IMPROVING STREAMFLOW PREDICTIONS WITH VISION TRANSFORMERS

Kshitij Tayal[1], **Arvind Renganathan**[2], and **Dan Lu**[1]

[1]Computational Sciences and Engineering Division, Oak Ridge National Laboratory
[2]Department of Computer Science and Engineering, University of Minnesota
tayalk@ornl.gov, renga@umn.edu, lud1@ornl.gov

ABSTRACT

Accurate streamflow prediction is crucial to understand climate impacts on water resources and develop effective adaption strategies. A global Long Short-Term Memory (LSTM) model, using data from multiple basins, can enhance streamflow prediction, yet acquiring detailed basin attributes remains a challenge. To overcome this, we introduce the Geo-ViT-LSTM model, a novel approach that enriches LSTM predictions by integrating basin attributes derived from remote sensing with a vision transformer. Applied to 531 basins across the United States (US), our method significantly outperforms existing models, showing an 11% increase in prediction accuracy. Geo-ViT-LSTM marks a significant advancement in land surface modeling, providing a more comprehensive and effective tool for managing water resources under climate change.

1 INTRODUCTION AND MOTIVATION

Climate change significantly impacts precipitation patterns and temperature, directly affecting streamflow volume and leading to challenges in ecosystem health and flood/drought risk management (Vaze et al., 2010). Thus, accurate predictive models of streamflow are critical for informed decision-making and effective adaptation strategies. Machine learning (ML) methods, particularly Long Short-Term Memory (LSTM) networks, have shown considerable promise in streamflow prediction by effectively capturing complex hydrological processes and modeling temporal dependencies and non-linear relationships (Kratzert et al., 2018; Konapala et al., 2020; Nearing et al., 2021). Recent research indicates that global LSTM models, trained on data from multiple river basins, can predict streamflow more accurately than basin-specific models (Kratzert et al., 2019a). This approach, which integrates sequential meteorological data and static catchment attributes like basin size, shape, soil type, and vegetation from diverse basins, is better equipped to handle various conditions, improving predictions in ungauged basins and during extreme events linked to climate change (Frame et al., 2022; Xie et al., 2021; Tayal et al., 2022). The effectiveness of these models in handling heterogeneity relies on a detailed understanding of each basin's attributes, as they critically influence hydrological responses to climatic inputs (Stein et al., 2021; Arsenault et al., 2023).

[1]This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan.

Currently, basin attributes (Addor et al., 2017) are mainly sourced from databases like the US Geological Survey's National Water Information System, which typically offer static, uniform values for each attribute across entire basins. This approach presents several challenges, including high uncertainty due to temporal changes, spatial heterogeneity, measurement errors, and missing data. Furthermore, not all river basins have a complete set of measured characteristics, hindering the development of a comprehensive global model and limiting the transferability of models between regions. Additionally, some crucial basin attributes for modeling streamflow response are often unknown or poorly understood. Addressing these limitations and effectively learning comprehensive basin attributes is vital for accurate streamflow prediction.

In this effort, we aim to advance streamflow prediction by enhancing basin attributes using terrain features, land cover, vegetation health, and water body perimeter derived from remote sensing (RS) data. Our premise is that RS can provide comprehensive, dynamic, and spatially complete characteristics across basins. Despite uncertainties in the RS data, their high spatial resolution allows for more precise characterization through machine learning (Shirmard et al., 2022). This enriched set of attributes can improve streamflow predictions and facilitate model transferability across different basins based on similarities in these characteristics. To integrate these enhancements, we introduce the Geo-ViT-LSTM approach, which augments the LSTM model by incorporating knowledge from RS via pre-trained Vision Transformers (ViT) (Zhang et al., 2021; Khan et al., 2022; Zhou et al., 2022), along with existing geomorphological static attributes. This approach leverages ViT to extract spatially informed context from RS data, enhancing the streamflow prediction with additional basin attributes. Geo-ViT-LSTM centers on two main aspects: (1) Multi-modal feature extraction, using ViT to transform RS imagery into spatially aware feature representations, highlighting key terrain, land cover, and vegetation health indicators, and (2) Integrated modeling, wherein the RS features are combined with static attributes and meteorological sequences within the LSTM model. This multi-modal integration not only enriches the input feature set with detailed basin-specific information but also bridges the gap between spatial basin characteristics and temporal meteorological data.

We validate the effectiveness of our Geo-ViT-LSTM method using observations from the CAMELS database, encompassing static basin attributes and sequential meteorological drivers across 531 basins in the US. Our experiments in streamflow prediction reveal that the inclusion of cross-basin knowledge derived from RS as additional features results in an 11% improvement compared to the next best method. This enhancement demonstrates the value of encoding and utilizing structured knowledge across different basins, highlighting the Geo-ViT-LSTM's potential in advancing hydrological predictions and other land surface modeling to understand the climate impacts better.

2 Proposed Approach

Let $\boldsymbol{X_i} = [\boldsymbol{x_i^1}, \boldsymbol{x_i^2}, \ldots, \boldsymbol{x_i^T}]$ be a matrix representing dynamical meteorological drivers such as precipitation, solar radiation, vapor pressure for the i^{th} basin, where $\boldsymbol{x_i^t} \in \mathbb{R}^{D_x}$ is the input vector at time $t \in [1, \ldots, T]$ with D_x dimensions. Correspondingly, $\boldsymbol{Y_i} = [y_i^1, y_i^2, \ldots, y_i^T]$ represents streamflow responses for the i^{th} basin, where each $y_i^t \in \mathbb{R}$ represents the streamflow at time t. Here, D_x is the dimensionality of the input features, T is the number of time steps, and N is the number of basins. Static attributes of a basin are represented by $\boldsymbol{s_i} \in \mathbb{R}^{D_s}$ and the RS image of a basin is represented by $\boldsymbol{I_i}$, and $\boldsymbol{z_i} \in \mathbb{R}^{D_z}$ is the latent vector of basin i, obtained using a ViT. The ViT segments $\boldsymbol{I_i}$ into equal-sized sub-images, which are linearly embedded and enhanced with positional embeddings to encode their spatial locations. These tokens are then processed through transformer encoder blocks, comprising layer normalization (LN), multi-head self-attention with a residual connection, followed by another LN, a multi-layer perceptron (MLP), and a second residual connection, sequentially connected. An MLP head aggregates these to incorporate global image information, which is then concatenated with existing static attributes and fed in a 32-hidden unit LSTM. This enables the LSTM gates and state updates to leverage the sequential precipitation/temperature signals alongside the spatial context of the basin derived from imagery and static attributes.

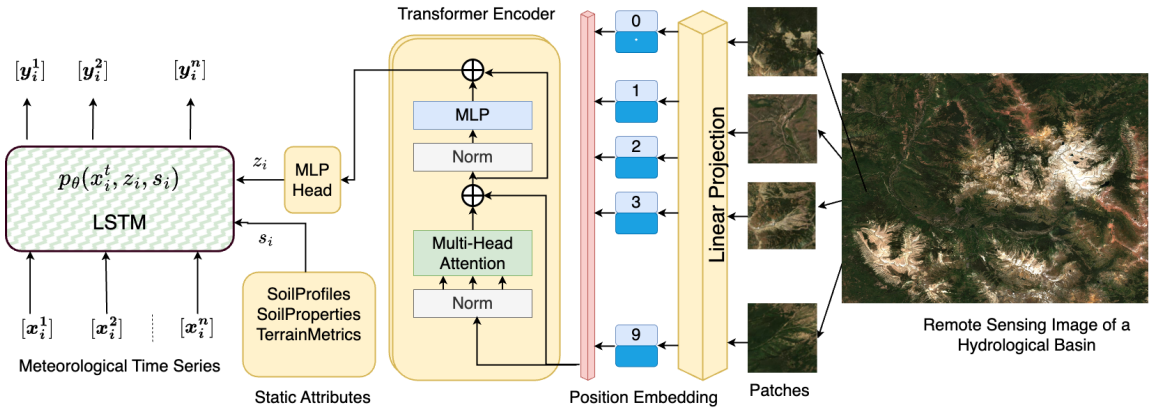
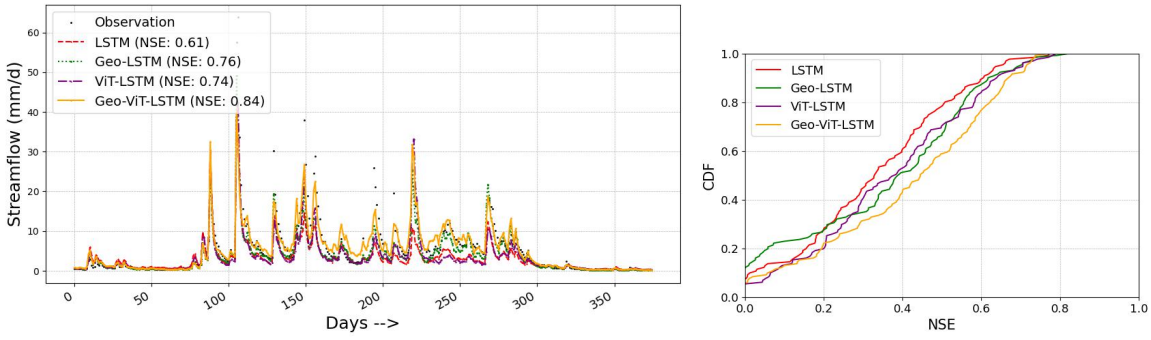


Figure 1: Schematic Illustration of Geo-ViT-LSTM: ViT encodes spatially complete representations of basins, static attributes capture geological features, while LSTM captures temporal dynamics.

Figure 1 illustrates our approach, where the model's parameters are learned by minimizing the Mean Squared Error (MSE) loss function. The MSE between the predictions of the Geo-ViT-LSTM for basin i and the corresponding ground truth values $\boldsymbol{Y_i}$ is calculated as MSE $= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i)^2$, where $\hat{\boldsymbol{Y}}_i = p_\theta(\boldsymbol{X}_i, z_i, s_i)$ represents the ViT-LSTM prediction for the inputs, with p_θ denoting the model parameterized by θ. We optimized our setup by freezing the transformer's base layers, incorporating pre-trained ImageNet weights, and training only the MLP head. This approach enables the model to capture latent features such as terrain attributes and land cover specifics via statistical summarization, providing essential information to the LSTM for accurate temporal modeling. As we will demonstrate in the experiments, this additional capability makes our model more accurate and generalizable. Recent research explores the combination of LSTM and transformer models (Tang et al., 2023); however, to the best of our knowledge, our work is the first to modulate the LSTM state with latent embeddings from transformer models.

3 Datasets Construction

In this study, we leverage CAMELS (Catchment Attributes and MEteorology for Large-sample Studies) Kratzert et al. (2019b) dataset for hydrological analysis. This dataset includes daily weather data, streamflow observations, and 27 basin attributes (see Appendix A.1 for a complete list) from 531 catchments across the US. The CAMELS's rich meteorological, streamflow, and catchment attribute data for these large, diverse basins provide a robust basis for our model's validation. We focused on data from October 2001 to September 2008 for training, October 1999 to September 2001 for validation, and October 1989 to September 1999 for testing, following the approach in Kratzert et al. (2019b). For all 531 basins, we utilized their shapefiles to download Sentinel-2 images via Google Earth Engine. We applied a cloud masking function to enhance image clarity using the QA60 band of Sentinel-2 images. This process effectively excludes clouds and shadows. Subsequently, we merged the cloud-masked images into a single composite image through a median operation to minimize anomalies. The size of these composite images varies, reflecting the different catchment sizes of the basins. Utilizing these RS images latent vectors $\boldsymbol{z_i}$ processed through a ViT and basin attributes $\boldsymbol{s_i}$, we implement our Geo-ViT-LSTM approach. This integration allows our model to leverage both detailed meteorological drivers and high-resolution spatial information, enhancing its predictive accuracy and generalizability across diverse hydrological settings, as demonstrated in the next section.



(a) (b)

Figure 2: (Left) Illustration of annual streamflow prediction across a basin. (Right) Cumulative distribution function (CDF) graph illustrating model comparisons through NSE (Nash-Sutcliffe Efficiency) metrics.

4 Experiments and Results

We benchmarked our approach against three baselines. The first baseline is **LSTM** (Xiang et al., 2020), which utilized only meteorological time-series inputs and excluded any static attributes. The second baseline is **Geo-LSTM** (Kratzert et al., 2019b), which concatenated the dynamic meteorological inputs with static catchment attributes. Our third baseline is **ViT-LSTM**, which incorporates only spatial context embeddings from RS combined with meteorological sequences before integration into the LSTM model. Finally, our proposed methodology is **Geo-ViT-LSTM** where we merged spatial context embeddings with the existing static catchment attributes and combined them with meteorological drivers before being fed into the LSTM, thereby enhancing the input feature set with characteristics from both catchment attributes and RS images. Across the 531 basins, all models were trained on 400 basins and tested on the remaining 131 basins. We implemented two testing strategies: Temporal Test, which involves testing on the 400 basins during the testing period, and Spatiotemporal Test, where we test on the 131 basins during the same testing period.

	Temporal Test		Spatiotemporal Test	
	RMSE	*NSE*	*RMSE*	*NSE*
LSTM	0.56	0.66	0.57	0.65
Geo-LSTM	0.51	0.72	0.54	0.68
ViT-LSTM	0.53	0.70	0.57	0.65
Geo-ViT-LSTM	**0.47**	**0.76**	**0.51**	**0.72**

Table 1: Evaluation of our Geo-ViT-LSTM model with three baselines for both temporal and spatiotemporal out-of-sample scenarios using Root Mean Squared Error (RMSE) and Nash-Sutcliffe Efficiency (NSE). The smaller RMSE and higher NSE is better.

The performance of the four models on the temporal and spatiotemporal out-of-sample testing is summarized in Table 1, which reports averaged RMSE and NSE values to measure the predictive accuracy. RMSE quantifies the average magnitude of errors between predicted and observed values, whereas NSE provides a normalized assessment of how well the predictions of the model match observed data, where a score of 1.0 signifies perfect prediction, and scores near 0 is akin to predicting the mean of observed data. The Temporal Test case evaluates the ability of the models to generalize to unseen time periods for basins included in the training data, while the Spatiotemporal Test case examines generalization both to new time periods and new basins excluded from training. Overall, our Geo-ViT-LSTM model achieved the best performance on both cases in terms of lower RMSE and higher NSE values, representing a 9% and 11% improvement over the next best method, respectively.

Figure 2a illustrates the annual streamflow prediction, which demonstrates the closest fitting of our Geo-ViT-LSTM model to the observations. Additionally, we have the following high-level findings from our analysis: (1) All trained models that combined additional attributes and meteorological sequences had supe-

rior performance compared to the LSTM, which was trained using only meteorological data. This reinforces our observation that the meteorological time series in itself is insufficient in discerning a basin, and we require additional attributes to handle heterogeneity that can modulate LSTM cell to output prediction, which is personalized to each basin. This confirms the value of additional attributes to improve the prediction. (2) All the models performed much better for the Temporal Test compared to the Spatiotemporal Test. This reinforces our observation that data is highly heterogeneous, and it can be challenging to build a geoscientific global model for all basins without additional context. (3) Both static attributes and spatial context attributes from the RS, when considered individually, fall short in addressing heterogeneity. However, when combined, they offer orthogonal information (e.g., static attributes capture basin topology and soil characteristics while spatial context captures terrain relationships within and around the basin), highlighting the complementary strengths that significantly enhance the uniqueness of the basin to enable more accurate streamflow modeling.

To further delve into the distribution of NSE scores within the test dataset, we utilize the Cumulative Distribution Function (CDF) plot. This method facilitates an intuitive depiction of how NSE scores are spread across different models. Specifically, the CDF plot illustrates the percentage of data points (represented on the y-axis) that possess NSE scores at or below a particular threshold (marked on the x-axis). As demonstrated in Figure 2b, our Geo-ViT-LSTM model's CDF curve lies beneath those of its counterparts, signifying that a larger segment of the test dataset achieved superior NSE scores with this model. We connect this success to a vision transformer, which effectively incorporates context attributes and enhances model accuracy by leveraging the spatial relationships inherent in the RS data.

5 Conclusions and Future Work

In this work, we propose a Geo-ViT-LSTM method that leverages the vision transformer's self-attention mechanism to learn spatial context attributes from RS data. This approach not only improves the accuracy of predictions in multi-basin models but also enhances our ability to manage water resources in an era of significant environmental change. Although we have shown improvements in streamflow modeling, the method is general and can be applied to many geoscience disciplines (Tsai et al., 2021) where static attributes are not available.

Looking ahead, we aim to extend the methodology for dynamic monitoring from multiple snapshots of RS images which will encompass the impact of seasonal variations, climate change, and human activities on basin characteristics. This approach will require advancement in transformer architecture, similar to (Dong et al., 2022; Li et al., 2022) and surrogate models (Furtney et al., 2022; Tayal et al., 2023), that will identify features across various scales and contexts and will improve our methodology by allowing a more accurate and timely reflection of the conditions influencing hydrological processes. Furthermore, we aim to broaden the application of our method to include other water- and carbon-related variables that depend on land surface attributes, thus improving predictions and deepening our understanding of Earth's climate dynamics and impacts.

References

Nans Addor et al. The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 2017.

Richard Arsenault, Jean-Luc Martel, Frédéric Brunet, François Brissette, and Juliane Mai. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, 27(1):139–157, 2023.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12124–12134, 2022.

Jonathan M Frame, Frederik Kratzert, Daniel Klotz, Martin Gauch, Guy Shalev, Oren Gilon, Logan M Qualls, Hoshin V Gupta, and Grey S Nearing. Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26(13):3377–3392, 2022.

JK Furtney, C Thielsen, W Fu, and Romain Le Goc. Surrogate models in rock and soil mechanics: Integrating numerical modeling and machine learning. *Rock Mechanics and Rock Engineering*, pp. 1–15, 2022.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Goutam Konapala, Shih-Chieh Kao, Scott L Painter, and Dan Lu. Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous us. *Environmental Research Letters*, 15(10):104022, 2020.

Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11): 6005–6022, 2018.

Frederik Kratzert, Daniel Klotz, Mathew Herrnegger, Alden K Sampson, Sepp Hochreiter, and Grey S Nearing. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55(12):11344–11354, 2019a.

Frederik Kratzert, Daniel Klotz, Guy Shalev, Günter Klambauer, Sepp Hochreiter, and Grey Nearing. 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, 2019b.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4804–4814, 2022.

Grey S Nearing, Frederik Kratzert, Alden Keefe Sampson, Craig S Pelissier, Daniel Klotz, Jonathan M Frame, Cristina Prieto, and Hoshin V Gupta. What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3):e2020WR028091, 2021.

Hojat Shirmard, Ehsan Farahbakhsh, R Dietmar Müller, and Rohitash Chandra. A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sensing of Environment*, 268: 112750, 2022.

Lina Stein, Martyn P Clark, Wouter JM Knoben, Francesca Pianosi, and Ross A Woods. How do climate and catchment attributes influence flood generating processes? a large-sample study for 671 catchments across the contiguous usa. *Water Resources Research*, 57(4):e2020WR028300, 2021.

Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13470–13479, 2023.

Kshitij Tayal, Xiaowei Jia, Rahul Ghosh, Jared Willard, Jordan Read, and Vipin Kumar. Invertibility aware integration of static and time-series data: An application to lake temperature modeling. In *Proceedings of the 2022 SIAM international conference on data mining (SDM)*, pp. 702–710. SIAM, 2022.

Kshitij Tayal, Arvind Renganathan, Rahul Ghosh, Xiaowei Jia, and Vipin Kumar. Koopman invertible autoencoder: Leveraging forward and backward dynamics for temporal modeling. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 588–597. IEEE, 2023.

Wen-Ping Tsai, Dapeng Feng, Ming Pan, Hylke Beck, Kathryn Lawson, Yuan Yang, Jiangtao Liu, and Chaopeng Shen. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature communications*, 12(1):5988, 2021.

J Vaze, DA Post, FHS Chiew, J-M Perraud, NR Viney, and J Teng. Climate non-stationarity–validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology*, 394(3-4):447–457, 2010.

Zhongrun Xiang, Jun Yan, and Ibrahim Demir. A rainfall-runoff model with lstm-based sequence-to-sequence learning. *Water resources research*, 56(1):e2019WR025326, 2020.

Kang Xie, Pan Liu, Jianyun Zhang, Dongyang Han, Guoqing Wang, and Chaopeng Shen. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. *Journal of Hydrology*, 603:127043, 2021.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5579–5588, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

A APPENDIX

A.1 BASIN CHARACTERISTICS:

In this section, we detail 27 characteristics of basins, categorized into three main groups: Climate, Soil Geology, and Geomorphology. These characteristics are important in understanding the hydrological and geological aspects of a basin.

Group	**Name (Full Form)**
Climate	Precipitation Mean (p mean) Potential Evapotranspiration Mean (pet mean) Precipitation Seasonality (p seasonality) Fraction of Precipitation as Snow (frac snow) Aridity Index (aridity) High Precipitation Frequency (high prec freq) High Precipitation Duration (high prec dur) Low Precipitation Frequency (low prec freq) Low Precipitation Duration (low prec dur)
Soil Geology	Fraction of Carbonate Rocks (carbonate rocks frac) Geological Permeability (geol permeability) Soil Depth according to Pelletier (soil depth pelletier) Soil Depth according to STATSGO (soil depth statsgo) Soil Porosity (soil porosity) Soil Conductivity (soil conductivity) Maximum Water Content (max water content) Sand Fraction (sand frac) Silt Fraction (silt frac) Clay Fraction (clay frac)
Geomorphology	Mean Elevation (elev mean) Mean Slope (slope mean) Area (measured using GAGES-II database) (area gages2) Fraction of Forest Cover (frac forest) Maximum Leaf Area Index (lai max) Leaf Area Index Difference (lai diff) Maximum Green Vegetation Fraction (gvf max) Green Vegetation Fraction Difference (gvf diff)

Table 2: Static Characteristics as defined in the referenced study by Addor et al., 2017.