A glimpse into the Unobserved: Runoff simulation for ungauged catchments with LSTMs

Frederik Kratzert[*] kratzert@ml.jku.at
Daniel Klotz[*] klotz@ml.jku.at
Mathew Herrnegger [†] mathew.herrnegger@boku.ac.at
Sepp Hochreiter[*] hochreit@ml.jku.at

Abstract

Runoff predictions of a river from meteorological inputs is a key task in the field of hydrology. However, current hydrological models require a substantial amount of parameter tuning on basis of historical records. If no historical runoff observations are available it is very challenging to produce good predictions. In this study we explore the capability of LSTMs for simulating the runoff for these ungauged cases. A single LSTM is trained to learn a general hydrological model from hundreds of catchments throughout the contiguous United States of America and evaluated against catchments not used during training. Our results suggest that LSTMs a) are able to learn a general hydrological model and b) in the majority of catchments outperform an established hydrological model, which was especially trained for these catchments.

1 Introduction

Runoff predictions from meteorological observations provide the basic information for the management of water resources, the design of hydropower plants and the planning of irrigation schemes. They also provide an important backbone to reduce the damages and casualties from floods, which are among the most frequent and destructive natural hazards [2]. Between 1980 and 2017 the World was affected by almost 6000 damaging flood events that claimed over 220000 lives and produced overall economic losses of USD 1007 billion [13]. Notwithstanding that the monetization of human losses is non-trivial, this monetary value can be put in perspective: It approximately correspond to the Gross Domestic Product (GDP) of Indonesia in 2017 (the 16th largest national economy in the world [16]).

Currently, most runoff predictions are based on (hydrological) models that require extensive parameter tuning on basis of historical runoff records. According to Worldbank [17] however, 80 % of the hydro-meteorological observation networks in middle to low-income countries are in a poor or declining state or inadequate to meet user needs. But, also in industrial, high-income nations the number of hydrological measurements are declining, e.g. in the USA 2632 stream gauges with 30 or more years of runoff records were discontinued between 1972 and 2006 [15]. Missing observations can make the task of model calibration challenging. Although more and more hydrological relevant data is becoming available due to space-borne remote sensing products, the opposite is the case for in-situ data [3]. Accordingly, forecasting the runoff of ungauged catchments (i.e. catchments without

[*]LIT AI Lab & Institute for Machine Learning, Johannes Kepler University Linz, Austria
[†]Institute of Water Management, Hydrology and Hydraulic Engineering, University of Natural Resources and Life Sciences, Vienna, Austria

Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NeuRIPS 2018), Montréal, Canada.

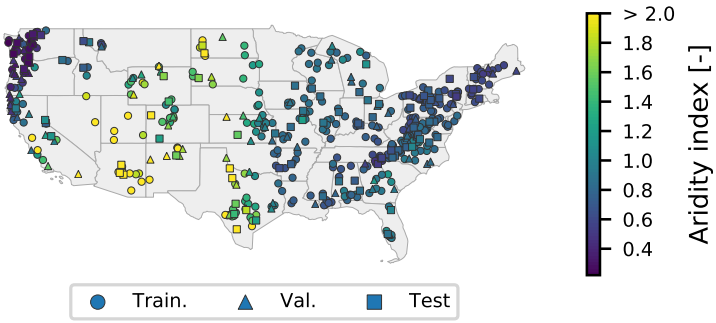

Figure 1: Location of the basins used in this study across the contiguous United States of America. The marker type shows exemplarily the data split for one of the 6 folds and the color of the marker depicts the aridity of the basin.

historic observations, where a catchment is the drainage area of a river) is seen as a key challenge in hydrology. The International Association of Hydrological Sciences even devoted an entire scientific decade to the to address the problem of ungauged catchments ("Predictions in Ungauged Basins" in the period 2003-2012). And, albeit this coordinated effort brought many advances, the central goal remains a challenge [7].

Recently, Kratzert et al. [9] have shown that LSTMs are well suited for rainfall-runoff modelling. In this study, we explore the capability of LSTMs for the task of predicting the runoff in ungauged catchments. We train a single LSTM to learn a general hydrological model of hundreds of catchments in various regions in the contiguous United States of America and assess if the model can simulate catchments not used in the training.

2 Case Study

2.1 Data and Setup

We use the publicly available CAMELS dataset [10] and the same 531 catchments (see Fig. 1) used by Newman et al. [12]. For each catchment approximately 35 years of catchment-aggregated daily meteorological observations (precipitation, min. and max. temperature, solar radiation and vapor pressure) as well as runoff records are available. A single day of runoff is predicted from the previous 365 days of meteorological observations. We hypothesize that this window is large enough to account for most long-term dependencies inherent in the system, such as e.g. snow accumulation and melting.

Additionally, we use a selection of the catchment attributes released by Addor et al. [1]. These attributes describe a wide range of characteristics for the given catchments (e.g. the corresponding climate and geography). We only use the set of attributes, that are not calculated from in-situ observations. This guarantees that the model can be used for catchments with missing measurement stations (for a comprehensive list of attributes see Appendix A).

We use k-fold cross-validation (with $k = 6$) to assess the capacity of the approach. The training set consists of four splits (353 catchments) and the validation and test set of one split each (which amounts to 89 catchments each). This setup is realized for two different settings: In the first setting, we only provide the meteorological time series data as input to the LSTM (hereafter: *Baseline LSTM*). In the second setting, we add the static catchments attributes as additional inputs for each time step (hereafter: *Global LSTM*). The comparison of both settings allows us to examine, if these additional features enable the LSTM to learn a "catchment-characteristic aware" hydrological model, which has the property to simulate a wider array of different hydrological responses.

Furthermore, we added two simple data-driven baselines. In the first one, we determine for each test set basin the most similar basin in the training set by using the minimal euclidean distance in the feature space of the catchment attributes. For this most similar basin we calculate the long-term mean discharge for each day of the year and use these values as prediction for the entire period of the test

basin (hereafter: *NN*). The second baseline is a simple multilayer perceptron (*MLP*) which uses the entire input sequence plus the catchment attributes as one long input vector. The MLP consists of a single layer with 45 hidden units to approximately match the number of learnable parameters of the Global LSTM.

The CAMELS dataset also includes a hydrological reference model (SAC-SMA + Snow-17, hereafter *SAC-SMA*, for details see [11]), against which we also compare. It can be seen as an upper benchmark, since it is a well-established model that was calibrated for each catchment specifically (using the first 15 years of available data).

2.2 Model

For the sake of simplicity we use a Vanilla LSTM [6, 4] with a single layer and 125 hidden neurons. The final runoff prediction is calculated by a dense layer from the output of the LSTM layer at the last time step. We use a batch size of 256, the mean-squared-error (MSE) as loss function and train the model using Adam optimizer [8] with a learning rate of $1 * 10^{-3}$ ($\beta_1 = 0.9, \beta_2 = 0.999$). We train the models for various epochs and select the final model by the lowest mean MSE of all catchments in the validation set. For the final evaluation we use the Nash-Sutcliffe-Efficiency (NSE), the most common measure in hydrology [5], which equals the R^2 of the observed and simulated discharge. Furthermore, because negative predictions of the runoff are physically implausible, we clip negative predictions to 0.

3 Results and Discussion

Figure 2 shows the results of our experiments, as well as the performance of the hydrological reference model. The two data-driven benchmarks perform (expectedly) bad, with both having negative average NSE. This means that using the mean of the runoff (assuming availability of observations) as a prediction would on average outperform both approaches. From Fig. 2 we can also see that adding catchment attributes to the meteorological inputs helps the LSTM in general to increase its performance (Comparison Baseline vs. Global LSTM): The Global LSTM (which receives additionally catchment attributes as inputs) has a higher median, higher mean, and a more skewed distribution towards better model performance in general. This is an expected result, since we assume that the additional features (i.e. the catchment attributes) provide additional information which can help the LSTM e.g. to cluster the basins internally and learn different hydrological behaviours for different catchment groups.

More interestingly however, is the fact that the Global Model (here we compare the Global Model*, for which the NSE values are calculate for the same period as the validation period of the SAC-SMA) has also a higher median (0.68) and higher 75th percentile (0.76) compared to the hydrological reference model, the SAC-SMA (0.65 and 0.74, respectively). This is a somewhat unexpected result for two reasons: (i) The SAC-SMA is a well established hydrological model that is used also by federal agencies, such as the United States National Weather Service. (ii) We compare a model (LSTM) that has never seen data of a specific catchment against a hydrological model that was especially calibrated for this catchment. And still in 293 out of the 531 catchments (55 %), the (Global) LSTM achieves higher accuracies compared to the SAC-SMA.

From Fig. 3b it can be seen that SAC-SMA is predominantly better in the regions of arid catchments (aridity > 1, see Fig. 1 for a reference). These basins also represent the lower outliers in the boxplot and explain, why the mean is slightly lower in the Global LSTM, compared the the SAC-SMA. Arid catchments are in general difficult to model, due to their hydrological behavior: During very long periods (sometimes even for multiple years), the river in these catchments dry-out because of too little rain (and/or in combination with a high potential evapotranspiration). Training a data-driven model, like LSTMs, on these periods provides very little information in the error signal. Runoff values above 0 give more information concerning the current underlying hydrological characteristics. For the case that no runoff is observed, in contrast, the interpretation of the underlying hydrological system becomes significantly more difficult. Simulating this special case of a threshold process is non-trivial, in general for all model types. For learning to predict the few days with actual runoff, very few data samples are effectively available. It might thus be beneficial to treat arid catchments similar to classification problems with unbalanced data and to oversample the data points for which a runoff signal greater zero exists. In contrast, classical hydrological models, like the SAC-SMA have

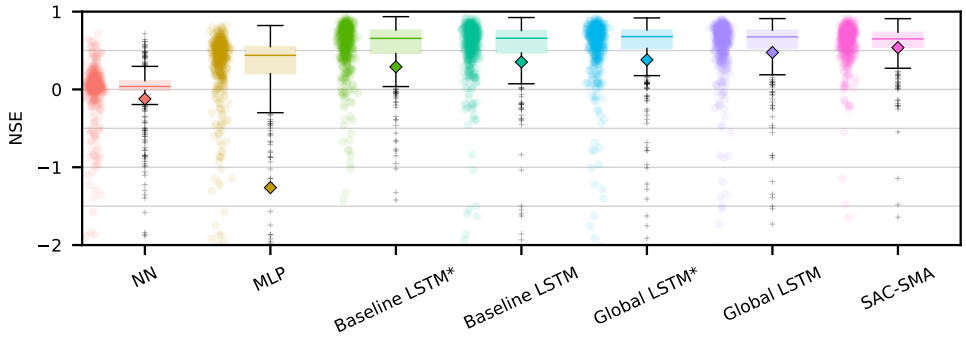

Figure 2: Boxplot of the model performances. Both LSTM variants with an asterisk (*) mean that the NSEs were only calculated for a part of the time series, which corresponds to the validation period of the SAC-SMA. The horizontal lines mark the median, the squares the mean, the box the 25th and 75th percentile and the black horizontal lines the 5th and 95th percentile, respectively. Underlying data is plotted at the side. Boxplot is capped at -2 for better clarity.

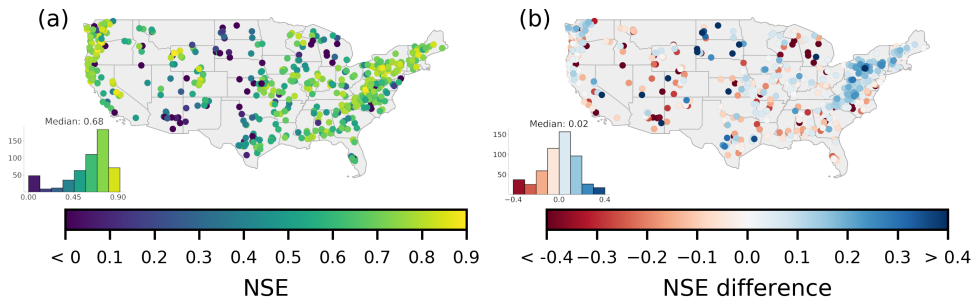

Figure 3: (a) Shows the NSE of each Basin of the Global LSTM*. (b) Shows the difference between the NSE of the Global LSTM* and the SAC-SMA hydrological reference model, where positive numbers (blue colors) mean the Global LSTM* outperforms the SAC-SMA and negative numbers (red colors) the opposite.

the advantage of (a) knowing basic physical laws like mass balance and (b) having the knowledge of these arid regions and their behaviour already implemented in the model routine. Generally speaking, it is however also worth mentioning that SAC-SMA also performs bad for said regions (see [11]).

Additionally, we can see that in areas with a high catchment density, especially the East and West coast, the LSTM seems to perform better (see Fig. 3b for a comparison between Global LSTM and SAC-SMA). This could imply that in the regions with a larger number of similar behaving catchments a beneficial cross learning (between catchments) takes place. If we abstract this result, it could mean that if we include data from catchments world-wide, we could increase the model performance due to more samples of all kind of catchment types in general.

4 Conclusion

Predictions in ungauged basins is a major challenge in the field of hydrology. In this study we presented a new data-driven approach, using a single LSTM with meteorological inputs and catchment attributes for hundreds of catchments. The so trained LSTM is able to achieve comparable model performance for unseen catchments compared to the well established SAC-SMA (which, in contrast, was especially calibrated for each of the catchments). Two future studies may follow this one: One study could consists of aggregating data from catchments all around the world and to train a single LSTM for the entire world. Another one could be to use ConvLSTMs [14] with gridded input data (from e.g. satellite products) in contrast to catchment-aggregated values used in this study.

References

[1] N. Addor, A. Newman, N. Mizukami, and M. Clark. Catchment attributes for large-sample studies, boulder, co, ucar/ncar, 2017.

[2] CRED and UNISDR. The human cost of weather-related disasters, 1995–2015. *United Nations, Geneva*, 2015.

[3] B. M. Fekete, R. D. Robarts, M. Kumagai, H.-P. Nachtnebel, E. Odada, and A. V. Zhulidov. Time for in situ renaissance. *Science*, 349(6249):685–686, 2015.

[4] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.

[5] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377 (1-2):80–91, 2009.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] M. Hrachowitz, H. Savenije, G. Blöschl, J. McDonnell, M. Sivapalan, J. Pomeroy, B. Arheimer, T. Blume, M. Clark, U. Ehret, et al. A decade of predictions in ungauged basins (pub)—a review. *Hydrological sciences journal*, 58(6):1198–1255, 2013.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Rainfall-runoff modelling using long short-term memory (lstm) networks. *Hydrol. Earth Syst. Sci. Discuss. in review*, 2018.

[10] A. Newman, K. Sampson, M. Clark, A. Bock, R. Viger, and D. Blodgett. A large-sample watershed-scale hydrometeorological dataset for the contiguous usa. *UCAR/NCAR, doi*, 10:D6MW2F4D, 2014.

[11] A. Newman, M. Clark, K. Sampson, A. Wood, L. Hay, A. Bock, R. Viger, D. Blodgett, L. Brekke, J. Arnold, et al. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous usa: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19(1):209–223, 2015.

[12] A. J. Newman, N. Mizukami, M. P. Clark, A. W. Wood, B. Nijssen, and G. Nearing. Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, 18(8):2215–2225, 2017.

[13] M. RE. Natural catastrophe know-how for risk management and research. *Available at http://natcatservice.munichre.com/ Accessed 19 Sep. 2018*, 2018.

[14] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[15] USGS. Streamgage history. *Available at https://water.usgs.gov/nsip/history.html. Accessed 19 Sep. 2018*, 2018.

[16] Worldbank. World development indicators. *Available at http://wdi.worldbank.org/tables. Accessed 3 July 2018*, 2017.

[17] Worldbank. Assessment of the state of hydrological services in developing countries. *Available at https://www.gfdrr.org/sites/default/files/publication/state-of-hydrological-services_web.pdf. Accessed 19 Sep. 2018*, 2018.

A List of catchment attributes used in this study

Table 1: Table of catchment attributes used in this experiments. Description taken from the data set [1]

Attribute	Description
p_mean	Mean daily precipitation.
pet_mean	Mean daily potential evapotranspiration.
aridity	Ratio of mean PET to mean precipitation.
p_seasonality	Seasonality and timing of precipitation. Estimated by representing annual precipitation and temperature as sin waves. Positive (negative) values indicate precipitation peaks during the summer (winter). Values of approx. 0 indicate uniform precipitation throughout the year.
frac_snow_daily	Fraction of precipitation falling on days with temperatures below $0°C$.
high_prec_freq	Frequency of high precipitation days (>= 5 times mean daily precipitation).
high_prec_dur	Average duration of high precipitation events (number of consecutive days with >= 5 times mean daily precipitation).
low_prec_freq	Frequency of dry days (< 1 mm/day).
low_prec_dur	Average duration of dry periods (number of consecutive days with precipitation < 1 mm/day).
gauge_x	Transformed x-coordinated in 3D-space from latidude and longitude.
gauge_y	Transformed y-coordinated in 3D-space from latidude and longitude.
gauge_z	Transformed z-coordinated in 3D-space from latidude and longitude.
elev_mean	Catchment mean elevation.
slope_mean	Catchment mean slope.
area_gages2	Catchment area.
forest_frac	Forest fraction.
lai_max	Maximum monthly mean of leaf area index.
lai_diff	Difference between the maximum and minimum mean of the leaf area index.
gvf_max	Maximum monthly mean of green vegetation fraction.
gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction.
soil_depth_pelletier	Depth to bedrock (maximum 50m).
soil_depth_statsgo	Soil depth (maximum 1.5m, layers marked as water and bedrock were excluded).
soil_porosity	Volumetric porosity.
soil_conductivity	Saturated hydraulic conductivity.
max_water_content	Maximum water content of the soil.
sand_frac	Fraction of sand in the soil.
silt_frac	Fraction of silt in the soil.
clay_frac	Fraction of clay in the soil.
carb_rocks_frac	Fraction of the catchment area characterized as "Carbonate sedimentary rocks".
geol_permeability	Surface permeability (log10).