Regionalization in a global hydrologic deep learning model: from physical descriptors to random vectors

Xiang Li[1], Ankush Khandelwal[2], Xiaowei Jia[3], Kelly Cutler[2], Rahul Ghosh[2], Arvind Renganathan[2], Shaoming Xu[2], John Nieber[1], Christopher Duffy[4], Michael Steinbach[2], Vipin Kumar[2]

[1]Department of Bioproducts and Biosystems Engineering, University of Minnesota Twin Cities, St.Paul, MN, USA
[2]Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA
[3]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA
[4]Department of Civil and Environmental Engineering, Pennsylvania State University, State College, PA, USA

Abstract

Streamflow prediction is a long-standing hydrologic problem. Development of models for streamflow prediction often requires incorporation of catchment physical descriptors to characterize the associated complex hydrological processes. Across different scales of catchments, these physical descriptors also allow models to extrapolate hydrologic information from one catchment to others, a process referred to as "regionalization". Recently, in gauged basin scenario, deep learning models have been shown to achieve state of the art regionalization performance by building a global hydrologic model. These models predict streamflow given catchment physical descriptors and weather forcing data. However, these physical descriptors are by their nature uncertain, sometimes incomplete, or even unavailable in certain cases, which limits the applicability of this approach. In this paper, we show that by assigning a vector of random values as a surrogate for catchment physical descriptors, we can achieve robust regionalization performance under gauged prediction scenario. Our results show that the deep learning model using our proposed random vector approach achieves a comparable and even marginally better predictive performance than the model using actual physical descriptors. The random vector approach yields robust performance under different data sparsity scenarios and deep learning model selections. Furthermore, our proposed random vector approach provides higher performance for regional modeling when physical descriptors are uncertain, or insufficient.

1 Introduction

In hydrology, streamflow prediction is essential for the forecast of water supply, floods, and droughts. It is a challenging task because of interacting hydrological processes (Beven, 1989, 1987; Freeze & Harlan, 1969; Freeze, 1974), spatial-varying parameter uncertainties (Keith Beven & Andrew Binley, 1992), and limited observations (Blöschl & Sivapalan, 1995). These challenges have motivated the advancement of hydrologic models from simple to complex. Encompassing more underlying hydrological processes, a complex hydrologic model includes more hydrologic parameters and detailed catchment physical descriptors to address the complexities (Beven, 2001, 2002) and associated scaling issues (McDonnell et al., 2007). But parameterizing such a complex hydrologic model for any individual catchment will become difficult when hydrologic data are unavailable. Thus, regionalization, which is defined as "how to extrapolate hydrologic information from one area to another?" (Blöschl & Sivapalan, 1995), specifies a research topic of modeling catchment runoff prediction using hydrologic information from multiple catchments, which will be given a brief background review in section 2.

Regionalization heavily relies on physical descriptors, such as, soil porosity, catchment elevation, etc. These physical descriptors account for hydrologic complexities and regional differences and are thus intensively used in regionalized hydrologic models, either process-based or data-driven.

Recently, Kratzert et al. (2019a) have presented a regionalized data-driven hydrologic model that greatly outperforms local process models. Specifically, they trained a single deep learning model (LSTM) for 531 basins in the US CAMELS (Catchment Attributes and Meteorology for Large Sample studies) dataset (Addor et al., 2017) and show that it is able to greatly outperform the well-established process-based models (e.g., SAC-SMA (Burnash, 1995), VIC (Liang et al., 1994), etc) that have been individually parameterized for each basin, and thus offer a better route to regionalization (Kratzert et al., 2019a).

Building such a model requires streamflow observation and weather forcings for many basins with diverse physical descriptors. It also relies upon the fact that all relevant basin physical descriptors are available and of high quality. Performance of such models may suffer if some of the descriptors are missing or are incorrect/uncertain. Our paper presents an approach where it is possible to build a data driven regionalized model even in the absence of any basin specific physical descriptors. It is able to use the weather forcing and streamflow data from a set of basins to build a global model without having any information about the physical descriptors of individual basins (For the background information of the global model, please see session 2). However the structure of this model is identical to the one used by Kratzert et al., as it only replaces the individual catchment physical descriptors by random vectors that simply provide a unique identity to each basin. Our results show that this approach provides at least as good global models as the ones produced using the knowledge of all available physical descriptors. But the performance is much better relative to the scenario where some of physical descriptors are missing and/or are incorrect/uncertain.

We note that the random vector and physical descriptor approaches are not in conflict and in fact give comparable results. In fact, for ungauged basins, Kratzert et al.'s model can be used (Kratzert et al., 2019b) and shows that physical descriptors serve as a bridge between gauged basins and ungauged basins. In our approach, the random vectors do not connect gauged basins and ungauged basins due to the lack of streamflow observation for the ungauged basins.

The paper is organized as follows. Section 2 introduces relevant background information, in particular the regionalization. Section 3 explains the details of the random vector method as well as the deep learning architecture involved. This section also explains the dataset and the set up of the experiment. The experiment includes an exhaustive analysis on the applicability of our proposed random vector methods under various data scarce situations and modeling structures. Section 4 lists our benchmarking results and the exhaustive analysis of the random vector applicability. Section 5 highlights scientific implications from our results and suggests a few future directions. Section 6 summarizes the scientific conclusions.

2 Background

Performing hydrologic prediction from multiple catchments, regionalization is closely related to the problem addressed in "prediction in ungauged basins" (PUB) (Sivapalan et al., 2003), and most literature uses"PUB" and "regionalization" interchangeably (Pagliero et al., 2019; de Lavenne et al., 2019; Choubin et al., 2019; Ecrepont et al., 2019; Zamoum & Souag-Gamane, 2019; Prieto et al., 2019; Guo et al., 2021; Alipour & Kibler, 2018). An underlying assumption behind regionalization is that similar basins have similar hydrologic behaviors. This implies that differences/similarities across catchments can be

classified into physical descriptors such as, climatology, geology, geomorphlogy, etc, with the assumption that incorporating these descriptors will improve streamflow prediction. In other words, hydrological behaviors as predicted from models for different catchments shall be based on similarities with regional information that is characterized by catchment physical descriptors. These approaches have been given a comprehensive review in particular for PUB (Guo et al., 2021; Samaniego et al., 2017; Beck et al., 2016) and can be grouped into model-dependent (process-driven) and model-independent (data-driven) methods, where 'model' denotes process-based models (Prieto et al., 2019).

Model-dependent methods give hydrologic predictions from process-based models. Information from the existing process-based hydrologic model is transferred to ungauged catchments based on certain criteria that link gauged to ungauged catchments. In practice, since those existing hydrologic models are calibrated to a specific catchment, this relies on some strategy of information transfer. A typical application of a model-dependent method implements a well-calibrated local hydrological process-based model and appropriate connections among catchments. In the review paper by (Guo et al., 2021), model-dependent methods can be classified into three categories: similarity based methods, regression based methods, and hydrological signature-based methods or some hybrid of each.

The model-independent approaches are data driven and do not rely on physical processes to simulate streamflow. Data driven methods learn how to predict streamflow from weather drivers and catchment physical descriptors directly without involving any hydrological process descriptions. Depending on either one or multiple catchments of data used, the data driven model will learn localized or regionalized hydrologic behaviors respectively. A local model is referred to as the model using hydrologic data from only one catchment. By contrast, when the hydrology data from multiple catchments are used and those catchments cover a wide range of all avialable hydrologic behaviors, the model is called a global model.

For data driven methods, one family is the neural network (Besaw et al., 2010; Hsu et al., 1995). Besaw built an artificial neural network on one catchment and transferred to another similar catchment without adaptation. It yielded unsatisfactory predictive performance (Besaw et al., 2010). In recent years, the Long Short-Term Memory (LSTM) networks (Hochreiter & Urgen Schmidhuber, 1997), one sub-family of neural networks, have shown burgeoning applicability in streamflow prediction tasks (Kratzert et al., 2018). LSTM based methods predict streamflow from antecedent weather drivers. In gauged scenarios, Kratzert et al. (2019a) have shown that using physical descriptors will train a universal global LSTM based model that outperforms process-based individual models given the same forcing data. One of the two versions of LSTM deveoped by Kratzert et al. provides additional physical interpretation, that is, basin similarities are preserved in the well trained machine learning (ML) model. Feng (Feng et al., 2020) embedded a global LSTM within a data integration framework (using predicted discharge from previous day) and found that it could marginally reduce prediction bias in regions with high flow autocorrelation. Frame showed that global LSTM outperforms the National Water Model (NWM) (Frame et al., 2020). In the poorly gauged scenarios, Ma (Ma et al., 2021) showed that fine tuning a global LSTM learned from data rich basins improved predictive performance in poorly gauged basins in contrast to local models learned solely from limited data.

It bears emphasis that regionalization approaches, either model-dependent and model-independent, rely heavily on physical descriptors. However, to obtain a satisfactory regionalization performance, physical descriptors need to be sufficient such that process complexities and associated scaling issues (Blöschl & Sivapalan, 1995) are encompassed. Otherwise, catchment scale prediction will be handicapped by the lack of sufficient information. For instance, modeling hydrological behaviors at the small scale can be accomplished by incorporating local processes with a few parameters. However, the incorporated processes and parameterization need to be adjusted, either made simpler or more

complex, to model hydrologic behaviors at a larger scale. The same adjustment also occurs when modeling hydrological behaviors between global scale and local scale, upstream and downstream. Accounting for these complexities and heterogeneities, sufficient physical descriptors must be involved. For example, Drost and Mudersbach found that merely incorporating landuse data with no additional physical descriptors provided little improvement to streamflow prediction and therefore may not benefit regionalization (Drost & Mudersbach, 2021). However, due to uniqueness of each catchment, such a complete characterization to resolve hydrologic complexity is difficult and challenging (Beven, 2020). This issue will be even more pronounced in applying models to data sparse regions where physical descriptors are limited, or even unavailable.

3 Methods

3.1 Long Short-Term Memory Network

Long short-term memory network (LSTM) (Hochreiter & Urgen Schmidhuber, 1997) is a special type of recurrent neural network designed especially for modeling time series predictions. Indeed, LSTM is the state-of-the-art deep learning model to predict streamflow (Kratzert et al., 2018, 2019a; Frame et al., 2020; Feng et al., 2020; Ma et al., 2021). In contrast to traditional recurrent neural network, LSTM avoids gradient vanishing or explosion (Bengio et al., 1994) and therefore preserves long term temporal dependencies for time series forecasting. This is achieved by using the gating architecture, which explicitly controls information flow and updates system hidden features. This memorizing mechanism and long term dependency allows LSTM to be well suited to model streamflow on a catchment scale. In particular, weather inputs feed and alter catchment response in various temporal scales. Although flooding season yields quick surface water response, the streamflow in winter periods under northern climate tends to have much longer response time because of involved snow and snowmelt processes. With the capability of the LSTM to account for long term dependency, it automatically learns these streamflow behaviors from data. Furthermore, it has been shown that some of the hidden features learned by the LSTM resemble snow processes (Kratzert et al., 2018).

An LSTM maps a sequence of time series input into the response variable. In this paper, we consider an LSTM based architecture that uses input features ($\mathbf{x}$) spanning T days to predict the observed discharge on the last day of the T-day window. The involved equations of an LSTM models are given below.

$$i\,[t] = \sigma(\mathbf{W_i} x\,[\mathbf{t}] + \mathbf{U_i} h\,[\mathbf{t-1}] + \mathbf{b_i}) \tag{1}$$
$$f[t] = \sigma(\mathbf{W_f} x\,[\mathbf{t}] + \mathbf{U_f} h\,[\mathbf{t-1}] + \mathbf{b_f}) \tag{2}$$
$$g\,[t] = tanh(\mathbf{W_g} x\,[\mathbf{t}] + \mathbf{U_g} h\,[\mathbf{t-1}] + \mathbf{b_g}) \tag{3}$$
$$o\,[t] = \sigma(\mathbf{W_o} x\,[\mathbf{t}] + \mathbf{U_o} h\,[\mathbf{t-1}] + \mathbf{b_o}) \tag{4}$$
$$c\,[t] = f[t] \odot c\,[t-1] + i\,[t] \odot g\,[t] \tag{5}$$
$$h\,[t] = o\,[t] \odot tanh(c\,[t]) \tag{6}$$

where $\sigma(\cdot)$ is sigmoid function, $tanh(\cdot)$ is the hyperbolic tangent function, and $\bigodot$ means element wise multiplication. $\mathbf{W}$, $\mathbf{U}$, $\mathbf{b}$ are model parameters, which will be learned during optimization. Other variables in equations represent basic computation units involved in the calculation. As gating variables, $i[t]$, $f[t]$, and $o[t]$ are input gate, forget gate, and output gate, respectively. They filter the information from the current and the previous time stamp, then combine them to update cell state $c[t]$. $c[t]$ underlines the intuition that motivates the LSTM design. $c[t]$ is maintained serially and embeds the temporal contextual information, which is characterized in $g[t]$, to then update the hidden representation $h[t]$. The stacked input x enters the LSTM sequentially and alters the information inherited from the previous timestamp. The previous information is stored in cell states $c[t]$ and hidden states $h[t]$, both of which characterizes the system memory. Cell

states $c[t]$ and hidden states $h[t]$ are initialized as zero vectors and then gradually modified until the final date in T-day time windows is reached. After a linear transformation, $x[t]$ infuses with previous hidden state $h[t-1]$ and then is non-linearly transformed in $i[t]$, $f[t]$, $g[t]$, and $o[t]$ via a corresponding activation function. The previous timestamp's cell state $c[t-1]$ is updated with $f(t)$ and then merges with an element-wise product of $i[t]$ and $g[t]$, which injects new information, to form a new cell state $c[t]$. After another hyperbolic tangent activation, this new cell state $c[t]$ merges with $o[t]$ and therefore updates the current hidden state $h[t]$. After the consecutive alteration of T time stamps, the final hidden state $h[T]$ is then transformed into the target variable, which in our case is streamflow.

In the context of regionalization based streamflow prediction, both dynamic weather variables and static catchment physical descriptors as formulated in equation 7:

$$Q_t = f(x^d, x^s) \tag{7}$$

where Q_t is streamflow, x^d is weather input vector, and x^s is a d-dimensional vector of physical descriptors. It bears emphasis that for a given catchment, x^s is assumed to be temporally static, while x_d is temporally dynamic. We assume catchment physical descriptors do not vary with the time. There are a number of ways in which physical descriptors can be incorporated in the LSTM architecture. In this paper, we consider three different models as illustrated in Figure 1. These models differ in terms of where x^s is added into the network. Specifically, in CT-LSTM physical descriptors are added before LSTM cell, whereas in EA-LSTM, they are used within the cell. Finally in FM-LSTM they are used in the last to modulate hidden states of the LSTM cell. Next, we describe these models in detail.

3.1.1 CT-LSTM

In CT-LSTM,at each timestamp, the dynamic weather input x^d is concatenated with the physical descriptors x^s to form the model input $x[t]$:

$$x[t] = \left[x^s, x^d[t]\right] \tag{8}$$

This model input enters the LSTM (equation 1 to 6), gets updated via the calculation of gates, and yields the final output - streamflow prediction. Through the calculation, physical descriptors are not placed within the LSTM cells or gates.

3.1.2 EA-LSTM

First proposed in (Kratzert et al., 2019a), EA-LSTM (Entity Aware LSTM) uses a modified version of LSTM where input gate takes physical descriptors as input instead of input features as previously shown in Equation 1. The key idea here is to explicitly empower the LSTM to customize its learning ability for catchment-wise adaptation.

$$i = \sigma(\mathbf{W_i} x^{\mathbf{s}} + \mathbf{b_i}) \tag{9}$$
$$f[t] = \sigma(\mathbf{W_f} x^{\mathbf{d}}[\mathbf{t}] + \mathbf{U_f} h[\mathbf{t-1}] + \mathbf{b_f}) \tag{10}$$
$$g[t] = tanh(\mathbf{W_g} x^{\mathbf{d}}[\mathbf{t}] + \mathbf{U_g} h[\mathbf{t-1}] + \mathbf{b_g}) \tag{11}$$
$$o[t] = \sigma(\mathbf{W_o} x^{\mathbf{d}}[\mathbf{t}] + \mathbf{U_o} h[\mathbf{t-1}] + \mathbf{b_o}) \tag{12}$$
$$c[t] = f[t] \odot c[t-1] + i[t] \odot g[t] \tag{13}$$
$$h[t] = o[t] \odot tanh(c[t]) \tag{14}$$

As illustrated in Figure 1b and also equations 9 to 14, x^s enters the LSTM via input gates, learns customized embedding (equation 9) for each basin, and updates the cell states recurrently at each timestamp. It therefore explicitly controls what modules in

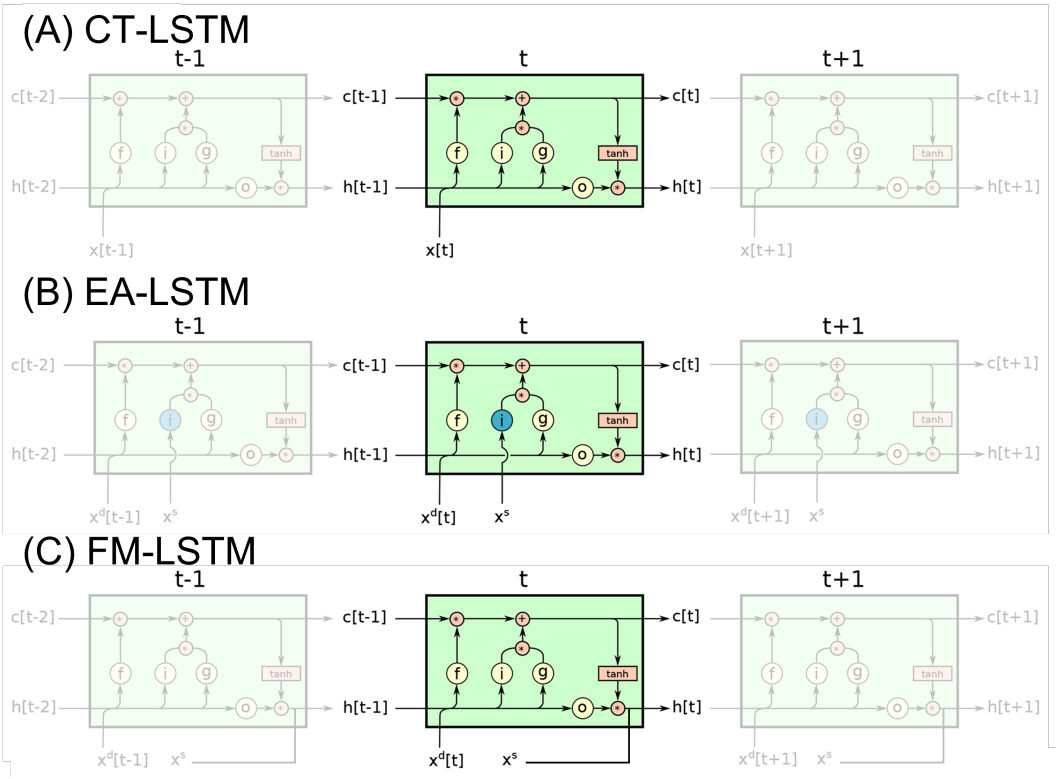


Figure 1: LSTM family illustration. Adapted from "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets" by Kratzert et al. (2019a), Hydrology and Earth System Sciences, 23, 5092 (Kratzert et al., 2019a)

LSTM respond to different catchments. This learned embedding will merge with other gates ($f[t]$, $g[t]$, $o[t]$), whose alteration are contributed by only dynamic weather inputs x^d. This separated role of x^s and x^d in EA-LSTM splits the contributions towards streamflow prediction from x^s in contrast to x^d. Additionally, the learned embedding affords an opportunity to examine cross-catchment response in a global model, which was shown to be close to the cross-catchment analysis using true basin characteristics (Kratzert et al., 2019a).

3.1.3 FM-LSTM

FM-LSTM uses the feature modulation concept which is becoming increasing popular in other areas such as meta-learning (add citations). The key idea here is to use a separate gate that takes static features as input and generates a modulation vector to modulate (adapt) the features learned by a traditional LSTM.

As illustrated in Figure 1C, x^s is mapped to an embedding layer customized for each basin (equation 20). This is then used to modulate the hidden states output (equation 21). x^s does not participate in the calculation in $i[t]$, $f[t]$, $g[t]$, $o[t]$, or $c[t]$.

$$i[t] = \sigma(\mathbf{W_i} x^{\mathbf{d}}[\mathbf{t}] + \mathbf{U_i} h[\mathbf{t-1}] + \mathbf{b_i}) \tag{15}$$
$$f[t] = \sigma(\mathbf{W_f} x^{\mathbf{d}}[\mathbf{t}] + \mathbf{U_f} h[\mathbf{t-1}] + \mathbf{b_f}) \tag{16}$$
$$g[t] = tanh(\mathbf{W_g} x^{\mathbf{d}}[\mathbf{t}] + \mathbf{U_g} h[\mathbf{t-1}] + \mathbf{b_g}) \tag{17}$$
$$o[t] = \sigma(\mathbf{W_o} x^{\mathbf{d}}[\mathbf{t}] + \mathbf{U_o} h[\mathbf{t-1}] + \mathbf{b_o}) \tag{18}$$
$$c[t] = f[t] \odot c[t-1] + i[t] \odot g[t] \tag{19}$$
$$p = \sigma(\mathbf{W_p} x^{\mathbf{s}} + \mathbf{b_p}) \tag{20}$$
$$h[t] = p \odot o[t] \odot tanh(c[t]) \tag{21}$$

3.1.4 General Requirement

Across these three models, we apply the same optimization algorithm (Adam optimizer (Kingma & Ba, 2017)) for training purposes to determine model parameters. Model parameters are learned from data and are thus continuously updated during training. The machine learning implementation also needs to specify hyper-parameters, which are set before training without learning from data. During training, hyper-parameters will not be updated. A few essential hyper-parameters include the look back period T and the dimension of hidden states $h[t]$. Adopting the previous work's specification (Kratzert et al., 2019a) of these hyper-parameters, we determine T to be 270 days and the dimension of hidden states to be 256. For the details on other hyper-parameters (e.g., learning rate, batch size), please read the Appendix B in Kratzert et al.'s paper (2019a).

The objective function is needed for training the deep learning model. To account for cross-catchment variance, which is not considered in the commonly used mean squared error option, we use a smooth-joint NSE function (Kratzert et al., 2019a). The smooth-joint NSE function is shown below.

$$NSE^* = \frac{1}{B} \sum_{b=1}^{B} \sum_{t=1}^{N} \frac{(Q_t^m - Q_t^o)^2}{(s(b) + \epsilon)^2} \tag{22}$$

where B is the number of catchments, N is the number of daily data (days) for one catchment, which is indexed by b. Q_t^m is the predicted discharge at timestamp $t (1 \leq t \leq N)$, while Q_t^m is the corresponding observed discharge. $s(b)$ is the standard deviation of the Q_t^o in basin b during training periods. ϵ is a constant term ($\epsilon = 0.1$) to avoid potential loss function explosion issue, which happens for catchments with extremely low $s(b)$.

3.2 Data

Our experiments use the continental hydrology dataset, CAMELS (Catchment Attributes and Meteorology for Large Sample studies) (Addor et al., 2017). The CAMELS data set contains continuous meteorologic input, observed streamflow data, and catchment dependent spatially varying but temporally physical descriptors. CAMELS encompasses a total of 671 watersheds across the contiguous US. Due to some watershed delineation errors (Addor et al., 2017), we followed the suggestion from Kratzert et al. (2019a) to select 531 basins whose watershed boundaries are confirmed to be correctly delineated without digital errors. Each watershed is supplied with observed discharge and climate forcing data from remote sensing products (Daymet, GLDAS, MAURER), climate models, and data assimilation with daily temporal resolution. Additionally, a corresponding hydrological model (SAC-SMA. Sacramento Soil Moisture Accounting model) is well calibrated for each watershed and its physical simulation is also available. Adopting such a wide distribution of watersheds, CAMELS provides a comprehensive and detailed physical description of watersheds. Selecting only a subset of those features as suggested by

Table 1: 27-d physical descriptors in CAMELS. The number in the brackets is the number of descriptors in the corresponding category.

Category	Physical descriptors
climate (9)	p_mean, pet_mean, aridity, p_seasonality, frac_snow_daily, high_prec_freq, high_prec_dur, low_prec_freq, low_prec_dur
Geomorphology(8)	elev_mean, slope_mean, area_gages2, forest_frac, lai_max, lai_diff, gvf_max, gvf_diff
Geology(10)	soil_depth_pelletier, soil_depth_statsgo, soil_porosity, soil_conductivity, max_water_content, sand_frac, silt_frac, clay_frac, carb_rocks_frac, geol_permeability

Kratzert et al. (2019a), we choose 27 physical descriptors from climatology, geomorphology and geology perspectives to characterize and discriminate across watersheds (Table 1). For details and physical meanings on those physical attributes, please see Table A1 in the Appendix.

These 27-d catchment physical descriptors are static vectors (x^s) characterizing each catchment. We selected meteorological data from an updated version of MAURER as model dynamic input (x^d), which are daily precipitation, daily minimum air temperature, daily maximum air temperature, average short-wave radiation, and vapor pressure. The observed discharge from USGS is our target variable (Q^O). Both daily meteorological weather inputs and discharge data cover a reasonably long record spanning from 1980 to 2014. The data for each catchment was partitioned into training and testing periods. Some experiments involved using a subset of training years or a subset of basins, therefore, we specify a standard test as training a global model using data from 531 basins with 20 years of data. Under this standard test, the training period starts from October 1st 1999 and ends on September 30th 2008. For a consistent evaluation, through all experiments, the testing period ranges between October 1st 1989 and September 30th 1999.

3.3 State of the Art

In terms of data-driven regionalization methods, CT-LSTM and EA-LSTM have been shown to perform very well for the streamflow prediction task (Kratzert et al., 2019a). In Figure 2, we show the testing NSE score for each catchment in the CAMELS dataset.

Table 2: State of the art LSTM based model. Mean and median refer to the summary statistics of the testing NSE scores across all 531 catchments in CAMELS.

Model	**Mean**	**Median**
Local LSTM	0.543	0.576
Global LSTM w/o static vectors	0.529	0.634
Global EA-lstm with 27-d descriptors	0.698	0.733

Local LSTM uses hydrologic data from only one catchment and does not need physical descriptors (x^s) to combine data from multiple catchments. Thus, for 531 catchments,

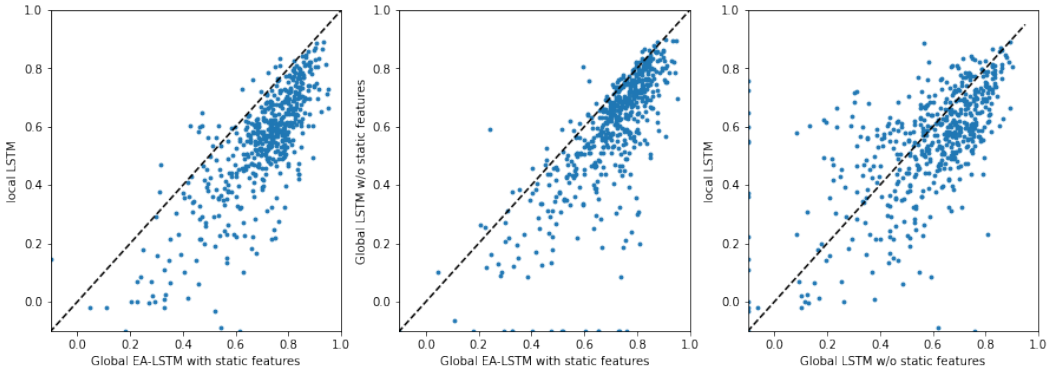


Figure 2: State of the art global regionalization performance using LSTM based deep learning architecture.

there are 531 Local LSTM models. On the other hand, global LSTM refers to a global model learned from the training data of 531 catchments. While the Global LSTM merges data from multiple catchments but does not use physical descriptors to adapt the network for different basins, the Global EA-LSTM with 27-d descriptors is also a global model trained and tested using all 531 catchments but it takes advantage of 27-d physical descriptors (Table 1) to perform robust regionalization. As shown in Table 2, both the mean and median of its NSE score is the highest (0.698 and 0.733 respectively) among the three model options. In this gauged prediction scenario, cross-catchment information sharing benefits global model performance. Further, inclusion of catchment physical descriptors indeed elevates predictive performance. These results have been previously shown by Kratzert et al. (2019a).

However, in practice, availability, uncertainty, and lack of completeness of those catchment physical descriptors might likely reduce the reliability of this global model performance. Concerns from these perspectives are our motivation and form this paper's objective as previously mentioned in section 1 and 2.

3.4 Proposed Approach

In this paper, our aim is to answer the question "How to perform regionalization when catchment physical descriptors are unavailable, uncertain, or of insufficient dimension?" To address this issue, we propose to assign a vector of random values as a surrogate for missing physical descriptors. Since a set of random vectors don't have any similarity structure (i.e. correlation between any two random vectors is zero), they are a suitable baseline to incorporate the fact we don't have any prior information on catchment similarity due to missing physical descriptors. By using these random vectors, we enable the deep learning network to account for heterogeneity in catchment responses while sharing data across multiple basins.

Furthermore, the proposed concept of using random vectors as a baseline can also be used to evaluate the efficacy of known catchment characteristics. In other words, the performance difference between using random vectors and actual characteristics can tell us about the quality of catchment characteristics. In section 4 (Results), we provide an extensive analysis of this concept in the context of streamflow prediction. In this paper, we consider two different strategies to create random vectors (Figure 3) as described below

(a) Gaussian vectors

(b) one-hot vectors

Figure 3: Random vector illustation

3.4.1 Gaussian Random Vectors

Figure 3a is a visual representation of the Gaussian vector (d-dimension) for all catchments. Random colors represent random numbers drawn from Gaussian distribution. In this strategy we assign d-dimensional vectors to each catchment where the vector values are drawn from a Gaussian distribution with zero mean and unit standard deviation. In other words, we randomly map each basin to a point in d-dimensional feature space.

3.4.2 One-hot Vectors

Figure 3b illustrates the one-hot vector representation. Each catchment is associated with a binary vector that is 1 for one dimension and is zero elsewhere. The dimension of the one-hot vectors equals the number of catchments. These one-hot vectors originated from the binary vectors used to encode categorical variables in regression, where in our case, the variable is catchment ID. There is one such one-hot binary vector for each basin and these vectors are orthogonal to each other. It bears emphasis that there's no freedom for the user to determine the dimension of the one-hot vector after the number of catchments in a global model is known. For k basins, the length of the one-hot vector for each basin is k. Although the one-hot vector does not involve random numbers, the randomness in this random vector assignment is from basin order. Regardless of how basins are sorted, one-hot vector assignment assures each basin will be assigned uniquely.

3.5 Experiments

To evaluate the applicability of our proposed random vectors method in regionalization, we first compared the predictive performance of a global model using random vectors (Gaussian or one-hot) against that using physically meaningful 27-d descriptors. In the CAMELS database, a global EA-LSTM was trained on 531 basins using 27-d physical descriptors (Kratzert et al., 2019a) to show the state-of-the-art predictive performance. Our proposed random vectors substitute the 27-d physical descriptors with random vectors, another global EA-LSTM was trained for comparison.

Machine learning models have uncertainties in model parameters after training. Initialized randomly, model parameters will often be optimized to different values during training. In simplistic terms, different model initializations will yield different models after training. Accounting for uncertainty, it has been shown that ensemble results from multiple model runs will facilitate the overall model performance (Kratzert et al., 2019a). Therefore, the final streamflow prediction in any experiment setting is from an ensemble mean of five model realizations. For instance, the prediction of the EA-LSTM using physical descriptors is an average of 5 model predictions, which are optimized from different initializations. Note that for the Gaussian vector experiment, the randomness originates from 2 sources, including model initializations and the Gaussian vector assignment. For each of the 5 runs, their Gaussian vectors are assigned with different values.

We then conduct an exhaustive analysis to investigate its applicability under other model settings, and different data richness scenarios (i.e., from short records to longer records, from fewer basins to many basins) with an intention to assess the generalizability of the random vector approach. We also explored practical implications of random vectors for understanding catchment modeling complexities for insufficient and uncertain catchment characteristics. Therefore, we compare the predictive performance between models using random vectors and physical descriptors under the scenarios described in sections 3.5.1 to 3.5.5.

For consistent model comparison, we're using the NSE score instead of RMSE (root mean squared error) to evaluate streamflow prediction. NSE is a metric suited particularly to evaluate hydrological predictions.

$$NSE = 1 - \frac{\Sigma_{t=1}^{T}(Q_t^m - Q_t^o)^2}{\Sigma_{t=1}^{T}(Q_t^o - \bar{Q}^o)^2} \tag{23}$$

Q^m is predicted discharge, Q^o is observed discharge, $\bar{Q}^o$ is the mean of observed discharge. A NSE score of 1 indicates a perfect time series prediction. Next, we describe different experiment settings used to analyze the random vector approach.

3.5.1 Number of Basins

An LSTM model is trained for a group of only k basins (k $<$ 531), which forms an insufficient global hydrologic dataset having relatively fewer basins. This selection of k basins aims to answer the question "Given only k basins without physical descriptors, will the proposed random vector strategy be applicable for regionalization?". We vary k from 10 to 50 to 100 and use the results from the aforementioned (Section 3.2) standard training and testing data periods as comparison.

To generate the data sets, we randomly select k basins as a group repetitively without replacement until all basins are selected. When the remaining basins cannot form a group with exactly the size k, those basins are either merged with the last group or form a stand-alone group as long as its order of magnitude approximates to k. For instance, when selecting 1a 0-basin group, we select 53 groups in total, and the last group contains 11 basins. Similarly, the last group (11th group) in the 50-basin group has 31 basins. The last group (fifth group) in 100-basin group has 131 basins.

3.5.2 Number of Training Years

Another perspective on data inadequacy is the number of training years. An LSTM model is trained for all 531 basins with a limited number of years. Varying the training years from 1 to 2 to 5 years, we sought the answer to this question "Given only a few years of training data, will the proposed random vector strategy be applicable for regionalization? " The LSTM using random vectors is tested against the 27-d physical vectors under three sparse data cases, which are 1 year of data (October 1st 2007 to September 30th 2008), 2 years of data (October 1st 2006 to September 30th 2008) and 5 years of data (October 1st 2003 to September 30th 2008). Model testing performance is evaluated during the same years (October 1st 1989 to September 1st 1999) as in the standard case for a consistent model comparison.

3.5.3 Alternative LSTM Architectures

Catchment physical descriptors can be integrated into LSTM in different ways, yielding different LSTM architectures. We compare the predictive performance between random vectors and 27-d physical descriptors under various model architectures. As mentioned in the section 3.1, we have three different LSTM based models: CT-LSTM, FM-LSTM, and EA-LSTM. Random vectors implementations are tested for all these three

models with an intention to answer this question"Under different model architectures, will the proposed random vector strategies be applicable for regionalization?"

3.5.4 Incomplete Characterization of Physical System

The aim of this experiment is to answer the question "Is the dimension of catchment physical descriptors sufficient for regionalization?" If catchments within the system are under-represented by the physical descriptors, how will the proposed random vector strategy benefit the model regionalization in this information deficient physical system? To test this deficiency, we define a physically underrepresented global system in CAMELS where only a subset of 27-d physical descriptors is used to distinguish basins. We compare the global LSTM using random vectors in contrast to the global model using these insufficiently informative descriptors. One extreme case is a system without any static catchment descriptors, which has been shown in the section 3.3 (Figure 2). Ignoring the model selection differences, we select EA-LSTM for this experiment because it explicitly modulates LSTM via static vectors. The global EA-LSTM using some subset of 27-d physical descriptors is trained and compared. EA-LSTM using 9-d climate features, 10-d geology features, and 8-d geomorphology features are trained separately and compared to the EA-LSTM using random vectors.

3.5.5 Uncertainties in Basin Characteristics

This experiment is designed to answer the question "Can random vectors represent uncertainties in the 27-d physical descriptors for catchments?" It was found that topological features and climate features are sensitive to added noise (Kratzert et al., 2019a). These sensitivities also indicate that they're uncertain because by using average values to represent a catchment the features simplify the spatial heterogeneity of watershed systems. By limiting the dimension of static vectors to 27-d, this simplification introduces uncertainty. We explore using random vectors to augment the 27-d physical descriptors for improving model performance and addressing feature uncertainty issues. These augmented static vectors are "mixed Gaussian vectors". Uncertainties are gradually represented by increasing the random vector dimension. To allow addressing uncertainties from each physical descriptor, we define a global system with 64-d, 128-d, 256-d, and 512-d vectors, all of which include the 27-d physical descriptors. For instance, for the 64-d features, besides the 27-d physical descriptors, 37-d (64-27=37) vectors are randomly drawn from the Gaussian distribution. At least one uncertainty dimension is reserved for each physical descriptor. For this experiment, the EA-LSTM was used solely as the model architecture because the CT-LSTM requires a larger number of parameters and thus increases computation burden and complicates model learning.

4 Results

The experiment section has outlined different model implementations. Each model implementation needs to specify the model architecture (either EA-LSTM, CT-LSTM, or FM-LSTM) and static vectors (x^s). Options for x^s include 27-d physical descriptors, random vectors, and mixing Gaussian vectors. For the simplicity of representing the results, we'll use acronyms to denote corresponding results of those experiments, that is, the combination of model architecture and x^s. These acronyms are shown in the table 3. Models for the incomplete physical systems are not given acronyms. This section is organized as follows. The first section (4.1) shows the comparison between EA-LSTM using random vectors (EG-d and EO) and EP. The following sections follow the experiment order listed in section 3.5. Section 4.2 presents the analysis on the impact of number of basins to the EA-LSTM using random vectors. Section 4.3 summarizes the number of years impact as another data inadequacy scenario. Section 4.4 presents the results of implementing random vectors for the model architectures outlined in Figure 1. This

Table 3: This acronym table denotes the acronyms of model implementations. Combinations of model architecture and x^s specifications are shown in their acronyms. The "d" in these notations represent the dimension of x^s, which is only needed to specify the models using Gaussian vectors. For instance, EG-512 means EA-LSTM model using 512-d Gaussian vectors. '*' means the corresponding models were not implemented.

x^s		EA-LSTM	CT-LSTM	FM-LSTM
27-d physical descriptors		EP	CP	FP
Random vectors	Gaussian d-dimension	EG-d	CG-d	FG-d
	One-hot	EO	CO	FO
Mixed Gaussian d-dimension vectors		EM-d	*	*

section also presents the analysis of data inadequacy impacts on CT-LSTM performance and comparison of random vectors across EA-LSTM and CT-LSTM. Section 4.5 presents the analysis of the regionalization performance improvement of random vectors in contrast to physically under-represented catchment systems, which attains the understanding towards the applicability of random vectors. Section 4.6 presents the applicability of random vectors to address physical descriptors uncertainties.

4.1 Effectiveness of Random vectors

We select EA-LSTM as a baseline model architecture. Note that the implementation of Gaussian vectors requires a specification of d, which is determined empirically. The cumulative density function plot of the NSE score, shown in Figure 4 suggests using 512 (black solid line) as the Gaussian vector dimension because its testing performance is optimal compared to others.

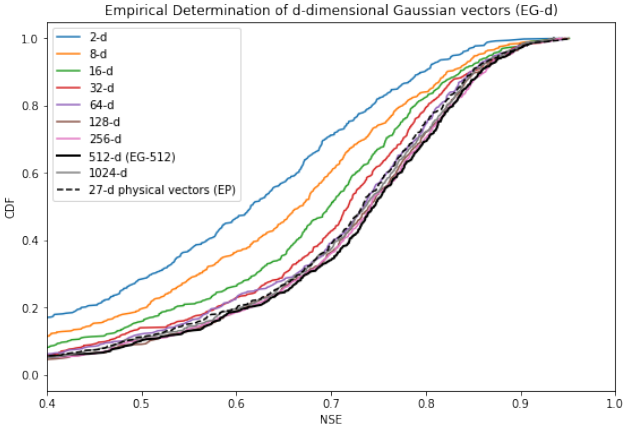
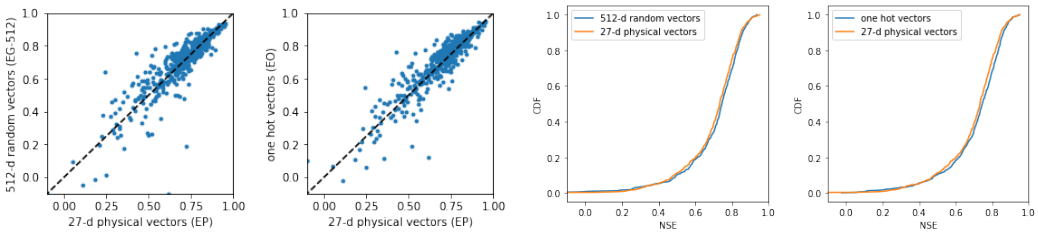


Figure 4: Cumulative density functions of the NSE score across different d Gaussian vectors for the EG-d. The X-axis is NSE score, which is truncated between 0.4 and 1 for a better illustration. The black dashed line represents the testing score corresponding to the EP. The black solid line corresponds to the EG-512, which yields the best performance in the EG-d.



(a) EG-512 versus EP (b) EO versus EP (c) cdf of EG-512 versus EP (d) cdf of EO versus EP

Figure 5: Performance comparison cross the EP, EG-512 and EO. Model architecture is EA-LSTM. Scatter plots are shown in a and b. Respectively, their cumulative density functions of the NSE are shown in c and d.

The scatter plot (Figure 5) shows the testing NSE of the EG-512 and the EO versus the EP respectively across all 531 basins. Among these results, testing NSE scores less than -0.1 are forced to be -0.1 for illustration purposes. For each scatter plot, a cumulative density function (cdf) plot of NSE is also given. The EG-512 scatter plot is slightly upper skewed, the cdf of the EG-512 is also slightly right skewed compared to the EP. Figure 5a and Figure 5c shows that EG-512 prediction performance is comparable and even slightly better than the EP. In Table 4, the mean and median of the EG-512 is 0.711 and 0.746, both of which yields more satisfactory results than EP. The same comparison between the EO and the EP also yields quite similar trend. The mean and median of NSE score for the EO is 0.707 and 0.745. EO reaches comparable and slightly better prediction performance than the EP.

Table 4: Performance comparison of the EA-LSTM using random vectors against physical descriptors. Statistical summaries across all 531 basins are in column 'mean' and 'median'

Catchment static vectors	Mean	Median
27-d physical vectors (EP)	0.698	0.733
512-d random vectors (EG-512)	0.711	0.746
one-hot vectors (EO)	0.707	0.745

As we can see, using random vectors we get performance comparable to using known physical descriptors. Furthermore, random vector approach leads to significantly better results when compared to other strategies that do not use known physical descriptors (i.e. Figure 2, building local models or trivial merging of data from multiple basins). Hence, the random vector approach is a viable solution when catchment characteristics are not available. This performance is evaluated using the standard setting (section 3.2, 10 years training data from 531 basins). Although such abundant training data shows slightly elevated testing performance, the proposed random vector method might still be inapplicable in data poor situations. To assess the impact of data sparsity, we conducted an exhaustive analysis on the different data inadequacy scenarios outlined in section 3.5.1 and 3.5.2.

4.2 Effect of number of basins

For this situation we're creating a data poor scenario where the training data consists of a limited set of basins. For the 53 groups of 10-basin groups, we compare the predictive performance using random vectors relative to the performance of the model using 27-d physical descriptor case given in the CAMELS dataset. This comparison is illustrated in Figure 6. The X-axis denotes one-hot vector idea and Gaussian vector (varying d). Each category shows a box plot of performance comparison across basins. Median (blue dots), 25th percentile and 75 percentiles (upper and lower box line) are shown for each box. Black hollow circles outside the upper and lower box lines are outliers outside the specified quantile range. The Y-axis is the NSE score improvement compared to the 27-d physical descriptors. The red line indicates the threshold for improved performance. A box plot whose NSE distribution is skewed to positive NSE score improvement indicates a general performance improvement in that random vector category. The 512-d Gaussian vectors show a visible performance improvement in contrast to the case of the 27-d descriptors. The one-hot vector is less productive by comparison.

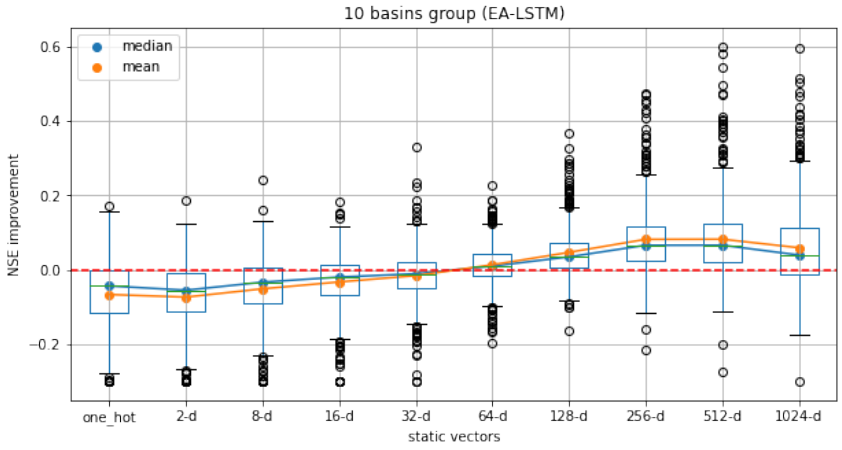


Figure 6: Random vectors implementation for 10-basin group EA-LSTM. Categories along the X-axis represent random vectors, including one-hot vectors (length of 10) and Gaussian vectors (dimension d varies from 2 to 1024). The Y-axis shows the NSE score improvement of the random vector in contrast to its corresponding EP, which is trained using the same basins. A zero NSE improvement indicates an improvement threshold marked by the red line. Within each category, 531 NSE improvement scores are distributed in the box plot where outliers exceeding 25th and 75th quantile are marked by black hollow circles.

For the 50-basin group and 100-basin group, the plot of NSE improvement is shown in Figure 7 except that we plot only the median of each case to provide succinct visualization. The red line also marks the performance improvement threshold. Table 5 summarizes the NSE score improvement for all cases. It shows a consistent performance improvement comparison. Regardless of how limited the number of basins, the Gaussian vector strategy (with an optimal dimension of either 256 or 512) outperforms the 27-d physical vectors. In particular, the performance improvement from the Gaussian vectors becomes saturated when d reaches 256 or 512. When the dimension of the Gaussian vector becomes a higher 1024-d, the performance improvement begins to degrade. In summary, we show that random vector approach shows robust performance even with fewer number of catchments in the dataset and hence can be used in situations where only few catchments are available.

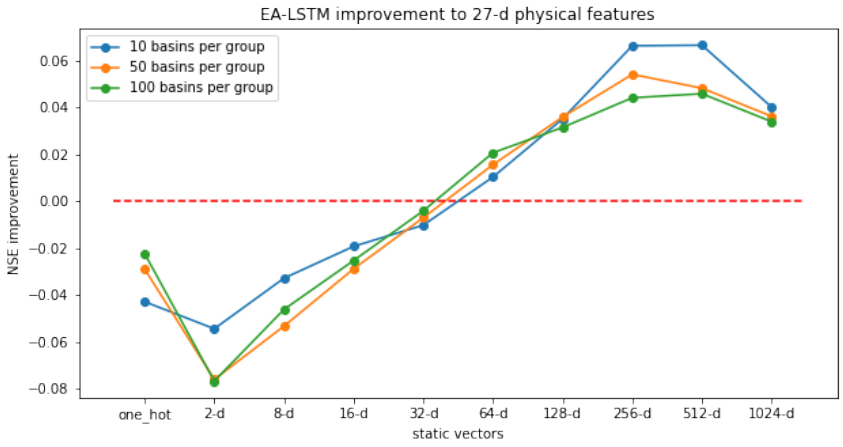


Figure 7: Random vectors implementation for k-basin group EA-LSTM. k varies from 10 to 50 to 100. The median in Figure 6 are blue lines. Within each random vector category as shown in X-axis, the median of NSE improvement score in contrast to EP for k basins is plotted. Orange dots are the 50-basin group while green color represents 100-basin group

Table 5: The performance improvement of EG-d and EO to EP

k-basin group		10	50	100	10	50	100
	d	mean			median		
Gaussian vector (EG-d)	2	-0.073	-0.085	-0.09	-0.054	-0.076	-0.077
	8	-0.051	-0.061	-0.058	-0.033	-0.053	-0.046
	16	-0.032	-0.031	-0.03	-0.019	-0.029	-0.025
	32	-0.015	-0.01	-0.008	-0.01	-0.007	-0.004
	64	0.014	0.018	0.025	0.01	0.016	0.021
	128	0.047	0.045	0.039	0.035	0.036	0.031
	256	**0.082**	**0.062**	**0.053**	**0.066**	**0.054**	**0.044**
	512	**0.082**	**0.055**	**0.053**	**0.066**	**0.048**	**0.046**
	1024	0.06	0.038	0.039	0.04	0.036	0.034
one-hot (EO)		-0.066	-0.035	-0.03	-0.043	-0.029	-0.022

4.3 Effect of number of training years

Here we limit the standard 10-year training data into periods of 1 year, 2 years and 5 years. Our previous empirical analysis indicates an optimal specification of d (Gaussian vector dimension) to be 512, so the implementation of basin random vectors includes either 512-d Gaussian vectors or one-hot vectors. In Figure 8, the dots and box portions above the red line (no NSE score difference) indicate that our proposed random vector strategy is more satisfactory. It shows that both strategies lead to prediction performance similar to the case utilizing 27-d physical descriptors. In particular, the EG-512 yields a more satisfactory performance than the EO. As shown in Table 6, a NSE score improvement (both in mean and median) is observed when implementing 512-d Gaussian vectors, while the NSE score improvement is only observed when using 5 years of training data when the one-hot vector strategy is applied. The results show that even when training data are limited, randomly assigned vectors are still able to learn as well as 27-d physical features.

Table 6: The impact of the number of training years on the performance improvement of random vectors for EA-LSTM. "Mean" and "Median" refer to statistics of the NSE score improvement in relative to EP. Positive numbers mean that random vectors yield better predictive performance.

Number of training years		1	2	5
Gaussian 512-d (EG-512)	mean	0.013	0.052	0.026
	median	0.009	0.041	0.019
one-hot (EO)	mean	-0.025	-0.003	0.015
	median	-0.023	-0.005	0.013

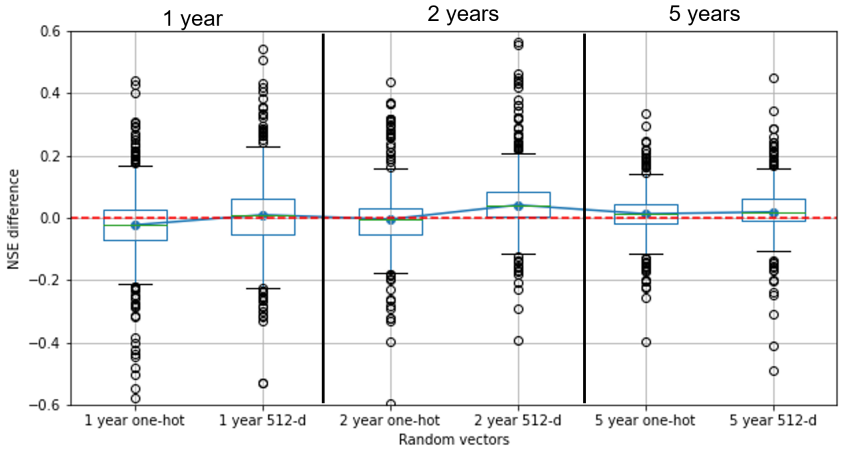


Figure 8: The impact of the number of training years on EA-LSTM. The Y-axis represents the NSE score difference between the corresponding category in X-axis and the predictive performance using 27-d physical descriptors (EP). The red line indicates performance improvement threshold.

4.4 Performance of alternative models

In this section, we address the performance difference between random vectors and 27-d physical descriptors for the CT-LSTM and FM-LSTM architectures (Figure 1). We compared the regionalization performance of random vectors across the EA-LSTM and the CT-LSTM to answer the question "Which random vector strategy is better suited for regionalization, Gaussian vectors or one-hot vectors?" We selected the CT-LSTM as another model architecture for an exhaustive analysis on the data inadequacy cases in terms of basin numbers.

From previous sections (section 4.1, 4.2, and 4.3), we've shown the efficacy of the random vectors in EA-LSTM in both data rich and data poor scenarios. Will that efficacy also be shown in other alternative models? We thus implemented random vectors in both CT-LSTM and FM-LSTM.

For the CT-LSTM, the Gaussian vector implementation needs to specify the optimal vector dimension d. Figure 9 shows that the CG-16 yields the most satisfactory performance among different Gaussian vector dimension options. Therefore, we empirically select 16 as the optimal Gaussian dimension to represent the CG-d performance (Figure 10c). Note that the optimal 16-d of the CG-d is less than the optimal 512-d of the EG-d. We'll explain this in the section 5.1 in "Discussion" section. Using 27-d physical descriptors, CP achieves performance comparable and slightly better than EP (Fig-

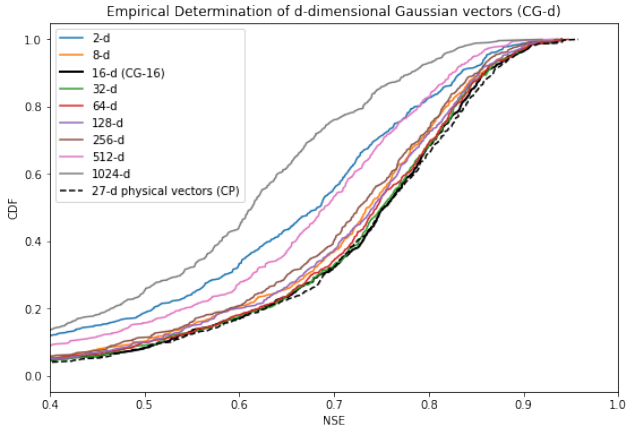
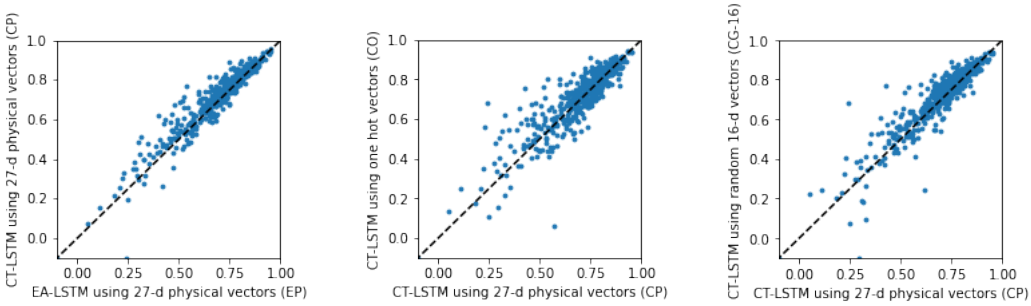


Figure 9: Cumulative density function plots of the NSE score across different d Gaussian vectors for the CT-LSTM. The X-axis is truncated between 0.4 and 1 for a better illustration. The black dashed line represents the testing score of the CP, the black solid line corresponds to the optimal 16-d performance among the Gauusian vector groups (CG-16).

Table 7: Random vector comparison cross different models

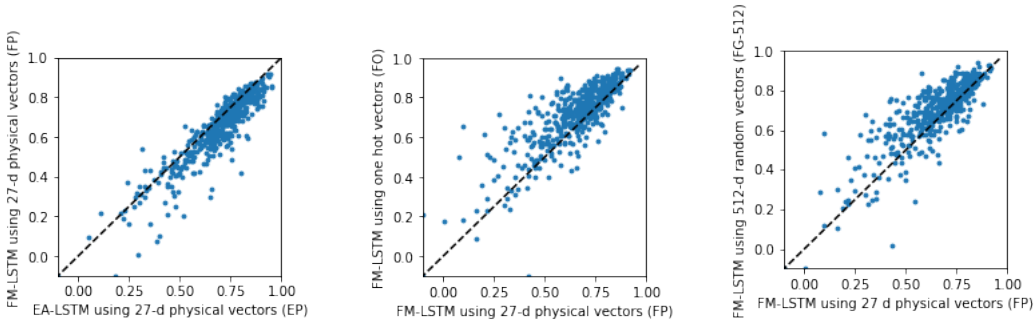
models	Mean	Median
EP	0.698	0.733
CP	0.715	0.744
CO	0.720	0.754
CG-16	0.717	0.752
FP	0.653	0.698
FO	0.716	0.746
FG-512	0.695	0.738



(a) CP versus EP (b) CO versus CP (c) CG-16 versus CP

Figure 10: Predicted performance comparison of a random vector implementation in CT-LSTM (CO and CG-16) in contrast to CT-LSTM using 27-d physical descriptors (CP)

ure 10a and Table 7). The median NSE score performance improves from 0.733 (CP) to 0.744 (CO). Random vector options (CO and CG-16) slightly outperform 27-d physical descriptors (CP). The median of testing NSE performance improves from 0.744 to 0.754 when using the one-hot vector strategy, while the CG-16 elevates the performance to 0.752.



(a) FP versus EP (b) FO versus FP (c) FG-512 versus FP

Figure 11: Predicted performance comparison of a random vector implementation in the FM-LSTM (FO and FG-512) in contrast to the FM-LSTM using 27-d physical descriptors (FP)

For the FM-LSTM, we specify the optimal Gaussian vector dimension as the same of the EA-LSTM because they share the similar model modulation strategy, that is, static vectors enter the LSTM separately from the dynamic weather inputs. Using 27-d physical descriptors, Figure 11a illustrates that the FP yields worse prediction performance compared to the EP. Even so, the FM-LSTM also attains benefits performance improvement from random vectors. Both one-hot vector and Gaussian 512-d vectors lead to significantly better predictive performance. In terms of the median, in contrast to the FP, FO elevates the performance from 0.698 to 0.746, while FG-512 improves the performance to 0.738. The one-hot vector benefits are more pronounced than those of 512-d Gaussian vectors in FM-LSTM.

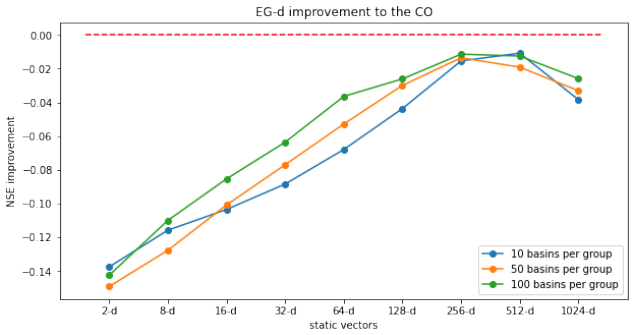


Figure 12: The performance of EG-d in contrast to the CO for k-basin group. k varies from 10 to 50 to 100. The Y-axis is the NSE difference score quantifying the performance of the EG-d (categories in X-axis) relative to the CO. The red line marks no NSE difference. Plotted points are median of the NSE difference across all basins. For points below the red line, they mean that the CO yields more satisfactory performance

Accordingly, under different modulations discussed so far (EA-LSTM, CT-LSTM, and FM-LSTM), the above experiments demonstrate that random vector strategies still prevail over 27-d physical vectors. The best random vector method for EA-LSTM and FM-LSTM is 512-d Gaussian vectors, while the best random vector strategy for CT-LSTM is one-hot vector. The preferable random vector strategy varies depending on modulation strategy. In a pursuit of model performance when utilizing random vectors, we need

to provide a practical solution to the question "When implementing random vectors to perform regionalization, shall I use Gaussian vectors or one-hot vectors?". To answer this question, we next compared the optimal random vector performance between the CT-LSTM and EA-LSTM. FM-LSTM is not considered because its performance is wore than EA-LSTM. Figure 12 shows the testing NSE difference from the various EG-d against the CO. Based on the previous result showing that the EO is not as good as EG-d, 'one-hot' on X-axis (EA-LSTM random vector strategy) is omitted. Figure 12 shows the median of the NSE difference for various selections of k basins. All points are below the performance threshold line, indicating that the CO slightly outperforms EG-d. When implementing the random vector strategy as a surrogate for missing physical descriptors, the best performance is obtained when applying CO.

Data abundance has always been an important factor impacting the machine learning model performance. To consolidate the argument that CT-LSTM with random vectors, especially one-hot vectors, yields better performance consistently under various data richness scenarios, we repeated the experiments outlined in section 3.5.1 for CT-LSTM. Training data are limited by the number of basins.

Table 8: The random vectors' improvement over 27-d physical features in the CT-LSTM. "Mean" and "Median" refer to statistics of NSE score improvement in relative to the CP. The most satisfactory performance is in bold font: 32-d Gaussian vector, 64-d Gaussian vector, and one-hot vector

k-basin group		10	50	100	531	10	50	100	531
	d	mean				median			
Gaussian vector (CG-d)	2	-0.079	-0.073	-0.074	-0.119	-0.075	-0.063	-0.063	-0.061
	8	-0.055	-0.031	-0.024	-0.026	-0.050	-0.028	-0.023	-0.016
	16	-0.037	-0.015	-0.007	**-0.004**	-0.037	-0.018	-0.010	**-0.004**
	32	**-0.022**	**-0.006**	**0.004**	**-0.009**	**-0.023**	**-0.008**	**0.001**	**-0.006**
	64	**-0.023**	**-0.004**	**0.002**	**-0.010**	**-0.021**	**-0.006**	**0.000**	**-0.008**
	128	-0.041	-0.019	-0.019	-0.005	-0.039	-0.016	-0.003	-0.015
	256	-0.074	-0.059	-0.033	-0.032	-0.072	-0.048	-0.026	-0.025
	512	-0.103	-0.125	-0.094	-0.074	-0.097	-0.114	-0.083	-0.057
	1024	-0.134	-0.188	-0.174	-0.143	-0.129	-0.191	-0.171	-0.124
one-hot (CO)		**-0.046**	**-0.007**	**0.005**	**-0.005**	**-0.042**	**-0.005**	**0.001**	**-0.004**

Figure 13 exhibits a box plot showing the NSE improvement for the 10-basin group using the CT-LSTM architecture. Any point above the red line (NSE score improvement threshold) indicates a performance improvement in contrast to 27-d physical descriptors. In the 10-basin group category, the optimal Gaussian d for the CG-d is lower than that of the EG-d. The optimal Gaussian vectors performance is comparable to that of one-hot vectors. To obtain a general insight, we varied k from 10 to 50 to 100 and therefore produced the following result in Figure 14 and Table 8.

Figure 14 shows the median of the NSE improvement using random vectors in CT-LSTM in contrast to the CP. Dots below the red line mean the prediction performance of the corresponding categories is worse than the CP. As the number of catchments available fro training increases, the one-hot vector strategy and optimal Gaussian vectors in CT-LSTM yields performance comparable to the CP. The optimal d for the CG-d is either 32 or 64, which is lower than the optimal 512-d in the EG-d. As also recognized in Figure 10, this discrepancy of optimal Gaussian d between the CT-LSTM and EA-LSTM can be explained by the number of parameters involved in these model architectures and we'll expand this discussion in section 5.1. We point out that these random vector strate-

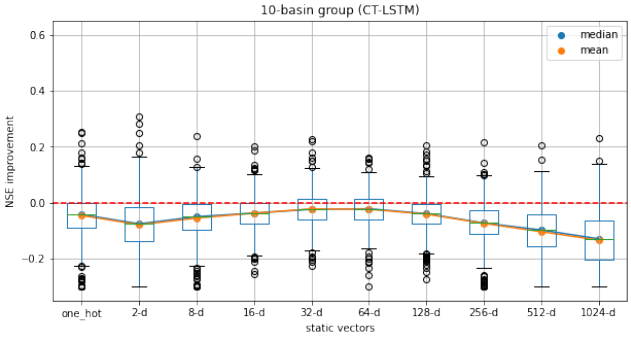


Figure 13: Impacts of random vectors on CT-LSTM for a 10-basin group. Categories on the X-axis represent random vectors, including one-hot vectors (length of 10) and Gaussian vectors (dimension d varies from 2 to 1024). The Y-axis show the NSE score improvement of the random vectors in contrast to the CP. A zero NSE improvement indicates no performance improvement marked by the red line.

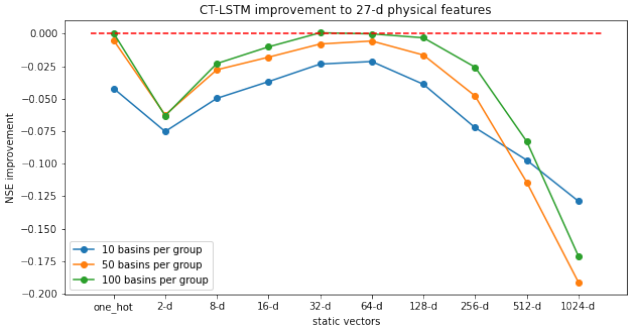


Figure 14: A random vector implementation for a k-basin group CT-LSTM. k varies from 10 to 50 to 100. Within each random vector category as shown in X-axis, the median of NSE improvement score in contrast to the CP for k basins is plotted. Blue dots are 10-basin group, orange dots are 50-basin group, while green color represents 100-basin group

gies are approximate to but do not marginally exceed the CP performance. In particular, the relative significant performance improvement occurs when using the one-hot vector in the 100-basin group but to a much lesser extent. Varying k from 10 to 50 to 100, as more catchments are involved until 531 basins are included, the one-hot vector is a preferable random vector strategy for CT-LSTM than Gaussian vectors.

4.5 Incompleteness of physical characteristics

So far we considered the scenario where physical descriptors are not available and assessed the performance of random vector approach. In this section, we consider a more common regionalization challenge where physical descriptors are incomplete. Those incomplete physical descriptors can only help regionalization in a limited degree. We might ask ourselves: "Are 27-d physical descriptors sufficient? What about 100 or 500-d descriptors?"

Catchment hydrologic models are formulated to resolve complexity and associated scaling issues in hydrological processes. Both issues will not be dealt without a comprehensive physical understanding. From a practical perspective, static physical descriptors (for instance, Table 1) can only characterize complex catchments to a limited dimension

because a sufficient catchment complexity characterization is challenging across scales. In the field scale, a hydrological model might characterize local hydrological processes completely, but the applicability of this locally built model to a larger basin might fail if the model is not adjusted, either simplified (reduce the number of parameters) or made complex (enrich physical parameters). Therefore, for the regionalization involving catchments at various scales, the question becomes "Are any given physical descriptors sufficient for modeling the complexity of catchments?" This question also implies another question "how many physical dimensions do we need for characterizing the complexities of streamflow generation processes" To answer these questions, we compared our ran-

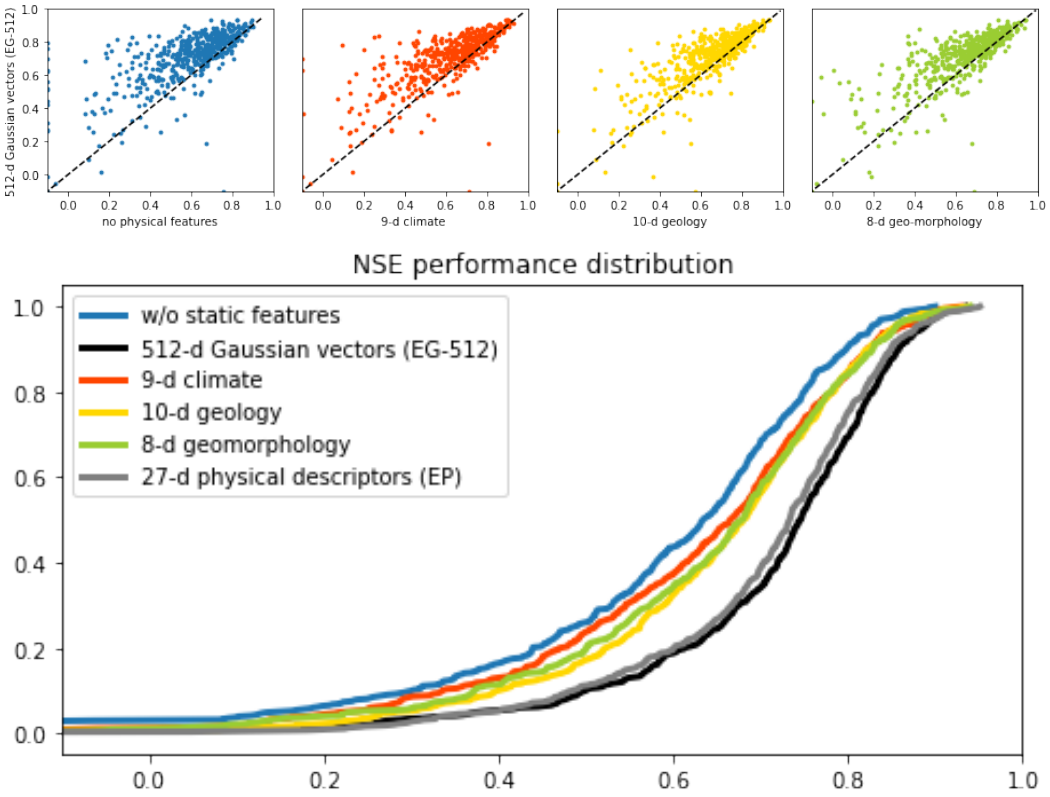


Figure 15: The performance of the EG-512 in contrast to EA-LSTM under a physically incomplete catchment system. Scatter plots show testing NSE scores comparison between the Y-axis and the X-axis. The y label is EG-512 and is fixed in the above 4 figures. The X-axis changes from a no physical feature system, a climate feature system, a geology feature system, to a geo-morphology feature system. The below cumulative density function plot collects NSE scores together for each category aforementioned. The black line is EG-512. We also plotted the benchmark performance with sufficient 27-d physical descriptors (grey solid line, EP) to remind the reader of the performance under a physically sufficient catchment system.

dom vector results to models utilizing incomplete sets of physical vectors. In Table 1, 27-d physical descriptors are categorized into three groups: climate, geology and geo-morphology. Among these, the descriptors of any single group are an under-representative description for basins. For instance, 9-d climate descriptors presumably characterize basins less informatively than 27-d physical descriptors. For this experiment, we choose the EA-LSTM as the model structure and use 512-d Gaussian vector as its optimal random vector strategy. Each one of the three descriptor subset groups leads to an EA-LSTM under a physically uninformative system since complexities are simplified and the system incurs in-

Table 9: Physical system completeness identification

catchment static vectors	Mean	Median
(0-d) Without static features	0.529	0.634
512-d Gaussian vectors (EG-512)	**0.707**	**0.745**
9-d climate features	0.611	0.665
10-d geology features	0.638	0.679
8-d geomorphology features	0.630	0.680
27-d physical descriptors (EP)	0.698	0.733

formation loss. For the extreme case where there are no physical descriptors present, the global model is a simple global LSTM without basin characteristics, results of which were shown early in section 3.3 (Figure 2).

In Figure 15, a distribution of scatters above the diagonal line (exactly equal performance from the methods indicated by axes) indicates that Gaussian 512-d vectors outperform all these physically incomplete conditions. This fact is better illustrated in the cumulative density function plot as the distribution of NSE scores is skewed to upper tail. Both its mean and median NSE scores are higher than any physically incomplete characterization (Table 9). Note that as shown earlier, the EG-512 case reaches comparable and slightly better performance than EP. This observation also implies that 27-d physical vectors are lacking additional physical characterizations.

4.6 Uncertainties in Basin Characteristics

Static vectors are deterministic representations of heterogeneous and complex physical systems. Multiple sources might contribute to uncertainties in static vectors. For example, one of them is the spatial simplification. Spatial dependent features are deterministic representations of catchments, such as, soil porosity, silt fraction, etc. Instead of quantifying these uncertainties, we explored another applicability of proposed random vectors utility and answer this question "Can we recognize these uncertainties in static vectors?" The mixed Gaussian vector is a combination of 27-d physical vectors and Gaussian vectors. Uncertainties are captured by those augmented Gaussian vectors. As shown in Figure 16, compared to the baseline performance, which is the EP, the mixed Gaussian vector (EM-d) yields better performance and achieves the maximal performance improvement at EM-512. On average, the NSE improvement is 0.018. From 64-d, to 1024-d, all of the EM-d results yield better performance (positive NSE score improvement statistics in Table 10). Augmenting the physical descriptors with the Gaussian vectors cap-

Table 10: NSE performance difference of the mixed Gaussian vectors (EM-d) and Gaussian vectors (EG-d) in contrast to 27-d physical vectors (EP). Positive scores mean that the EP yielded worse predictive performance

static vector dimension (d)		64	128	256	**512**	1024
Gaussian vectors (EG-d)	mean	-0.007	0.010	0.010	**0.009**	0.002
	median	-0.006	0.002	0.004	**0.007**	0.000
Mixed Gaussian vectors (EM-d)	mean	0.009	0.013	0.014	**0.018**	0.009
	median	0.004	0.006	0.007	**0.011**	0.008

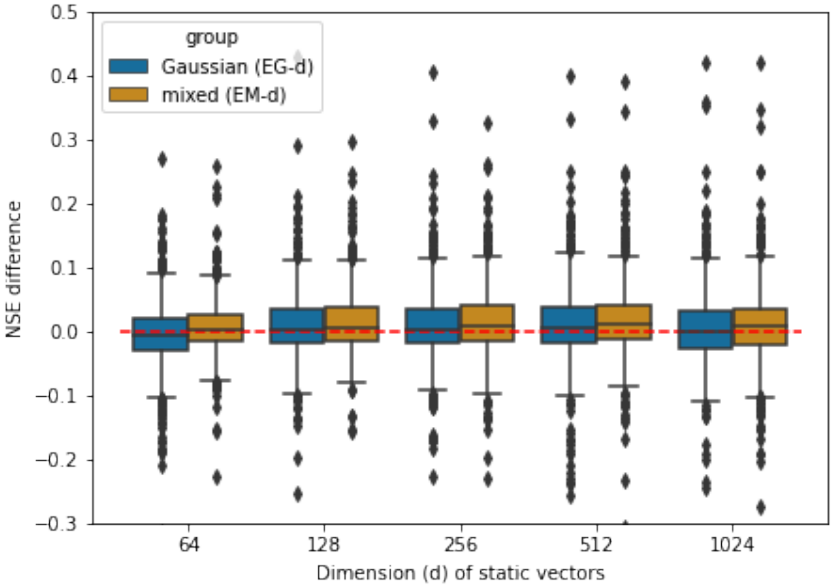


Figure 16: Comparison of the performance between Gaussian vectors (EG-d, blue box) and mixed Gaussian vectors (EM-d, flaxen box). The X-axis is the dimension of the static vector (from 64 to 1024), while the Y-axis shows the NSE difference in contrast to the EP. The red line specifies the performance improvement threshold. Box portions above the red line indicate performance improvement.

turing uncertainties, the incremental performance improvement verifies the presence of uncertainties in physical vectors. When the uncertainties of catchment physical systems are explained by these random vectors, the predictive performance also improves. In contrast to a pure random Gaussian system, given the same d Gaussian dimension (blue and flaxen box in the same X-axis category), the EM-d also marginally improves the NSE score. In Table 10, the NSE improvement in 'Mixed Gaussian vectors' are consistently more pronounced than 'Gaussian vectors' across varying static vector dimension d. The mixed Gaussian vector leads to marginally better global model performance compared to either pure random Gaussian system or pure physical system. As such, it suggests that when physical descriptors are augmented with randomized uncertainties, it supports and benefits regionalization.

5 Discussion

The recently developed LSTM based models have pioneered and advanced the scientific frontier of model-independent (data-driven) regionalization methods. LSTM based models leverage both meteorological data x^d and catchment static vectors x^s to learn a universal global hydrologic model and achieve state-of-the-art performance (Feng et al., 2019; Kratzert et al., 2019a). In particular, catchment physical descriptors are of great value for extrapolating hydrologic information across basins, so the derived deep learning model using these descriptors learn cross-basin similarities internally as feature embeddings (Kratzert et al., 2019a). This is the focus in section 3.3. However, dependence on physical descriptors for regionalization might be potentially problematic because those descriptors are mostly uncertain, sometimes incomplete, or even unavailable in certain regions. These issues hamper the applicability of LSTM based models. Our proposed random vector approach provides a viable solution.

5.1 Unavailable catchment characteristics

Results in section 4.1 show that the proposed random vector method achieves comparable and slightly better performance than the state-of-the-art model (Figure 5, Table 4). In other words, without any knowledge of physical descriptors, the global LSTM based model using random vectors successfully learns universal hydrologic behavior and sustains benchmark streamflow prediction performance. These random vectors retain practical feasibility without having to obtain any physical descriptions of basins. This is arguably the significant scientific contribution of this paper.

The exhaustive analysis from section 4.2, 4.3, and part of 4.4 verifies the applicability of employing random vectors in data scarce regions. For a limited number of basins (Figure 6, 7, 13, 14, Table 5, 8) and a few years of training data (Figure 8), two situations which restrict hydrologic extrapolation across catchments, random vectors are still viable for hydrologic regionalization.

Functionalizing random vectors as static vectors (x^s) that represent each basin, the LSTM family models actually modulate each basin from this characterization. For each catchment, the LSTM model will modulate its internal computation and mapping across neurons. That is, for a given weather input x^d, the global model is aware of which basin the x^d data originates from and thus modulates how streamflow shall be predicted in a different way in contrast to other basins. Yet this modulation extent varies in different model selections. CT-LSTM concatenates static vectors with weather drivers at each timestamp and thus performs the strongest modulation because this catchment awareness is passed through all gates in the LSTM. Merely feeding static vectors into the input gate, the EA-LSTM does not modulate the network as well as the CT-LSTM but it ensures x^s is involved in the temporal context update (memorizing and forgetting). In contrast, the FM-LSTM performs the weakest modulation since the x^s is only involved for updating the hidden representation, which is the last step in an LSTM update cycle before proceeding to next timestamp. Across those three different LSTM model selections with various catchment modulation degree, random vectors consistently perform as well as, if not better than, physical descriptors for learning across basins (Figure 11, Table 7). Indeed, regardless of the impact of x^s on catchment modulation, random vectors enhance the LSTM's ability to learn across basins to a similar or even better extent than what the 27-d catchment physical descriptors are capable of performing.This insight and discovery actually has a broad and significant implication for hydrologists to examine the value of x^s (either physical descriptors or random vectors) that were brought in for modeling catchment complexities.

Typically, using either model-dependent or model-independent methods, hydrologists will consider only physical descriptors as static vectors in addition to dynamic input and streamflow observation for building hydrologic models. This is a scientifically intuitive and classic approach because physical descriptors accompany the development of hydrological models to explain rainfall-runoff behaviors. Catchment physical descriptors are essential for not only underlining physical hydrologic processes in individual basins but also transferring hydrologic knowledge for regional modeling. However, these characteristics are somewhat problematic in terms of the uncertainties in them and potentially missing physical rules. Thus, a more relevant hydrologic question to ask is:"Are catchment physical descriptors sufficient to model streamflow generation complexities? If not, how many dimensions do we need?"

By definition, the needed dimension for characterizing catchment complexity is the dimension of the static vectors. Selected from the CAMELS dataset with prior catchment understanding (Kratzert et al., 2019a), physical descriptors are 27-d static vectors (Table 1). Without any catchment information, the implementation of the Gaussian random vector needs to specify its dimension d, which is empirically obtained. The optimal d is different between EA-LSTM and CT-LSTM. For EA-LSTM, the optimal d is

either 256 or 512 (Figure 4, 7, 6), while for CT-LSTM, it is in the range of 16 to 64 (Figure 9, 13, 14). EA-LSTM needs a higher dimension of static vectors to perform regionalization than the CT-LSTM. We explain this difference by the amount of trainable machine learning parameters. The increasing number of trainable model parameters of CT-LSTM hinders the training processes. For CT-LSTM, an increased static input will expand the concatenated input $x[t]$ dimension (Equation 8), which in turn enlarges the dimension of the transformation matrices $\mathbf{W_i}$, $\mathbf{W_f}$, $\mathbf{W_g}$, $\mathbf{W_o}$. By contrast, the static input (x^s) dimension only impacts EA-LSTM's input gate dimension ($\mathbf{W_i}$). Consequently, given the same x^s dimension augmentation, the parameter increment of CT-LSTM is four times the increase of the number of parameters in EA-LSTM. A higher d-dimension Gaussian vector CT-LSTM becomes more difficult to optimize than that for EA-LSTM considering the number of machine learning parameters involved.

Although the optimal d differs between the CT-LSTM and the EA-LSTM, the performance saturation trend is identical. As illustrated by the results between section 4.1 and 4.4, when expanding static vector dimension, the predictive performance saturates at a certain point and then deteriorates. This pattern indicates a presence of the optimal d, which cannot be too large or too small. In particular, the optimal d is universal regardless of the number of basins involved (Figure 6, 7, 13, 14). Thus, it suggests implications for addressing catchment modeling complexities, which are often entangled with associated scaling issues between catchments as one of the Two Clouds in hydrology (Beven, 1987). Hydrological models need to either be simplified or made more complex to account for scaling transformations between catchments that have different complexities. This can be done by reducing or increasing the number of parameters, which can also be reasonably interpreted as the dimensionality of static vectors. A recognized optimal d illustrates that the level of an appropriate scale for regionalization exists and the involved complexities exceed what the physical descriptors can provide.

Random vectors comparable regionalization predictive performance also implies the uniqueness of catchment systems and randomness in modeling systems. Both the Gaussian vectors or the one-hot vectors map catchments into a high dimension space and preserves their uniqueness. The Gaussian vector characterizes catchments as statistically independent from each other. In the space characterized by the one-hot vector, catchments become orthogonal to each other. Although these random vectors do not quantify catchment similarities, they assure catchments are different from each other in a consistent way. This suggests that preserving the uniqueness of catchments improves regional modeling in a deep learning framework, which reflect a recently raised hydrologic concern – When essential catchment characteristics are not well understood or defined and thus not even included in catchment physical descriptors, how could a derived deep learning model perform satisfactory regionalization performance (Beven, 2020). Although not explicitly defining catchment characteristics, our proposed random vector can be interpreted as non-physical descriptors characterizing the uniqueness of catchments. Catchment systems are composed of linked components representing the functional relationships between weather inputs and streamflow. The uniqueness of catchments further suggests the uniqueness of those individual functions. Additionally, random vectors also support the randomness of catchment system. The stochasticity represented by the random vectors is indicative of the randomness in hydrologic processes. Arguably, catchment system involves organized complexities where complexity exists in a similar way as randomness (Nearing et al., 2020; Dooge, 1986; Weinberg, 2001). The deep learning framework leverages this random complexity for streamflow prediction.

Additionally, modeling hydrologic complexities suggests a necessity for understanding catchment similarities in the context of regionalization. An underlying assumption behind the regionalization method is that physical descriptors define basin similarities/differences, which characterize similar/different catchment hydrologic processes. This assumption leads to a selection of 27-d physical descriptors in LSTM methods, but the optimal per-

formance in random vectors indicates that the similarities across catchments are only partially represented by 27-d physical descriptors. In other words, the similarities encoded in catchment physical descriptors represent insufficient similarities in hydrologic processes. To depict catchment similarities, these current 27-d physical descriptors need to be enriched such that they provide a characterization sufficient enough to denote cross-catchment hydrologic behavior similarities.

Our results are delivered in a deep learning framework. The random vector approach exhibits the strong modeling capacity of deep learning and shows a potential solution to involving complexity into a deep learning model without explicitly incorporating hydrologic processes. This approach does not add physical process understanding into the model architecture; instead, it is developed purely from a data driven perspective. We hypothesize that an appropriate dimension that accommodates catchment complexities exists and allows deep learning models to automatically distinguish cross basin similarities and therefore benefits regional modeling.

5.2 Incomplete and uncertain catchment characteristics

Results from section 4.5 and 4.6 provide preliminary solutions for performing regionalization when physical descriptors are incomplete, or uncertain respectively. Incomplete catchment physical descriptors constrain modeling complexities and thus downgrade regional modeling. Compared to the performance with incomplete features (climate features, geology features, or geomorphology features) and to the performance without any physical descriptors, the predictive performance of the random Gaussian vectors method significantly outperforms in those scenarios. Random Gaussian vectors enable deep learning models to learn complexities more sufficiently than those physically limited descriptors. This insight has practical utility for determining the sufficiency of physical descriptors in the real world, which is challenging considering the uncertainties and complexities in hydrologic processes. When LSTM models using a specified set of physical descriptors are outperformed by random vectors, it demonstrates that those given physical descriptors are not able to resolve catchment complexities and thus suggests a need to complement them with missing features for regional modeling. For instance, as a direct illustration, Figure 4 and Table 4 suggests that 27-d physical descriptors partially address hydrologic complexities and need a certain degree of feature augmentation.

Kratzert et al. (2019a) pointed out that these 27-d physical features are intrinsically uncertain since spatial heterogeneities are simplified as spatial averages and therefore lose certain regional information. To address the uncertainties introduced by the global system enclosed by 27-d physical descriptors rather than individual features, we propose to concatenate physical descriptors with additional Gaussian vectors as a preliminary solution. The added Gaussian vectors do not specify which features are uncertain, but instead, they represent the uncertainties for the whole system. A mixed static feature consists of 27-d physical vectors and Gaussian vectors. Results (Figure 16, Table 10) show that the mixed static features can account for some degree of uncertainties in the physical descriptors. This peak performance is realized by 512-d, which implies the presence of a large degree number of uncertainties (485 dimensions of Gaussian vectors (485 = 512-27)) in the physical hydrology system. Another indication from the results in section 4.6 is that mixed static vectors always outperform pure Gaussian vectors. Given the same dimension of static vectors, the information contained in 27-d physical features improve regional modeling. In contrast to a pure random system formed by all dimensions of Gaussian vectors, we hypothesize that mixed static vectors introduce ordered information and physically similarities, and thus benefit regionalization.

5.3 Limitations and future direction

Although the predictive performance of random vectors proves to be comparable to 27-d physical descriptors, we want to emphasize that this result is limited to gauged prediction. The deep learning model has to have training data of the basin to predict, so the scope of this research cannot not be expanded to PUB. Therefore, recognizing this limitation, it merits future research to leverage the complexity modeling capacity found in random vectors into PUB.

Our ability to model catchment complexities depends on the dimension of the random vector. Although we show the presence of an optimal d, which recognizes the existence of physical processes that are not characterized, we do not provide further quantitative interpretations of the optimal d. How to utilize the observation that the optimald of EA-LSTM is 512? In future studies it will be important to identify physical processes that are not captured by physical charactersitics (e.g., variable recession characteristics (Beven, 2020)) and adapt machine learning models to resolve them.

The results focus on 531 basins in the United States. Their catchment area exhibits a wide range between 4 to 1980 square kilometers, which indicates strong spatial heterogeneities across catchments. An interesting hypothesis to test is that a heterogeneous catchment prefers high dimension Gaussian vectors to account for model complexities. To test this hypothesis it will be necessary to obtain the data from catchments expressing different levels of heterogeneities. Because catchments are naturally heterogeneous, this test will require the use of synthetic data generated by physically based hydrological models. It is hypothesized that a collection of homogeneous catchment will require fewer static vectors while a collection of more complex catchment will require many static vectors. The synthetic data set will represent a system of catchments with a controlled level of heterogeneities, which will allow an opportunity to investigate how heterogeneous and homogeneous catchment systems differentiate hydrologic regionalization and modeling complexities.

Random vectors characterize a system of basins as unique positions in high dimensional space. The only physically distinctive information involved becomes weather inputs and associated catchment responses. This insight suggests the possibility of learning catchment similarities from weather inputs and is thus closely related to the inverse modeling problem, a field where machine learning is also advancing (Ongie et al., 2020). It therefore merits future research for an improvement in unveiling catchment characterization mysteries in a physically consistent way, likely inferred from weather inputs and catchment responses.

Our discovery has strong generalizable implications for other applications in water related or science problems. Regionalization can be conceptualized in a broader concept, that is, each local entity contributes to learn a regional or global model where cross entity information sharing benefits the predictive performance. In the context of streamflow prediction, an entity is a catchment. For water science, an entity can also be a reservoir, lake, stream, etc. The target variable might vary depending on specific problems to solve where each problem may require a different set of entity descriptors. Mathematically, entities can be approximate functions in identical formulations with varying parameters. The benefit of random vectors in modeling regional complexities merits further research to demonstrate their practical applicability. We expect further research can test our proposed random vector approach to solve general regionalization problems across disciplines.

6 Conclusion

In this work we showed that random vectors can be used for hydrologic regionalization when catchment physical descriptors are not available. Random vector based hy-

drologic regionalization shows robust performance even under data sparsity and different model strategies. This method can also identify if any given physical descriptors are sufficient to account for rainfall runoff complexities. In summary, the scientific contributions of this paper are:

- The random vector method was proposed and used for regionaliation in the absence of explicit physical descriptors.
- Random vectors show robust performance even under different data sparsity scenarios and different LSTM based model selection.
- Random vectors can improve streamflow prediction when basin characteristics are insufficient and uncertain. Thus, random vectors have a practical usage in determining if any given physical features are sufficient.

We also investigated scientific implications of the dimension of random vectors. This provides useful insights for the development of hydrologic models to address the model complexity and associated scaling issues.

Appendix A physical descriptor description (CAMELS)

Table A1: 27-d physical descriptors in CAMELS. Descriptions are from (Addor et al., 2017)

Category	Physical descriptors	Description
climate (9)	p_mean	Mean daily precipitation
	pet_mean	Mean daily potential evapotranspiration.
	aridity	Ratio of mean PET to mean precipitation.
	p_seasonality	Seasonality and timing of precipitation. Estimated by representing annual precipitation and temperature as sine waves. Positive (negative) values indicate precipitation peaks during the summer (winter). Values of approx. 0 indicate uniform precipitation throughout the year.
	frac_snow_daily	Fraction of precipitation falling on days with temperatures below 0 .
	high_prec_freq	Frequency of high-precipitation days ($\geq$ 5 times mean daily precipitation).
	high_prec_dur	Average duration of high-precipitation events (number of consecutive days with $\geq$ 5 times mean daily precipitation).
	low_prec_freq	Frequency of dry days ($<$ 1 mm d^{-1}).
	low_prec_dur	Average duration of dry periods (number of consecutive days with precipitation $<$ 1 mm d^{-1}.
Geomorphology(8)	elev_mean	Catchment mean elevation.
	slope_mean	Catchment mean slope.
	area_gages2	Catchment area.
	forest_frac	Forest fraction.
	lai_max	Maximum monthly mean of leaf area index.
	lai_diff	Difference between the max. and min. mean of the leaf area index.
	gvf_max	Maximum monthly mean of green vegetation fraction.
	gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction.
Geology(10)	soil_depth_pelletier	Depth to bedrock (maximum 50 m).
	soil_depth_statsgo	Soil depth (maximum 1.5 m).
	soil_porosity	Volumetric porosity.
	soil_conductivity	Saturated hydraulic conductivity.
	max_water_content	Maximum water content of the soil.
	sand_frac	Fraction of sand in the soil.
	silt_frac	Fraction of silt in the soil.
	clay_frac	Fraction of clay in the soil.
	carb_rocks_frac	Fraction of the catchment area characterized as "Carbonate sedimentary rocks".
	geol_permeability	Surface permeability (log10).

Acknowledgement. This work was funded by the NSF HDR Grant: NSF Award 1934721. J.L. Nieber's effort on this project was partially supported by the USDA National Institute of Food and Agriculture, Hatch/Multistate project MN 12- 109. Access to computing facilities was provided by the Minnesota Supercomputing Institute (`https://www.msi.umn.edu/`). Both the CAMELS data (`https://doi.org/10.5065/D6G73C3Q`) and the extended Maurer forcing data (`https://doi.org/10.4211/hs.17c896843cf940339c3c3496d0c1c077`) are publically available. The code to reproduce our work is available at (`https://github.com/lixx5000/Global-deep-learning-regionalization-from-physical-descriptors-to-random-vectors`).

Author contributions. XL and AK had the idea for Gaussian vectors. VK had the idea for one hot vectors. All the authors were involved in the discussion of experiments design and results, which were mainly led by XL and AK. XL conducted all the experiments and analyzed the results, and together with AK. XL, AK, XJ, KC, JN, CD, MS, VK worked on the manuscripts. XL wrote the original draft and led the editing. JN, CD supervised the manuscript from the hydrologist perspective. AK, KC, XJ, MS, VK supervised the manuscript from the computer scientist perspective.

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. doi: 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. Retrieved from `https://doi.org/10.1016/j.jhydrol.2017.12.019` doi: 10.1016/j.jhydrol.2017.12.019

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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR018247` doi: https://doi.org/10.1002/2015WR018247

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157-166. doi: 10.1109/72.279181

Besaw, L. E., Rizzo, D. M., Bierman, P. R., & Hackett, W. R. (2010). Advances in ungauged streamflow prediction using artificial neural networks. *Journal of Hydrology*, *386*(1-4), 27–37. Retrieved from `http://dx.doi.org/10.1016/j.jhydrol.2010.02.037` doi: 10.1016/j.jhydrol.2010.02.037

Beven, K. (1987). Towards a new paradigm in hydrology. *Water for the future. Proc. Rome symposium, 1987*(164), 393–403.

Beven, K. (1989). CHANGING IDEAS IN HYDROLOGY- THE CASE OF PHYSICALLY-BASED MODELS. , *105*, 157–172.

Beven, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, *5*(1), 1–12. doi: 10.5194/hess-5-1-2001

Beven, K. (2002). Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. *Hydrological Processes*, *16*(2), 189–206. doi: 10.1002/hyp.343

Beven, K. (2020). Deep learning, hydrological processes and the uniqueness of place. *Hydrological Processes*, *34*(16), 3608–3613. doi: 10.1002/hyp.13805

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3-4), 251–290. doi: 10.1002/hyp.3360090305

Burnash, R. J. C. (1995). The NWS river forecast system–catchment modeling.

Computer models of watershed hydrology, 311–366.

Choubin, B., Solaimani, K., Rezanezhad, F., Habibnejad Roshan, M., Malekian, A., & Shamshirband, S. (2019). Streamflow regionalization using a similarity approach in ungauged basins: Application of the geo-environmental signatures in the Karkheh River Basin, Iran. *Catena*, *182*(June), 104128. Retrieved from `https://doi.org/10.1016/j.catena.2019.104128` doi: 10.1016/j.catena.2019.104128

de Lavenne, A., Andréassian, V., Thirel, G., Ramos, M. H., & Perrin, C. (2019). A Regularization Approach to Improve the Sequential Calibration of a Semidistributed Hydrological Model. *Water Resources Research*, *55*(11), 8821–8839. doi: 10.1029/2018WR024266

Dooge, J. C. I. (1986). Looking for hydrologic laws. *Water Resources Research*, *22*(9S), 46S-58S. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR022i09Sp0046S` doi: https://doi.org/10.1029/WR022i09Sp0046S

Drost, N. F. W.-A. A., S., & Mudersbach, C. (2021). The impact of land cover data on rainfall-runoff prediction using an entity-aware-lstm.. doi: https://doi.org/10.5194/egusphere-egu21-1136,2021.

Ecrepont, S., Cudennec, C., Anctil, F., & Jaffrézic, A. (2019). PUB in Québec: A robust geomorphology-based deconvolution-reconvolution framework for the spatial transposition of hydrographs. *Journal of Hydrology*, *570*(January), 378–392. doi: 10.1016/j.jhydrol.2018.12.052

Feng, D., Fang, K., & Shen, C. (2019). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. , 1–49. Retrieved from `http://arxiv.org/abs/1912.08949`

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), 1–24. doi: 10.1029/2019WR026793

Frame, J., Nearing, G., Kratzert, F., Raney, A., & Rahman, M. (2020, Jul). *Post processing the u.s. national water model with a long short-term memory network.* EarthArXiv. Retrieved from `eartharxiv.org/4xhac` doi: 10.31223/osf.io/4xhac

Freeze, R. A. (1974). Streamflow generation. *Reviews of Geophysics*, *12*(4), 627–647. doi: 10.1029/RG012i004p00627

Freeze, R. A., & Harlan, R. (1969). BLUEPRINT FOR A PHYSICALLY-BASED, DIGITALLY-SIMULATED HYDROLOGIC RESPONSE MODEL. *Journal of Hydrology*, *9*, 237–258.

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary Reviews: Water*, *8*(1), 1–32. doi: 10.1002/wat2.1487

Hochreiter, S., & Urgen Schmidhuber, J. (1997). Long Shortterm Memory. *Neural Computation*, *9*(8), 17351780.

Hsu, K. Gupta, H. V., & Sorooshian, S. (1995). Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*, *31*(10), 2517–2530. doi: 10.1029/95WR01955

Keith Beven, & Andrew Binley. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. Retrieved from `http://dx.doi.org/10.1002/hyp.3360060305`

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization.*

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall – runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*, 6005–6022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019b). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344-11354. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR026065` doi: https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019a). 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. doi: 10.5194/hess-23-5089-2019

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, *99*(D7), 14415-14428. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JD00483` doi: https://doi.org/10.1029/94JD00483

Ma, K., Feng, D., Lawson, K., Tsai, W., Liang, C., Huang, X., ... Shen, C. (2021). Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, 1–26. doi: 10.1029/2020wr028600

McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., ... Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), 1–6. doi: 10.1029/2006WR005467

Nearing, G. S., Kratzert, F., Sampson, A. K., Craig, S., Frame, J. M., Klotz, D., & Gupta, H. V. (2020). What Role Does Hydrological Science Play in the Age of Machine Learning ? *Water Resources Research*, 1–17. doi: 10.31223/osf.io/3sx6g

Ongie, G., Jalal, A., Metzler, C. A., Baraniuk, R. G., Dimakis, A. G., & Willett, R. (2020). *Deep learning techniques for inverse problems in imaging.*

Pagliero, L., Bouraoui, F., Diels, J., Willems, P., & McIntyre, N. (2019). Investigating regionalization techniques for large-scale hydrological modelling. *Journal of Hydrology*, *570*(September 2017), 220–235. Retrieved from `https://doi.org/10.1016/j.jhydrol.2018.12.071` doi: 10.1016/j.jhydrol.2018.12.071

Prieto, C., Le Vine, N., Kavetski, D., García, E., & Medina, R. (2019). Flow Prediction in Ungauged Catchments Using Probabilistic Random Forests Regionalization and New Statistical Adequacy Tests. *Water Resources Research*, *55*(5), 4364–4392. doi: 10.1029/2018WR023254

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., ... Attinger, S. (2017). Toward seamless hydrologic predictions across scales. *Hydrology and earth system sciences discussions*, *2017*(89), 4323–4346. doi: 10.5194/hess-2017-89

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., ... Zehe, E. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. doi: 10.1623/hysj.48.6.857.51421

Weinberg, G. M. (2001). *An introduction to general systems thinking (silver anniversary ed.).* USA: Dorset House Publishing Co., Inc.

Zamoum, S., & Souag-Gamane, D. (2019). Monthly streamflow estimation in ungauged catchments of northern Algeria using regionalization of conceptual model parameters. *Arabian Journal of Geosciences*, *12*(11). doi: 10.1007/s12517-019-4487-9