CS230

Streamflow Forecasting with Deep Learning Methods

Rahul Dhakal
Department of
Environmental Engineering
rrdhakal@stanford.edu

Josue Fonseca
Department of
Energy Resources Engineering
josuesdf@stanford.edu

Rohan Pasalkar
SPCD
pasalkar@stanford.edu

Abstract

With increased access to large datasets from remote sensing products, Machine Learning scientists in recent years have shown that the use of artificial recurrent neural network like the Long Short-Term Memory networks (LSTMs) are ideal in learning long-term hydrological dependencies which could be generalized for multiple watershed basins. However, with large training datasets common in Hydrology, training LSTMs and their gate (usually 3 to 4) parameters can be very computationally expensive. In this paper, we trained an Adapted Gated Recurrent Network (A-GRU)- which is a GRU adapted to train the static and dynamic input features separately with 2 gates- on a large publicly available dataset with 531 basins over a 19 years period . Our hypothesis was that the A-GRU model would be as efficient as LSTMs in learning the long-term hydrologic dependencies on fewer parameters with the added simplicity of fewer gates and lower computational cost. Our result confirmed this hypothesis with A-GRU showing similar Nash Sutcliffe efficiency (NSE) metric compared to the baseline LSTM model and faster learning requiring lesser iterations on epochs.

1 Introduction

Given the accelerating effects of climate change in the hydrologic cycle and rapid population expansion in the floodplains, precision risk modeling of inland floods is more important and challenging than ever today. Majority of streams around the world are either ungauged or poorly gauged, which is why the rainfall-streamflow modeling in ungauged watershed basins still remains an important challenge in Hydrological forecasting. The traditional models used to forecast the streamflow and floods such as the Sacramento Soil Moisture Accounting [4] (SAC-SMA) are models highly based on heuristic physical processes and are calibrated empirically to fit the hydrological signatures of particular basin(s).

In contrast to the physical ones, data-driven models do not suffer from this limitation of regionalization and can be generalized over a wide range of basins more effectively. In recent years, papers like [4] have shown that the use of artificial recurrent neural networks like the Long Short-Term Memory networks (LSTMs) are ideal in learning long-term hydrological dependencies which could be generalized for multiple watershed basins with improved accuracy of predictions. In this paper, we propose a different architecture based on another commonly used recurrent neural network- Gated Recurrent Units (GRU) which shares the same advantage of having memory cells, but with lesser parameters to train compared to LSTM and therefore- lower computation cost. The input dataset to our model was fetched from a large publicly available repository called Catchment Attributes and Meteorology for

CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017)

Large-Sample Studies (CAMELS) [2] which contains 19 years of hydrometeorological time-series data and static basin attributes for 531 basins across the United States. This is fed into our neural network which has been adapted to take in the time-series and static values separately (trained with separate parameters), and from here on will be referred to as the Adapted GRU, or A-GRU. The model then outputs one single volumetric streamflow value at each time-step which will be compared against the ground truth values also available in CAMELS.

2 Related work

Most recently, [4] compared the results of the conventional hydrological models like SAC-SMA to their LSTM architecture using the Nash Sutcliffe efficiency (NSE) metric - explained in [3] - and had an improvement in the overall streamflow predictions. We discovered that most of the previous approaches of Machine Learning in Hydrology have focused on different versions of LSTM and have reached the same conclusion. The version that [4] used is called Entity-Aware (EA) LSTM where input features are separated into dynamic ($x_d[t]$), and static / time-independent (x_s) vectors (further explained in section 3.2) in the EA-LSTM cells. In each of these cells, input gate is controlled by x_s , whereas $x_d[t]$ controls information in the forget gate and memory gate as illustrated in figure 1. Overall, the authors found that EA-LSTM provides an improved NSE of 0.71 (median of all studied basins) compared to SAC-SMA's 0.64 NSE. We used the Git Repository of EA-LSTM for our baseline model and comparison with the proposed A-GRU, which will be explained further in section 5.

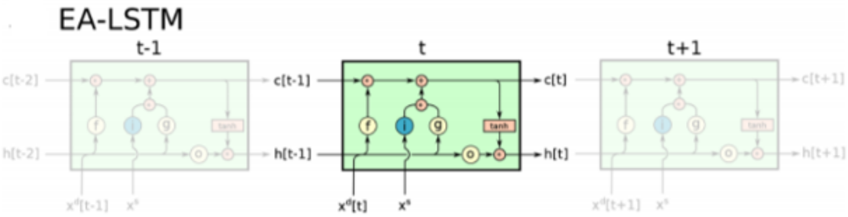


Figure 1: *Visualization of the Entity-Aware-LSTM (EA-LSTM) cell as defined by [4], where $i[t]$, $f[t]$, and $o[t]$ are the input gate, forget gate, and output gate, respectively and $g[t]$ is the cell memory.*

3 Dataset and features

3.1 Dataset description

The data used in this project was found in publicly available Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS) data set curated by the National Center for Atmospheric Research. The CAMELS data set comprises the data collected for 671 basins/ catchments in the U.S. spanning from 1989 to 2008 time-frame, 531 of which will be used for this project. This dataset in A-GRU is divided the same way as our benchmark EA-LSTM model. It was separated into the training and validation sets by [4] as: Training set: 531 basins, 1999 to 2008 (9 years); Development/Evaluation set: 531 basins, 1989 to 1999 (10 years).

3.2 Relevant dataset features and their basic statistics

The CAMELS data is broadly classified into 2 different datasets. First is the hydrometeorological time series introduced in [2]. The second data set is the static catchment attributes introduced in [1, 5]. The input features will consist of 5 time series daily meteorological forcing data: maximum air temperature, minimum air temperature, precipitation, solar radiation, and vapor pressure. Additionally, at each time step, the meteorological inputs will be augmented with 27 different static catchment data ranging from slope, area, elevation, to geological permeability.

In summary, the available input data of interest consists of 32 features: 5 meteorological time-series ($x_d[t]$), augmented with 27 static catchment attributes (x_s) at each time step (daily) for 30 years, in 531 different basins. The full list of features in $x_d[t]$ and x_s are shown in appendix A. As for the ground truth or observed output data, we will use the daily streamflow values over the 20 years time period also found in the CAMELS dataset, as collected by United States Geological Services (USGS).

4 Methods

4.1 Model architecture

The proposed A-GRU architecture is modified to take in two input vectors $x_d[t]$ and x_s at each time step. x_s will be used to augment data at each time step of $x_d[t]$. Our intuition was that an A-GRU would provide the same memory capability of an LSTM in learning long-term hydrological dependencies without explicitly requiring a forget gate and output gate. Another important distinction of A-GRU would be that both $x_d[t]$ and x_s inputs will control information flow on all gates, as opposed to our baseline EA-LSTM where the two input vectors each control different gates (section 2). Our proposed A-GRU model is illustrated in figure 2.

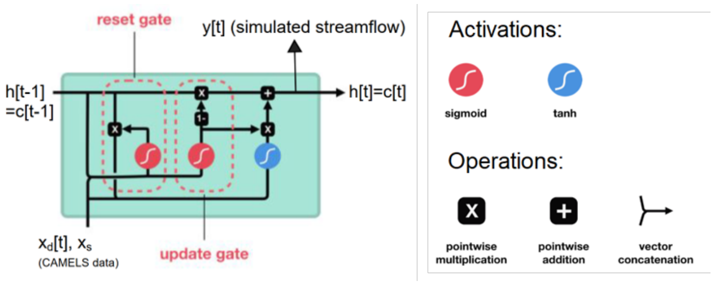


Figure 2: *Visualization of the proposed Adapted-GRU showing the data flow across different gates and activations.*

The eqs. (1) to (5) describe the forward pass through the A-GRU, where i is an input gate, which does not change over time for a basin, x_s are the static inputs and $x_d[t]$ are the dynamic inputs (e.g., meteorological forcings) at time step t $(1 \leqslant t \leqslant T)$. Similarly, $g[t]$ is the cell input, $h[t]$ is the recurrent input, and $c[t]$ is the cell state. One major difference we have in A-GRU compared to EA-LSTM is that it has the reset gate instead of the forget gate, and we no longer use the output gate.

$$i = \sigma(W_i x_s + b_i) \tag{1}$$

$$r[t] = \sigma(W_r x_d[t] + U_r h[t-1] + b_r) \tag{2}$$

$$g[t] = tanh(W_g x_d[t] + U_g h[t-1] + b_g) \tag{3}$$

$$c[t] = (1 - r[t]) * c[t-1] + r[t] * i * g[t] \tag{4}$$

$$h[t] = c[t] \tag{5}$$

4.2 Loss function and evaluation metric

The Loss function used at each time step of A-GRU cell is the Nash-Sutcliffe Efficiency Metric (NSE) [3] as shown in equation 6.B represents the number of basins, N is the number of samples (days) per basin. $\hat{y_n}$ is the prediction of sample n, and y_n is the observation. $s(b)$ is the standard deviation of the discharge in basin b, and ϵ is just a small positive term to prevent zero division.

$$\mathcal{L}_{NSE} = \frac{1}{B} \sum_{b=1}^{B} \sum_{n=1}^{N} \frac{(\hat{y_n} - y_n)^2}{(s(b) + \epsilon)^2} \tag{6}$$

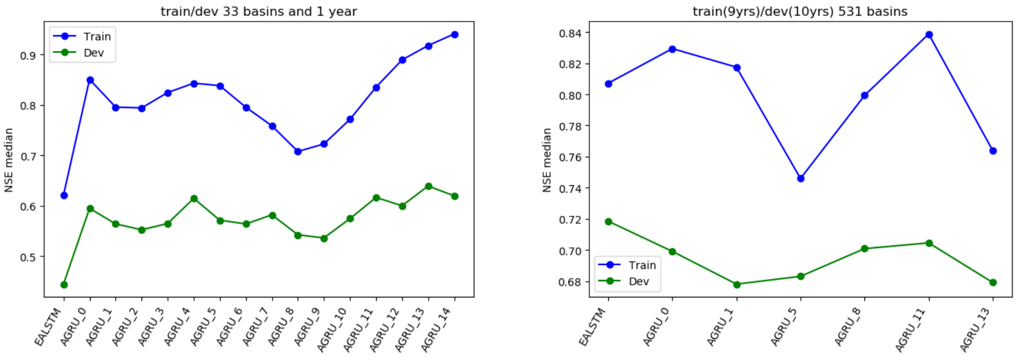
NSE is a commonly used number to evaluate physical models in Hydrology. To compare the model results, we used the NSE shown in equation 7 for each basin and reported the median score across all 531 basins as our evaluation metric. Good predictions have NSE closer to 1.

$$NSE = 1 - \frac{\sum_{n=1}^{N}(\hat{y_n} - y_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{7}$$

5 Experiments/Results/Discussion

5.1 Hyperparameter experiments

In order to tune the hyperparameters of A-GRU in our local machines, we trained a smaller subset of 33 basins over 1 year period and validated it over a different 1 year period. The various combinations of hyperparameters were based on our intuition of working with the model and the baseline, which is reported in appendix B, along with the results. We ran the model on the entire dataset (531 basins over 19 years) using the most promising hyperparameter combinations based on the NSE scores of 33 basins' training and evaluation. Figure 3a illustrates the evaluation metrics (Median NSE scores of 33 basins) for each hyperparameter combination on both training and validation dataset of 33 basins. Using this plot we were able to identify the hyperparameter sets with the least bias-variance gaps, which we then used to tune the hyperparameters of the entire dataset. Results of the latter run on the whole dataset can be seen in figure 3b.



***(a)** NSE median values of the 33 basins using different hyperparameter combinations (EA-LSTM vs. A-GRU)*

***(b)** NSE median values of the entire dataset with selected hyperparameters.*

***Figure 3:** NSE median values as metric. Comparison between training set and development set on **(a)** small and **(b)** whole dataset*

5.2 Results and discussion

As can be seen from figure 3a, AGRU's development set median NSE score came very close (within 98%) to EA-LSTM's median NSE score, but was not able to outperform it. Figure 4 shows the results of running A-GRU and EA-LSTM across 531 basins over the available time-period in statistical curves: Probability Distribution Function (PDF) and Cumulative Distribution Function (CDF). One can note that hyperparameter set 8 of A-GRU results have one of the best NSE distribution with higher probable values. Recall that this set had the lowest variance when analysing only 33 basins on figure 3a.

Furthermore, figure 5 shows our justification for invoking early-stopping on our hyperparameter set 8 of A-GRU. We see that at epoch 15, set 8 yields a much better median NSE score over the dataset compared to epoch 30, which is what the baseline EA-LSTM is tuned to. Hence, with a combination of hyperparameter tuning and early stopping, we were able to get a much better performance out of A-GRU which is comparable to EA-LSTM results. Please refer to appendix B for all our intermediate results of hyperparameter tuning efforts.

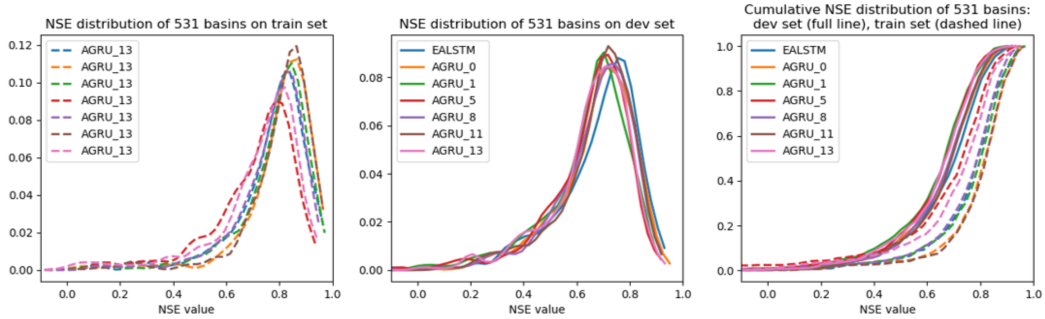


Figure 4: *Statistical analyses (probability distribution and cumulative distribution functions) of selected hyperparameter sets for A-GRU as well as the baseline EA-LSTM.*

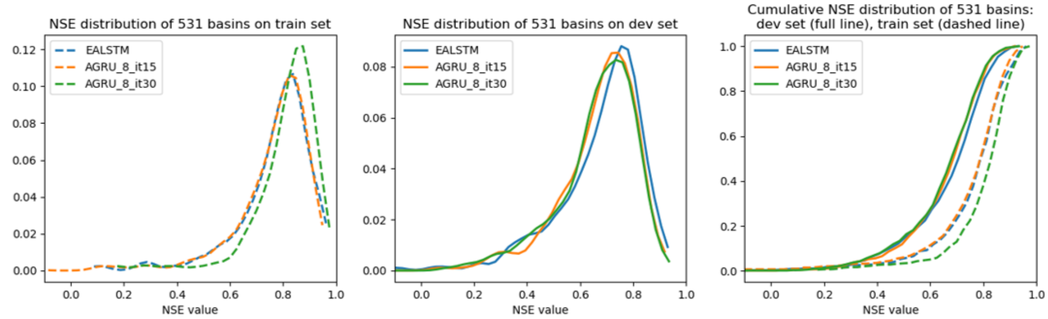


Figure 5: *Statistical analyses of epochs 15 and 30 of hyperparameter set 8 in A-GRU compared to EA-LSTM baseline.*

6 Conclusion/Future Work

With the increasing digitization of hydrological data with remote sensing products, computational cost will become crucial in future hydrological data-driven endeavors. Although researchers have shown that LSTM yields promising results compared to the physical heuristic models, it has high computational cost associated with its large number of learnable parameters and gates. In this paper, we were able to conclude that a simpler Recurrent Neural network model like the Adapted-GRU was able to perform almost as well as the EA-LSTM baseline (based on NSE score metric) with lesser learnable parameters,epochs, and overall runtime. Moving forward, we believe reducing the learning rate and increasing the regularization efforts in our model could increase its performance even further, possibly surpassing the baseline EA-LSTM.

7 Contributions

All team members with different academic/professional backgrounds contributed to each project task uniquely in delivering the final products.

References

[1] N. Addor, A. J. Newman, N. Mizukami, and M. P. Clark. The camels data set: catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.*, 21:5293–5313, 2017.

[2] N. Addor, A. J. Newman, N. Mizukami, and M. P. Clark. Catchment attributes for large-sample studies, 2017.

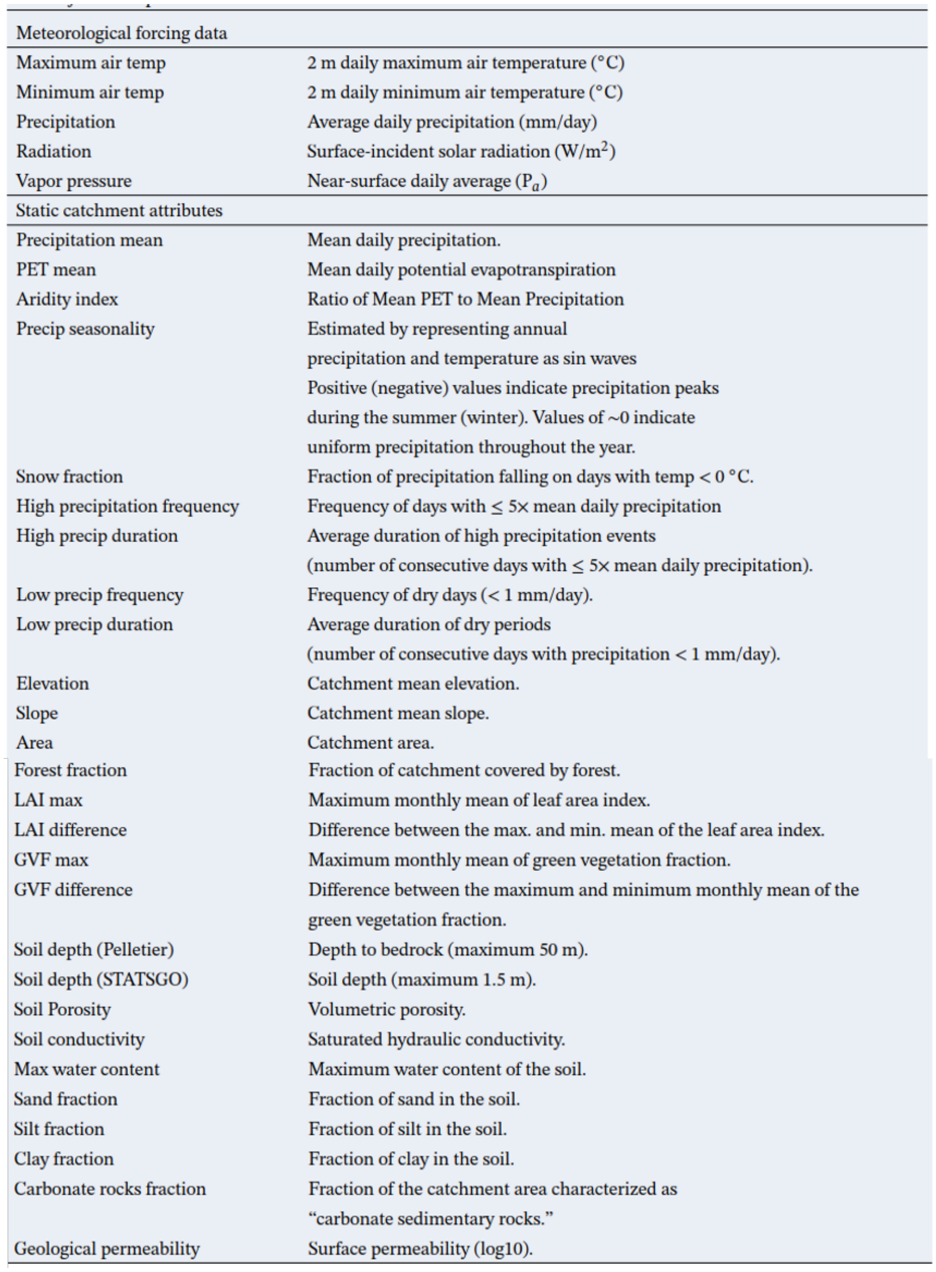
[3] Hoshin Gupta and Harald Kling. On typical range, sensitivity, and normalization of mean squared error and nash-sutcliffe efficiency type metrics. *Water Resources Research - WATER RESOUR RES*, 47, 10 2011.

[4] Frederik Kratzert, Daniel Klotz, Guy Shalev, Günter Klambauer, Sepp Hochreiter, and Grey Nearing. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology & Earth System Sciences*, 23(12), 2019.

[5] A. Newman, K. Sampson, M. P. Clark, A. Bock, R. J. Viger, and D. Blodgett. A large-sample watershed-scale hydrometeorological dataset for the contiguous usa, 2014.

A Appendix: input data

The figure 6 reports the complete list of 32 input features (both static and dynamic) at each time step of A-GRU.

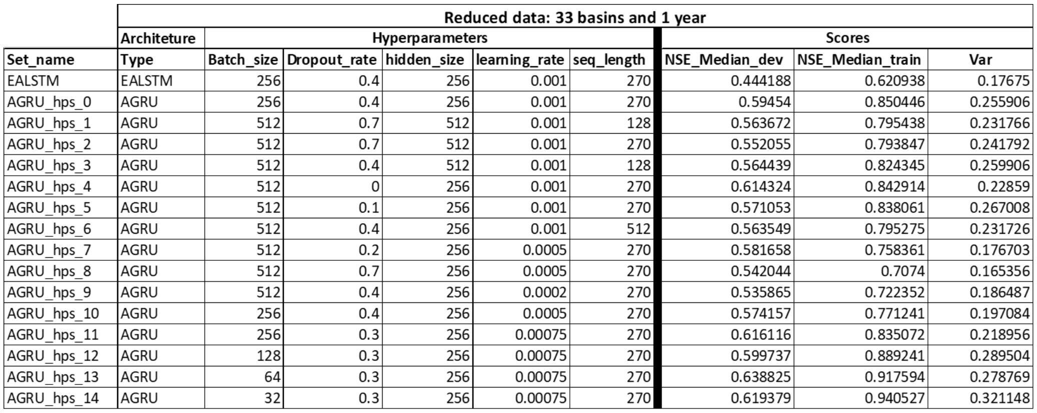


Meteorological forcing data	
Maximum air temp	2 m daily maximum air temperature (°C)
Minimum air temp	2 m daily minimum air temperature (°C)
Precipitation	Average daily precipitation (mm/day)
Radiation	Surface-incident solar radiation (W/m^2)
Vapor pressure	Near-surface daily average (P_a)
Static catchment attributes	
Precipitation mean	Mean daily precipitation.
PET mean	Mean daily potential evapotranspiration
Aridity index	Ratio of Mean PET to Mean Precipitation
Precip seasonality	Estimated by representing annual precipitation and temperature as sin waves Positive (negative) values indicate precipitation peaks during the summer (winter). Values of ~0 indicate uniform precipitation throughout the year.
Snow fraction	Fraction of precipitation falling on days with temp < 0 °C.
High precipitation frequency	Frequency of days with ≤ 5× mean daily precipitation
High precip duration	Average duration of high precipitation events (number of consecutive days with ≤ 5× mean daily precipitation).
Low precip frequency	Frequency of dry days (< 1 mm/day).
Low precip duration	Average duration of dry periods (number of consecutive days with precipitation < 1 mm/day).
Elevation	Catchment mean elevation.
Slope	Catchment mean slope.
Area	Catchment area.
Forest fraction	Fraction of catchment covered by forest.
LAI max	Maximum monthly mean of leaf area index.
LAI difference	Difference between the max. and min. mean of the leaf area index.
GVF max	Maximum monthly mean of green vegetation fraction.
GVF difference	Difference between the maximum and minimum monthly mean of the green vegetation fraction.
Soil depth (Pelletier)	Depth to bedrock (maximum 50 m).
Soil depth (STATSGO)	Soil depth (maximum 1.5 m).
Soil Porosity	Volumetric porosity.
Soil conductivity	Saturated hydraulic conductivity.
Max water content	Maximum water content of the soil.
Sand fraction	Fraction of sand in the soil.
Silt fraction	Fraction of silt in the soil.
Clay fraction	Fraction of clay in the soil.
Carbonate rocks fraction	Fraction of the catchment area characterized as "carbonate sedimentary rocks."
Geological permeability	Surface permeability (log10).

Figure 6: *Table with input features to the network.*

B Appendix: Hyperparameters tuning

The figure 7 reports the hyperparameters set used on the small set with A-GRU network. It also has the hyperparameters used by [4] on EA-LSTM network. Note that only some set were used for the whole dataset.



	Reduced data: 33 basins and 1 year								
	Architeture	Hyperparameters					Scores		
Set_name	Type	Batch_size	Dropout_rate	hidden_size	learning_rate	seq_length	NSE_Median_dev	NSE_Median_train	Var
EALSTM	EALSTM	256	0.4	256	0.001	270	0.444188	0.620938	0.17675
AGRU_hps_0	AGRU	256	0.4	256	0.001	270	0.59454	0.850446	0.255906
AGRU_hps_1	AGRU	512	0.7	512	0.001	128	0.563672	0.795438	0.231766
AGRU_hps_2	AGRU	512	0.7	512	0.001	270	0.552055	0.793847	0.241792
AGRU_hps_3	AGRU	512	0.4	512	0.001	128	0.564439	0.824345	0.259906
AGRU_hps_4	AGRU	512	0	256	0.001	270	0.614324	0.842914	0.22859
AGRU_hps_5	AGRU	512	0.1	256	0.001	270	0.571053	0.838061	0.267008
AGRU_hps_6	AGRU	512	0.4	256	0.001	512	0.563549	0.795275	0.231726
AGRU_hps_7	AGRU	512	0.2	256	0.0005	270	0.581658	0.758361	0.176703
AGRU_hps_8	AGRU	512	0.7	256	0.0005	270	0.542044	0.7074	0.165356
AGRU_hps_9	AGRU	512	0.4	256	0.0002	270	0.535865	0.722352	0.186487
AGRU_hps_10	AGRU	256	0.4	256	0.0005	270	0.574157	0.771241	0.197084
AGRU_hps_11	AGRU	256	0.3	256	0.00075	270	0.616116	0.835072	0.218956
AGRU_hps_12	AGRU	128	0.3	256	0.00075	270	0.599737	0.889241	0.289504
AGRU_hps_13	AGRU	64	0.3	256	0.00075	270	0.638825	0.917594	0.278769
AGRU_hps_14	AGRU	32	0.3	256	0.00075	270	0.619379	0.940527	0.321148

Figure 7: *Table with hyperparameters values for each tested set on small data. The scores of figure 3a are also reported.*

The figure 8 illustrates the NSE distributions, using the small set, on the used basins.

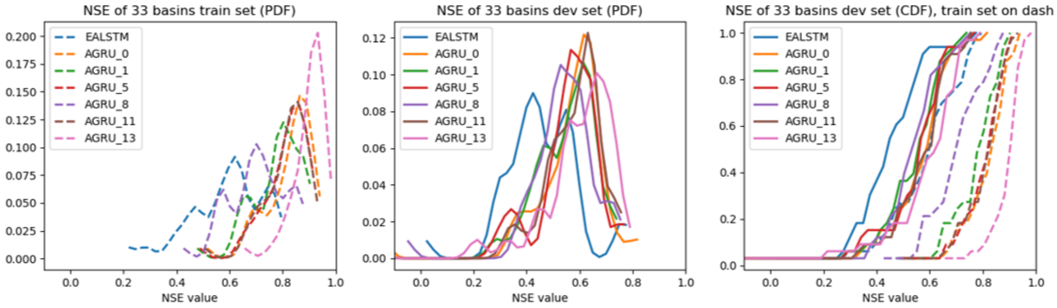


Figure 8: *Statistical analyses (probability distribution and cumulative distribution functions) of selected hyperparameter sets for A-GRU as well as the baseline EA-LSTM on the small set.*