Joan Sias
Final Project Report for C204
Dec. 8, 2023

Title: Application of machine learning to prediction of the streamflow recession exponent as a means to evaluate two alternative exponent-estimation procedures.

Introduction

An important theme in hydrologic research is regionalization of streamflow model parameters. To regionalize a hydrologic model parameter, one looks for significant regressions between empirical watershed attributes one the one hand, and – on the other - hydrologic model parameters determined for gaged locations through model calibration. If a strong relationship is found, then the possibility exists for successful parameterization of a hydrologic model for an ungaged watershed based on the regression model. Botterill and McMillan 2023 (hereafter, BM23) use a connected neural network model to "learn" streamflow simulation model parameters from a combination of geophysical data and hydro-meteorological time series data. They performed this training for watersheds in the CAMELs dataset (Addor et al. 2017). They produced 16 machine-learned ("ML") parameters for each of 670 watersheds in the continental United States. Using a new subjective procedure (Sias, unpublished) I have - for sixty-one of these watersheds - independently estimated two characteristic recession parameters that I will call beta and alpha. Alpha and beta together parameterize a classic non-linear recession model. When b=1, the model is considered log-linear, which is to say, negative-exponential, and α is then equates with the well-known linear recession coefficient. In principle beta should depend only on geology and, as such, should be temporally invariant. The parameter alpha on the other hand may - depending on how it is estimated - be temporally variant. My procedure attempts to estimate temporally invariant characteristic values of both alpha and beta. The parameters alpha and beta areinteresting to me because of their potential usefulness as a hydrologic simulation model parameters. Using the same geophysical feature variables that BM22 used in their model training, I will train a new ML model to predict beta. I will perform this training/testing twice, using using my beta as the target in one training and the BM22 beta in the second training. I hypothesize that the ML model trained on my beta values will produce better test performance than that of a ML model trained on the BM22 beta values.

Statement of Hypothesis.

H0 (null): The BM23 estimated value for the CAMELS watershed-specific recession exponent beta (hereafter denoted b_bm)will be better predicted by a machine learning model than will be the equivalent parameter that I have estimated using an unpublished procedure. My estimate of beta is hereafter denoted b_js.

H1 (alternative): A machine learning model will produce lower prediction error when the model is fit to b_js than when fitted to b_bm.

For both models, the independent data is identical.

Data

The CAMELS data set (Addor et al, 2017) is the source of feature data for 671 watersheds distributed across the continental United States. URLs for specific CAMELS files are as follows.

1. https://gdex.ucar.edu/dataset/camels/file/camels_clim.txt

2. https://gdex.ucar.edu/dataset/camels/file/camels_topo.txt

3. https://gdex.ucar.edu/dataset/camels/file/camels_geol.txt

4. https://gdex.ucar.edu/dataset/camels/file/camels_hydro.txt

5. https://github.com/mcmillanhk/HydroML/blob/master/data/extra_sigs/gw_array.csv

Botterill and McMillan2023 is the source of most of the target variables in this study. These target variables and four additional variables are listed in Table 1. Though this report is focused on estimate of the parameter beta, I include additional targets in Table 1 for reasons that these are present in the jupyter notebook created for this project. The units for my a4_js parameter are different from the units for a_bm. Nevertheless, both are expressed in terms of specific discharge, so the correlation should not depend on the difference in units. The parameter b_* is dimensionless and is insensitive to time increment and units of discharge, so the parameters b_js and b_bm are directly comparable. The b_value is related to concavity of the recession model. When beta and α are fitted to a watershed, the parameters are interdependent. This confers a rather large uncertainty on each parameter, since – over a limited range - a change in one can be compensated for by a change in the other. It is desirable to have a single parameter that "summarizes" the two parameters. Toward this end, I computed for each watershed the recession slope of the non-linear fitted model computed at the mean discharge for all the watersheds in the sample. I call this parameter Kest. More detail on how this parameter was estimated is given below.

Table 1. Target Parameters.

Source[1]	Name	Code	Dimension
BM21	RecessionParametersAlpha	a_bm	$(m^{1-b}d^{b-2})$
BM21	RecessionParametersBeta	b_bm	dimensionless
BM21	BaseflowRecessionK	K_bm	(1/d)
BM21	FirstRecessionSlope	fK_bm	(1/d)
BM21	MidRecessionSlope	mK_bm	(1/d)
JS	RecessionParametersAlpha	a4_js	$(mm^{1-b}d^{b-2})$
BM21	RecessionParametersBeta	b_js	dimensionless
JS	Estimated K	Kest	1/d
JS	Estimated K	Kest_js	1/d

[1] Botterill and McMillan (2023)

A signifant part of the effort for this project was to increase my sample size of CAMELS gages for which I have calibrated the parameters a4_js and b_js. Because the climate in the PNW is favorable for determination of characteristic recession parameters (as I have determined through my research), and because I most of the CAMELS gages that I had already calibrated prior to the start of this project are located in the PNW, I decided to focus my analysis on CAMELS gages located in OR, WA, MT, and ID and having at least 250 mm of precipitation in the year 1950. Calendar year 1950 precipitation for each watershed was obtained from data set known as GAGESII. The gages in CAMELS are in fact a subset of GAGESII. I have identified 83 CAMELS gages that meet my selection criteria. Of these I had 28 calibrated at the start of this project, and I have calibrated an additional 32 gages to bring the sample size up to 60. Potentially there are 23 more CAMELS gages that meet my selectin criteria and that I could add to increase my sample size.

Theory indicates that the exponent b should be related to geology. It should not depend on climate. My research (not published) informs me that a watershed will exhibit a characteristic b-value only if the climate is conducive to the expression of this characteristic (geologically-determined) value. The climate in Oregon and Washington is optimal for expression of characteristic b-value.

Data sets created.

I created four data sets by aggregated CAMELS data, BM21 parameters (b_bm; a_bm), a parameter Kest derived from such data, and Sias parameters (b_js; a4_js, Kest_js).

1. dat652 652 records from the CAMELS database, including all variables from camels_topo.txt, camels_geo.txt, camels_clim.txt. This also includes the parameters b_bm and a_bm. These two latter parameters were obtained from BM21. I added to this data set a parameter Kest. This parameter is described elsewhere in this report. This data set does not have any of my calibrated parameters. This has 19 climate, topological, and geological variables, as well as 13 hydrologic variables. These 13 hydrologic variables are sourced from both CAMELS and BM23.
2. Dat652_q. This is identical to dat652, except that it includes all hydrologic variables.
3. Dat60. This is identical to dat652, except that it includes only those CAMELS records for watersheds that I have calibrated (N=60). This dataset thus includes the parameters b_js, a4_js, and Kest_js.
4. Dat60_sg. This is identical to dat60, except that it lacks geological variables.

I added synthetic parameters Kest and Kest_js to my data sets. This parameter Kest uses the mean value of the q_mean parameter and the (alpha, beta) recession parameters to obtain an "estimated" linear recession coefficient for each watershed The rationale for adding this is to have a single parameter that corresponds meaningful to the dual-recession parameters (alpha, beta).

$$Kest\_js[k] = a\_bm[k]\ \ Qmean^{b\_js[k]} \qquad \text{for} \quad 1 \le k \le 60$$

$$Kest[k] = a\_bm[k]\ Qmean^{b\_bm[k]} \qquad \text{for} \quad 1 \le k \le 652$$

where Qmean is the average value of q_mean across all watersheds in the sample, and k is a watershed index.

Data clean-up.

The data element "geol_2nd_class" in the CAMELS geology file had a missing value for 120 records. Therefore, this data element was eliminated from consideration. Nineteen records were eliminated due to missing values for desirable feature or target variables, leaving a final sample size of 652 in the two larger data sets.

The CAMELS dataset provides six parameters related to geology. I eliminated two of the geology parameters due to a large number of records having missing values for these parameters (i.e., ("geol_2nd_class" and "glim_2nd_class_frac").. One of the remaining four parameters is descriptive ("geol_1st_class"), having values such as "Siliciclastic sedimentary rocks","Metamorphics", and "Basic volcanic rocks." Therefore I used one-hot encoding to represent this variable.

Transformations. On the basis of histograms of each target and feature variables, I chose to take the logarithm of the watershed area. The watershed area was approximately gaussian after the transformation. For the training data in the small sample (N=61), I was able to justify only this transformation. In the training data for the large sample (N=652) I identified several variables that were appropriate for log transformation, and three that were appropriate for square-root transformation. Other variables would probably have benefitted from bucketizing, but I did not pursue this option.

My primary target variable is the b exponent. I also constructed models for predicting a_bm, a4_js, mK_bm, fK_bm, and Kest. I am not reporting here on results for these secondary targets. I include information about these secondary targets because they appear in the project code.

Modelling

I constructed a pipeline to implement Ridge Regression with variable transformations) where appropriate(i.e., log, square root), and with all variables standardized prior to model fitting. Because I had one variable that was desriptive-categorical, I used one-hot encoding to represent this variable. Because my dataset included numerical and non-numerical data, I used the ColumnTransformer constructor to produce a preprocessor model. I formed a separate preprocessor model for the target variables and for the feature variables. The preprocessors were also specific to the data set. Therefore, I elaborated eight distinct preprocessors. Once the preprocessors were completed, I was able utilize the preprocessors to efficiently contruct these ski-kit learn models: LinearRegression model, a Ridge regression model (with α a hyperparameter), and SopportVectorMachine

I constructed a pipeline to implement Ridge Regression. I inspected historgrams of the training set feature variables in order to assess whether any of the variables were good candidates for log or square-root transformation. For the two data sets with 652 records, I identified several variables for log transformation, and several for square-root transformation. I did not explore whether "bucketizing" and of the variables would improve the distribution (making the data look more Gaussian). In the case of the dat60 training data set, I identified watershed drainage area as the only variable suitable for transformation. I applied a log transformation to this variable. The preprocessing code performs transformations specific to the training and test set for each dat set (thus, I have six different preprocessors).. The preprocessors all standardize all target and feature variables. Because I had one variable that was descriptive-categorical, I used one-hot encoding to represent this variable. 6Because my dataset included numerical and non-numerical data, I used the ColumnTransformer constructor to produce a preprocessor model. I formed a separate preprocessor model for the target variable, and for the feature variables. The preprocessors were also specific to the data set., I was able utilize the preprocessors to efficiently construct these ski-kit learn models: LinearRegression, Ridge regression (with α a hyperparameter), and SupportVectorMachine

Results

For each model, I produced scattergrams of predicted and observed values of the target variable. The code produces Pearson's r, squared, root mean squared error (RMSE), and an REC graph for each target variable. Varying the ridge regression hyperparameter α over the α=0.1 to α=1 produced marginal difference in the results. Therefore I present only results for α=1.0. Figure 1 shows the REC for three pairs of target variables. These results a produced with the dat60 sample.s

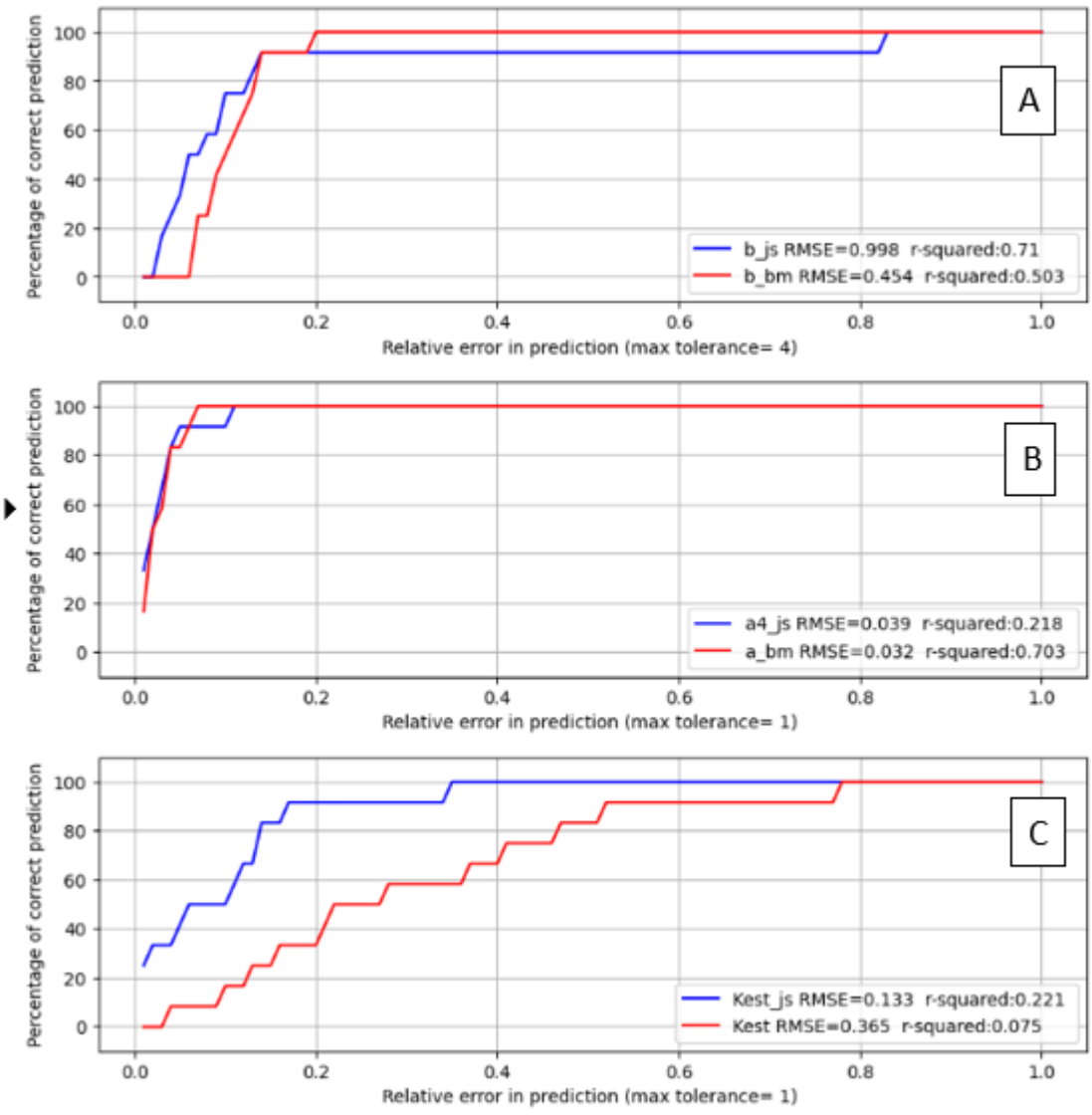


Figure 1. Generalization error measure with a trained linear regression model having sixteen (non-hydrologic) feature variables: Test-sample REC curves for two target variables. These results pertain to the dat60 data set. This is the only data set in which the BM23 and my parameters can be compared through application of the machine learning algorithms. A. b_js versus bm. B. a4_js versus a_bm. C. Kest_js versus Kest.

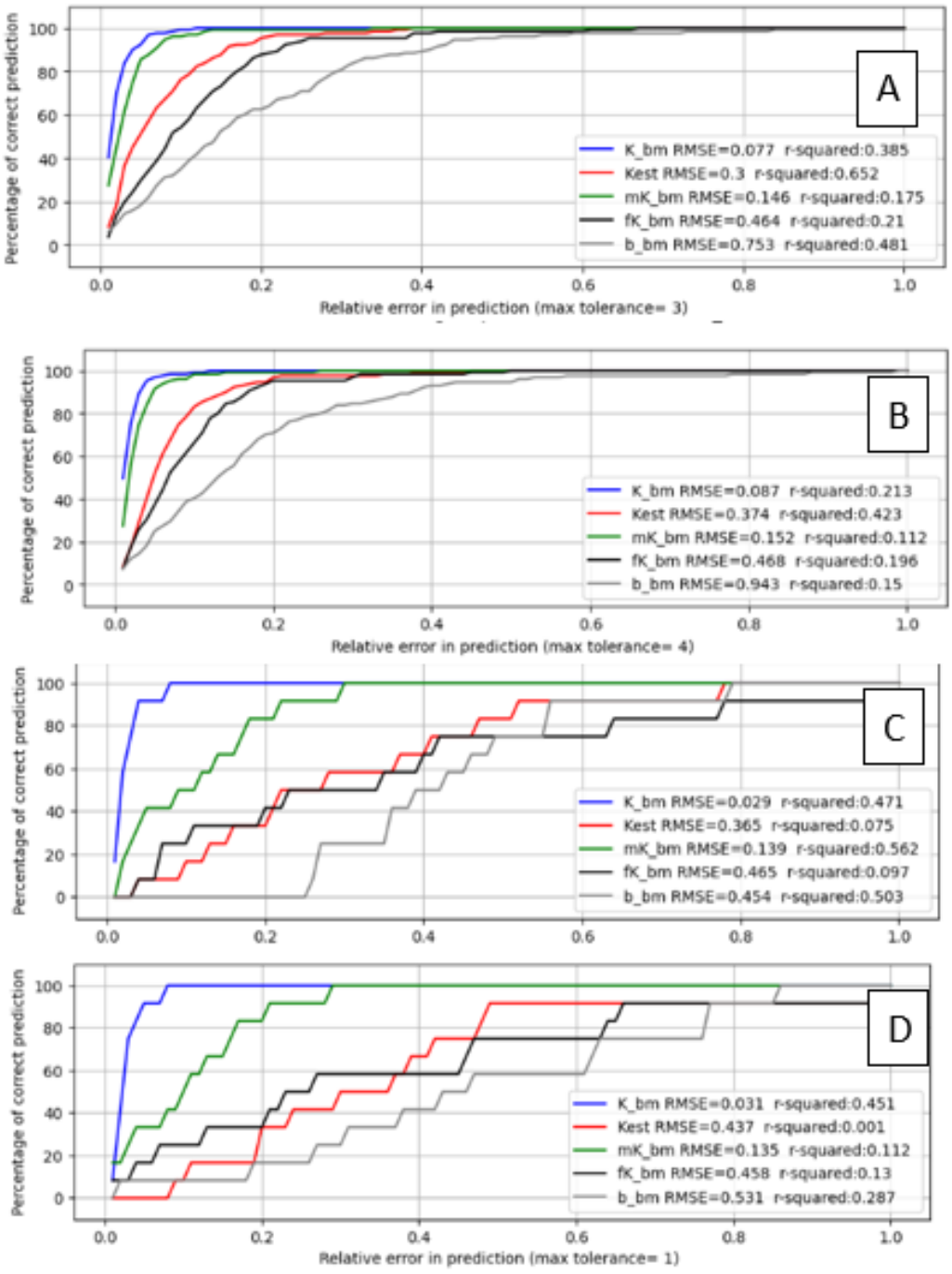


Figure 2. Generalization error measure with a trained linear regression model having twenty (i.e., sixteen non-hydrologic plus four hydrologic) feature variables and 652 watersheds before splitting. Test-sample REC curves for five different target variables. The four panels differ in the data set use. A. dat652_q. B. dat652. C. dat60. D. dat60_sg.

Discussion

From Figure 1 we can conclude that my recession parameters are similarly well if not slightly better estimated by machine learning as are the corresponding BM23 parameters. This result is encouraging.

Figure 2 shows a decline of model results as the feature variables become increasingly reduce. Removing the hydrologic variables had little effect (panel A versus B). The deterioration from panel B to C is due to an more than 90 percent reduction in sample size. The smaller deterioration in signal from panel C to D is due to elimination of geologic variables within the feature variables. In general the linear recession variables are much better estimated than the non-linear recession exponent (b_bm).

Conclusions

I have described a machine learning model applied in order to compare different estimates of streamflow recession parameters. Specifically, I focused on the exponent of the recession slope power law first described in Brutsaert and Neiber, 1977. The b_js parameter tends to have larger RMSE than the b_bm parameter, suggesting that the H1 hypothesis should be rejected. However, the REC shows that in most models a higher proportion of test sample data points have lower error when the model is predicting b_js.

In future developments of this model, I will explore whether the model prediction error is reduced by the following measures.

1. Increase the sample size.
2. Expand samples to include greater diversity of climate zones.
3. Use binning to transform feature variables that are non-Gaussian but not amenable to the log or square-root transformation.

My github site.

https://github.com/jcsias/jcsias.github.io/blob/main/project.md

Orientation to jupyterlab notebook for this project.

Begin by running all cells.

Scroll down to find the cell that allows you to choose the data set and set the random_state hyperparameter. Rerun the entire notebook if you change the data set or the hyperparameter.

To fit the model and predict the target variable(s), go to the last cell in the notebook. At the top of this cell you have the option to change the model before executing the cell.

The available target variables differ between the data sets (the dat652 and dat652_q do not have the *_js target variables). This in turn alters the specific figures that will be produced when the cell is executed.

https://github.com/jcsias/jcsias.github.io

Below is a link to my jupyter notebook in Google Colab.

The name of my notebook is C204_JSias.ipynb

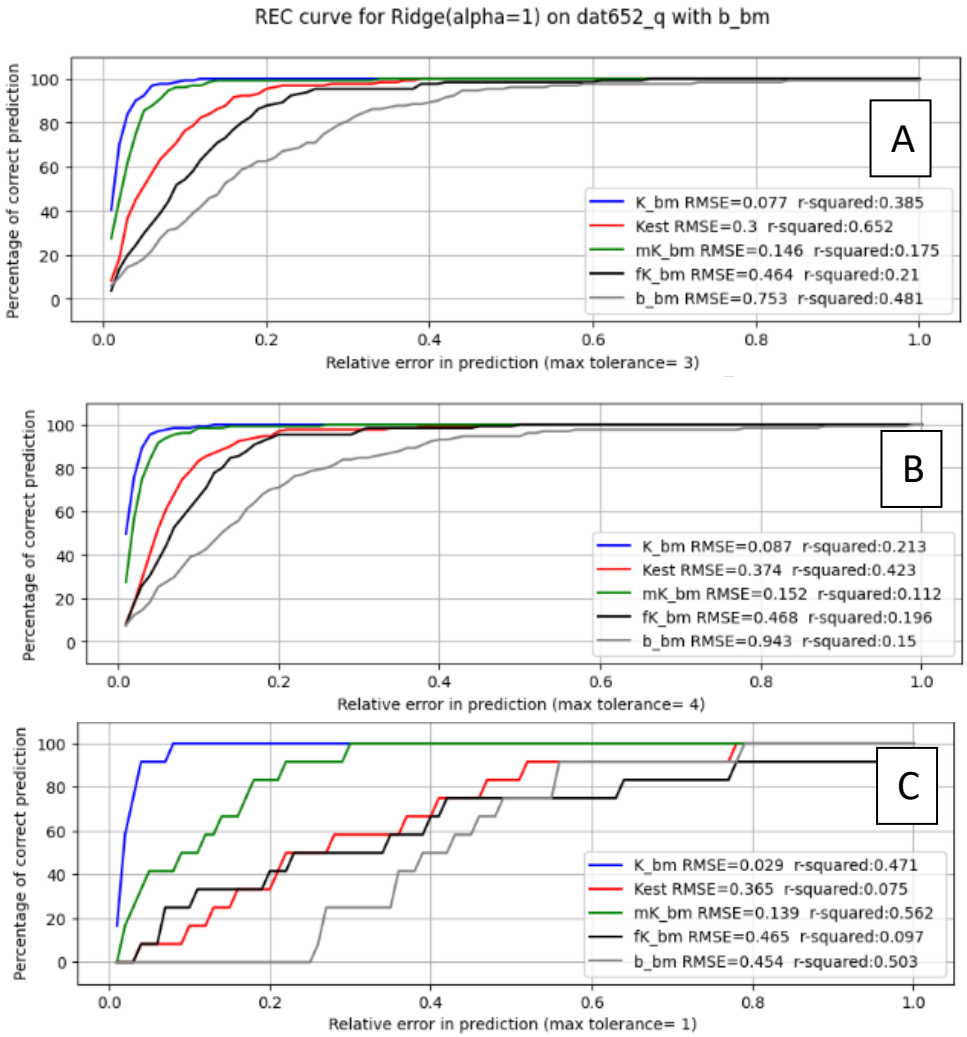
I will probably be sending a fresher link between now and the 11:59 pm deadline.

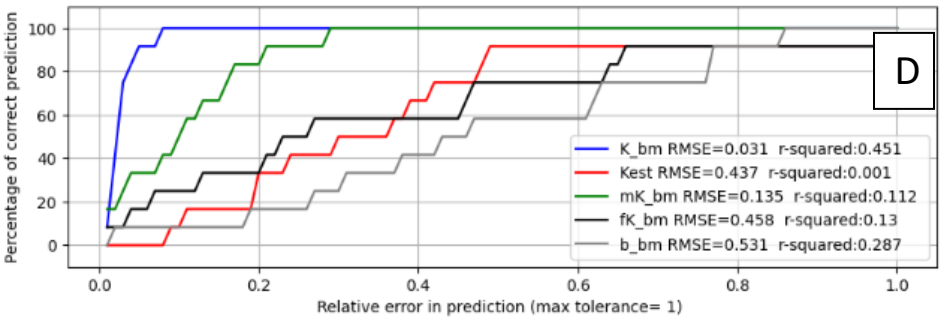
https://colab.research.google.com/drive/1ReJTHNBA_bbiKBmR3_D0ZaWH7MouxVVT?usp=sharing

References

Addor, Nans, et al. "The CAMELS data set: catchment attributes and meteorology for large-sample studies." Hydrology and Earth System Sciences 21.10 (2017): 5293-5313.

Botterill, T. E., & McMillan, H. K. (2023). Using machine learning to identify hydrologic signatures with an encoder–decoder framework. Water Resources Research, 59, e2022WR033091. https://doi. org/10.1029/2022WR033091




Percentage of correct prediction
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
Relative error in prediction (max tolerance= 1)
D
K_bm RMSE=0.031 r-squared:0.451
Kest RMSE=0.437 r-squared:0.001
mK_bm RMSE=0.135 r-squared:0.112
fK_bm RMSE=0.458 r-squared:0.13
b_bm RMSE=0.531 r-squared:0.287