of data available for each level III ecoregion varied by a significant amount, and not all of them contained enough data to create viable local models. Thus we selected a subset of level III ecoregions to serve as our regions of interest (ROIs). For soil moisture we selected the six largest level III ecoregions (8.3.5, 9.3.3, 9.4.1, 9.4.2, 10.1.5, and 10.2.4) with at least 50 pixels. In this streamflow experiment, as the number of reference basins in level III ecoregions are inadequate, we included non-reference basins to increase sample numbers to 2773 basins. 12 level III ecoregions containing at least 60 USGS basins are selected (5.3.1, 8.1.7, 8.2.3, 8.2.4, 8.3.1, 8.3.4, 8.3.5, 8.4.1, 8.4.2, 8.5.3, 9.2.3, and 9.4.2).

We compared two scenarios where (a) data size was not controlled (hence the sizes of the datasets were only limited by availability of data), and (b) data size was controlled (so that the homogeneous training data and the heterogeneous data were of roughly the same size). For each scenario and ROI (e.g., ecoregion 8.3.5 without data size controlled), we trained four models:

1. The “local” model was trained on data only from within the ROI (e.g., data from ecoregion 8.3.5).
2. The “local + close neighbors” model was trained on data from all close neighbors of the ROI, equivalent to the entire level II ecoregion containing the ROI (e.g., data from ecoregion 8.3).
3. The “local + far neighbors” model was trained using all the neighbors in the same level-1 ecoregion as the ROI, excluding the “close neighbors” (e.g., data from ecoregions 8.3.5, 8.1, 8.2, 8.4, 8.5).

| Table 1 | Conversion Between the Experimental Sub-Regions and EPA Ecoregions |
|---------|---------------------------------------------------------------|
| New ID  | EPA ID | New ID | EPA ID | New ID | EPA ID |
| A       | 5.2, 5.3 | G      | 8.4    | M      | 10.1   |
| B       | 6.2     | H      | 8.5    | N      | 10.2   |
| C       | 7.1     | I      | 9.2    | O      | 11.1   |
| D       | 8.1     | J      | 9.3    | P      | 12.1   |
| E       | 8.2     | K      | 9.4    | Q      | 13.1, 13.2 |
| F       | 8.3     | L      | 9.5, 9.6 | R      | 14.3, 15.4 |

New ID
EPA ID
New ID
EPA ID
New ID
EPA ID
A
5.2, 5.3
G
8.4
M
10.1
B
6.2
H
8.5
N
10.2
C
7.1
I
9.2
O
11.1
D
8.1
J
9.3
P
12.1
E
8.2
K
9.4
Q
13.1, 13.2
F
8.3
L
9.5, 9.6
R
14.3, 15.4

We then compared two scenarios: a single LSTM model trained with data from all 18 sub-regions (1 global model), and (b) individual models for each sub-region trained only with data from that sub-region (18 local models). In the testing phase, for each sub-region we compared the predictions from the global model and from that sub-region’s corresponding local model. More specifically, the global model was tested on the same pixels (for soil moisture) or gages (for streamflow) inside each sub-region as the corresponding local model. The same comparison was also conducted using the 2-digit Hydrologic Unit Code (HUC2) divisions to ensure our conclusions are robust.
4. The “local + dissimilar” model was trained using all of the ecoregions that were dissimilar to the ROI (e.g., data from ecoregion 8.3.5 and all areas outside of ecoregion 8).

In this first scenario where training size was not controlled, the models were trained using all of the data in the ecoregions that were available to them. The numbers of pixels and basins inside each experimental region are listed in tables S1 and S2 in the Supporting Information S1. Figure S6 in Supporting Information S1 presents the maps of local, close neighbors, far neighbors, and dissimilar regions for all selected ROIs.

As mentioned earlier, to help disentangle the impacts of “more data” and “more dissimilar data,” we trained an additional four models for each ROI where the amount of added training data was controlled. Here, the data points fulfilling the criteria for addition beyond the “local” scenario were resampled so that the “local + close neighbors,” “local + far neighbors,” and “local + dissimilar” datasets each had the same amount of added data. This modification was performed for the soil moisture data, as the pixels are approximately evenly and continuously spatially distributed, making it straightforward to uniformly sub-sample data from the close, far, and dissimilar regions. For streamflow, it is more difficult to obtain a representative size-controlled sub-sample than for the soil moisture case because there are far more streamflow gages (especially reference ones) than soil moisture grid points. We also had to include non-reference basins, which contain more noise due to human impacts. Consequently, there would be a larger extent of variance between possible sub-samples, and we only presented this experiment for three ecoregions (8.3.1, 8.3.4, and 8.3.5) with relatively large sampling size in Supporting Information S1.

2.3.3. Evaluation of Model

Trained models were evaluated for temporal extrapolation inside each ROI, on identical pixels or basins. Soil moisture models were trained from 2015/04/01 to 2016/03/31 and tested from 2016/04/01 to 2018/04/01; streamflow models were trained from 1979/01/01 to 2009/01/01 and tested from 2010/01/01 to 2019/01/01. To evaluate soil moisture models, we calculated the correlation coefficient and RMSE between the observations and predictions for each pixel in a region during the testing period. For streamflow predictions, correlation was also calculated, but the Nash Sutcliffe model efficiency coefficient (NSE) was calculated instead of RMSE, to be in line with previous hydrologic literature. In both cases, the larger the value of the metric, the better a model performs. It is worth noting that all of the error metrics reported in the manuscript without specific labels are testing error – that is, calculated during the testing period.

3. Results and Discussion

3.1. Global Versus Local Experiments

The global versus local experiments compared unification (training a single model on the entire data set) to stratification (dividing data by region and separately building models for each individual region). Metrics resulting from these experiments are plotted in Figure 2 for soil moisture and Figure 3 for streamflow. Note that not all regions had sufficient pixels (for soil moisture) or basins (for streamflow) for analysis, so the specific regions investigated will differ (see §2.3.1 for details).

For the soil moisture problem, the global model significantly outperformed local models. The median RMSE was smaller for the global model than the local model, while the median correlation was larger for each region (Figure 2). To test the statistical significance between local models and the global model, we used the Wilcoxon signed-rank test, as we could not assume normality of the metrics. We conducted this for each region individually as well as for the entire CONUS by pooling the local predictions from each region together, and the results (p-value and testing sample size) are shown in Table S3 in Supporting Information S1. All of the p-values were small; the largest value was under 0.009 and most were orders of magnitude smaller. Aggregating all the tested pixels, the average test RMSE values for the global and local models were respectively 0.32 and 0.38, while corresponding correlations were 0.82 and 0.75. Global model prediction had a smaller testing RMSE than the local model for 87% of pixels, and higher correlation for 95% of pixels. This clearly demonstrates that for soil moisture, the global model consistently and significantly (both in a practical and statistical sense) outperformed the local models. Our additional experiments (Figure S1 in the Supporting Information S1) showed that the changes due to hyperparameters (hidden size varied from 256 to 16 while stopping epoch from 500 to 100) are minor in contrast.
to the differences between global and local models. Given all the different hyperparameter settings, none of the local models approached the performance of the global models.

The streamflow experiment suggests a similar conclusion. Within each region, the median NSE value (calculated over all basins in the region) for the global model was also higher than that for the local model (Figure 3). It should be noted however that in region K, even though the median NSE was higher, the global model’s error variability was so large that in practice the local model would be preferred. As with soil moisture, we used the Wilcoxon signed-rank test to measure the statistical significance (Table S3 in Supporting Information S1). Only regions K and Q had p-values larger than 0.01 (note that region Q only had a sample size of 7 basins). The overall median correlations for the global and local models were 0.84 and 0.79 respectively, while the corresponding NSE values were 0.73 and 0.65. NSE for the global model was higher than for the local model in 81% of the basins and correlation was higher for 84% of the basins. Like the soil moisture models, these streamflow modeling results showed that the global model generally had higher quality than the stratified models. Similar to soil moisture, altering hyperparameters did change our conclusions (Figure S3 in the Supporting Information S1).

Besides, it is worth mentioning that our experiments on the HUC2 regions gave qualitatively the same conclusions (Figures S7 and S8 in the Supporting Information S1).

One reason that could explain this advantage is that a global-scale model has the opportunity to see a much wider range of forcings and responses, as well as more combinations of attributes. For example, some northern SMAP pixels would normally be frozen during winter, and the local model would fail to predict soil moisture when the ground froze unusually late or thawed early, as Figure 4a shows. The global model learned about soil moisture dynamics during warm springs and winters (highlighted in Figure 4a) from other pixels and could apply that knowledge to this pixel, while the local model could not. For another example, Figure 4b shows a pixel inside ecoregion G (8.4). This pixel has winter wheat as the major land use but ecoregion G overall does not – the

Figure 2. Result of global versus local experiments on soil moisture models. Testing performance inside regions of interest (ROIs) are compared between the global model (trained with all Soil Moisture Active and Passive pixels over the conterminous United States) and local models (trained with pixels inside the ROI). Upper panel shows root-mean-square error; lower panel shows correlation.
majority of winter wheat agriculture is inside ecoregion K (10.4). As a result, the local model was not adequately trained to predict winter soil moisture patterns (highlighted in Figure 4b), while the global model alleviated this issue.

Analogous examples can be found in the streamflow experiments which highlight the capability of the global model in predicting hydrograph peaks compared to local models across the entire CONUS, for example, Figures 4c and 4d. In addition, on snow-dominated basins (Figure 4e), local models seemed to mis-calculate snow accumulation and over-predict the spring streamflow due to snowmelt. This advantage of the global model may be simply due to the fact that it has the opportunity to see more extreme events from combining all regions. For each ecoregion, by definition, rare events are rare, and they may be poorly represented in the local training data. However, the global model could absorb and transfer the knowledge of responses to extreme events between regions. Therefore, there is a synergistic effect in pooling data together from different regions.

These results are materially different from earlier results mentioned above (Beck et al., 2020, also personal communication about this result), which indicate that local calibration at the site of interest outperformed large-scale regionalized parameters. In this scenario, the more traditional calibration method struggled to simultaneously accommodate the different error sources at different basins, while the large-capacity DL models worked well.

3.2. Similar Versus Dissimilar Experiments

3.2.1. Data Set Size Not Controlled

As described in methods, to clarify if similar or dissimilar data bring in the most benefit, we identified “close,” “far,” and “dissimilar” neighbors based on ecoregion stratification, and examine their impacts on model
performance inside the ROI. For SMAP soil moisture prediction in each chosen ROI, we saw that RMSE and correlation monotonically improved as we added increasingly diverse data to the “baseline” local model (the model trained using only data from the ROI), with the best performance being achieved by the most heterogeneous data set (local + dissimilar; Figure 5). The improvement was less pronounced for the drier western regions (10.1.5, 10.2.4) as compared to the wetter eastern regions, where soil moisture has larger fluctuations. After evaluating statistical significance, we saw that in these wetter regions, all of the pairwise comparisons were significant, with p-values much lower than the 0.01 significance threshold (Table S4 in Supporting Information S1). For the two drier regions (10.1.5, 10.2.4), a few of the comparisons were not statistically significant at this small sample size. However, for correlation, all comparisons involving “local + dissimilar” were significant, showing that adding in the data from other Level I ecoregions did not hurt performance (as conventional wisdom might suggest). Rather, it actually helped the most.

Figure 4. Example time series of soil moisture and streamflow simulation, comparing the global and local model. Upper panel: soil moisture experiment, and yellow circle highlighted events demonstrated in §3.1; lower panel: streamflow experiment.
Figure 5. Performance metrics for soil moisture similar versus dissimilar experiments without training data set size controlled. Upper panel shows root-mean-square error; lower panel shows correlation.

Figure 6. Performance metrics for streamflow similar versus dissimilar experiments without training data set size controlled. Upper panel shows correlation; lower panel shows Nash Sutcliffe model efficiency coefficient.
For streamflow we observed a similar general trend in that a more diversified training set improved predictions, but the effect was smaller than for soil moisture and not as monotonic (Figure 6). Due to the smaller effect size and the small sample size within each region, most, but not all comparisons were statistically significant at the 0.01 level (Table S5 in Supporting Information S1). However, when the ROIs were pooled together for hypothesis testing (last line of Panels A and B), they showed unambiguously that the differences were statistically significant, implying that overall, diversity helped improve predictions. Our numerical experiments using different hyperparameters also reported similar results (Figures S2 and S4 in the Supporting Information S1).

There were some exceptions to this trend, however. Upon closer inspection, we noted that for some cases, NSE dropped from “local + close” to “local + far” data (regions 8.4.1, 8.4.2, 9.2.3), suggesting that in those cases, the dissimilar training set may have introduced additional bias to the model (Figure 6). Furthermore, where the LSTM models performed poorly (e.g., region 9.4.2), including diverse training regions did not improve the model performance. Large errors tended to be associated with large basin areas, which may have been due to a variety of factors including that (a) the sub-basins were heterogeneous and there was not enough data for the local model to learn this heterogeneity, (b) the watershed boundaries were unclear, or (c) cross-basin groundwater flow (which was not part of the model) had a larger impact than anticipated. Besides, it is worth noting that in general, “local + dissimilar” contained more samples compared to “local + far”, and “local + close” had the fewest sample numbers. The numbers of pixels and basins inside each experimental region are listed in Tables S1 and S2 in the Supporting Information S1.

These observations suggest that one needs to prioritize the collection of enough local data to build a local model with reasonably good performance. After that, additional improvements can be obtained from data collected outside the ROI with preference toward heterogeneous data, as it may provide a regularizing effect and help guard against overfitting. If the local model underfits though, the heterogeneous data may not help. This conclusion is further supported by the experiments presented in Figures S12 and S13 in the Supporting Information S1, where models trained only on dissimilar regions had worse performance compared to models trained on local regions. It is worth repeating, however, that while this was the case for “close” versus “far” regions, comparisons between “close” versus “dissimilar” always showed significantly improved predictions.

3.2.2. Data Set Size Controlled

Our data-size-controlled experiments, which was designed to further clarify the significance of “more diverse data” as opposed to “more data” by maintaining the same sample size for all training sets, show that differences in performance between the alternatives was significantly dampened but still noticeable (Figure 7). Due to the small sample size per region, most pairwise tests were still significant, but the fraction of insignificant tests was larger than for the case without sample size controlled (Table S6 in the Supporting Information S1). When all the data were pooled, it was clear that the improvement of “local + far” over “local + close” was significant, as was the improvement of “local + dissimilar” over “local + close.”

Interestingly, there was no evidence to suggest a meaningful difference between “local + dissimilar” versus “local + far” (Table S6 in the Supporting Information S1). The “local + dissimilar” was better in some cases (10.1.5, 10.2.4) while “local + far” was better in some other cases (9.3.3 and 9.4.1). With similar amounts of data, the “far” data set may have been more informative in some cases, possibly because these examples clarified the impacts of fine-grained differences in some input properties. The implication here is that when seeking to enrich the training set by adding more heterogeneous examples, we do not have to search too far away from the region of interest, unless doing so would allow us to enlarge the training size substantially.

For the streamflow experiment, we controlled sample size by randomly selecting subsets from “far” and “dissimilar” basins. We found that there was a large variation in the performance between the different sub-samples, but the data synergy effect remained statistically sound. Three size-controlled cases (8.3.1, 8.3.4, 8.3.5) with relative large sample sizes are presented in Figures S9, S10, and S11 in the Supporting Information S1. For these three cases, models with controlled size present a similar but dampened pattern compared to uncontrolled ones, similar to what we observed in the soil moisture experiment. For example, for 8.3.1, out of the five random “local + far” sub-samples, four of them incurred better NSE values than “local + close.” Hence our conclusion remained robust for the streamflow case.
These results allow us to reject the notion that the “far” and “dissimilar” data points were of lower value for building a model at any given ROI. Combined with the uncontrolled data experiments, we saw that both quantity and diversity in data helped to improve model quality, with the former showing a larger effect. It is worth clarifying that these experiments do not suggest that any dissimilar sample will improve the local model performance. Assuming there are certain out-of-region samples that could further assist a robust model that is adequately trained, a more diverse data set is more likely to capture those “helpful” samples. This experiment highlighted the advantage of diverse data (not simply dissimilar data) over more homogeneous data, further supporting (but not proving) the hypothesis that heterogeneity in data has a regularizing effect that could reduce overfitting.

Overall, the experiments together show an inherent benefit of allowing more heterogeneous training data to deep learning models in hydrology – not only do heterogeneous inputs appear to help the model, but heterogeneous datasets are also naturally much more plentiful, permitting us to amass much larger datasets. This observation liberates us from the need to use small, stratified datasets when applying deep learning in hydrology and (in our opinion) should not be understated.

4. Discussion

There are several (not mutually exclusive) explanations for the data synergy effect. Besides training diversity, one is that heterogeneous data may provide a regularizing effect that reduces overfitting. Another is that a deep learning model may use its internal representations to construct a multi-level model that captures similarities among regions (e.g., the main effect) as well as region-specific differences, as discussed earlier. If the latter case were true, it would suggest that deep networks extract the common part of the data and build a basic soil moisture dynamics model, knowing, for example, that soil moisture rises when rainfall occurs and declines when rainfall ceases. The model can also be specialized to predict different response curves as modulated by different soil and land use characteristics. When data comes from more different regions, it is easier for the model to discern the most basic, fundamental responses, whereas data from similar regions may have had more commonalities overall.
but not all of them were fundamental. However, both local data and heterogeneous data are necessary for DL models to learn robust hydrologic responses. The data synergy effect encourages pooling data together, rather than choosing one over another. An auxiliary experiment is shown in Figure S12 in the Supporting Information S1, where DL models could learn the general pattern (evidenced by high correlation but high RMSE) from the non-local samples, or detailed dynamics (suggested by low correlation but low RMSE) from local samples. Nevertheless, using both local and non-local together led to the best performance.

The data synergy effect seemed to be less pronounced with streamflow predictions than soil moisture predictions. One potential explanation is that streamflow involves more latent processes with rainfall-runoff modeling, for example, the input representations for geology (aquifer laying and transmissivity), and the stream networks were too simplified. Due to these unknown and potentially confounding factors, it would be more difficult for the network to extract the true multilevel model. This situation would not be unique to streamflow prediction, and may also apply to stream temperature modeling (Rahmani et al., 2020), water chemistry (Zhi et al., 2020), and other hydrologic problems. It can be said that most geoscientific variables, to some extent, have latent variables or parameters that we cannot fully describe. Also, when large amounts of data exist locally (e.g., we have a high density of gauges with long records), we would expect the benefits of dissimilar data to wane accordingly. Hence, we caution against generalizing data synergy in the absolute sense to all stratification schemes and to all problems. However, our results suggest that pooling big data together is certainly one option that can be tried to improve performance for other hydrological puzzles.

There are important implications of the data synergy effect for climate change impact assessment. Many regions expect to see warmer climate and more frequent extreme events (Lee et al., 2021). As these shifts continue to occur, the response to future events of a basin may have already been witnessed in the historical records of other basins, for example, its southern, warmer neighbors. There is a higher likelihood that predicting the response to future extreme events is an act of interpolation rather than extrapolation, if we use a large data set consisting of many heterogeneous regions. Hydrologists have long adopted strategies such as “trading space for time” (Singh et al., 2011). DL models are well suited to tap into such synergistic effect almost effortlessly.

The data synergy effect is consistent with the data scaling relationship observed when DL is used to efficiently parameterize process-based models (Tsai et al., 2021). Although arising from different contexts, they both suggest pooling data together leads to beneficial effects. There has been repeated calls for hydrologic studies, and geoscientific studies in general, to transcend the uniqueness of places (Sivapalan, 2006; Wagener et al., 2020). It appears this objective can be achieved by machine learning, potentially via automatically built multilevel models, without human supervision. We would like to explicitly note that we have not “proven” the multilevel theory or the regularization theory, although both support our observed experimental results. Additional study of the network parameters themselves would be needed to confirm either theory. Future effort could devise visualization approaches to understand how this was accomplished and what commonality was extracted (Shen et al., 2018).

We also want to explicitly state that this work does not discourage hydrologic classification. Classification is a highly effective and illuminating tool to “provide an organizing principle, create a common language, guide modeling and measurement efforts” (Wagener et al., 2007). In our experiment, the input training data contains labels extracted from several classification frameworks, including Hydrologic Landscape Regions (Wolock, 2003), generalized geologic maps (Reed & Bush, 2001), Potential Natural Vegetation (Schmidt, 2002), and ecoregions. Excluding those regional labels will not affect the conclusion of this manuscript (Figure S14 in the Supporting Information S1). How to utilize these classification frameworks to assist training of DL models would be an interesting topic requiring future study, and is beyond the scope of this work. Nevertheless, the implication of data synergy here is that, for the purpose of making better predictions with DL, it would be worthwhile to collect larger and more heterogeneous data that are not confined to a small region of interest. While the notion of large-sample hydrology has been publicized (Addor et al., 2020), our work systematically and quantitatively examines the benefit of data synergy. We also did not study prediction in ungauged basins (PUB). All of our models were tested on the local basins in the training set, which allowed us to answer the questions we raised. However, since the global model exceeded locally calibrated models, there is a high chance that the global model will have equivalent or better results than a regional model for the case of PUB, where the generalizability of the model is more important.
5. Conclusion

In this study, we studied the data synergy effects in predicting soil moisture and streamflow using LSTM networks, and concluded that both more data and diverse data are independently helpful in improving model performance. On a practical level, these data synergy effects provide guidance for future data set construction and processing: unless we fundamentally lack critical inputs, we should not assume stratification is the best approach. Rather, we should try to compile a large data set from diverse domains, and attempt a unified model. If a data collection budget is limited, we should first collect enough local data to build a robust model with reasonable performance, but then may benefit from collecting data from modestly heterogeneous sources. While it cannot be guaranteed that performance will be better – for problems where a DL model itself performs very poorly or there are critical variables that are not known, stratification may nonetheless be useful – our experiences suggest it is likely that a more diverse data set will lead to a more robust and more accurate model. Meanwhile, if we only have a small data set and have to build a machine learning model specifically only using this data set, we should not expect the model to provide optimal predictions or capture universal relationships. In the case of truly heterogeneous inputs that are not comparable, other approaches such as transfer learning are applicable and may be more helpful (Ma et al., 2020).

Notably, among all the experiments we tried here, there were no cases in which the model performed worse after training with data from regions outside that of primary interest. This suggests that DL models’ performance is not compromised by additional information, even when they appear to be unrelated. In fact, as the similar versus dissimilar experiment shows, dissimilar ecoregions can bring in more knowledge as compared to the similar ones. The exact mechanism by which DL models accomplish this is not yet known, but we hypothesize that it may be related to multilevel models. Additionally, allowing more heterogeneous data by default makes eligible a much greater amount of potential model training data, which could be an important reason why big data machine learning techniques improve performance. Hence, we conclude that both the effects of data quantity and characteristics due to heterogeneity are important for DL models.

The data synergy effect of DL models could provide a vital pathway toward more accurate estimation of climate change impacts. Allowing heterogeneous training data will inherently permit the use of more training data. Large training datasets collected from diverse regions naturally provide more samples of extreme events and responses that resemble future scenarios. In summary, models that can easily leverage the data synergy effect may be able to better predict the future.

Data Availability Statement

All data used in this study are available from public sources, including forcing data from gridMET (https://doi.org/10.1002/joc.3413), land surface characteristics (including soil texture from ISRIC-WISE (https://www.isric.org/projects/world-inventory-soil-emission-potentials-wise), land cover from NLCD (https://www.mrlc.gov/data/nlcd-2016-land-cover-conus), and NDVI (https://ecocast.arc.nasa.gov/data/pub/gimms/3g.v1/)), basin attribute data (https://doi.org/10.3133/70046617), SMAP measurements, and streamflow data (https://www.mrlc.gov/data/nlcd-2016-land-cover-conus). The LSTM code can be downloaded from the open-source repository (https://doi.org/10.5281/zenodo.4068602).

Acknowledgments

K. Fang, D. Feng, and C. Shen were primarily supported by the Biological and Environmental Research program from the U.S. Department of Energy under contract DE-SC0016605. C. Shen and K. Lawson were also partially supported by Google AI Impacts Challenge Grant 1904-57775. C. Shen and K. Lawson 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 integrity of research at The Pennsylvania State University.

References

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology, 33*(1), 121–131. https://doi.org/10.1002/joc.3413

Abdalla, E. M. H., Pons, V., Stovin, V., De-Ville, S., Fassman-Beck, E., Alfredsen, K., & Muthanna, T. M. (2021). Evaluating different machine learning methods to simulate runoff from extensive green roofs. *Hydrology and Earth System Sciences Discussions, 1–24.* https://doi.org/10.5194/hess-2021-124

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal, 65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

Archfield, S. A., Pugliese, A., Castellarin, A., Skjøten, J. O., & Kiang, J. E. (2013). Topological and canonical kriging for design flood prediction in ungaged catchments: An improvement over a traditional regional regression approach? *Hydrology and Earth System Sciences, 17*(4), 1575–1588. https://doi.org/10.5194/hess-17-1575-2013

Barrett, S., Harlina, D., Jones, J., Hann, W., Frame, C., Hamilton, D., et al. (2010). Interagency fire regime condition class guide book version 3.0. USDA Forest Service, US Department of Interior, and The Nature Conservancy.
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

Berghuis, W. R., Woods, R. A., Hutton, C. J., & Sivapalan, M. (2016). Dominant flood generating mechanisms across the United States. *Geophysical Research Letters*, 43(9), 4382–4390. https://doi.org/10.1002/2016GL069790

Bierk, K., Rathjens, H., Allen, P. M., & Arnold, J. G. (2015). Development and evaluation of bankfull hydraulic geometry relationships for the phyiographic regions of the United States. *Journal of the American Water Resources Association*, 51(3), 842–858. https://doi.org/10.1111/jawr.12282

Brunner, M. I., Furrer, R., Sikorska, A. E., Viviroli, D., Seibert, J., & Favre, A.-C. (2018). Synthetic design hydographs for ungauged catchments: A comparison of regionalization methods. *Stochastic Environmental Research and Risk Assessment*, 32(7), 1993–2023. https://doi.org/10.1007/s00477-018-1523-3

Burn, D. H., Zini, Z., & Kowalchuk, M. (1997). Regionalization of catchments for regional flood frequency analysis. *Journal of Hydrologic Engineering*, 2(2), 76–82.

Castro, J. M., & Jackson, P. L. (2001). Bankfull discharge recurrence intervals and regional hydraulic geometry relationships: Patterns in the Pacific Northwest, USA. *Journal of the American Water Resources Association*, 37(5), 1249–1262. https://doi.org/10.1111/j.1752-1688.2001.tb03636.x

Duan, S., Ulrich, P., & Shu, L. (2020). Using convolutional neural networks for streamflow projection in California. *Frontiers in Water*, 2, 28. https://doi.org/10.3389/fwater.2020.00028

Falcone, J. A. (2011). GAGES-II: Geospatial attributes of gages for evaluating streamflow (Report). https://doi.org/10.3133/70046617

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. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US: Storage-streamflow correlation spectrum. *Water Resources Research*, 53(9), 8064–8083. https://doi.org/10.1002/2016WR020283

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*. 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

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

Fennessey, N. M., & Vogel, R. M. (1996). Regional models of potential evaporation and reference evapotranspiration for the northeast USA. *Journal of Hydrology*, 184(3), 337–354. https://doi.org/10.1016/0022-1694(95)02980-X

Frame, J. M., Kratzert, F., Runey, A., II, Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-processing the National water model with long short-term memory networks for streamflow predictions and model diagnostics. *Journal of the American Water Resources Association*, 57(6), 885–905. https://doi.org/10.1111/jawr.12686

Gaug, M., Kratzert, F., Klotz, D., Nearing, G. Lin, I., & Hochreiter, S. (2021). Rainfall-runoff prediction at multiple timescales with a single long short-term memory network. *Hydrology and Earth System Sciences*, 25(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

Gaug, M., Mai, J., & Lin, J. (2021). The proper care and feeding of camels: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, 135, 104926.

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *WIREs Water*, 8(1), e1487. https://doi.org/10.1002/wat2.1487

Hu, S., Liu, D., & Mu, L. (2021). Prediction of Yangtze river streamflow based on deep learning network with El Niño–southern oscillation. *Scientific Reports* (Nature Publishing Group), 11, 17338. https://doi.org/10.1038/s41598-021-90964-3

Herath, H. M. V. V., Chadalawada, J., & Babovic, V. (2021). Hydrologically informed machine learning for rainfall–runoff modelling: Towards distributed modelling. *Hydrology and Earth System Sciences*, 25(8), 4373–4401. https://doi.org/10.5194/hess-25-4373-2021

Higgins, S., Topouzi, M., Andrade-Cabrer, C., O’Dwyer, C., Darby, S., & Finn, D. (2018). Achieving data synergy: The socio-technical processes of handling data. In C. Foulds, & R. Robson (Eds.), *Advancing energy policy: Lessons on the integration of social sciences and humanities* (pp. 63–81). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-99097-2_5

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* (Publisher: MIT Press), 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hogue, T. S., Bastidas, L., Gupta, H., Soroshehian, S., Mitchell, K., & Emmerich, W. (2005). Evaluation and transferability of the NOAH land surface model in semiarid environments. *Journal of Hydrometeorology*, 6(1), 68–84. https://doi.org/10.1175/JHM-402.1

Kratzert, F., Klotz, D., Schalev, G., Klamboe, 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

Kumar, R., Livneh, B., & Samaniego, L. (2013). Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resources Research*, 49(9), 5700–5714. https://doi.org/10.1002/wrcr.20431

Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, 49(1), 360–379. https://doi.org/10.1002/wrr2012195

Lee, J., Marotzek, J., Bala, O., Gao, L., Corti, S., Dunne, J., et al. (2021). Future global climate: Scenario based projections and near-term information. *Climate change*. Li, D., Marshall, L., Liang, Z., Sharma, A., & Zhou, Y. (2021). Characterizing distributed hydrological model residual errors using a probabilistic long short-term memory network. *Journal of Hydrology*, 603, 126888. https://doi.org/10.1016/j.jhydrol.2021.126888

Li, Y., Grimaud, S., Pauwels, V. R., & Walker, J. P. (2018). Hydrologic model calibration using remotely sensed soil moisture and discharge measurements: The impact on predictions at gauged and ungauged locations. *Journal of Hydrology*, 557, 897–909. https://doi.org/10.1016/j.jhydrol.2018.01.013

Liu, J., Rahmani, F., Lawson, K., & Shen, C. (2022). A multiscale deep learning model for soil moisture integrating satellite and in-situ data. *Geophysical Research Letters*. (in press).
Zhang, J., Zhu, Y., Zhang, X., Ye, M., & Yang, J. (2018). Developing a long short-term memory (lstm) based model for predicting water table depth in agricultural areas. *Journal of Hydrology, 561*, 918–929. https://doi.org/10.1016/j.jhydrol.2018.04.065

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., & Li, L. (2020). From hydrometeorology to water quality: Can a deep learning model learn the dynamics of dissolved oxygen at the continental scale? https://doi.org/10.1002/essoar.10504429.1

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 and Technology, 55*(4), 2357–2368. https://doi.org/10.1021/acs.est.0c06783