On the Assessment of Several Regression-Based Machine Learning Approaches in Runoff Prediction

Sadegh Sadeghi Tabas
Department of Civil Engineering
CUID: C93338799
sadeghs@clemson.edu

Abstract

Runoff prediction from meteorological observations provides the basic information for the management of water resources, the design of hydro-power plants and the planning of irrigation schemes. They also act as a backbone in reducing the damages and casualties from floods, which are among the most frequent and destructive natural hazards. Various approaches exist, ranging from physically based over conceptual to fully data-driven models. In this research a number of data driven (machine learning) methods including vanilla linear regression, lasso and ridge regression, multi-layer perceptron (MLP) and support vector machine (SVM) have been implemented to predict sequential flow rate values based on a set of collected runoff factors in Stevens Creek basin, located in SC, USA. This study has focused on the effects of input data characteristics on model performance (sequential data), therefore, the type of input data, the amount of input data, and the correlation of the data series has been considered. The data sets are gathered from CAMELS database which provided by Newman et. al (2015) and is publicly available at the *UCAR website*. In order to predict runoff at time step (t), meteorological forcing data at time step (t) as well as antecedent day runoff (t-1) introduced as inputs to the models. To evaluate the mentioned methods, the available dataset divided to two periods of time (train and test periods) and the employed prediction models were validated and tested using NSE, KGE and TRMSE criteria. The findings of this study suggest the potential of applying data-driven models in the field of hydrological runoff prediction.

1 Introduction

1.1 Problem Specification

Watershed simulations help understand the past and current state of rainfall-runoff processes in basins and provide a way to explore the implications of management and planning decisions and imposed changes (such as land use change, climate change). In complex environmental systems such as the coastal plain watersheds, very significant modeling efforts have gone through lumped and physical based simulations (Sadeghi-Tabas et al., 2017; Samadi et al., 2020, 2017; Samadi and Meadows, 2017). Indeed, in pursuit of improved accuracy, there is a plethora of hydrological tools that have been enhanced and implemented for complex rainfall-runoff simulations. In the realm of process-based hydrologic modeling, time-series machine learning (ML) or data-driven algorithms have been recently utilized to improve short- and long-term streamflow predictions at various scales and domains (e.g., Fang et al., 2017; Kratzert et al., 2018; Shen, 2018; Shen et al., 2018; Wu et al., 2018).

Data-driven algorithms attempt to estimate the mapping function (f) from the input variables (x) to numerical or continuous output variables (y) and to understand time-series natural temporal ordering

across long-term records. These methods can directly learn patterns hidden in time-series data, without requiring manually-designed features or making strong physical assumptions (Kasiviswanathan et al., 2016; LeCun et al., 2015; Schmidhuber, 2015) and are good at handling large datasets with high dimensionality and heterogeneous feature types. Many studies have demonstrated that ML models can outperform other state-of-the-art techniques in hydrologic simulation (Yang et al., 2020).

1.2 Related Studies

To give a few selective examples, Sivapragasam and Muttil, (2005) studied the application of SVM to extend the development of rating curves at three gauging stations in Washington, USA. Their results indicated that SVM was better suited for rating curves extrapolation compared to widely used logarithmic method and higher order polynomial fitting method. Sadler et al., (2018) applied RF methods for the coastal urban stormwater prediction in Virginia, USA. They used quality-controlled, crowd-sourced street flooding reports ranging from 1 to 159 per storm event for 45 storm events to train and evaluate RF models. Their results showed that RF performed better than Poisson regression at predicting the number of flood reports and had a lower false negative rate. Bui et al., (2016) proposed a new artificial intelligence approach based on neural fuzzy inference system and metaheuristic optimization for flood susceptibility modeling (MONF) over the Tuong Duong district in Central Vietnam. They found that MONF outperformed other machine learning algorithms including Multi-layer Perceptron (MLP) and Decision Tree for flood susceptibility mapping.

Despite the expanded use of ML models in streamflow simulation, few studies have used advanced data-driven methods to model streamflow within complex hydrological environments. The closest work may be the statistical analysis of streamflow records in the United States to compute the probabilities of high and low flow events in the past several decades along with the projected changes in the coming decades (Asadieh and Krakauer, 2017; Campbell et al., 2011; Hidalgo et al., 2009).

1.3 Motivation and Novelty

In this study we are going to predict daily runoff for Stevens Creek basin, SC, USA, using various state-of-the-art ML methods as data-driven approaches. Data-driven approaches can be appropriate tools for runoff simulation due to the complexity of modeling rainfall-runoff processes, which makes using a physical model difficult. This study investigated and compared five different ML algorithms including vanilla Linear Regression, Lasso Regression, Ridge Regression, Multilayer Perceptron (MLP), and Support Vector Machine (SVM) in runoff prediction. Proposed modeling approach is divided into multiple stages. In order to impute the missing values with an accurate estimate, first stage performs pre-processing of the available datasets. In the second stage, all data-driven models are trained and saved for inference in the next stage. And finally in the last stage the trained models validated and tested for runoff prediction using various performance assessment criteria.

2 Methodology

2.1 Case study: Stevens Creek Basin

In order to investigate the application of ML methods in runoff prediction the Stevens Creek Basin located in South Carolina, US (USGS Guage ID: 02196000) has been selected as study area. The catchment drains 350 squared kilometers and the underlying data is retrieved from the CAMELS data set which is provided by Newman et al. (2015) and Addor et al. (2017) and is publicly available at the *UCAR website*[1]. The data set contains catchment meteorological forcing data and observed discharge at the daily timescale. The catchment meteorological forcing data consists of precipitation, shortwave downward radiation, maximum and minimum temperature, and vapor pressure attributes. In this research all meteorological forcing data as well as antecedent day runoff considered as input to each ML method.

[1] `https://ral.ucar.edu/solutions/products/camels`

2.2 Machine Learning Models

Machine learning algorithms are powerful models that first consider the learning styles that an algorithm can adopt. These approaches can be divided into 3 broad categories (i) supervised learning, (ii) unsupervised learning, and (iii) reinforcement learning. Supervised learning is useful in cases where a property (label) is available for a certain dataset (training set) but is missing and needs to be predicted for other instances. Unsupervised learning is useful in cases where the challenge is to discover implicit relationships in a given unlabeled dataset (items are not pre-assigned). Reinforcement learning falls between these 2 extremes; there is some form of feedback available for each predictive step or action, but no precise label or error message. In this study, we focused on using several supervised learning models as regression-based approaches to simulate daily runoff values. The machine learning methods implemented in this study are including Vanilla Linear Regression, Ridge Regression, Lasso Regression, Multilayer Perceptron (MLP) and Support Vector Machine (SVM). Because of maximum page limitation the model description is not presented in this report as they discussed in-detail in our class. The schematic figure of the framework of our modeling approach is presented in Figure 1.

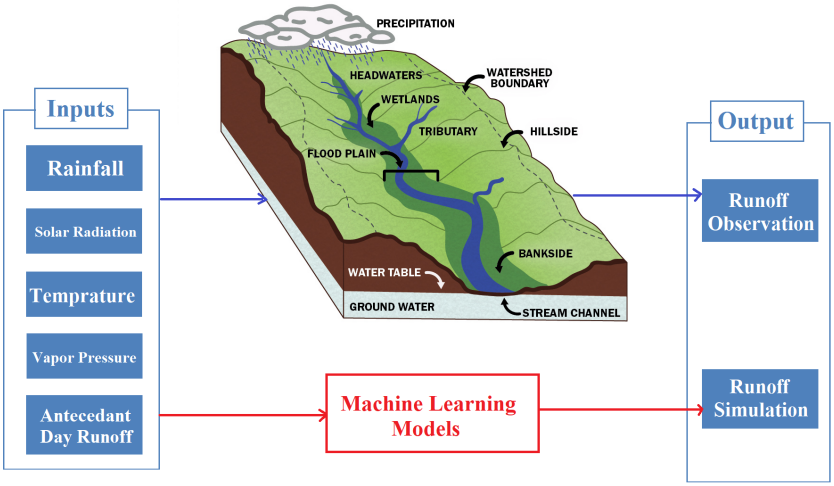

Figure 1: The schematic framework of the modeling approach for runoff prediction

2.3 Performance Assessment Criteria

Because no one evaluation metric can fully capture the consistency, reliability, accuracy, and precision of a streamflow model, it was necessary to use a variety of performance metrics for model benchmarking (Hoshin Vijai Gupta et al., 1998). The metrics for model evaluation are the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) which emphesizes on high flows, the three decompositions following Hoshin V Gupta et al. (2009) which are the correlation coefficient of the observed and simulated discharge (r), the variance bias (α) and the total volume bias (β). These three metrics are combined and presented in the Kling-Gupta efficiency (KGE; Gupta et al., 2009) metric presented in equation 2. And the third criterion emphasizes low flow simulation errors using the Box-Cox transformed root-mean-square error (TRMSE; Box and Cox, 1964) presented in equation 3. in these equation Q_s is the simulated runoff, Q_o observed runoff and $\lambda = 0.3$.

$$NSE = 1 - \frac{\sum_{t=1}^{t}(Q_s^t - Q_o^t)^2}{\sum_{t=1}^{t}(Q_o^t - \bar{Q_o})^2} \tag{1}$$

$$KGE = 1 - \sqrt{(cc-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{2}$$

$$TRMSE = \sqrt{\sum_{t=1}^{n} \frac{(Q_s^t - Q_o^t)^2}{n}}, \quad where: \quad Q = \frac{(1+Q)^\lambda - 1}{\lambda} \tag{3}$$

3 Results and Discussion

3.1 Preprocessing

The dataset retieved from CAMELS database for the study area and some missing values imputed to have a consistent and reliable dataset for modeling process. Based on related studies in the area of runoff prediction using ML methods, antecedent values of input variables could be considered to represent the catchment characteristics in a data-driven model (Solomatine and Xue, 2004; Tongal and Booij, 2018). To do so, in this study the most significant input combination of precipitation (P), maximum and minimum temperatures (Tmax and Tmin), solar radiation (Srad), vapor pressure (VP) in addition to the antecedent day streamflow (Q_{t-1}) considered for runoff generation. After a number of experiments on the training dataset aimed at finding the relevant inputs, the following common input structure was selected:

$$Q_t^s = f(P_t, T_t^{max}, T_t^{min}, Srad_t, VP_t, Q_{t-1}) \tag{4}$$

Prior to the model development, normalization was applied to all input variables as:

$$X_{norm} = \frac{X_o - \mu_x}{\sigma_x} \tag{5}$$

where X_norm and X_o indicate the scaled and original data, $_x$ and $_x$ represent the mean and standard deviation of the original data, respectively.

3.2 Model Setup and Hyperparameter Configuration

As of today, a widespread calibration strategy for the ML models is to subdivide the data into two parts, referred to as training and test datasets. The first split is used to derive the parametrization of the models (calibration in the context of hydrologic simulation) and the remainder of the data to diagnose the actual performance (validation in the context of hydrologic simulation). We used around 24 years of the available data from Jan 01, 1984 to June 30, 2008 as training data and the following data from July 1, 2008 to Dec 3, 2009 selected as the independent test period (based on our TA suggestion). The training period is so important to get a good actual result because a small division of dataset as training period would lead to a high risk of overfitting. The optimal model parameters were selected by maximizing the objective function (Mean Squared Error; MSE). The optimal value of parameters were determined by a trial and error process and also by examining several hydrological studies such as Shortridge et al. (2016), Deka (2014), Kumar et al. (2016), Naghibi et al. (2017), Worland et al. (2018) and Tongal and Booij (2018). Rectified linear unit (ReLU) and radial basis function selected as transform and kernel functions for MLP and SVM models respectively. Then, the optimal model parameters were kept, and the models were used to simulate runoff for the test period and different model assessment criteria calculated.

3.3 Model Assessment

Figure 2 illustrates the daily predicted runoff by each model as well as the observation values. As the precipitation plays a pivotal role in generation of runoff, its observation values plotted from top to bottom in the second y axis (gray time series).

The visualization of streamflow prediction shows the MLP and SVM models performed significantly better than the others (Table 1). The two red circles shows how the MLP model simulated the high flows which leads to flood events. Also the blue circle shows the models simulation for the low flows, as it is obvious the employed machine learning approaches have some issues with simulation of the low flows. Also, various metrics are performed and presented in table 1 for performance assessment.

Based on the results presented in table 1 it is obvious that there is no significant difference among the three linear regression models' performance but based on the table 1 we can say that lasso model provided the weakest results. It is well-known that machine learning models could fail in simulating runoff from only meteorological variables in the absence of antecedent streamflow values. The main reason for this could be low and lagged relationships between runoff and meteorological variables (Tongal and Booij, 2018). To overcome this inefficiency, this study propose two solution for future

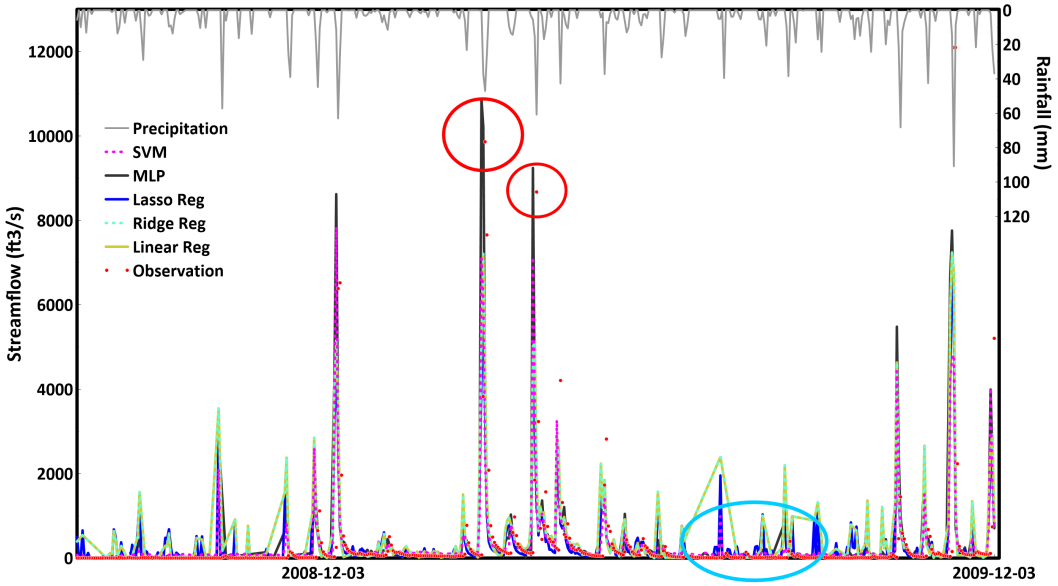

Figure 2: The observed and simulated runoff by various ML methods during test period.

Table 1: The calculated performance criteria for the simulated runoff in the test period

Models	NSE	TRMSE	KGE
Linnear Regression	0.65	6.52	0.68
Lasso Regression	0.62	8.62	0.58
Ridge Regression	0.65	6.52	0.68
MLP	**0.8**	**6.19**	**0.84**
SVM	0.72	7.5	0.67

works. As we know the employed machine learning approaches have a memory-less structure so they are not able to store information of previous time steps for prediction at the current time step (that's why they need the antecedent day information which playing a role as memory), so to overcome this issue use of recurrent neural networks (RNNs) is suggested for the rainfall-runoff modeling problem (as they are memory-based learning methods). Also future works should focus on separating runoff into different components such as base flow and surface flow to improve simulation and forecasting capabilities of machine learning models.

4 Conclusion

This study employed various supervised ML regression-based techniques for simulation rainfall-runoff processes a complex watershed located in SC, USA. Supervised machine learning means we have examples (rows) with input and output variables (columns). The algorithm uses function approximation to map inputs to outputs on specific prediction task in such a way that it has skill. The results showed daily runoff simulation is improved with ML regression supervised learning algorithms. Generally we can say that ML Models are able to predict runoff from meteorological observations with satisfactory accuracy. But we need to keep in mind the Application of Machine Learning methods in runoff modeling can not replace physical modelling approach, but strongly complement and enrich it. Performance assessment of ML methods showed MLP and SVM algorithms have a better performance in runoff prediction comparing with the others. Also, the results showed that there is no significant difference among the linear models' performance (Vanilla, Lasso and Ridge regression methods) but we can say that Lasso model provided the weakest results. As it mentioned in the previous section It is well-known that machine learning models could fail in simulating runoff from only meteorological variables in the absence of antecedent runoff values and the reason for that is low and lagged relationships between streamflow and meteorological variables. Thus, to overcome

this inefficiency, future works should focus on separating streamflow into different components and make use of memory-based learning techniques.

Code and Data Availability

For more information please refere to this github[2] repository which contain the whole package including model programming, datasets and the modeling results.

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. *Hydrology and Earth System Sciences (HESS)*, 21(10):5293–5313, 2017.

[2] B. Asadieh and N. Y. Krakauer. Global change in streamflow extremes under climate change over the 21st century. *Hydrology and Earth System Sciences*, 21(11):5863, 2017.

[3] G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[4] D. T. Bui, B. Pradhan, H. Nampak, Q.-T. Bui, Q.-A. Tran, and Q.-P. Nguyen. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using gis. *Journal of Hydrology*, 540:317–330, 2016.

[5] J. L. Campbell, C. T. Driscoll, A. Pourmokhtarian, and K. Hayhoe. Streamflow responses to past and projected future changes in climate at the hubbard brook experimental forest, new hampshire, united states. *Water Resources Research*, 47(2), 2011.

[6] P. C. Deka et al. Support vector machine applications in the field of hydrology: a review. *Applied soft computing*, 19:372–386, 2014.

[7] K. Fang, C. Shen, D. Kifer, and X. Yang. Prolongation of smap to spatiotemporally seamless coverage of continental us using a deep learning neural network. *Geophysical Research Letters*, 44(21):11–030, 2017.

[8] H. V. Gupta, S. Sorooshian, and P. O. Yapo. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763, 1998.

[9] H. G. Hidalgo, T. Das, M. D. Dettinger, D. R. Cayan, D. W. Pierce, T. P. Barnett, G. Bala, A. Mirin, A. W. Wood, C. Bonfils, et al. Detection and attribution of streamflow timing changes to climate change in the western united states. *Journal of Climate*, 22(13):3838–3855, 2009.

[10] K. Kasiviswanathan, J. He, K. Sudheer, and J.-H. Tay. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. *Journal of Hydrology*, 536:161–173, 2016.

[11] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018.

[12] D. Kumar, A. Pandey, N. Sharma, and W.-A. Flügel. Daily suspended sediment simulation using machine learning approach. *Catena*, 138:77–90, 2016.

[13] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[14] Z. Liu, W. Xu, J. Feng, S. Palaiahnakote, T. Lu, et al. Context-aware attention lstm network for flood prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1301–1306. IEEE, 2018.

[15] S. A. Naghibi, K. Ahmadi, and A. Daneshi. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management*, 31(9):2761–2775, 2017.

[16] J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970.

[2] `https://github.com/sadeghitabas/CPSC8420-Advanced-Machine-Learning`

[17] 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, 2015.

[18] S. Sadeghi-Tabas, S. Samadi, B. Zahabiyoun, et al. Application of bayesian algorithm in continuous streamflow modeling of a mountain watershed. *European Water*, 57:101–108, 2017.

[19] J. Sadler, J. Goodall, M. Morsy, and K. Spencer. Modeling urban coastal flood severity from crowd-sourced flood reports using poisson regression and random forest. *Journal of hydrology*, 559:43–55, 2018.

[20] S. Samadi and M. Meadows. The transferability of terrestrial water balance components under uncertainty and nonstationarity: A case study of the coastal plain watershed in the southeastern usa. *River Research and Applications*, 33(5):796–808, 2017.

[21] S. Samadi, M. Pourreza-Bilondi, C. Wilson, and D. Hitchcock. Bayesian model averaging with fixed and flexible priors: Theory, concepts, and calibration experiments for rainfall-runoff modeling. *Journal of Advances in Modeling Earth Systems*, 12(7):e2019MS001924, 2020.

[22] S. Samadi, D. Tufford, and G. Carbone. Assessing parameter uncertainty of a semi-distributed hydrology model for a shallow aquifer dominated environmental system. *JAWRA Journal of the American Water Resources Association*, 53(6):1368–1389, 2017.

[23] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[24] C. Shen. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, 54(11):8558–8593, 2018.

[25] C. Shen, E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.-J. Chang, S. Ganguly, K.-L. Hsu, D. Kifer, Z. Fang, et al. Hess opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences (Online)*, 22(11), 2018.

[26] J. E. Shortridge, S. D. Guikema, and B. F. Zaitchik. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology & Earth System Sciences*, 20(7), 2016.

[27] C. Sivapragasam and N. Muttil. Discharge rating curve extension–a new approach. *Water Resources Management*, 19(5):505–520, 2005.

[28] D. P. Solomatine and Y. Xue. M5 model trees and neural networks: application to flood forecasting in the upper reach of the huai river in china. *Journal of Hydrologic Engineering*, 9(6):491–501, 2004.

[29] H. Tongal and M. J. Booij. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of hydrology*, 564:266–282, 2018.

[30] S. C. Worland, W. H. Farmer, and J. E. Kiang. Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental Modelling & Software*, 101:169–182, 2018.

[31] S. Yang, D. Yang, J. Chen, J. Santisirisomboon, W. Lu, and B. Zhao. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. *Journal of Hydrology*, 590:125206, 2020.