

Water Resources Research®

RESEARCH ARTICLE

10.1029/2024WR039008

Special Collection:

Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Key Points:

- We propose a hybrid framework where the catchment dynamics are learned by *HydroLSTM* and the spatial regionalization by *Random Forest*
- The *Random Forest* approach learns distributions of catchment attributes aligned with distinct dynamical behavioral classes across the continental US
- We show that the “*Regional HydroLSTM*” state tracks “*potential*” streamflow, while the output gate corrects it using temporal context

Supporting Information:

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

Correspondence to:

L. A. De la Fuente, A. Bennett, H. V. Gupta and L. E. Condon,
ldelafue@arizona.edu;
andrbenn@arizona.edu;
hoshin@arizona.edu;
lecondon@arizona.edu

Citation:

De la Fuente, L. A., Bennett, A., Gupta, H. V., & Condon, L. E. (2025). A HydroLSTM-based machine-learning approach to discovering regionalized representations of catchment dynamics. *Water Resources Research*, 61, e2024WR039008. <https://doi.org/10.1029/2024WR039008>

Received 23 SEP 2024

Accepted 29 JUL 2025

© 2025. The Author(s).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](#), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

A HydroLSTM-Based Machine-Learning Approach to Discovering Regionalized Representations of Catchment Dynamics

Luis A. De la Fuente^{1,2} , Andrew Bennett¹ , Hoshin V. Gupta¹ , and Laura E. Condon¹ 

¹Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA, ²Department of Earth, Environmental and Resource Sciences, University of Texas at El Paso, El Paso, TX, USA

Abstract Finding similarities between model parameters across different catchments has proved to be challenging. Existing approaches struggle due to catchment heterogeneity and non-linear dynamics. In particular, attempts to correlate catchment attributes with hydrological responses have failed due to interdependencies among variables and consequent equifinality. Machine Learning (ML), particularly the Long Short-Term Memory (LSTM) approach, has demonstrated strong predictive and spatial regionalization performance. However, understanding the nature of the regionalization relationships remains difficult. This study proposes a novel approach to partially decouple learning the representation of (a) catchment dynamics by using the *HydroLSTM* architecture and (b) spatial regionalization relationships by using a *Random Forest* (RF) clustering approach to learn the relationships between the catchment attributes and dynamics. This coupled approach, called *Regional HydroLSTM*, learns a representation of “*potential streamflow*” using a single cell-state, while the output gate corrects it to correspond to the temporal context of the current hydrologic regime. RF clusters mediate the relationship between catchment attributes and dynamics, allowing identification of spatially consistent hydrological regions, thereby providing insight into the factors driving spatial and temporal hydrological variability. Results suggest that by combining complementary architectures, we can enhance the interpretability of regional machine learning models in hydrology, offering a new perspective on the “*catchment classification*” problem. We conclude that an improved understanding of the underlying nature of hydrologic systems can be achieved by careful design of ML architectures to target the specific things we are seeking to learn from the data.

Plain Language Summary This study addresses the challenge of defining similarity/differences in behavioral dynamics across different hydro-geo-climatic regions. This task, a process known as “*regionalization*,” has been complicated by many factors, including non-uniformity in the inputs and high correlation between catchment attributes. By coupling two Machine Learning (ML) representations, specifically a single-cell-state version of *HydroLSTM* to identify hydrologic behaviors, and a Random Forest to define regions with similar behavioral dynamics, we can classify catchments into regional groups. Parameters associated with these regions facilitate understanding of streamflow prediction at gauged locations. We find that the cell-state of *HydroLSTM* model tracks information about “*potential streamflow*” while the output gate modulates this value based on hydro-climatic context. Meanwhile, the *Random Forest* model learns to associate observed watershed characteristics with specific hydrological behaviors, thereby clarifying the controlling factors in each region. This method improves the interpretability of ML models in hydrology while offering a new perspective on catchment classification. We conclude that careful design of an ML framework based on clearly identified aspects of interest can lead to a better understanding of the underlying nature of hydrologic systems.

1. Introduction

Our understanding of catchment hydrology (encoded into models) is based on sparse, imperfect, and spatio-temporally discontinuous data. These limitations affect the nature of our knowledge and the development of models. To fill the gap, we seek regionalization relationships that generalize across space and time. However, faced with heterogeneous watersheds (e.g., land-surface properties) and complex non-linearity in dynamics (Oudin et al., 2008; Sivapalan, 2003), such efforts remain limited (Guo et al., 2021) compared with other recent advancements.

Author Contributions:

Conceptualization: Luis A. De la Fuente, Hoshin V. Gupta, Laura E. Condon

Formal analysis: Luis A. De la Fuente, Laura E. Condon

Funding acquisition: Laura E. Condon

Investigation: Luis A. De la Fuente, Hoshin V. Gupta, Laura E. Condon

Methodology: Luis A. De la Fuente, Hoshin V. Gupta, Laura E. Condon

Software: Luis A. De la Fuente

Supervision: Andrew Bennett, Hoshin V. Gupta, Laura E. Condon

Validation: Luis A. De la Fuente

Visualization: Luis A. De la Fuente, Laura E. Condon

Writing – original draft: Luis A. De la Fuente

Writing – review & editing: Andrew Bennett, Hoshin V. Gupta, Laura E. Condon

An important goal in hydrology is to develop regionalization techniques that enable hydrologic predictions in ungauged basins. Of course, such relationships should provide the right answers for the right reasons, so as to improve our understanding of hydrology. Achieving this goal requires us to characterize the spatiotemporal nature of factors that drive the hydrological responses of interest, thereby resulting in a more comprehensive understanding of the underlying data-generating processes.

Several studies have attempted to relate observable catchment attributes (such as size, shape, soil, and topographic properties, etc.) to the rainfall-runoff and other relationships that characterize their input-state-output dynamics. Some of these approaches seek to characterize catchments based only on information encoded by catchment attributes and forcings, while others are based on characterization of the catchment responses or associated model parameters (regression-based regional analysis) with the goal of mapping from one to the other (He et al., 2011). Both approaches arguably overlook the core issue: the need for a robust way to characterize and classify the underlying catchment dynamics. For this reason, identifying key descriptors that can be used to understand the spatial variation in these dynamics remains crucial (Yadav et al., 2007). Limitations in our ability to do so are, arguably, the main reason that previous regionalization approaches have only achieved limited success.

For example, attempts to relate catchment model parameters to attributes, via regression, often face severe equifinality (Guo et al., 2021). A major reason is that, typically, the model parameters and catchment attributes are highly correlated (Blöschl, 2005; Kuczera & Mroczkowski, 1998). Such interdependence complicates attempts to interpret the regressed relationships because different combinations of parameters and attributes can give similar results which finally complicate the application of such relationships to unobserved catchments.

An alternative is to relate catchment attributes to hydrological signatures extracted from the streamflow data (Yadav et al., 2007). Such signatures add a needed degree of interpretability to regionalization investigations (Guo et al., 2021). However, it remains unclear which signatures encode the most relevant and useful information to distinguish between catchment types. For instance, the “flashiness” (a measure of how quickly a hydrograph shifts from baseflow to precipitation-generated peak values) of a steep catchment with low precipitation intensity, can be the same as for a flatter catchment with high precipitation intensity. So, while hydrological signatures aid in the classification of dynamic catchment responses, the precise link between attributes and signatures remains uncertain (Guo et al., 2021). Instead, we need a method to classify the functional natures of the dynamical systems that convert the inputs, properties, and antecedent conditions into streamflow. In doing so, we must acknowledge that the models that approximate this functionality are imperfect due to their many assumptions and simplifications. Overall, this complicates the problem of understanding what constitutes a good characterization of the factors that determine similarities and differences in the dynamics of various catchments. Note that this does not mean hydrological models are doomed—it just requires us to relax the underlying assumptions, while keeping the models as simple as possible. Since this has been demonstrated to be a difficult task (Guo et al., 2021), new methods must be explored.

Recent developments in machine learning (ML) have provided a powerful strategy for the modeling of complex input-output relationships, showing good spatiotemporal performance in the geosciences (including hydrology). Models based on the Long Short-Term Memory architecture (LSTM, Hochreiter & Schmidhuber, 1997) have received considerable attention due to their excellent predictive performance (Kratzert et al., 2019a, 2019b, 2019c; Sabzipour et al., 2023). Further, several studies have reported good out-of-sample (generalization) performance for such models (Arsenault et al., 2023; Kratzert et al., 2019a, 2019b, 2019c). For instance, Ma et al. (2021) demonstrated that an LSTM model trained only on US data can be successfully applied to catchments around the world (data from which was not used during training).

Despite this level of predictive and generalization performance, ML-based models (such as LSTMs) can suffer from a relatively low degree of interpretability due the very large numbers of cell-states and parameters used to achieve such performance (Adadi & Berrada, 2018; Xu & Liang, 2021). Some efforts to “interpret” what such models are doing “under the hood” have investigated the sensitivities of model outputs to input perturbations (e.g., Addor et al., 2018; Kratzert et al., 2019a, 2019b, 2019c) or through the use interpretability method as SHAP values (Roth, 1988). However, such methods do not reveal much about the climatic conditions and local characteristics (catchment attributes) under which a particular model could or could not be successfully applied to another catchment. Meyer and Pebesma (2021) were able to define the area of applicability but the hydrological reasons for such areas are still unknown. Moreover, the internal complexity of traditional ML-based models

makes it difficult to identify the essential differences in input-state-output dynamics across catchments, which considerably limits our understanding of the hydrological functioning implemented in such ML models.

One important reason for the excellent performance of ML-based models is that they are trained on large data sets in a manner such that the representation (model architecture) is required to achieve good spatiotemporal generalization performance via a single modeling step. The input-output data (that encode information about catchment dynamics) and the catchment attributes (that encode information about material and geometric properties, etc., of the system) from many locations are used simultaneously to train a single “*regional*” model, with the goal of achieving good spatiotemporal predictive performance across the entire region represented by the data. Using this strategy, Kratzert et al. (2024) showed that a single, general, representation can be achieved that performs better than location-specific models trained individually to each catchment.

Because regional models are trained simultaneously across both space and time, the mechanisms/relationships attributable to each of these dimensions can be difficult to disentangle. Efforts to investigate this issue include Kratzert et al. (2019a, 2019b, 2019c) who reported that while an “*Entity-Aware*” LSTM (EA-LSTM; in which catchment attributes were made available only to the input gates) outperformed common process-based models, an LSTM model trained with attributes made available to all of the gates performed much better. Due to this approach, it remains unclear how the information regarding catchment attributes is exploited by the model to determine the appropriate streamflow response.

Our recent work (De la Fuente et al., 2024) has focused on the issue of (hydrological) interpretability. Therein, we developed a modified LSTM architecture, named *HydroLSTM* (see Section 3.1), that uses a parsimonious representation of the system state (only a few cell-states per catchment) and trainable gate-weights (parameters) to encode information regarding the dominant processes governing the dynamical input-state-output behaviors of a catchment. An interesting finding was that the trained gate-weight sequences (shape generated by the temporal distribution of weights) showed patterns that are sensitive to spatially varying catchment attributes such as aridity. The fact that the pattern (general characteristics present in the data) of gate-weight sequences, which control the models input-state-output dynamics, differ from one hydro-geo-climatic region to another is a good indication that these weights are likely to be correlated with information encoded within the catchment attributes. This in turn should help us to decouple the space/time modeling problem into two parts, one that characterizes the location-specific representation of system dynamics (using *HydroLSTM*) and another that characterizes how these dynamics vary across space (using differences in the weight sequences).

Achieving this decoupling is facilitated by the creation of an intermediate “*latent*” space, whereby the catchment attributes (which vary in space) are used to distinguish between dynamical system properties at different locations. Examples of the latent space strategy are found in image generation, where *Generative Adversarial Networks* (Goodfellow et al., 2014) and Autoencoders (Kramer, 1991) can generate different realizations that are consistent with some underlying data-generating process simply by modifying the values of the (latent) factors that make up this space. In hydrology, Yang and Chui (2023) learned a latent space representation of catchments and hydrological model instances (structural form and parameter value realizations) that enabled them to map between model instances and catchments based on similarity in the latent space, thereby obviating the need for computationally intensive optimization search.

In this study, we approach the problem of hydrologic regionalization by explicitly decoupling the problem into two components so that ML-based architectures can be carefully specified for each component in a manner that is properly suited to the particular task. We use the *HydroLSTM* architecture to represent system input-state-output dynamics at each spatial location, conditional on a classification of the “*type*” of catchment dynamic. The latter classification is achieved by learning a latent space representation, which can then be queried by conditioning on the information provided by each observed catchment, using the *Random Forest* architecture (RF, Breiman, 2001). Both of these component representations are readily “*interpretable*,” which enables us to study (a) how catchment dynamics vary across (cluster in) space and (b) what factors help to explain these differences/similarities, thereby resulting in a better understanding of the nature of the underlying data-generating process.

Accordingly, the main contributions of this work are (a) a meaningful strategy for identifying distinguishable sub-classes within a broad class of dynamical systems, wherein different types of process dominate the dynamics of each sub-class, and (b) an interpretable basis for determining/predicting which type (sub-class) of dynamics can be expected to occur at any given spatial location based on local (static) catchment characteristics. Together these

provide the basis for hydrological regionalization and for understanding the factors contributing to hydrological variability in space and time, and consequently the basis for a possible solution to the prediction in ungauged basins problem (Di Prinzio et al., 2011; Kanishka & Eldho, 2017, 2020). While developed in the context of catchment hydrology, the approach is general and may find applicability in numerous fields.

In Section 2, we provide a high-level explanation of our two-component approach to constructing a hybrid ML-based representation that facilitates interpretability. Section 3 presents specific details of the methodology, discusses the data used for our studies, and explains the experimental design. Section 4 presents the results and interpretations thereof, while Section 5 presents a broader discussion of implications. Finally, Section 6 summarizes our conclusions and provides comments regarding future directions.

2. Two-Component Hybrid-ML Representational Approach

The *HydroLSTM* architecture (De la Fuente et al., 2024) exploits the strengths of the LSTM-based sequence-to-sequence modeling strategy while providing (a) a parsimonious representation of the system state and dynamics and (b) gate-weight sequences that are hydrologically interpretable when trained locally. De la Fuente et al. (2024) showed that a single cell-state version of *HydroLSTM* can achieve acceptable performance when trained locally at any specific location (two or three cells could increase accuracy in some cases) that is similar to that of a standard LSTM having larger numbers of cell-states (locally trained too). This parsimony is achieved (single cell-state), in part, by increasing the representational complexity of the gates. By integrating (convolving) over weighted values of past-lagged input data, the *HydroLSTM* learns how the behaviors of the gates should vary with historical hydrological context and encodes this information as a characteristic pattern in the weight sequences. By shifting the representation of “complexity” from the number of cell-states to time-lagged weighting sequences (convolution filters) in the gates, the *HydroLSTM* gains overall interpretability, because these weight sequences can be shown to encode information about the dominant processes that characterize the dynamics of a catchment. For these reasons, we selected the *HydroLSTM* as the component for representing local system dynamics; more details and figures of the *HydroLSTM* architecture can be found in De la Fuente et al. (2024).

To map between catchment attributes and system dynamics, we require a method that can detect and classify (cluster) locations based on similarity/differences and predict values for the *HydroLSTM* gate-weight sequences associated with each spatial location. Suitable candidates for this include Support Vector Machines (Cortes & Vapnik, 1995), decision trees, and the Random Forest (RF, Breiman, 2001). Of these, we selected the RF architecture due to its demonstrated good performance, and also because of its relative ease of interpretability using techniques such as feature importance, tree analysis, and partial dependence plots. While other ML-based representations, such as LSTM and neural networks might be capable of better predictive performance, their use would make it much more difficult to investigate the major reasons/factors contributing to such performance.

Of course, the resulting gain in interpretability comes at a cost, which is the increased complexity in the joint training of the coupled components of the hybrid model. Whereas the *HydroLSTM* uses backpropagation for optimizing the gate-weights, training of the Random Forest involves trying all possible splitting configurations at each branch of a tree to find the best one, a task that cannot be achieved via backpropagation. Further, inherent uncertainty associated with the training of the *HydroLSTM* gate-weights (due in part to random initialization) can complicate the task of training the RF model.

To overcome this training difficulty, we developed an iterative sequential learning approach (Figure 1; details in Section 3.4) that is, in a sense, analogous to expectation maximization (Dempster et al., 1977). The *HydroLSTM* component is first trained locally (catchment by catchment) to find gate-weight sequences that, due to equifinality, may be far apart in the parameter space even though corresponding to similar input-state-output dynamics. This reflects that each catchment data set does not, on its own, provide enough information to properly constrain (regularize) the inference problem such that similar gate-weight sequences correspond to similar system dynamics. The clustering process, provided by the RF component is used to counter this tendency toward equifinality, by replacing the gate-weight sequences at every catchment location associated with an RF cluster (leaf node) with their expectation taken over all members of that cluster (Figure 1a).

In practice, we train each catchment-specific *HydroLSTM* for a small number of epochs using backpropagation (starting at some randomly initiated values for the gate-weights), train an RF model to predict the gate-weight

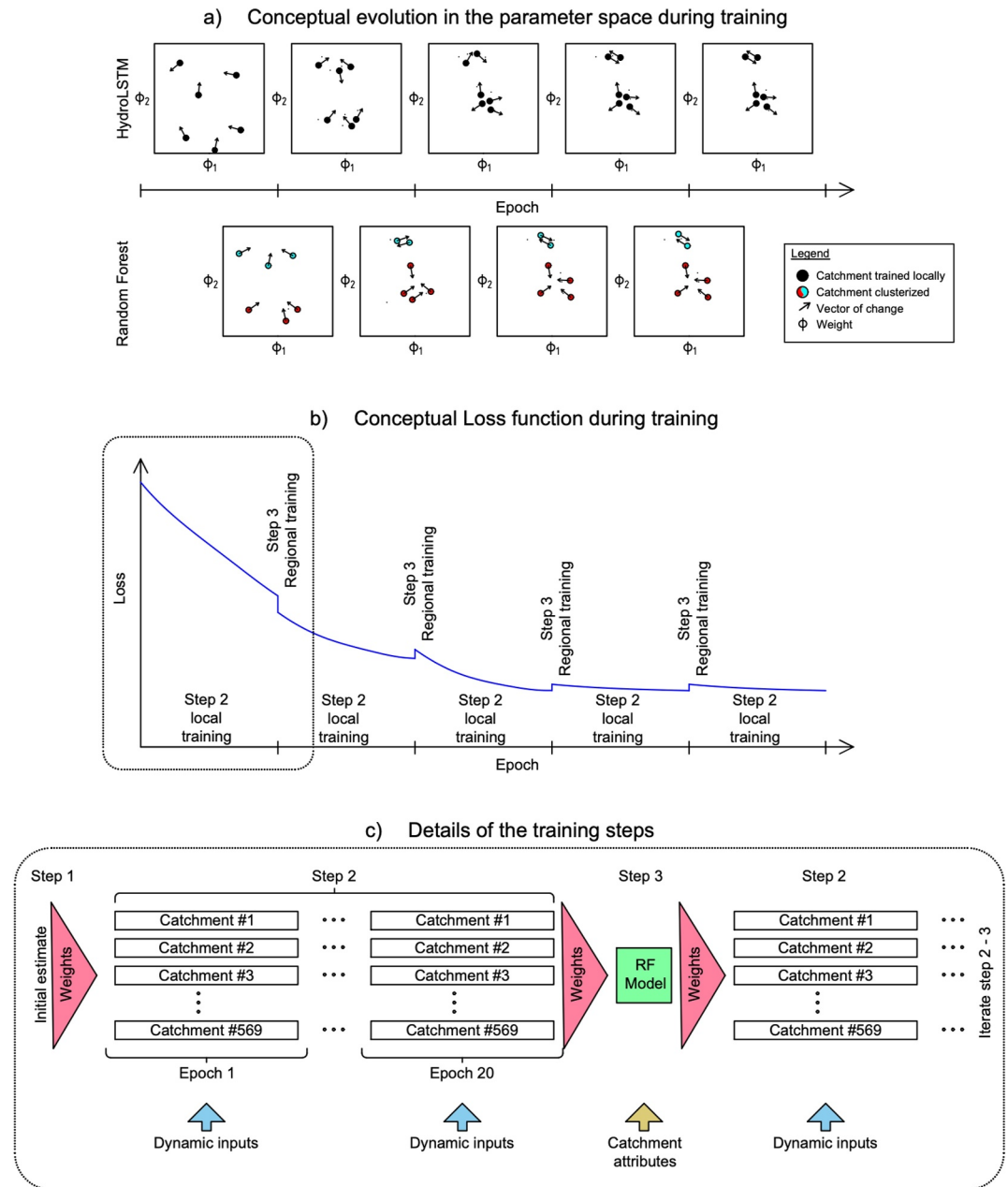


Figure 1. Schematic representation for the training of the *regional HydroLSTM*. (a) Graphic representation of the divergent behavior when the weights are trained locally versus grouping when they are trained regionally. (b) Loss evolution during local and regional training. At the beginning of the training, the regional training decreases the loss. At the beginning of the next training stage, the loss increases (due to being over constrained). (c) independent training of each catchment during local training (parallelism not implemented). Weight sequences are shared only within the RF model.

sequences from catchment attributes, take the expectations of these sequences for each catchment cluster determined by the RF, use those expectations to initialize the weights for a subsequent round of *HydroLSTM* training, and iterate the overall process to convergence. Our results (Section 4) demonstrate that the predictive loss function progressively improves as shown in Figure 1b, and that the gate-weight sequences tend toward clusters indicating similarity between catchments. In this sense, the RF acts as a filter to reduce equifinality and learns a mapping that enables regionalization of the *HydroLSTM* model.

3. Method

This section outlines the methods used in this study. The hybrid ML model combining *HydroLSTM* and RF components is referred to as the *Regional HydroLSTM*.

3.1. Brief Description of the *HydroLSTM* Architecture

The architecture of the *HydroLSTM* is a modification of the original LSTM (Hochreiter & Schmidhuber, 1997) that incorporates two main changes. In the original LSTM, the four gates (f, g, i, o) that operate on the cell-states (c) are informed by the system input data from the current time step and by the LSTM prediction (h) made at the previous time step. To account for the time-sequential history of system inputs when estimating the cell-state at the current time step, the LSTM recursively processes a pre-determined sequence length of the past time series data of those inputs.

In contrast, the gates of the *HydroLSTM* (De la Fuente et al., 2024) are provided with access to a pre-determined sequence length of the time series of those past forcing values weighted by unique values applied to each day (effectively a fixed pattern of attention weights). The temporal pattern of each such gate-weight sequence describes a characteristic shape that is interpretable. As in the LSTM, the gates also incorporate the prediction from the previous time step. The *HydroLSTM* cell-state is continuously updated from the beginning to the end of the entire available time series of system inputs in a manner similar to conceptual hydrological models. This enables the cell-state to track the state of the system over its entire (available) history. The reason for this modification to the gating architecture is that it enables the model to directly learn which input forcings to attend to (are important/relevant) when making the current step prediction by adjusting the weights assigned to those forcings. For more details about the LSTM and *HydroLSTM* architectures, please refer to the original paper.

3.2. Brief Description of the *Random Forest* Architecture

The Random Forest architecture (Breiman, 2001) is based on the use of decision trees. A decision tree is a set of rules that progressively bifurcates the input space in a hierarchical manner. At each level, the threshold used to further split the space is estimated by analyzing all the possible splits and selecting the one that maximizes target similarity within each group. This process continues until some pre-specified maximum number of splits has been performed, or when each subspace contains a pre-specified minimum number of elements. Because this process can be biased by the particular realization of data used, the Random Forest constructs an ensemble of such trees, where each tree is provided with access to a random subset of input variables (a procedure known as bootstrap aggregation, or bagging), such that decisions made using the ensemble are considerably less biased. The method of selecting split thresholds to maximize target similarity acts to group the inputs into clusters that generate similar outputs, making this architecture a suitable candidate for regionalization based on catchment attributes.

3.3. Description of the Hybrid *Regional HydroLSTM* Architecture

As mentioned earlier, the hybrid *Regional HydroLSTM* architecture achieves a reasonable balance of predictive performance with high interpretability by separating the tasks of “spatial” and “temporal” generalization into two components. Specifically, the *HydroLSTM* architecture (Section 3.1) learns a representation of system dynamics, while the RF architecture (Section 3.2) learns a representation of the regionalization relationship. This is made possible by the fact that *HydroLSTM* can provide good predictive performance when trained locally catchment-by-catchment while being sufficiently parsimonious that a relatively high level of interpretability can be achieved by examining the weighting sequences learned for the gates (see De La Fuente et al. (2024) for more details).

The RF component exploits this fact to discover regional relationships between those weighting sequences and observed catchment attributes. To do so, it uses data from all of the available catchments simultaneously to learn a decision tree ensemble that predicts (as output features at each location) values for the weight sequences of the *HydroLSTM* gates from the observed catchment attributes (input features). Accordingly, the RF approach constructs “clusters” of catchments within which the weight sequences have similar shape, resulting in a higher degree of interpretability than is typically achieved via other ML approaches. As mentioned earlier, biases in the individual decision trees comprising the “Forest” tend to cancel when ensembles are created using different sampled subsets of the inputs (Breiman, 2001). The RF component was implemented using the scikit-learn library (Pedregosa et al., 2011).

3.4. Method for Training Regional HydroLSTM Architecture

The training steps for the model development proceed as follows (Figure 1c):

- A. Initial Training of the *HydroLSTM* Model: For each available catchment, we determine the “optimal” single-cell-state *HydroLSTM* architecture as follows
 - a. *Determine the Optimal Number of Lags (past number of days)*: First, we train 20 randomly initialized single-cell-state *HydroLSTM* models for each case of $N_{Lag} = 4, 8, 16, 32, 64, 128, 256, \text{ and } 512$ days (eight possibilities), where N_{Lag} is the number of lagged data time-steps used by the *HydroLSTM* gates, for a total of $20 \times 8 = 160$ trained models. Each model is trained for 512 epochs. From this set, we identify the model with the best “*selection period*” (commonly called validation period in data science) performance and choose the value of N_{Lag} associated with that model to be the “*optimal*” number of lags (N_{Lag}^{Opt}).
 - b. *Determine an Initial Estimate of the Gate-Weight Sequences*: Next, we take the 20 trained models associated with $N_{Lag} = N_{Lag}^{Opt}$ in the previous step. For each trained model, if $N_{Lag} < 512$, we pad out its precipitation and potential evapotranspiration gate-weight sequences with zeros so that each of those sequences (whether associated with precipitation or potential evapotranspiration) now consists of 513 values (where $Lag = 0$ corresponds to the current time step). By concatenating these sequences, including the 513 lagged-precipitation weights, the 513 lagged-potential-evaporation weights, plus the weight associated with the predicted flow value at the previous time-step (thereby providing indirect information about the “state” of the system) plus the constant bias term, we obtain an overall sequence of 1,030 gate-weight sequence values ($513 + 513 + 2 + 2$) for each of the 20 models. We then average these gate-weight sequences over the corresponding 20 models and pass those averaged gate-weight sequences (one sequence per gate per catchment) to the next step.
- B. Retraining of the *HydroLSTM* Model: For each catchment, we retrain its corresponding *HydroLSTM* model for 20 epochs, using its corresponding cluster-averaged gate-weight sequences obtained in the previous step for model initialization. We pass the re-trained weight sequences so obtained (from epoch 20) to the next step.
- C. Weight Sequence Regionalization Using the *Random Forest* Algorithm: Construct the data-matrix of inputs (17 catchment attributes) and outputs (Step 2 estimates for the weight sequences) required to train the RF decision tree ensemble (Section 3.2). After training the RF model with bootstrapping, use the 17 observed attributes for each catchment to generate RF-based predictive estimates of the (retrained) gate-weight sequences, and pass these estimates to the next step.
- D. Iterate to Convergence: Repeat Steps Two through Three several times, using the predictive estimates of the (retrained) gate-weight sequences obtained via Step 3 as the initial estimates for Step Two. In our experiments, we found that 15 iterations of Steps 2 and 3 led to stable results.

The final result is a coupled *Regional HydroLSTM* model that uses the RF procedure to generate predictive estimates of the gate-weight sequence values for each individual-catchment *HydroLSTM*-based model based on its catchment attribute data. Accordingly, the RF component performs the “*regionalization*” function of the model, while the *HydroLSTM* component models the input-state-output dynamics at each catchment.

Our choice to use the RF architecture for catchment regionalization, rather than a Neural Network architecture as is more common (Botterill & McMillan, 2023; Feng et al., 2022; Kratzert et al., 2019a, 2019b, 2019c; Toth, 2013), is to take advantage of the ease by which the RF results can be interpreted, thereby providing insights into how catchment properties contribute to the estimates the weight sequence for each catchment.

3.5. Some Algorithmic Details

For simplicity of exposition, some details were omitted in Section 3.4. One of these is the linear “head” layer commonly used at the last stage of a neural network. The *HydroLSTM* uses a linear regression layer with two parameters per catchment (slope and intercept) that are used to project the output of a *HydroLSTM* neuron into the target space (streamflow). Given that those parameters are specific for each catchment they must also be regionalized. For that purpose, an additional RF model was trained on Step 3 using the same procedure and inputs as described in Section 3.3.

Further, given that *HydroLSTM* is trained locally for each catchment, the inputs and outputs are normalized locally, using “*mean normalization*” (value minus mean divided by the range), where the values of the mean, minimum, and maximum for each catchment are stored for later use to de-normalize the prediction. While this

approach differs from other regional models, we want the model to discriminate between catchments only through the weight sequences, so as to maximize the interpretability of the model.

Training of the *HydroLSTM*-based models for each catchment was performed using the SmoothL1 loss function (Girshick, 2015), which behaves as an L1 norm for errors higher than 0.5 and as an L2 norm for errors lower than 0.5 (meaning it is less sensitive to outliers). Training of the RF regionalization component used the L2 norm to determine the optimal decision tree splits. The KGE metric (Gupta et al., 2009) was used for the selection of the best *Regional HydroLSTM* model and as the basis for model performance evaluation in the performance analysis presented later.

3.6. Data Set

The models in this study used three types of data, meteorological forcings, catchment attributes, and observed streamflow. Several suitable data sets are available, including CAMELS (Addor et al., 2017b), CAMELS-CL (Alvarez-Garretton et al., 2018), and CARAVAN (Kratzert et al., 2023), etc.

We chose CAMELS for the experiments reported herein because it has been widely used for similar investigations in the USA. The CAMELS data set consists of five dynamic variables and 56 catchment attributes, available for 671 catchments across the Continental US (CONUS). For consistency with our previous study (De La Fuente et al., 2024), we used the meteorological forcings provided by the Maurer data set (Maurer et al., 2002) available within CAMELS. This data runs from 01/01/1980 to 31/12/2008 for 671 catchments. Of these 671 catchments, only 569 have data available for consistent partitioning of the data into training, selection, and evaluation periods (see Section 3.7 below). Therefore, only the data from these 569 catchments was used.

To maintain a relatively parsimonious *HydroLSTM* model, we used only daily precipitation and potential evapotranspiration (calculated using the Hargreaves and Samani (1985) equation) as input meteorological forcings. This decision helps to ensure that the model inputs tend to be relatively independent, which would not happen if (e.g.) temperature, radiation, and vapor pressure were used as input drivers.

From the 59 available catchment attributes, we selected only the 50 having numerical values. Because our goal is to “learn” the dynamics of the system via the *HydroLSTM* architecture, we then further dropped those associated with forcing or streamflow dynamics (such as the average duration of high precipitation events, and streamflow precipitation elasticity), leaving only 29 attributes. Finally, we removed attributes that are used to calculate others, or that are highly correlated (thereby containing equivalent information), such as aridity (annual mean precipitation/annual mean potential evapotranspiration), latitude, and longitude, leaving the 17 catchment variables presented in Table 1.

3.7. Data Splitting Used in Model Development

The data set was partitioned into three consecutive periods for model training, selection, and evaluation:

- *Training*: The model training period consists of ~19 water years, from 10/29/1981 to 30/09/2000. Several days prior to 10/29/1981 were used to construct the lagged data structures required by the *HydroLSTM* gates to predict streamflow for the first day in the training period.
- *Selection (validation)*: The selection period consists of four water years, from 10/01/2000 to 09/30/2004. This period was used to select the value of N_{Lag} in Section 3.3.
- *Evaluation (testing)*: The evaluation period, running from 10/01/2004 to 12/31/2008, is never used in the model development (training and selection) process. Unless otherwise mentioned, all performance results reported in this paper are based on this period.

3.8. Experimental Design

We designed three experiments to explore the weight sequences that *HydroLSTM* learns. Moreover, we explored the connections between those sequences and the catchment attributes when we trained a *Regional HydroLSTM*. The experiments are summarized as follows:

- *Experiment 1*. *HydroLSTM* is trained locally over each of the catchments to extract the weights as described in Step One, Section 3.4. The goal of the experiment is to demonstrate the existence of “weight sequences” and their relationship with attributes.

Table 1
Catchment Attributes Selected From the CAMELS Data Set

Classification	Attribute and abbreviation
Topography	<ul style="list-style-type: none"> • Mean elevation (elev_mean) [m] • Mean slope (slope_mean) [m/km] • Area (area_gage2) [km²]
Climate indices	<ul style="list-style-type: none"> • Mean annual precipitation (p_mean) [mm/day] • Mean annual potential evapotranspiration (pet_mean) [mm/day]
Land cover characteristic	<ul style="list-style-type: none"> • Forest fraction (forest_frac) [unitless] • Maximum monthly mean of the leaf area index (lai_max) [unitless] • Difference between the maximum and minimum monthly mean of the leaf area index (lai_diff) [unitless] • Maximum monthly mean of the green vegetation fraction (gvf_max) [unitless] • Difference between the maximum and minimum monthly mean of the green vegetation fraction (gvf_diff) [unitless]
Soil characteristic	<ul style="list-style-type: none"> • Depth to bedrock (soil_depth_pelletier) [m] • Saturated hydraulic conductivity (soil_conductivity) [cm/hr] • Mean sand fraction (sand_frac) [%] • Mean clay fraction (clay_frac) [%] • Fraction of the top 1.5 m marked as water (water_frac) [%]
Geological characteristic	<ul style="list-style-type: none"> • Fraction of the area characterized as “Carbonate sedimentary rocks” (carb_rocks_frac) [unitless] • Subsurface permeability (log10) (geol_permeability) [m²]

- *Experiment 2. Regional HydroLSTM* is trained following Steps 2–4. The goal is to show that by coupling the RF with a simple ML architecture such as *HydroLSTM*, we can discover relationships between the attributes and catchment dynamics. For that, we first evaluate the overall performance to ensure the representativeness of the results by use of the Cumulative Density Function (CDF), which allows us to compare distributions of performance.
- *Experiment 3.* By using the weight sequences learned by the RF component of the *Regional HydroLSTM*, we extract the best single tree from the RF model (we name this the reduced RF model) for the specific goal of simplification to obtain insight into the nature of the dynamics encoded by the weights. Two types of reduced models are used in this experiment:
 - A reduced RF model that targets individual weights of the 1,030 weights learned by the *Regional HydroLSTM* model. This allows us to analyze how the feature importance changes between the weights.
 - A reduced RF model that predicts only the four weights that capture most of the variability (days with higher range and more distinguishing shapes). This model is used to explore the overall decisions and clusters learned by the *Regional HydroLSTM* model.

4. Results

The results of the experiments are presented in four stages. First (Section 4.1), we examine the gate-weight sequence obtained initially in Step 1 (Section 3.3) when the *HydroLSTM* model is trained “locally” at each catchment. This analysis shows that the weights sequences associated with the *HydroLSTM* gates exhibit (even before the RF-based regionalization procedure is implemented) spatial distribution patterns that are consistent with hydrologic understanding. In a sense, these represent “prior knowledge,” established before implementing the regionalization procedure that learns from the entire catchment data set.

Next (Section 4.2), we compare the “locally” and “regionally” trained model results, to ensure that the latter maintains a similar level of skill to the former. We also compare the results against a well-vetted, and state-of-the-art performance benchmark.

Then (Section 4.3), we analyze the relationship between catchment attributes and sequence weights that are learned by the *Regional HydroLSTM* model. This exploration helps us to understand what the RF-based regionalization can reveal about how hydrological concepts are encoded into the gate-weight sequences, and its relationship with catchment attributes.

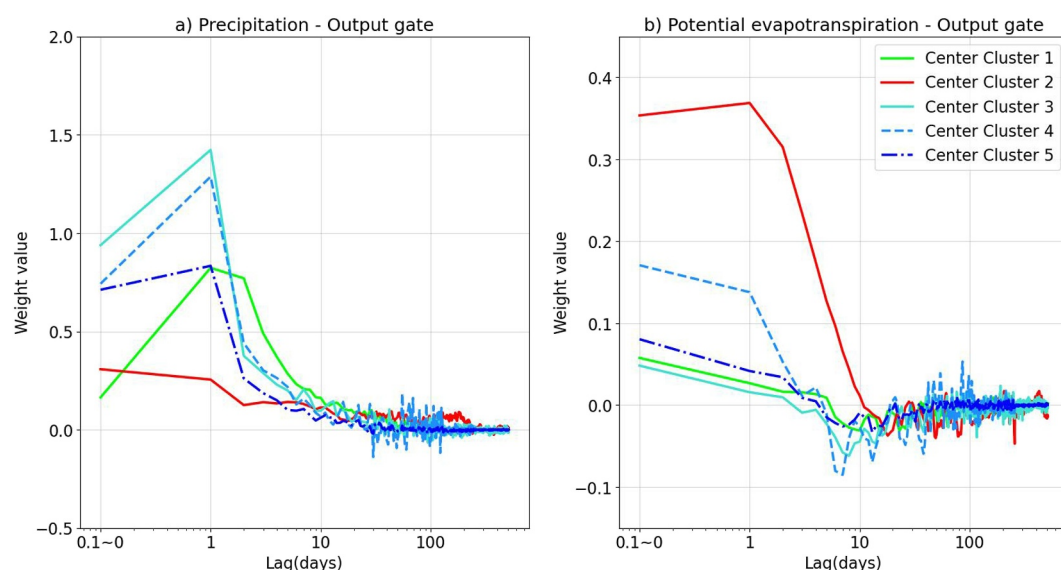


Figure 2. Averaged weight sequences for each of the 5 clusters identified using the K-means algorithm applied to the output gate weight sequence. Even though each catchment was trained locally (local model), common shapes are present. Red color could be associated with snow-dominated catchments, blues with recent rainfall-dominated catchments, and green with historical rainfall-dominated catchments (De La Fuente et al., 2024; Jiang et al., 2022).

Finally (Section 4.4), we analyze the behaviors of the *Regional HydroLSTM* cell-states and associated output gates for each catchment, to examine what this can reveal about catchment dynamics and corresponding interpretability of the regionally trained *HydroLSTM* models.

4.1. Evidence That the Gate-Weight Sequences of Locally-Trained *HydroLSTM* Models Encode Hydro-Climatically Relevant Information

We examine the initial gate-weight sequence obtained in Step One (Section 3.3) when the *HydroLSTM* model is trained “locally” at each catchment, to explore whether the weights present in the *HydroLSTM* model (prior to regionalization) follow a spatial distribution that is consistent with our hydrologic understanding.

After training the *HydroLSTM* model locally at each of the 569 catchments in the data set (Section 3.3, Step 1), we applied the K-Means algorithm (MacQueen, 1967) to all the 569 forcing learned gate-weight sequences per catchment (consisting of $4,104 = 4 \times 1,026$ weights associated with the 4 gates \times 513 weights \times 2 variables). We grouped them into five clusters having statistically distinguishable weight sequences. To select the optimal number of clusters, we used the Silhouette score (Rousseeuw, 1987). For each cluster, we then average the associated weight sequences across catchments in that cluster and display these averaged weight sequences in Figure 2.

The results show clear characteristic shapes that distinguish the hydro-dynamical behaviors associated with each cluster. For example, Cluster 1 (green line) is associated with larger positive non-zero weight values for precipitation at time lags 0 to 5 (Figure 2a), indicating significant influence of the current and recent precipitation on the operation of the output gate, but potential evapotranspiration has only a low influence (relatively small weight values; Figure 2b), indicating the dominant role played by available water over available energy. In contrast, Cluster 2 (red line) shows a significant influence of the past 10 days of potential evapotranspiration (Figure 2b) but with a relatively low impact of precipitation (Figure 2a), indicating the dominant role played by available energy. Clusters 3 and 5 (different shades of blue) have behaviors similar to Cluster 1 but with different and somewhat stronger degrees of reactivity to the strength of current and recent precipitation. Finally, Cluster 4 (turquoise) is associated with the relatively strong influence of both available water (current and recent precipitation) and available energy (current and recent potential evapotranspiration).

Overall, the differences in weight sequence patterns indicate that catchments in the different clusters react differently to the same hydroclimatic (input) forcings. Note that catchments with higher weight values associated

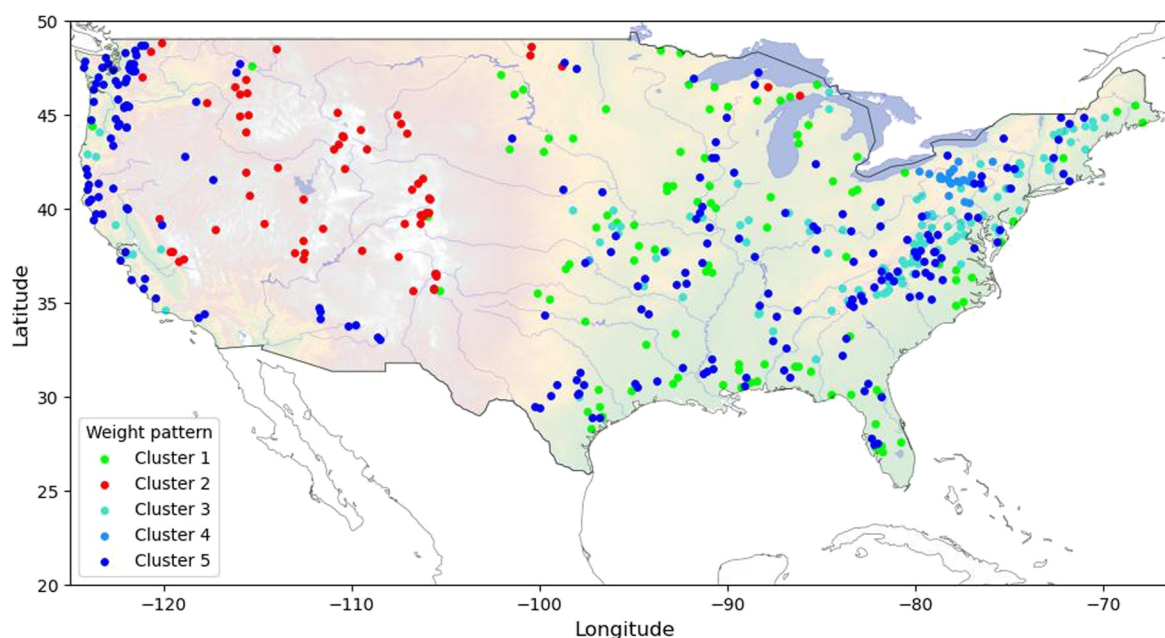


Figure 3. Spatial distribution (local model) of the catchment clusters obtained by using the K-means algorithm to group the weight sequences obtained by training the *HydroLSTM* locally at each catchment. The results indicate that the clusters are associated with hydro-geo-climatic regions having hydrologically similar behavioral shapes.

with lag zero and lag one precipitation can be considered to be quickly reacting, indicating a “flashier” runoff response in those catchments.

Certainly, given that a Silhouette score close to zero was obtained when defining the number of clusters, apparently, there is considerable overlap between the clusters, indicating that the behaviors depicted in Figure 2 are not unique, and combinations can be present in each catchment, as is commonly observed in hydrologic signatures. Nevertheless, the existence of such weight sequence is indicative that information regarding underlying hydrological processes is being encoded into them. Additional figures presented in Supporting Information S1 reveal that distinct weight sequences are also present in the forget and input gates, providing support for the hypothesis that such sequences can be treated as hydrological informative.

Further support for this hypothesis is presented in Figure 3, where we see that the spatial distributions of catchments associated with the five clusters are associated with different geo-hydro-climatic regions. Cluster 2 (red dots) is associated with the Rocky Mountains where available energy dominates the process of streamflow generation via snow accumulation and melt. This behavior is consistent with the dominant influence of “potential evapotranspiration” (an informative surrogate for available energy) in that region (Figure 2b). Clusters 3, 4, and 5 (blue dots) are associated with the Appalachian Mountains (East) and Cascade Range (West), where the catchments have steep slopes and streamflow generation is “flashy” and predominantly associated with liquid precipitation, consistent with the weight sequence shown in Figure 2a. Finally, the catchments associated with Cluster 1 (green dots), while being more geographically dispersed, are generally located in the Great Plains and the Coastal regions, where catchments respond less quickly to precipitation (Figure 2a).

This analysis suggests that the learned *HydroLSTM* model weights encode information that is consistent with our expectation of the spatial distribution of hydrological behaviors. Of course, the evidence presented so far is only “suggestive,” as other factors that were not considered here will also play roles in streamflow generation. Further investigation would require analysis of regional models (presented in the next section) in which catchment attributes serve as surrogate variables for similarity. This way a regional model can manage a deeper and more exhaustive search for dominant hydrological signatures associated with different geospatial regions.

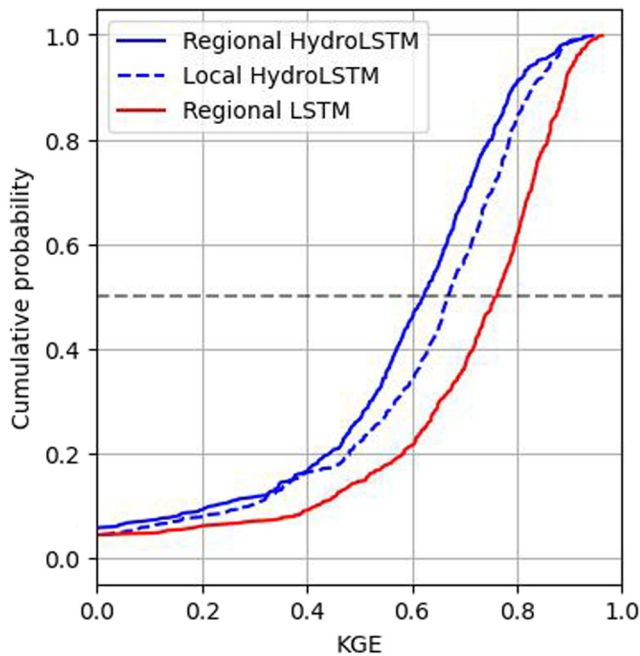


Figure 4. Cumulative density function of evaluation period KGE performance for the 569 catchments. Due to having only one cell-state, the *Regional HydroLSTM* is not able to surpass the *Regional LSTM* in terms of performance. However, other benefits exist that are present in the next section.

4.2. Overall Performance of the *Regional HydroLSTM* Model

Our overall goal is to understand the spatial distribution of dynamical behaviors determined by the *Regional HydroLSTM* model. However, before we do this, we evaluate its performance against a standard benchmark model to ensure sufficient representativeness of the results. For that, we examine the evaluation period KGE performance of the trained *Regional HydroLSTM* model obtained at the end of Step Five (Section 3.3). As a benchmark basis for comparison, we use the state-of-the-art *Regional LSTM* model reported and shown by Kratzert et al. (2019a, 2019b, 2019c) to compare well with traditional lumped process-based models. To ensure consistency and fairness, we retrained their *Regional LSTM* using the same training period data used in this study (Section 3.6), using the hyperparameters reported in their paper, and using their published *NeuralHydrology* library (Kratzert et al., 2022).

The CDFs of KGE metrics for each case are presented in Figure 4—the locally trained *HydroLSTM* models (blue dashed line), the *Regional HydroLSTM* model (blue solid line), and the *Regional LSTM* model (red line, external benchmark). The local *HydroLSTM* and *Regional HydroLSTM* (both single cell-states) models do not perform as well as the benchmark multi-cell-state *Regional LSTM* model. This is to be expected given the lower capacity of the *HydroLSTM*-based models due to their considerably lower complexity (only one cell-state). Further, the median KGE performance for the *Regional HydroLSTM* model is approximately 0.05 worse (median) than for the locally trained *HydroLSTM* models, indicating that solving the regionalization (spatial generalization) problem is more difficult than simply learning the local dynamics of each catchment (temporal generalization).

To place this in context, it is common for regionalized hydrological models to achieve lower performance than locally trained models, because it is easier to learn a location-specific dynamic relationship than to learn a spatially generalizable one. The reasons for this are many, but an important one is that the total information content in our data set is much denser with respect to time than space. For instance, a regional model has as many as 6,911 days of data per catchment from which to extract a temporal function, while having only 569 different catchments from which to extract a spatial relationship. Another reason is that the catchment attributes may not contain sufficient information to successfully characterize the spatial variations associated with catchment dynamics. That happens because these aggregated variables are, in general, created to be humanly “interpretable,” meaning that they do not preserve all the possible information available. Moreover, there is an important amount of information about our catchments that is not encoded by the static attributes, therefore limiting what can be achieved via regionalization.

The approximately 0.15 lower KGE performance of the *Regional HydroLSTM* model compared with the *Regional LSTM* model can also be understood as indicative of considerable room for further improvement by increasing its “capacity.” Recall that in the development of the *Regional HydroLSTM* architecture we sacrificed model capacity (fewer degrees of freedom, represented by the number of cell-states, and fewer input variables) to gain enhanced interpretability. The *Regional HydroLSTM* uses only one cell-state, versus 256 used by the *Regional LSTM*. Further, the *Regional LSTM* model is provided with access to maximum and minimum temperatures, while the *Regional HydroLSTM* is provided with potential evapotranspiration estimates derived from those temperatures. Arguably, the extra model capacity enabled by access to more input variables and cell-states allows the benchmark *Regional LSTM* model to potentially learn more complex relationships than the more parsimonious *Regional HydroLSTM* architecture.

This argument suggests that the capacity, and hence performance, of the *Regional HydroLSTM* could be enhanced by increasing the number of cell-states. Because this study is focused on insights that can be gained by maintaining ease of interpretability of the regional model with a reasonably good performance, we leave such investigation for future research. Note that augmentation of model capacity would have to be pursued with care, as higher performance would most likely be accompanied by increased levels of equifinality associated with the

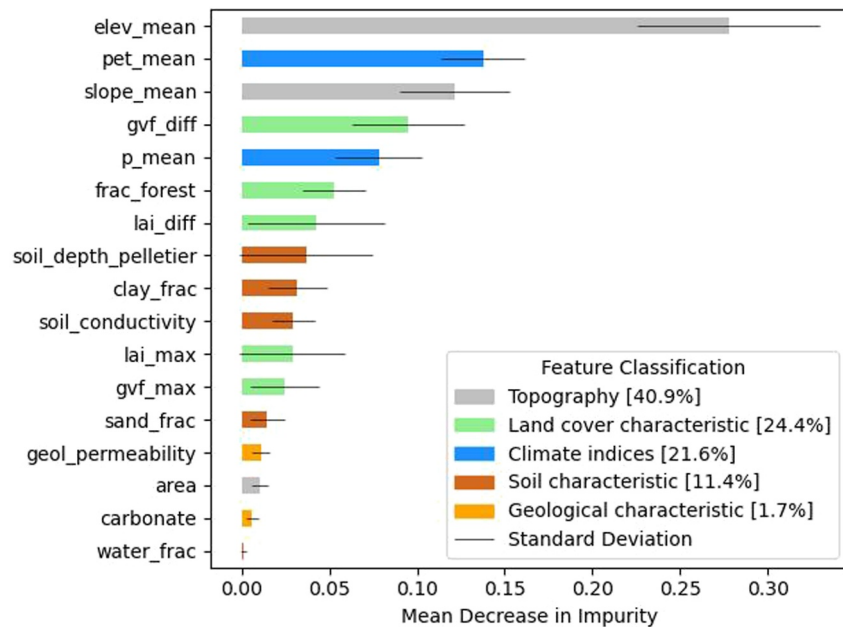


Figure 5. Overall feature importance (regional *HydroLSTM*) as determined by the RF regionalization model. Topographic features dominate the overall clustering of weight sequences, which is consistent with how important elevation and slope in many hydrological processes are.

learned weight sequences, and consequent decline in degree of interpretability. For now, however, we are interested in the insights gleaned from using a minimal number of maximally interpretable components.

4.3. Evaluating Spatial Distribution in the Regional *HydroLSTM* Model

We examine the “*feature clusters*” that are learned by the *Regional HydroLSTM* model. Our goal is to understand what the RF-based regionalization component of the model can tell us about how hydrological concepts are encoded into the gate-weight sequences, and how these sequences at each catchment are related to the associated catchment attributes. This analysis takes advantage of interpretability tools that are available for the RF architecture; in particular, we make use of “*feature importance plots*,” “*partial dependence plots*,” and a visualization of the “*best tree*” (Molnar, 2022).

4.3.1. Assessing the Regional Dependence of System Dynamics on Catchment Attributes

The regionalization component of the *Regional HydroLSTM* model has been trained to generate spatially varying predictive estimates of the gate-weight sequence at each catchment based on observed data regarding the 17 catchment attributes at that location. Because this component uses an RF-based decision tree architecture, we can use Gini Importance (Gini, 1997) to examine the relative importance of each catchment attribute in determining the nature of the gate-weight sequence. This feature importance method is intrinsic to the algorithm because it is evaluated as the change in variance during the splitting selection (training). This makes it an easy and fast option to analyze the importance.

The results of the feature importance analysis are presented in Figure 5. The plot shows that by far the most important attribute is “elevation,” which is consistent with the fact that the catchment clusters shown in Figure 3 (which were generated without a regionalization method) correlate well with topographic elevation across the US. Arguably, this should not be surprising given that the dominant controls imposed by various hydro-climatic variables (such as precipitation intensity, mean temperature, wind speed, relative humidity, etc.) correlate strongly with geospatial coordinates (including elevation, latitude, and longitude), which has been shown to be a good segmentation variable for explaining dominant surface-level meteorologic covariations (e.g., Minder et al., 2010). While “*elevation*” has, in general, no direct hydrological meaning per se, it can serve as a surrogate attribute that is informative about how the complexity of hydro-climatic processes can be related to geospatial

location. Further, hydrologically relevant land surface properties such as land use, soil type, and geological characteristics tend also to be strongly correlated with elevation. However, if elevation were to somehow change (e.g. due to subsidence or geological uplift) this would not necessarily result in corresponding changes to hydrologically relevant variables such as soil type etc. Accordingly, this “*high importance*” assigned to “*elevation*” by the RF-based model should be treated (interpreted) with caution.

An interesting (perhaps surprising) finding, indicated by Figure 5, is that whereas the attributes related to land cover (24.4%), climatology (21.6%), and soil type (11.4%) are assigned moderate to high levels of importance, the attributes related to geological characteristics (Figure 5, legend) are assigned relatively low importance (1.7%). A possible explanation is that land cover, climatology, and soil type tend to more directly affect the high-frequency (quick-response) dynamics of streamflow, whereas geological characteristics tend to be more directly associated with the lower-frequency longer-term dynamics associated with baseflow recessions and regional groundwater configurations. So, while geological attributes can be expected to play an important role in determining baseflow recession dynamics and should therefore be expected to be an “*important*” variable for many catchments, the fact that the RF-based analysis did not detect this relationship could (perhaps) instead be an indication of poor informativeness (relatively high uncertainty) in the data related to such variables. While the low feature importance of an attribute could be an actual consequence of its low relative importance, it could also arise due to high relative uncertainty in the data associated with that attribute. Another reason could be associated with the loss function used in the training, which could be highlighting high streamflow values and, in consequence, attributes related to them. Arguably, a concerted effort should be devoted to disentangling the long-term effects of groundwater dynamics from the shorter-term quick-response dynamics associated with precipitation and evaporation responses of the streamflow, and how these different dynamics are related to observable catchment attributes.

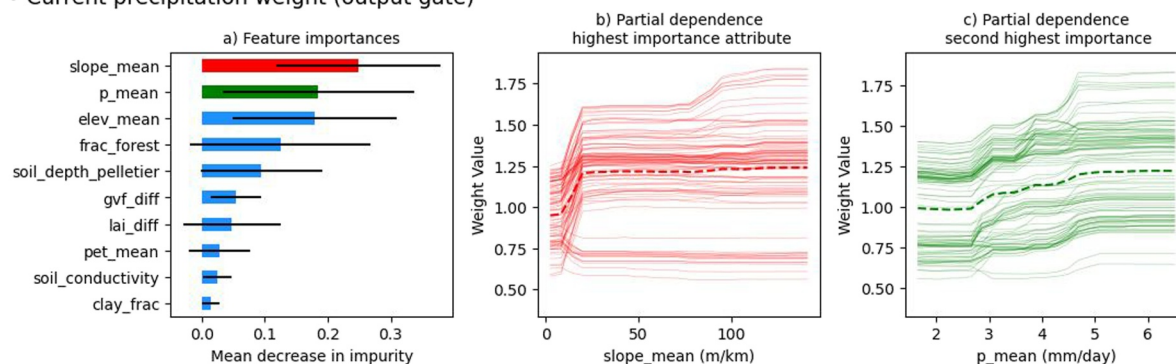
In this regard, it should be noted that the *feature importance* estimates for many of the attributes in Figure 5 have relatively large uncertainties (standard deviations; indicated by black lines at the centers of the bars), indicating that the estimated importance can change drastically between different trees in the forest. One possible reason is that there is a high degree of covariance (shared importance) between attributes, meaning that when one attribute is selected first it has high importance and another one has low, and vice-versa. Another possible cause is differences in the importance of the attributes over different catchments, but given that the number of catchments (569) far exceeds the number of the attributes (17), the removal of an attribute during the creation of a tree (due to the bagging operation) will have a larger impact on feature importance than the removal of catchments. Accordingly, the high uncertainty (variability in importance over trees in the ensemble) is likely mainly due to collinearity (shared information) in the attributes. This suggests that pruning the set of attributes would be beneficial, which we explore next.

4.3.2. Pruning as an Interpretability Tool

We examine a reduced number of attributes (inputs) to understand feature importance for specific weights of the output gate. The goal here is to simplify the relationship between attributes and weights to provide clearer insights into the nature of the regionalization relationship learned by the RF-based model. For this analysis, we selected the “*most relevant*” weights learned by the *Regional HydroLSTM* model and then trained separate RF models (called reduced RF models) to predict the values of each of those weights (one at a time) using the reduced set of attributes. From Figure 2, we infer that the weights having relatively high variability and differentiation between weight sequences correspond to precipitation from the current (lag 0) and potential evapotranspiration from immediately previous (lag 1) days. Further, Figure 5 suggests that seven of the attributes (*area*, *lai_max*, *gvf_diff*, *sand_frac*, *water_frac*, *carbonate*, and *geol_permeability*) can likely be removed without significant deterioration in model performance. With this reduced subset of two weights and ten attributes, we constructed “*reduced RF*” models to predict weights from attributes.

Figure 6 shows plots of feature importance (left column) and plots of “*Partial Dependence*” for the two aforementioned weights (figures for other weights are presented in Text S3 in Supporting Information S1). Note, from Figures 6a and 6d, that the sets of “*most important attributes*,” and their order, are different from the weights shown in Figure 5, indicating sensitivity to different attributes. This differing sensitivity is desirable, given that we seek to use the attributes to explain the differences in gate-weight sequences (hydrological signature) in the *HydroLSTM* model.

• Current precipitation weight (output gate)



• Previous PET weight (output gate)

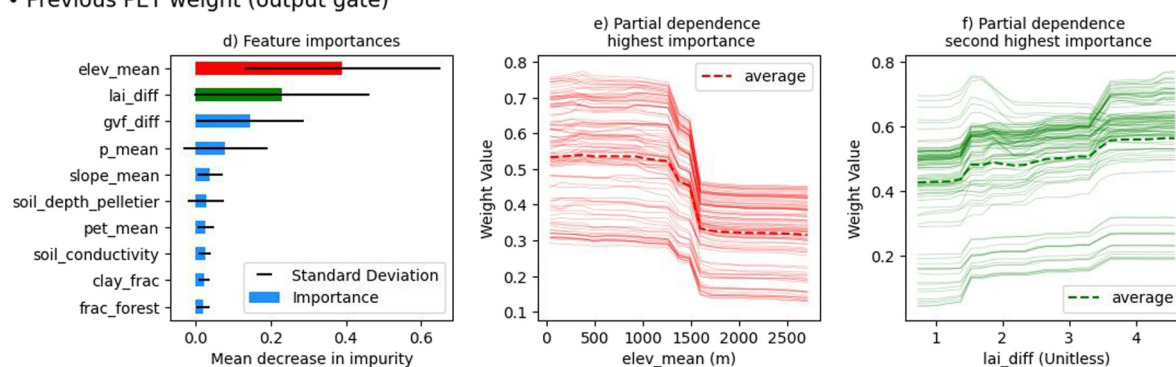


Figure 6. Feature importance and partial dependence (*Regional HydroLSTM*) plots for two selected weights of the output gate. Panels (a), (b), and (c) for current Precipitation weight. Panels (d), (e), and (f) for previous Potential Evapotranspiration weight. The RF model captured different relationships for the weights, which highlights the ability of the weight sequence to capture different hydrologic signatures.

These plots indicate that the weight of current precipitation is strongly dependent on *mean slope* and *daily mean precipitation*. Figure 6b indicates that the weight on current precipitation (associated with “flashiness”) increases rapidly with slope until 20 m/km, and thereafter changes only slightly, indicating that other factors are relevant for catchments with higher slopes. This is consistent with Saharia et al. (2021), who reported a positive correlation between flashiness and *mean slope*, with an S shape in the log scale. This correlation between climatic attributes and flashiness was reported by Gannon et al. (2022) to have an overall importance of around 30%, which is similar to the importance value found in our study (sum of *p_mean* and *pet_mean*, Figure 6a). Li et al. (2023) reported a similar finding but additionally found *p_mean* (Figure 6c) to be the most positively correlated attribute. In summary, the weight associated with current precipitation is consistent with findings reported in the literature mentioned before, thereby supporting our hypothesis that the RF model is successful at relating catchment attributes to the gate-weight sequences of the *HydroLSTM* model.

Similarly, Figure 6d shows that the weight on the previous day's potential evapotranspiration correlates strongly with “*mean elevation*” and “*mean leaf area index*.” Both of these attributes are informative about vegetation type, which controls evapotranspiration. The correlation between elevation and vegetation type is well-known to be determined by the altitudinal zonation of variables such as temperature, humidity, solar radiation, etc. (Dauenbire, 1943). It is, therefore, encouraging that the reduced RF model determined this attribute to be the most important one. Moreover, Jung et al. (2013) found that elevation has a negative effect on runoff sensitivity in high-elevation regions; again, this behavior is captured by the reduced RF model (Figure 6e). Jung et al. (2013) also reported a positive correlation with elevation for lowlands around the Mississippi River, which might correspond with the positive slope of some of the lines in the partial dependence plot for elevations lower than 500m.

Turning now to the second most important attribute, many relationships have been reported in the literature that use some kind of vegetation index and/or leaf area index to relate potential evapotranspiration to actual evapotranspiration (Leuning et al., 2008; L. Wang et al., 2021; Yan et al., 2012). In all such relationships, a

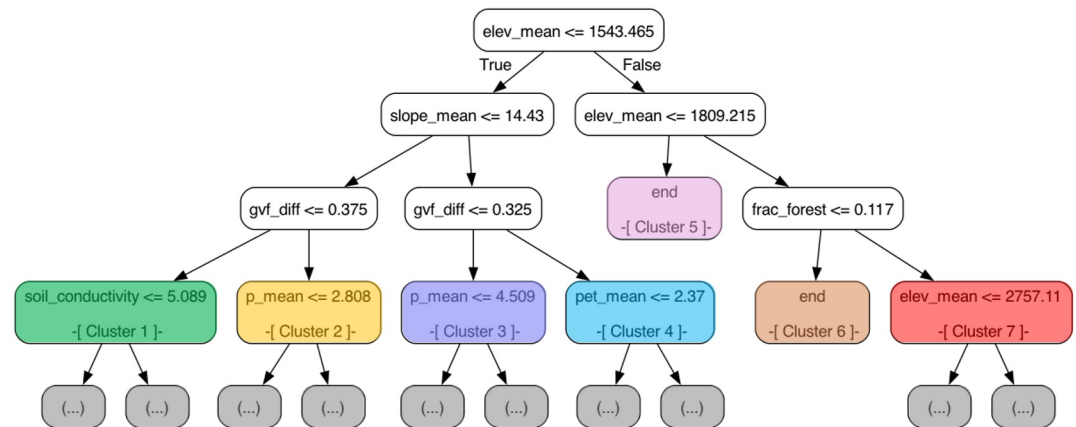


Figure 7. The “best” decision tree, shown truncated to the first three levels (level zero at the top). Leaf nodes are color-coded to correspond to the clusters of catchment locations depicted in Figure 8. Gray-colored locations correspond to branches with deeper levels (for or more).

positive (linear or quasilinear) correlation is represented, consistent with (Figure 6f). Overall, these results suggest that the regionalization model represents the dependence between catchment attributes and the stream-flow generation process in a realistic manner.

4.3.3. Assessing Clusters Associated With a Reduced RF Model

In the previous section, the analysis was focused on predicting a single weight at a time (one reduced RF model predicts only one of the weights). Here, we develop another reduced RF model to simultaneously predict the four output-gate weights explaining most of the variability, by using the same 10 attributes from the previous section. Specifically, those corresponding weights are: (a) precipitation at the current time (lag_0), (b) precipitation at the previous time (lag_1), (c) precipitation at 10 days in the past (lag_{10}), and (d) potential evapotranspiration at the previous time (lag_1); see Figure 2. Our goal is to examine the decision tree that best describes the relationship between attributes and weights. With this best tree, we can analyze the corresponding clusters that are created by each branch. These clusters are then presented using a map to show their spatial distribution.

Figure 7 shows the “best” decision tree calculated using the splits that minimize the error between the mean and target in each terminal node and at each level. For ease of visualization and analysis, we show only a truncated version of the tree (three levels and only the threshold on each split) that results in a relatively small number of clusters (terminal nodes). Although it is certainly true that a tree can suffer from overfitting in the splitting value and the level when a particular attribute is used, analysis of the tree can reveal a lot about the “grouping” i.e. occurring. For example:

1. The first split (level 0) is based on elevation ($elev_mean$). It occurs at a relatively high elevation (1,543 m), after which the branch for higher elevation splits twice again to indicate partitions at even higher elevations (1,809 and 2,703 m). This indicates that unique behaviors are related to different (high) elevation bands, consistent with the dynamics of snowmelt-dominated catchments.

To be clear, this choice of elevation to determine the first split does not mean that elevation is a “causal” attribute—instead, it serves as a surrogate for other (non-represented) characteristics such as mean temperature. For the same reason, other trees in the ensemble might instead use alternative surrogate variables for this purpose, such as slope (see left branch at level 1) which is typically highly correlated with elevation.

2. At the second and third levels, vegetation and climatic characteristics emerge as important determinants (explanatory variables), consistent with our understanding of the main drivers of runoff generation and with the overall feature importance of the RF-based model (Figure 5).

The catchment locations that correspond to each of the 7 color-coded clusters (leaf-nodes) depicted in Figure 7 are visualized on the map shown in Figure 8 to illustrate their high degree of geo-spatial consistency. The table

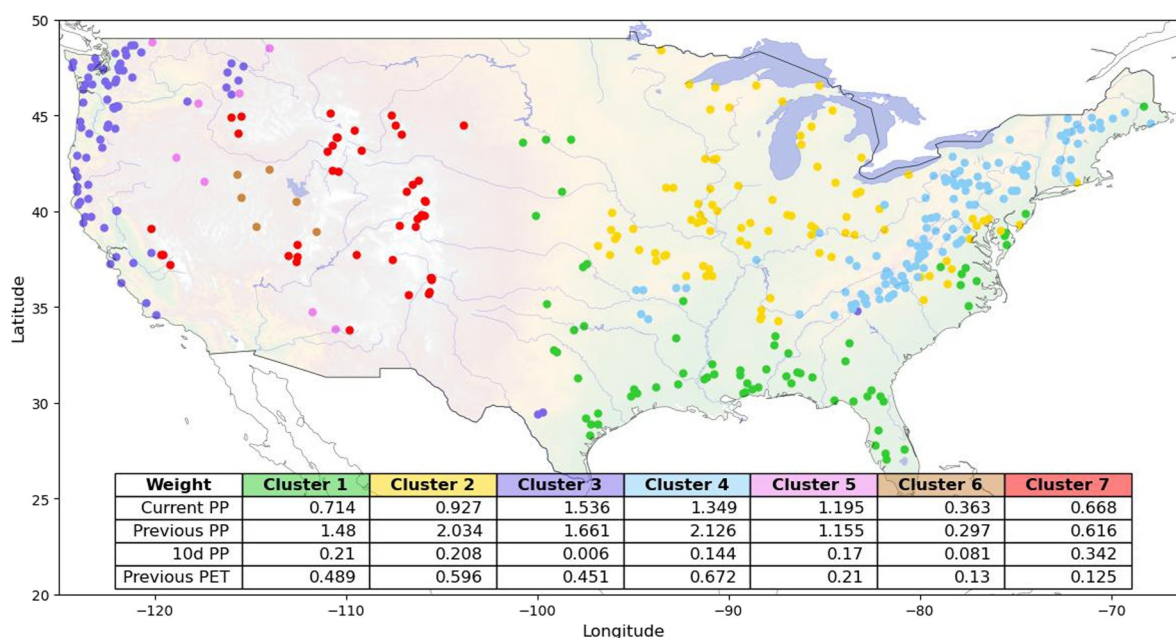


Figure 8. Clusters created by the “best” tree of the reduced RF model that was trained to predict *HydroLSTM* weight-sequence from catchment attributes. Color coding corresponds to the leaf nodes in Figure 7. The inset table presents the 4 most diverse output gate weights associated with each cluster (PP: precipitation, PET: potential evapotranspiration, Current: time = 0, previous: time = -1, 10d: 10 days in the past).

attached to Figure 8 lists the values associated with the aforementioned four output-gate weights determined for the corresponding clusters. From this we see that:

Lower Elevation Clusters ($elev_mean \leq 1,543$ m):

1. Cluster 1 (green) is associated mainly with the Deep South, Florida, and the region between the Appalachian Mountains and the Atlantic Ocean. The weight sequence indicates that this region has relatively low streamflow flashiness, but with moderate magnitudes for the weights assigned to the previous and 10 days precipitation weight. This corresponds to a relatively higher influence of “past” precipitation and/or longer residence times for water in the catchment.
2. Cluster 2 (orange) has similar a weight sequence to Cluster 1 but with higher values for precipitation weights at the current (lag_0) and previous (lag_1) times. This indicates a slightly higher degree of flashiness.
3. Cluster 3 (purple) is associated mainly with the Northwestern part of the Rocky Mountains and Sierra Nevada, where slopes and precipitation magnitudes/intensities are relatively high. Consistent with this, there is a larger weight assigned to current precipitation and a lower weight assigned to precipitation input 10 days in the past.
4. Cluster 4 (light blue) is associated mainly with the Appalachian Mountains and their extension up into New England. It appears to be similar to cluster 3 (dark blue), but from the tree, we see that it is associated with a higher green vegetation fraction difference (gvf_diff), which corresponds to higher weights assigned to precipitation at the previous time and 10 days in the past, and to potential evapotranspiration at the previous time. That seems consistent with the effects that denser vegetation fractions should have on the streamflow generation process.

Higher Elevation Clusters ($elev_mean > 1,543$ m):

5. Cluster 5 (pink) is the lowest of the three aforementioned “high” elevation zones ($1,543$ m $< elev_mean \leq 1,809$ m), and seems to correspond to a transition zone between Cluster 3 (a flashy area) and Cluster 7 (snowmelt dominated).
6. Cluster 6 (brown) is also at a high elevation ($1,809$ m $< elev_mean \leq 2,757$ m) but is associated with a lower forest fraction than Cluster 7 (red). The precipitation weights associated with all of the temporal scales (current to 10 days in the past) are relatively low, indicating a low water yield compared with all other clusters. This may be related to the high number of endorheic areas in Nevada and Utah.
7. Cluster 7 (red) is dynamically quite different from the others. It has the largest weight associated with precipitation 10 days in the past, indicating longer catchment memory. This is likely due to information regarding

“snowpack accumulation and melt” (corresponding to available water) being encoded via this weight. At the same time, the weight associated with potential evapotranspiration at the previous time is relatively low. While this potentially conflicts with our understanding of how temperature determines snowmelt, it could instead indicate low amounts of actual evapotranspiration. More insights about that situation are presented in the next section.

In terms of the overall spatial distribution of the clusters, they are pretty similar to others that have been found (Jiang et al., 2022; Wu et al., 2021). However, our clusters exhibit less noise due to the regularization implemented by the RF model. Moreover, our strategy is quite different given that we applied a similarity assessment to the behavioral dynamics as represented by the weight sequences, and the attributes are only used as surrogate variables to represent them in space.

4.4. Assessing Information Encoded in the Cell-State Dynamics

In our previous work (De La Fuente et al., 2024) that proposed and tested the *HydroLSTM* architecture, we reported that despite working with only a single cell-state (state variable) per catchment, we found no consistency in the behavioral patterns of the dynamical evolution of the states. We concluded that the relationship between the values of the cell-state and the output gate was not sufficiently constrained to ensure unique behavior. Here, we revisit that issue and examine whether the additional regularizing constraints imposed by the process of regionalization (via the Regional *HydroLSTM* model) result in a more uniquely determined representation of the cell-state. If so, this would help to improve interpretability by co-relating the cell-state dynamics with hydrologic knowledge about expected catchment states (e.g., snow water equivalent, soil moisture, and groundwater storage).

In Figure 9, we plot cell-state and output-gate trajectories for two representative catchments known to have quite different hydrological behaviors, following the regions defined by Jiang et al. (2022). The first (upper panel; South Fork of Williams Fork near Leal, CO) is located in a snowmelt-dominated region. The second (lower panel; Leaf River, near Collins, MS) is in a rainfall-dominated region.

In terms of the cell-state, Figure 9 presents two different behaviors:

1. In the snowmelt-dominated basin (Figure 9a), the cell-state (red line) does not correlate well with expected snow accumulation during the winter season (December–February), and instead is “high” during the summer season (June–August). So rather than encoding for storage of available water (i.e., snow water equivalent, SWE) it seems to instead be encoding for available energy (see correspondence to the mean temperature trajectory; orange line). During the winter, when the temperature tends to be below zero, the cell-state remains relatively constant and close to zero, while during the summer it approaches its maximum possible value (close to one). Further, the period of changing in cell-state corresponds mainly with the Spring onset of snow melt and the rising limb of the hydrograph (dashed line).
2. In the rainfall-dominated catchment (Figure 9b), the cell-state (red line) is strongly correlated with streamflow (dashed line), and is relatively insensitive to temperature.

An interesting thing to note about the *HydroLSTM* architecture is that it can encode information related to accumulation and release dynamics in two different ways. One of these is through the “cell-state” that encodes for longer-term storage regarding the overall effects of everything that has happened in the past. The other is through the gates, which we name “gate-states” that encodes for how a specific event, or a series of events, impacts the present. In other words, the gate-state stores “contextual” information from the past, whereas the cell-state uses the summary of the past to define the “current” state. That means that the hydrological idea of water and/or energy accumulation could be expressed as the sum of all the past events in the gate or as the value stored in the state variable. Which representation is used in the *HydroLSTM* architecture will probably depend on the complexity of the relationships and the ease of training.

In the case of the output gate, Figure 9 shows how it constrains the release of the cell-state:

1. In the case of the snowmelt-dominated basin (Figure 9a), the cell-state (red line) approaches 1 in the May–June period, and the output gate is accordingly drained relatively quickly. However, after this period the output gate “closes” (eventually reaching zero) and thereby restricts almost completely the production of streamflow.

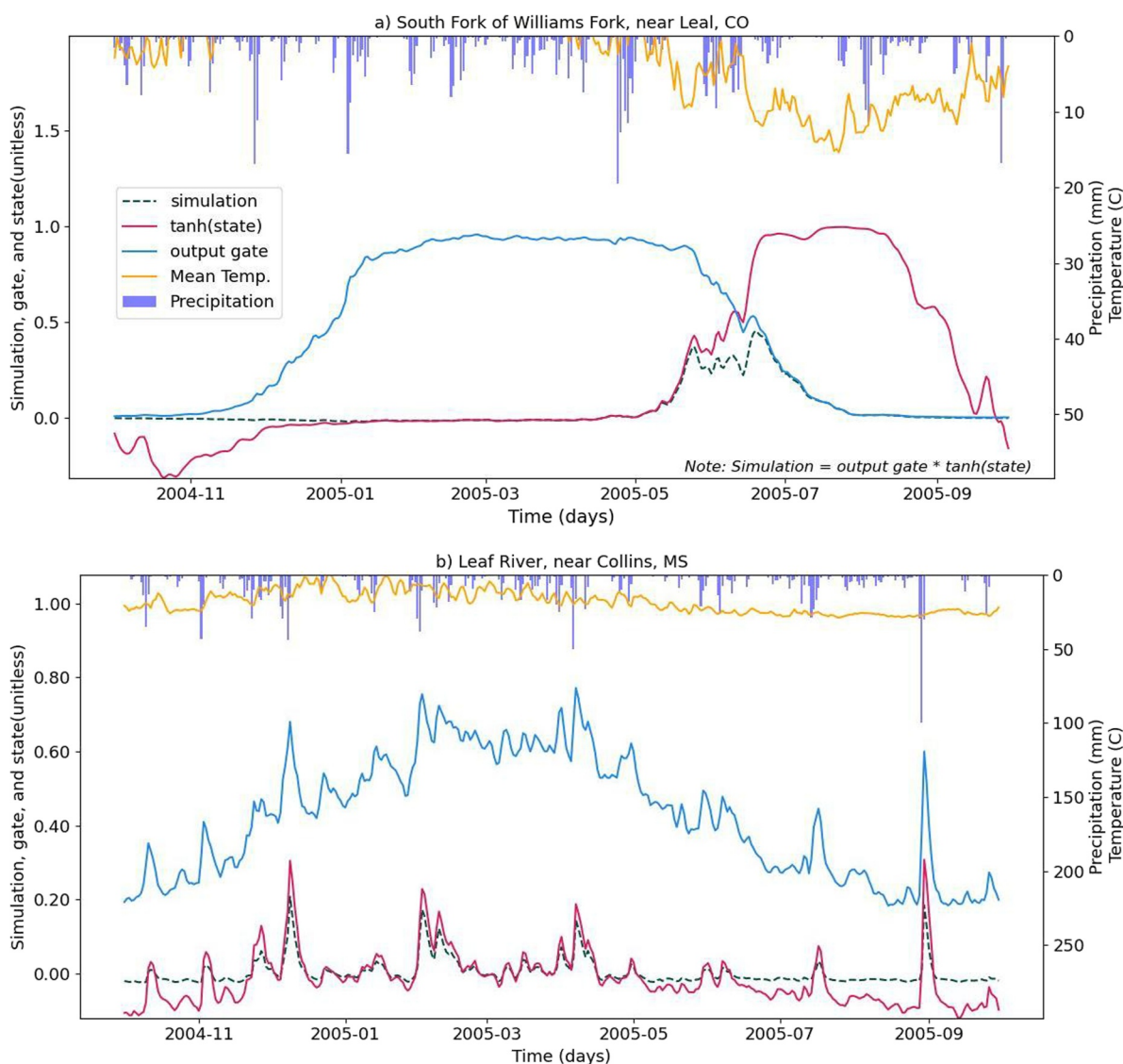


Figure 9. Hydrographs for two hydrological different catchments. (a) Snowmelt dominated: output gate tracks water availability, and the cell-state the potential streamflow. (b) Rainfall-dominated: the output gate corrects the state variable when it is below the baseflow. More cases are clustered in Figure A1.

2. In the case of the rainfall-dominated catchment (Figure 9b), the cell-state tends to be larger during precipitation events than the actual simulated output, and the output gate restricts water release probably when the catchment is not fully saturated, thereby decreasing streamflow. During summer, the state variable tends to be negative, however the output gate corrects the simulation by imposing a minimum value of streamflow in the river (“baseflow”)

From these two examples, we see that the cell-state and gate-state trajectories determined by the *Regional HydroLSTM* display interpretable behaviors that are consistent with hydrological knowledge. Similar behaviors were observed for other catchments in the same cluster, especially for those corresponding to Clusters 5, 6, and 7 (Appendix A). This suggests that the regionalization procedure helps to regularize the model development process. However, contrary to “hydrological” intuition, because the cell-state is not constrained to be positive-valued, it does not necessarily encode for actual available (accumulated) storage of water in the catchment. Instead, the “cell-state” and the “gate-state” of the *Regional HydroLSTM* model jointly encode useful and interpretable information.

To further understand this, consider a simple example in which an extreme precipitation event has occurred between 4 and 5 days in the past. We can think of the relevant information as being the volume of precipitated water and the time at which the event occurred. In the *HydroLSTM*, information about “volume” can potentially be reflected in summary fashion by the cell-state (as aggregated exponentially-decayed weighted information about the entire past history of incoming precipitation), or by the gate-state as some attention-weighted sum of precipitation values occurring during the recent past. Meanwhile, information about timing, which may be associated with residence times that determine how quickly/slowly water is routed through the system to arrive at the gauge, can be reflected by the temporal pattern of (attention) weights that control the values of the gate-states.

In our particular case (Figure 9), the *Regional HydroLSTM* results suggest that volume-related information has been mainly encoded via the output gating mechanism. While this might seem counterintuitive from a conventional hydrological-modeling point of view—where we typically conceive of “mass/volume” of water as being stored in Markovian cell-states (the accumulated contents of bucket-like storage elements)—we should not forget that the so-called “*antecedent precipitation index*” used in the Curve Number (CN, Soil Conservation Service, 1986) method also has a time-honored role in hydrological representation as a way to mediate (set the context for) how strongly a catchment will respond to precipitation inputs. Similarly, in auto-regressive (ARX) time-series modeling (Bolzern et al., 1982), it is not unusual for several past-lagged weighted system inputs (the “X” in ARX; e.g., precipitation, potential evaporation, etc.) and outputs to appear in the (typically linear) representation that determines how strongly the system responds to current inputs; note that the ARX formulation does not involve “cell-states” per se, unless you think of the past lagged outputs as informational surrogates for past cell-states (Young, 2015).

In that regard, if we think in terms of the cell-state being mediated by a nonlinearly-draining reservoir (i.e., streamflow $Q = k \times S$ where S is storage and the outflow rate $k = f(S, X)$ where $f(\cdot)$ is a non-linear function and X represents other contextually informative variables; see Wang and Gupta (2024), then we can imagine that a neural network used to model/predict the value of “ k ” would need to be provided with information about both the current cell-state S and other variables X that are helpful in determining how rapidly the tank should be drained, thereby influencing the residence time, and hence temporal pattern of the system response. Meanwhile, the evidence suggests that the Markovian cell-state in our study (Figure 9) can be thought of as representing a kind of “*potential streamflow*” (i.e., an “*informational quantity*” rather than an actual “*volumetric*” quantity), which hypothesis is supported by the fact that it stays close to zero during the colder winter periods and it has higher values during the summer (Figure 9a) for this catchment in which snowmelt-dominates the process of streamflow generation. In contrast, for the rainfall-dominated catchment (Figure 9b), both the cell-state and the output-gate-state track the streamflow dynamic, consistent with a situation in which both “*states*” are informative about the volume of accumulated water (storage) in the system. However, the cell-state continues to be more closely tied to the idea of “*potential streamflow*.”

5. Discussion

We focus our discussion on two main topics; (a) catchment attributes and (b) the memory concept used in our approach. We consider these to be two relevant elements that differentiate our approach from other approaches.

5.1. Catchment Attributes as a Proxy for Dynamic Classification

The task of mapping between system properties (catchment attributes) and system dynamics (catchment behaviors) has proven to be a difficult task in hydrology. Part of this can be explained as being due to the high variability in importance of the catchment attributes (Figure 5), due to which many combinations of catchment attributes can result in similar hydrological responses. However, the findings obtained using our approach (that separates the representation into components representing dynamics and spatial regionalization) lead us to suggest that we can view spatial regions as functioning as a kind of “*latent space representation*” that acts to link attributes with dynamics. This would explain why (in many cases) spatial proximity of catchments tends to be the most reliable basis for regionalization (our results suggest that this is true *within* a cluster; Figure 8), while small changes in space can sometimes result in completely different responses (our results suggest that this is due to moving from one cluster to another). In this sense, our RF-based regionalization model differs from other methods because this approach explicitly creates hydrological dynamic similarity and then these regions are mapped to catchment attribute similarity. This suggests a new approach to the exploration of catchment dynamics, allowing

us to better understand the underlying data-generating process and how an ML model can encode such relationships through the use of catchment attributes. These results align with current research about the clustering role of attributes in regional models, which supports our separation hypothesis (Heudorfer et al., 2025; Yu et al., 2024).

A similar approach could be applied to physically-based (PB) models for generating regional models. However, the numerous assumptions inherent in PB structures may lead to errors that could affect the regionalization process. This complicates the task of determining whether any inaccuracies in regionalization stem from the process itself or the PB model. As a result, learning the regionalization tends to be more straightforward when employing a machine learning model. Nonetheless, significant advancements have been achieved using the differential approach (Feng et al., 2022), leading to the HVB model performing comparably to the LSTM model.

However, catchment attributes should be used carefully. In many cases, they act as surrogates for other dominant characteristics, such as in the case of mean elevation, which is highly correlated with long-term mean temperature (not included as an attribute in the data set), vegetation, slope, etc. This suggests that the reason that our ML models could be found to be weakly sensitive to changes in the actual drivers is because of not properly discriminating between static and dynamic variables.

Further issues with the use of catchment attributes could eventually emerge when we apply ML models to analyze nonstationary scenarios such as climate change. We know only that catchment attributes are (a) correlated with dynamical catchment behaviors in the long term, and (b) not correlated with daily forcings in the very short term. However, the nature of this relationship at intermediate time scales is not clear. For example, consider an extended drought scenario, during which we expect that some of the vegetation will die, while some will survive—but how rapidly this change will occur and how the catchment dynamics will respond is (in general) poorly understood. Moreover, since catchment attributes are not independent (e.g., mean elevation is correlated with mean slope; forest fraction, leaf area index, and green vegetation fraction are correlated; clay fraction is correlated with soil conductivity, etc.), we do not necessarily capture all of the relevant variability or trend when changing one attribute at a time (e.g., mean temperature in the case of climate change). Further, we could be losing information about important inflection points or predicting unrealistic scenarios. Because these comments are valid for any ML model using catchment attributes, the ability to disentangle the importance of attributes and dynamics to prediction is fundamental to understanding model limitations. Our *Regional HydroLSTM* study can be viewed as a small step in that direction.

A remarkable property of catchment attributes is how well we can do with a rather small number of them (17 in the RF model, and 10 in the reduced RF model), indicating that most of the relevant information can be represented using only a small number of (carefully chosen) attributes. Conversely, if only a few attributes can serve to characterize the catchment dynamics well, discovering what information is missing can be a difficult task. Note that other sources of information that might better characterize catchment dynamics have not been included in the analysis. Potentially informative examples might include attributes such as topographic index, depth to rock at the gauge, and river network information. Another option might be to increase the spatial resolutions of attributes that we have already included. For instance, if 50% of the catchment is covered by forest, this could result in different dynamic behaviors depending on whether this percentage is distributed closer to the outlet/gauge or to the headwater regions. Similar analyses can be performed for the dynamic variables—for example, statistics of their spatial distribution should be incorporated as additional information. Of course, while such information (including the addition of new dynamic variables) could help to improve performance, it might not improve the interpretability of our models. In general, we should probably explore such aspects while trying to ensure that we are learning the right (meaningful) relationships between attributes and dynamics.

5.2. The Interaction of Two Types of Memory

Another interesting aspect of the *HydroLSTM* architecture is the two types of “memory” (cell-state and gate-state) that interact to determine the behavior of a single cell-state of the *HydroLSTM* representation. The cell-state evolves through time in Markovian fashion. As such, it uses the same equations as in the original LSTM, where the cell-state is the most direct source of information regarding the past history of events (the gates are informed indirectly through the use of the cell state), while the system inputs at the current time step determine the values of the gates (they use the information store in the cell-state of the previous time step). However, in the *HydroLSTM* representation, the values of the gates at each time step are *also* determined by the past history of

inputs (not just the information from the current time step). So the behavioral dynamics of the *HydroLSTM* actually depends on two kinds of information—the cell-states that track the long-term history of the system in a truly Markovian fashion (similar to the LSTM but with an effectively approaching-infinite sequence length), and the gate-states that are determined by a shorter-term recent window on the history of the system (in this study chosen to be 513 time steps). Note that it is primarily the nature of this shorter-term contextual memory that determines the spatial differences in the dynamical response of the clusters in the *Regional HydroLSTM*.

One way to reconcile these two seemingly disparate types of memories is to think of the cell-state as informationally representing the “*potential streamflow*” that would be generated in the absence of mediating context—for the snow-melt dominated catchment that mediating context is temperature, which prevents flow from being generated when temperatures are below freezing (Figure 9a). Meanwhile, the gate-state can be thought of as representing “*conductivity*” and acts to informationally represent when “*potential streamflow*” cannot be achieved and must be corrected (as the output gate accumulation decreases, restricting the melting of the snowpack). For the non-snow catchment (Figure 9b), where freezing does not happen (temperatures always remain above 0°C), this “*potential streamflow*” and the gate-state (indicating “*conductivity*” given some water availability) are more closely aligned, and the latter acts to influence both the volume (especially during summer) and timing of release from storage.

Overall, this discussion relates to the issue of how to obtain interpretable information from models based on ML architectures. Of course, our understanding remains at an early stage, and the process of learning from hybrid modeling efforts will require ongoing thought and debate regarding whether it makes sense to modify our interpretations, or modify the ML architectures to match our prior conceptualizations regarding how to architecturally represent any given hydrological system, or both. Wherever path is chosen, this study has shown that the interpretability of a carefully crafted and parsimonious ML-based model is possible and that further improvements in this direction should be pursued.

6. Conclusions

The task of “*regionalization*” has proven to be a difficult problem, and it is possible that no “*unique/correct*” solution exists. Nonetheless, new methods such as Machine Learning have helped to advance predictive and generalization performance, and have the potential to improve our ability to learn (extract hydrological knowledge) from the data. We have proposed a hybrid ML-based architecture that enables regional “*interpretable*” streamflow model development, by imposing an explicit separation between the problems of modeling local input-state-output catchment dynamics, and the problem of learning spatial similarities/distinctions in how such dynamics are mediated by various observable catchment characteristics. This imposed regularization of the problem helps to maximize our ability to interpret what is being learned by our regional model.

Our implementation, named *Regional HydroLSTM*, is a coupling of the *HydroLSTM* and *Random Forest* architectures. We enforce relative parsimony by constraining the *HydroLSTM* component to use only a single cell-state, which facilitates hydrological interpretability of the factors determining the dynamics of catchment state-output response to climatic drivers. Meanwhile, the *Random Forest* component facilitates interpretable discovery of the role that catchment attributes play in determining how these dynamics vary with location, and enables a kind of similarity-based catchment classification system of different kinds of catchment dynamics, akin to a representation based in different kinds of process dominance (Guse et al., 2016; Sivakumar, 2008). We suggest that this approach (or something similar) may help better to understand the long-standing “*catchment classification*” problem (Sawicz et al., 2014; Sivakumar & Singh, 2012; Wagener et al., 2007), and improve our ability to make “*understanding-based*” (as opposed to black-box based) predictions in ungauged basins. To our knowledge, the *Regional HydroLSTM* modeling approach represents the first demonstration of how architectural regularization can be used to learn dynamical similarities between catchments, which understanding can then be used to advance streamflow prediction using a simple *parsimonious* model architecture representing catchment behavioral dynamics.

Certainly, there remains room for improvement. Future work should explore the (performance) benefits of allowing more than one cell-state to track relevant informational quantities (features) that should obey Markovian dynamics while finding ways to minimize the growth of equifinality and retain a useful level of hydrological interpretability. This is a “*balancing act*” that will require careful study—for example, the desirable number of cell-states could conceivably vary with catchment location, and it remains an open question as to whether this desirable number can be predicted based on information extractable from the observed static characteristics and/or summary properties extracted from the observed climate drivers and system responses.

In conclusion, we argue that the scientific benefits of hydrological model “interpretability” can be achieved using ML-based architectures. We just need to pay careful attention to implementing regularization strategies that impose inductive biases on the learning problem through the selection (and parsimonious implementation) of architectural components that make hydrological sense.

Appendix A: Dynamic Behavior in Some Clusters

Figure A1

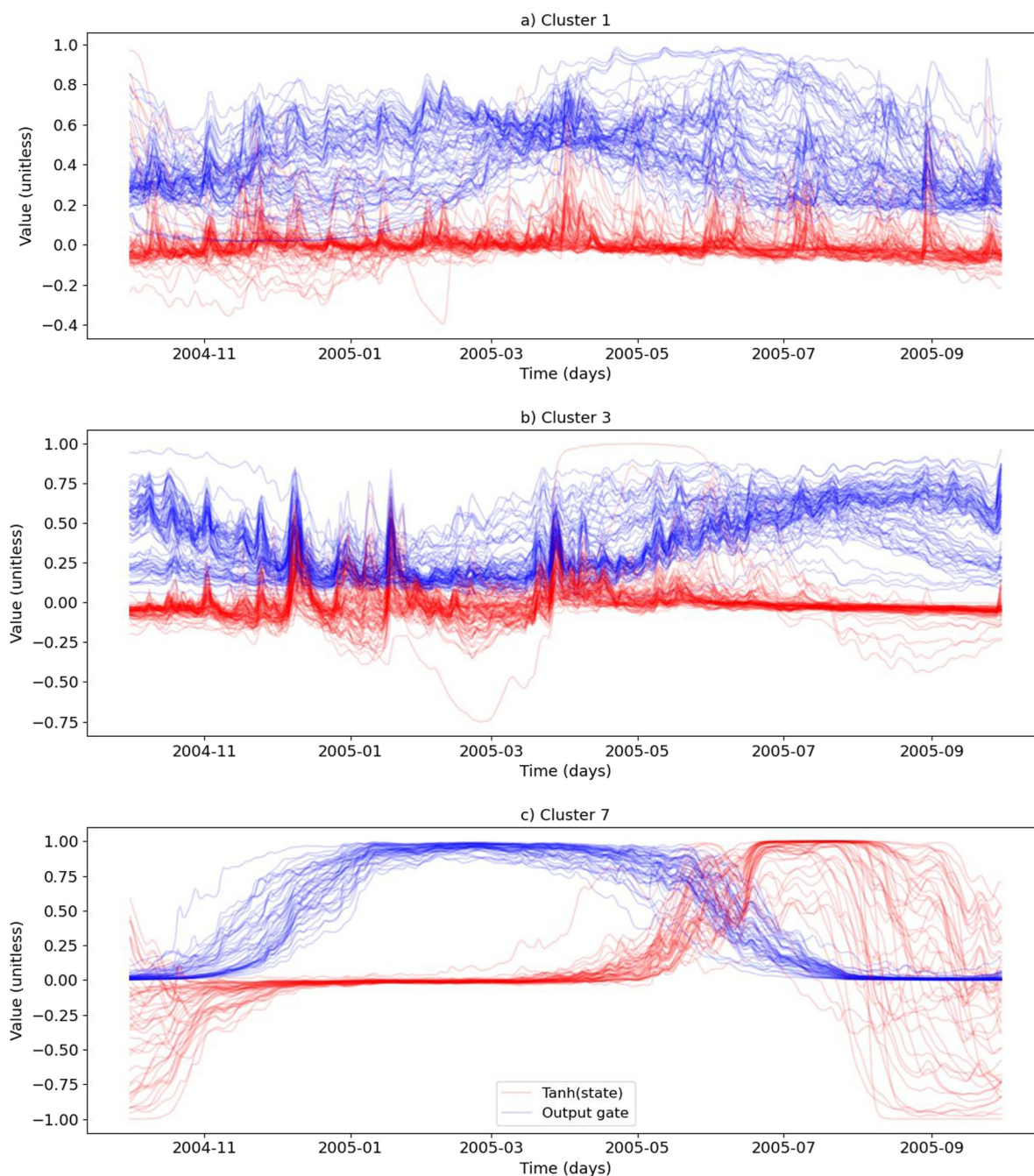


Figure A1. Behaviors of the state and output gate for all the catchments in specific clusters. (a) Cluster 1 shows high variability in the output gate which is an indication that subclusters coexist (two or more). (b) Cluster 3 presents a strong behavior and a weak second behavior in the output gate (two subclusters). (c) Cluster 7 indicates that the behavior in the output gate is unique but probably subcluster could be created in the state behavior.

Data Availability Statement

The CAMELS data set is a merge of two data sets, the time series (Newman et al., 2014) and catchment attributes (Addor et al., 2017a), and they are freely available. The codes to run the models and the Jupyter Notebook used to analyze the results are freely available at Zenodo (De La Fuente, 2025).

Acknowledgments

The authors would like to acknowledge the funding of NSF and ANID, and the feedback received from the Condon Lab team and the reviewers. This research has been supported by the Division of Earth Sciences (Grant 1945195), the Innovation and Technology Ecosystems (Grant 2134892), and the Comisión Nacional de Investigación Científica y Tecnológica (Grant Becas Chile, 2022). We acknowledge the use of AI tools in checking the grammar and fluency of the manuscript.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, 54(11), 8792–8812. <https://doi.org/10.1029/2018WR022606>
- Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017a). Catchment attributes for large-sample studies [Dataset]. *UCAR/NCAR*. <https://doi.org/10.5065/D6G73C3Q>
- Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017b). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 21(10), 5293–5313. <https://doi.org/10.5194/hess-21-5293-2017>
- Alvarez-Garretón, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., et al. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies – Chile dataset. *Hydrology and Earth System Sciences*, 22(11), 5817–5846. <https://doi.org/10.5194/hess-22-5817-2018>
- Arsenault, R., Martel, J.-L., Brunet, F., Brisette, F., & Mai, J. (2023). 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. <https://doi.org/10.5194/hess-27-139-2023>
- Blöschl, G. (2005). Rainfall-runoff modeling of ungauged catchments. In M. G. Anderson & J. J. McDonnell (Eds.), *Encyclopedia of hydrological sciences* (1st ed.). Wiley. <https://doi.org/10.1002/0470848944.hsa140>
- Bolzern, P., Fronza, G., & Guariso, G. (1982). Stochastic flood predictors: Experience in a small basin. In *Developments in water science* (Vol. 17, pp. 530–537). Elsevier. [https://doi.org/10.1016/S0167-5648\(08\)70737-8](https://doi.org/10.1016/S0167-5648(08)70737-8)
- Botterill, T. E., & McMillan, H. K. (2023). Using machine learning to identify hydrologic signatures with an Encoder–Decoder framework. *Water Resources Research*, 59(3), e2022WR033091. <https://doi.org/10.1029/2022WR033091>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Daubenmire, R. F. (1943). Vegetational zonation in the Rocky Mountains. *The Botanical Review*, 9(6), 325–393. <https://doi.org/10.1007/BF02872481>
- De La Fuente, L. A. (2025). ldelafue/regionalHydroLSTM: Regional HydroLSTM v1.0.0 (version v1.0.0) [Computer software]. *Zenodo*. <https://doi.org/10.5281/ZENODO.15265955>
- De La Fuente, L. A., Ehsani, M. R., Gupta, H. V., & Condon, L. E. (2024). Toward interpretable LSTM-based modeling of hydrological systems. *Hydrology and Earth System Sciences*, 28(4), 945–971. <https://doi.org/10.5194/hess-28-945-2024>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1–38. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Di Prinzio, M., Castellarin, A., & Toth, E. (2011). Data-driven catchment classification: Application to the pub problem. *Hydrology and Earth System Sciences*, 15(6), 1921–1935. <https://doi.org/10.5194/hess-15-1921-2011>
- Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research*, 58(10), e2022WR032404. <https://doi.org/10.1029/2022WR032404>
- Gannon, J. P., Kelleher, C., & Zimmer, M. (2022). Controls on watershed flashiness across the continental US. *Journal of Hydrology*, 609, 127713. <https://doi.org/10.1016/j.jhydrol.2022.127713>
- Gini, C. (1997). Concentration and dependency ratios. *Rivista Di Politica Economica*, 87(8–9), 769–790.
- Girshick, R. (2015). Fast R-CNN. <https://doi.org/10.48550/ARXIV.1504.08083>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afcf3-Paper.pdf
- Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *WIREs Water*, 8(1), e1487. <https://doi.org/10.1002/wat2.1487>
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- Guse, B., Pfannerstill, M., Strauch, M., Reusser, D. E., Lüdtke, S., Volk, M., et al. (2016). On characterizing the temporal dominance patterns of model parameters and processes. *Hydrological Processes*, 30(13), 2255–2270. <https://doi.org/10.1002/hyp.10764>
- Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, 1(2), 96–99. <https://doi.org/10.13031/2013.26773>
- He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, 15(11), 3539–3553. <https://doi.org/10.5194/hess-15-3539-2011>
- Heudorfer, B., Gupta, H. V., & Loritz, R. (2025). Are deep learning models in hydrology entity aware? *Geophysical Research Letters*, 52(6), e2024GL113036. <https://doi.org/10.1029/2024GL113036>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Jiang, S., Zheng, Y., Wang, C., & Babovic, V. (2022). Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. *Water Resources Research*, 58(1), e2021WR030185. <https://doi.org/10.1029/2021WR030185>
- Jung, I. W., Chang, H., & Riskey, J. (2013). Effects of runoff sensitivity and catchment characteristics on regional actual evapotranspiration trends in the conterminous US. *Environmental Research Letters*, 8(4), 044002. <https://doi.org/10.1088/1748-9326/8/4/044002>
- Kanishka, G., & Eldho, T. I. (2017). Watershed classification using isomap technique and hydrometeorological attributes. *Journal of Hydrologic Engineering*, 22(10), 04017040. [https://doi.org/10.1061/\(ASCE\)JE.1943-5584.0001562](https://doi.org/10.1061/(ASCE)JE.1943-5584.0001562)

- Kanishka, G., & Eldho, T. I. (2020). Streamflow estimation in ungauged basins using watershed classification and regionalization techniques. *Journal of Earth System Science*, 129(1), 186. <https://doi.org/10.1007/s12040-020-01451-8>
- Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2), 233–243. <https://doi.org/10.1002/aic.690370209>
- Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS opinions: Never train an LSTM on a single basin. *Catchment hydrology/Modelling approaches*. Preprint. <https://doi.org/10.5194/hess-2023-275>
- Kratzert, F., Gauch, M., Nearing, G., & Klotz, D. (2022). NeuralHydrology—A Python library for Deep Learning research in hydrology. *Journal of Open Source Software*, 7(71), 4050. <https://doi.org/10.21105/joss.04050>
- Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019a). NeuralHydrology – Interpreting LSTMs in hydrology. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700, pp. 347–362). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_19
- 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. <https://doi.org/10.1029/2019WR026065>
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019c). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12), 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>
- Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., et al. (2023). Caravan—A global community dataset for large-sample hydrology. *Scientific Data*, 10(1), 61. <https://doi.org/10.1038/s41597-023-01975-w>
- Kuczera, G., & Mroczkowski, M. (1998). Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Water Resources Research*, 34(6), 1481–1489. <https://doi.org/10.1029/98WR00496>
- Leuning, R., Zhang, Y. Q., Rajaud, A., Cleugh, H., & Tu, K. (2008). A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman-Monteith equation. *Water Resources Research*, 44(10), 2007WR006562. <https://doi.org/10.1029/2007WR006562>
- Li, Z., Gao, S., Chen, M., Zhang, J., Gourley, J. J., Wen, Y., et al. (2023). Introducing flashiness-intensity-duration-frequency (F-IDF): A new metric to quantify flash flood intensity. *Geophysical Research Letters*, 50(23), e2023GL104992. <https://doi.org/10.1029/2023GL104992>
- Ma, K., Feng, D., Lawson, K., Tsai, W., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, 57(5), e2020WR028600. <https://doi.org/10.1029/2020WR028600>
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations (Vol. 1, pp. 281–297). Retrieved from https://www.google.com/books/edition/Proceedings_of_the_Fifth_Berkeley_Sympos/IC4Ku_7dBFUC
- Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., & Nijssen, B. (2002). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of Climate*, 15(22), 3237–3251. [https://doi.org/10.1175/1520-0442\(2002\)015<3237:ALTHBD>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2)
- Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9), 1620–1633. <https://doi.org/10.1111/2041-210X.13650>
- Minder, J. R., Mote, P. W., & Lundquist, J. D. (2010). Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains. *Journal of Geophysical Research*, 115(D14), 2009JD013493. <https://doi.org/10.1029/2009JD013493>
- Molnar, C. (2022). Interpretable machine learning (2nd ed.). Retrieved from <https://christophm.github.io/interpretable-ml-book>
- Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R., Blodgett, D., et al. (2014). A large-sample watershed-scale hydrometeorological dataset for the contiguous USA (p. approximately 2.5 GB) [Dataset]. *UCAR/NCAR - GDEX*. <https://doi.org/10.5065/D6MW2F4D>
- Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungauged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resources Research*, 44(3), 2007WR006240. <https://doi.org/10.1029/2007WR006240>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Roth, A. E. (Ed.) (1988). *The Shapley value: Essays in Honor of Lloyd S. Shapley* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511528446>
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Sabzipour, B., Arsenaault, R., Troin, M., Martel, J.-L., Brissette, F., Brunet, F., & Mai, J. (2023). Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment. *Journal of Hydrology*, 627, 130380. <https://doi.org/10.1016/j.jhydrol.2023.130380>
- Saharia, M., Kirstetter, P., Vergara, H., Gourley, J. J., Emmanuel, I., & Andrieu, H. (2021). On the impact of rainfall spatial variability, geomorphology, and climatology on flash floods. *Water Resources Research*, 57(9), e2020WR029124. <https://doi.org/10.1029/2020WR029124>
- Sawicz, K. A., Kelleher, C., Wagener, T., Troch, P., Sivapalan, M., & Carrillo, G. (2014). Characterizing hydrologic change through catchment classification. *Hydrology and Earth System Sciences*, 18(1), 273–285. <https://doi.org/10.5194/hess-18-273-2014>
- Sivakumar, B. (2008). Dominant processes concept, model simplification and classification framework in catchment hydrology. *Stochastic Environmental Research and Risk Assessment*, 22(6), 737–748. <https://doi.org/10.1007/s00477-007-0183-5>
- Sivakumar, B., & Singh, V. P. (2012). Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. *Hydrology and Earth System Sciences*, 16(11), 4119–4131. <https://doi.org/10.5194/hess-16-4119-2012>
- Sivapalan, M. (2003). Prediction in ungauged basins: A grand challenge for theoretical hydrology. *Hydrological Processes*, 17(15), 3163–3170. <https://doi.org/10.1002/hyp.5155>
- Soil Conservation Service. (1986). Urban hydrology for small watersheds TR-55.
- Toth, E. (2013). Catchment classification based on characterisation of streamflow and precipitation time series. *Hydrology and Earth System Sciences*, 17(3), 1149–1159. <https://doi.org/10.5194/hess-17-1149-2013>
- Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass*, 1(4), 901–931. <https://doi.org/10.1111/j.1749-8198.2007.00039.x>
- Wang, L., Liu, Z., Guo, J., Wang, Y., Ma, J., Yu, S., et al. (2021). Estimate canopy transpiration in larch plantations via the interactions among reference evapotranspiration, leaf area index, and soil moisture. *Forest Ecology and Management*, 481, 118749. <https://doi.org/10.1016/j.foreco.2020.118749>
- Wang, Y., & Gupta, H. V. (2024). A mass-conserving-perceptron for machine-learning-based modeling of geoscientific systems. *Water Resources Research*, 60(4), e2023WR036461. <https://doi.org/10.1029/2023WR036461>

- Wu, S., Zhao, J., Wang, H., & Sivapalan, M. (2021). Regional patterns and physical controls of streamflow generation across the conterminous United States. *Water Resources Research*, 57(6), e2020WR028086. <https://doi.org/10.1029/2020WR028086>
- Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. *WIREs Water*, 8(5), e1533. <https://doi.org/10.1002/wat2.1533>
- Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, 30(8), 1756–1774. <https://doi.org/10.1016/j.advwatres.2007.01.005>
- Yan, H., Wang, S. Q., Billesbach, D., Oechel, W., Zhang, J. H., Meyers, T., et al. (2012). Global estimation of evapotranspiration using a leaf area index-based surface energy and water balance model. *Remote Sensing of Environment*, 124, 581–595. <https://doi.org/10.1016/j.rse.2012.06.004>
- Yang, Y., & Chui, T. F. M. (2023). Profiling and pairing catchments and hydrological models with latent factor model. *Water Resources Research*, 59(6), e2022WR033684. <https://doi.org/10.1029/2022WR033684>
- Young, P. C. (2015). Refined instrumental variable estimation: Maximum likelihood optimization of a unified Box–Jenkins model. *Automatica*, 52, 35–46. <https://doi.org/10.1016/j.automatica.2014.10.126>
- Yu, Q., Jiang, L., Schneider, R., Zheng, Y., & Liu, J. (2024). Deciphering the mechanism of better predictions of regional LSTM models in ungauged basins. *Water Resources Research*, 60(7), e2023WR035876. <https://doi.org/10.1029/2023WR035876>