Knowledge-guided Self-supervised Learning for Estimating River-Basin Characteristics

Rahul Ghosh  
University of Minnesota  
ghosh128@umn.edu,

Arvind Renganathan  
University of Minnesota  
renga016@umn.edu,

Xiang Li  
University of Minnesota  
lixx50000@umn.edu,

John Neiber  
University of Minnesota  
neiber@umn.edu,

Ankush Khandelwal  
University of Minnesota  
khand035@umn.edu,

Christopher Duffy  
Penn State University  
cxd11@psu.edu,

Xiaowei Jia  
University of Pittsburgh  
xiaowei@pitt.edu,

Vipin Kumar  
University of Minnesota  
kumar001@umn.edu

Abstract

Machine Learning is being extensively used in hydrology, especially streamflow prediction of basins/watersheds. Basin characteristics are essential for modeling the rainfall-runoff response of these watersheds and therefore data-driven methods must take into account this ancillary characteristics data. However there are several limitations, namely uncertainty in the measured characteristics, partially missing characteristics for some of the basins or unknown characteristics that may not be present in the known measured set. In this paper we present an inverse model that uses a knowledge-guided self-supervised learning algorithm to infer basin characteristics using the meteorological drivers and streamflow response data. We evaluate our model on the the CAMELS dataset and the results validate its ability to reduce measurement uncertainty, impute missing characteristics, and identify unknown characteristics.

1 Introduction

Streamflow prediction is important for understanding hydrology cycles, water supply management, flood mapping, and other operational decisions such as reservoir release. Recently, machine learning (ML) is increasingly being used to solve this task. Compared to traditional hydrological models that are often calibrated on each single specific catchment, machine learning models have been shown to benefit from a large cross section of diverse training data, because knowledge can be transferred across sites (Kratzert et al. 2019). In particular, machine learning models are able to learn the entire mapping from the meteorological drivers to streamflow directly (Razavi and Coulibaly 2013). However, there are inherent characteristics of river-basins that govern the relationship between input meteorological drivers and streamflow response. For example, for the same amount of rainfall, two different basins will have very different streamflow values depending on the basin characteristics (e.g., soil property, slope, and vegetation). Thus, the data-driven methods must take into account this ancillary basin characteristics to distinguish basins and to effectively model the rainfall-runoff relationship. Figure 1a shows the diagrammatic representation of this data-driven forward model.

Figure 1: (a) Data-driven forward model which uses the meteorological drivers ($X^i_t$) and the basin characteristics ($z_i$) to predict the streamflow ($y^i_t$) at a given point in time; (b) The inverse model estimates the characteristics ($z_i$) of a basin by capturing the relationship between the meteorological drivers ($X^i_t$) and its streamflow response ($y^i_t$)

In the scientific application of streamflow modeling, these basin characteristics are available to the data-driven models in the form of indices. As pointed out by (Beven 2020), most catchment indices are only surrogate variables of the true basin characteristics because of spatial heterogeneities and thus can lead to several challenges. First, there often exist much uncertainty in hydrological measurement which in turn causes corruption in basin characteristics. Second, the full set of basin characteristics may not be measured across all the river basins resulting in incompleteness basin characteristics. Finally, there may be some basin characteristics which are essential in modeling the rainfall-runoff response relation but maybe completely unknown, not well understood, or not present in the available set of basin characteristics. Thus, the ability to infer these time-invariant basin characteristics from the time-varying meteorological and streamflow data is essential in terms of both model prediction and hydrological process understanding.

As shown in Figure 1a, the basin characteristics ($z$) act as modulating variables that adjusts the machine learning network to capture a specific rainfall-runoff relation. In this paper we raise a question: can we estimate these time-invariant basin characteristics by capturing the relation between the meteorological driver and streamflow response, as shown in Figure 1b? To answer this question, we design an inverse model that uses a knowledge-guided self-supervised learning algorithm and the limited data on basin characteristics to generate new data representations. Our objective is to use...
2 Method

2.1 Data Description

CAMELS (Catchment Attributes and MEteorology for Large sample Studies) provides a dataset encompassing a total of 671 watersheds/basins across the contiguous US where each basin is supplied with observed streamflow discharge \( y_i \) and multi-variate meteorological drivers data \( X_i \) from ground observations and remote sensing products at a daily scale. Meteorological inputs are daily precipitation, daily minimum air temperature, daily maximum air temperature, average short-wave radiation, and vapor pressure. Both daily meteorological weather inputs and discharge data cover a reasonably long records spanning from 1980 to 2014. In addition, each basin \( i \) is characterized from climatology, geomorphology and geology perspectives by 27 physical features \( Z_i \) as summarized in Figure 2. These 27 features are relatively stable over time and we assume them to be static in this study. As suggested by (Kratzert et al. 2019), in our study we selected 531 out of the total 671 basins, which removes watersheds whose boundaries are likely mis-delineated and therefore causes unwanted spatial heterogeneity issues. Accounting data quality controls from multiple watersheds, only a segment of those records are used for both training and testing. The training period starts from October 1st 1999 and ends on September 30th 2008. The testing period ranges between October 1st 1989 and September 30th 1999. This selection of year starting and ending follows the definition of water year in hydrology community. October 1st marks the first day in a water year.

Figure 2: 27-d physical descriptors along with their sub categories

2.2 Architecture

The inverse-mapping task is to predict the physical attributes \( Z_i \) for a basin, given its daily streamflow discharge \( y_i \) and meteorological driver \( X_i \) data. Basin characteristics remain constant through time and to estimate them we propose a deep representation learning architecture that generates time invariant and source specific embeddings from time varying meteorological driver and streamflow discharge data. Specifically, for a time-window \( a \) of length \( W \), we use a Bidirectional LSTM based sequence encoder which takes in the concatenated input consisting of the drivers and response \( \{X_i^a; y_i^a\} \) of a basin \( i \). The embeddings of the given basin \( i \) for the time-window \( a \) is given by the hidden representations of the last time-steps from the LSTM layers, as shown in Figure 3a. In our study, we set the window length to be one year.

Since the characteristics of a basin \( Z_i \) remain constant through time and are different from the characteristics of another basin, we build a self-supervised learning approach to extract the representation using a triplet loss. The intuition is to use this method to extract embeddings that are different across different basins but stay invariant in a single basin. In the triplet loss, we create \{anchor, positive, negative\} triplet, and then use the triplet loss to minimize the distance between the anchor and positive and maximize the distance between anchor and negative samples. Specifically, for an anchor time-series sample \( (a_i) \) from a basin \( i \), we define a positive sample \( (p_i) \) as another time-series sample from the same basin and a negative sample \( (n_i) \) as a time-series sample from another basin. The LSTM encoder generates a triplet of embeddings \( \{h_{a_i}, h_{p_i}, h_{n_i}\} \). To force the distance between embeddings from the same basin to be smaller compared to embeddings from the different basins, the triplet loss is defined as follows,

\[
E_{Triplet} = \max(D(h_{a_i}, h_{p_i}) - D(h_{a_i}, h_{n_i}) + \alpha, 0),
\]

where \( D \) is the distance function (Euclidean in our study) and \( \alpha \) is the minimum distance margin, i.e., \( D(h_{a_i}, h_{p_i}) + \alpha > D(h_{a_i}, h_{n_i}) \).

In addition, we regularize the representation learning architecture by adding a reconstruction loss \( E_{reconstruction} \) to preserve the key information from input data. Given the
generated embedding, we add a downstream task to predict static characteristics $Z_i$ from learned embeddings. We can define another loss $L_{inverse}$ by comparing the predicted characteristics and ground-truth characteristics that are available for certain basins. Hence, we create the final loss function as:

$$L = \lambda_1 L_{Triplet} + \lambda_2 L_{Reconstruction} + \lambda_3 L_{Inverse},$$  \hspace{1cm} (2)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters to control the weight of three loss terms.

3 Experiments

Our proposed method will be evaluated in the context of three tasks: 1) reducing measurement uncertainty, 2) imputing missing characteristics, and 3) identifying important characteristics that are missing from available characteristics $Z_i$.
Figure 5: (a) Experiment setting; (b) Scatter plot of the reconstructed characteristics (y-axis) vs true characteristics (x-axis) for the testing basins. We can observe that representation learning model significantly reduces measurement error in characteristics by an average of RMSE:0.54. Moreover, the model trained using corrupted values (in green) shows similar performance to the model trained using true values (in red), which shows the robustness of our method.

3.2 Impute missing characteristics

The inverse model trained on multiple basins can potentially estimate characteristics when they are missing for some basins. The 531 available basins in the CAMELS dataset is split into two groups, training basins, and testing basins. We train our model on training basins during training years and predict on testing basins during testing years. Figure 5a shows the quadrant for visualizing testing, training basins and testing, training Years. To evaluate our predictions, we compare the predicted value with the true values for all the testing basins. Figure 5b shows that for most characteristics, the model shows good performance when trained using 400 basins to predict characteristics for the remaining 131 basins.

3.3 Identify unknown characteristics

Representation Learning has the potential to identity some time invariant characteristics that may be missing from available characteristics. We train the inverse model without using any knowledge of available characteristics as a constraint. Triplet and reconstruction loss are used for model training. To empirically demonstrate the characteristics captured by the learned embeddings, we calculate the pairwise-euclidean distance between two basins using learned embeddings (Figure 6b) and compare them with the distances computed using their 27d physical characteristics (Figure ??). From the figure we observe similar patterns in both the distance matrices which shows that representation learning model generates embeddings that contains meaningful similarity structure between basins.

4 Conclusion

In this work, we show that basin characteristics can be estimated from the rainfall-runoff data. The inverse model is able to learn generalizable patterns that is robust under measurement uncertainty, imputation of missing values and identifying unknown characteristics of basins. Self-supervised representation learning based on triplet loss show promise in identifying unknown characteristics. In the future work, the proposed representation learning architecture can be used as a source aware modulation in a meta-learning framework (Vuorio et al. 2019).

5 Acknowledgement

This work was funded by the NSF awards 1838159 and 1739191. Access to computing facilities was provided by the Minnesota Supercomputing Institute.
References

[Beven 2020] Beven, K. 2020. Deep learning, hydrological processes and the uniqueness of place. *Hydrological Processes* 34(16):3608–3613.

[Ding et al. 2015] Ding, S.; Lin, L.; Wang, G.; and Chao, H. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* 48(10):2993–3003.

[Kratzert et al. 2019] Kratzert, F.; Klotz, D.; Shalev, G.; Klambauer, G.; Hochreiter, S.; and Nearing, G. 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences* 23(12):5089–5110.

[Newman et al. 2015] Newman, A. J.; Clark, M. P.; Craig, J.; Nijssen, B.; Wood, A.; Gutmann, E.; Mizukami, N.; Brekke, L.; and Arnold, J. R. 2015. Gridded ensemble precipitation and temperature estimates for the contiguous united states. *Journal of Hydrology and Meteorology* 16(6):2481–2500.

[Razavi and Coulibaly 2013] Razavi, T., and Coulibaly, P. 2013. Streamflow prediction in ungauged basins: review of regionalization methods. *Journal of Hydrologic Engineering* 18(8):958–975.

[Vuorio et al. 2019] Vuorio, R.; Sun, S.-H.; Hu, H.; and Lim, J. J. 2019. Multimodal model-agnostic meta-learning via task-aware modulation. *arXiv preprint arXiv:1910.13616*. 