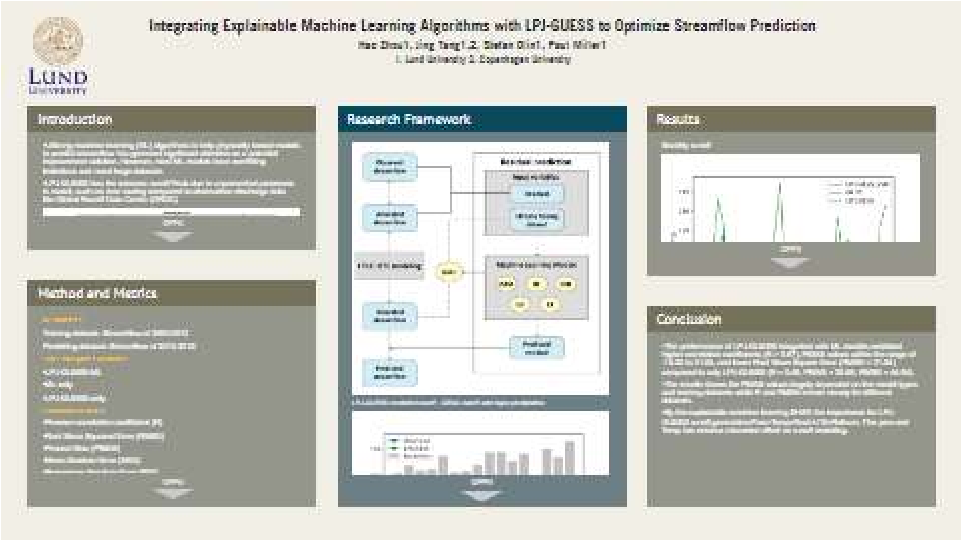
Integrating Explainable Machine Learning Algorithms with LPJ-GUESS to Optimize Streamflow Prediction



Hao Zhou1, Jing Tang1,2, Stefan Olin1, Paul Miller1

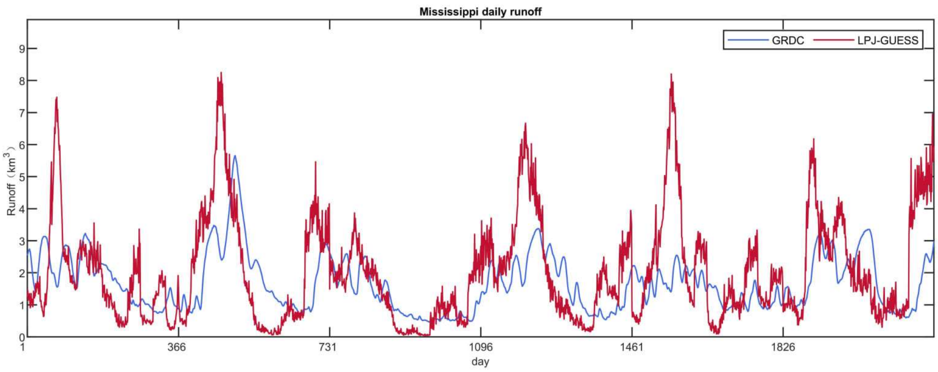
1. Lund University 2. Copenhagen University

PRESENTED AT:

INTRODUCTION

•Utilizing machine learning (ML) algorithms to help physically based models to predict streamflow has garnered significant attention as a potential improvement solution. However, most ML models have overfitting limitations and need large datasets

•LPJ-GUESS has the advance runoff Peak due to unpresented processes in model, such as river routing compared to observation discharge data like Global Runoff Data Centre (GRDC)



METHOD AND METRICS

ML datasets:

Training dataset: Streamflow of 2000-2012

Predicting dataset: Streamflow of 2013-2015

Three categories of Model:

•LPJ-GUESS-ML

•ML-only

•LPJ-GUESS-only

Evaluation metrics：

•Pearson correlation coefficient (R)

•Root Mean Squared Error (RMSE)

•Percent Bias (PBIAS)

•Mean Absolute Error (MAE)

•Percentage Absolute Error (PAE)

Variables:

Precipitation(Prec)

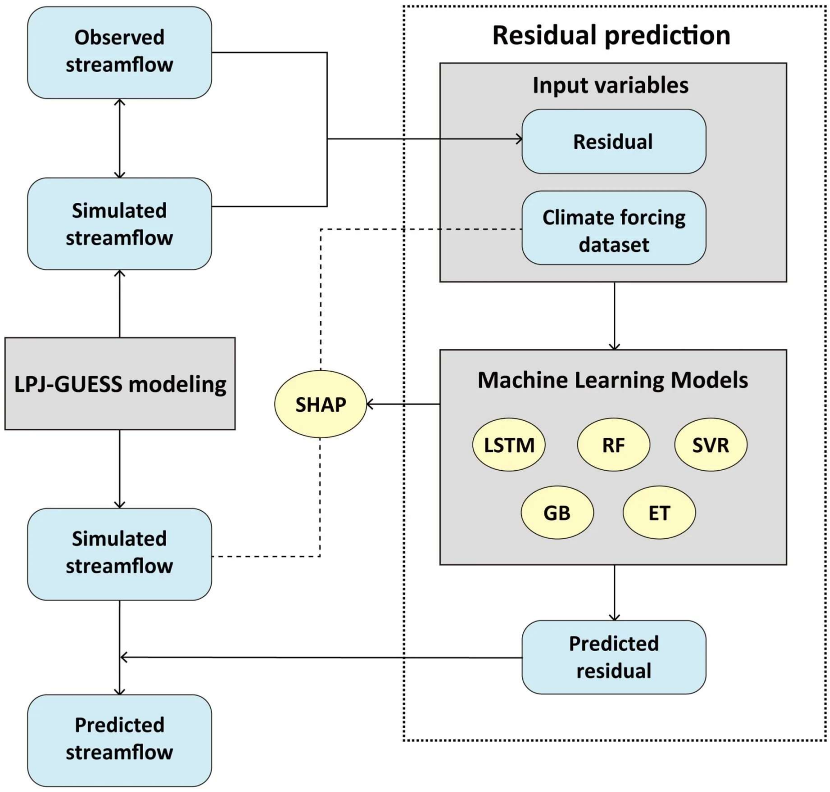
Temperature(Temp)

Radiation(Rad)

Wind speed(U10)

Relative humidity(Relhum)

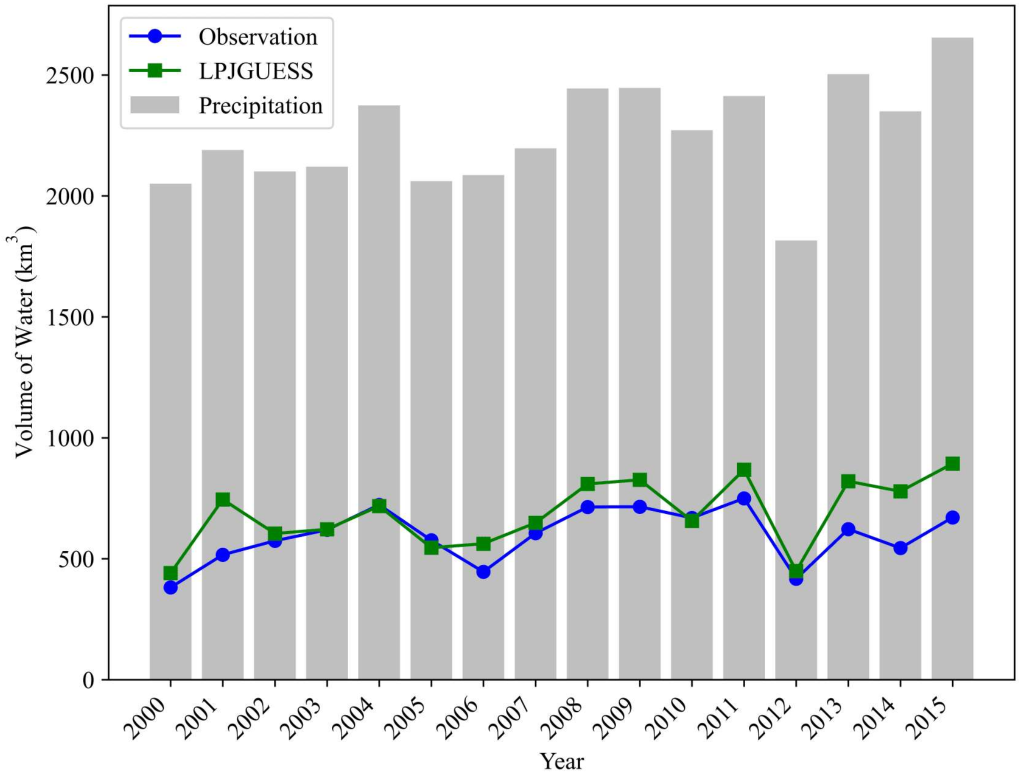
RESEARCH FRAMEWORK



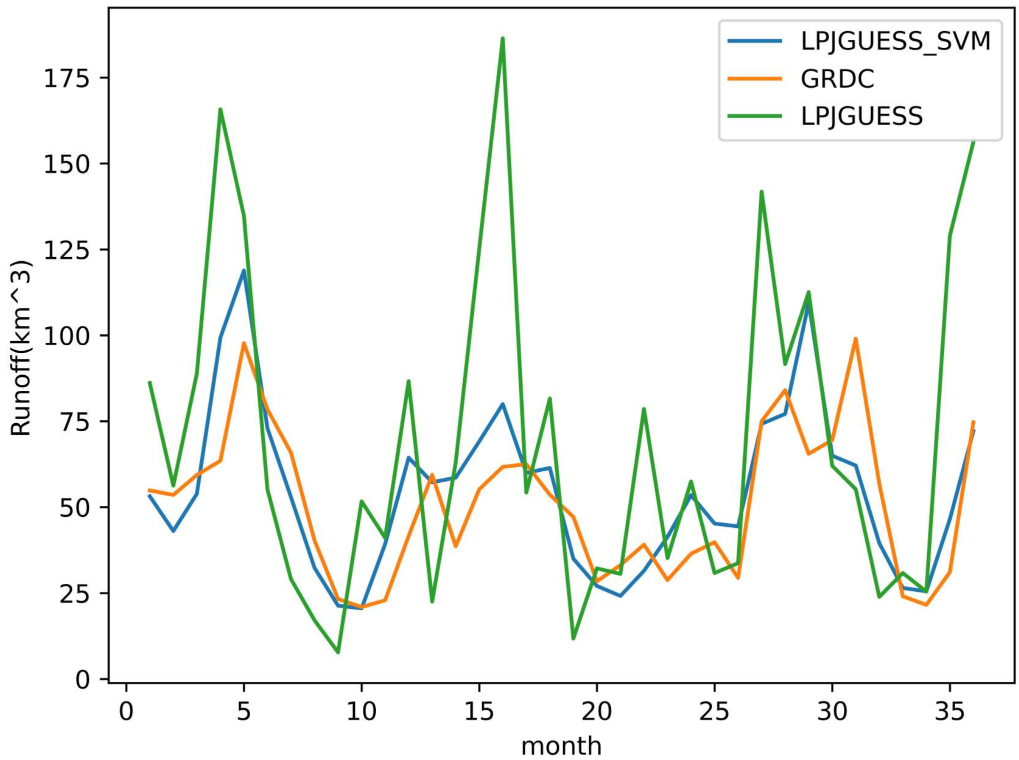
In this approach, ML models were employed to predict the residual errors of LPJ-GUESS using physically related model climate forcing variables such as temperature, precipitation, radiation, etc. We selected the Mississippi catchment in the United States as the study area for streamflow prediction, employing three different configurations: LPJ-GUESS integrated with ML models, only LPJ-GUESS, and only ML models. Furthermore, the explainable machine learning method SHAP was used to identify the driving factors and their interacted effect of LPJ-GUESS modelled runoff.

RESULTS

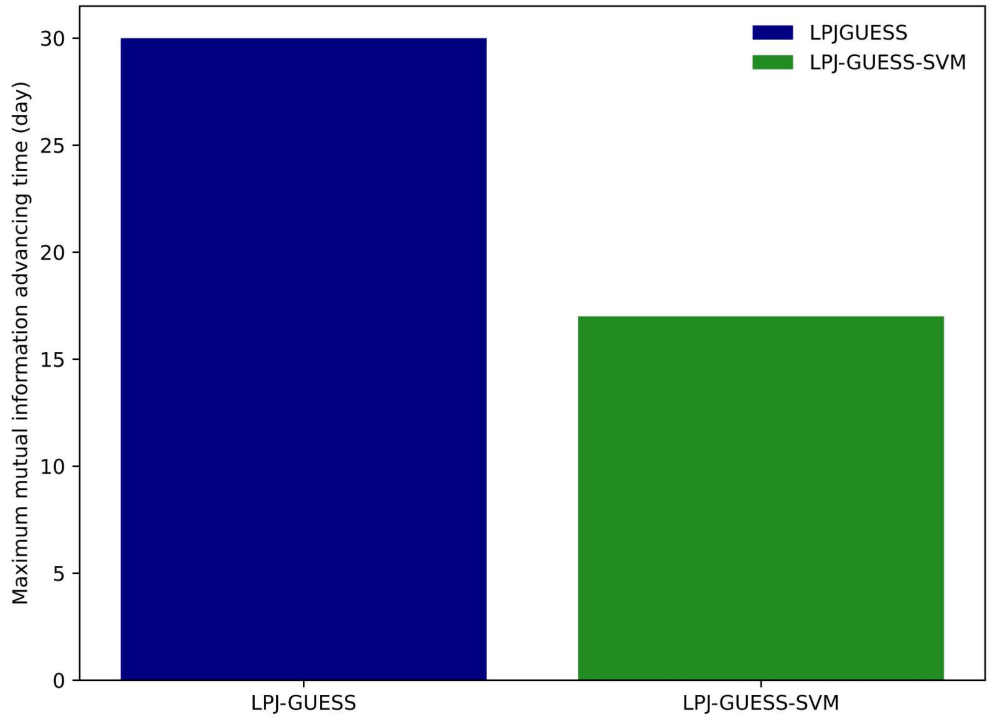
LPJ-GUESS modelled runoff , GRDC runoff and input precipiation



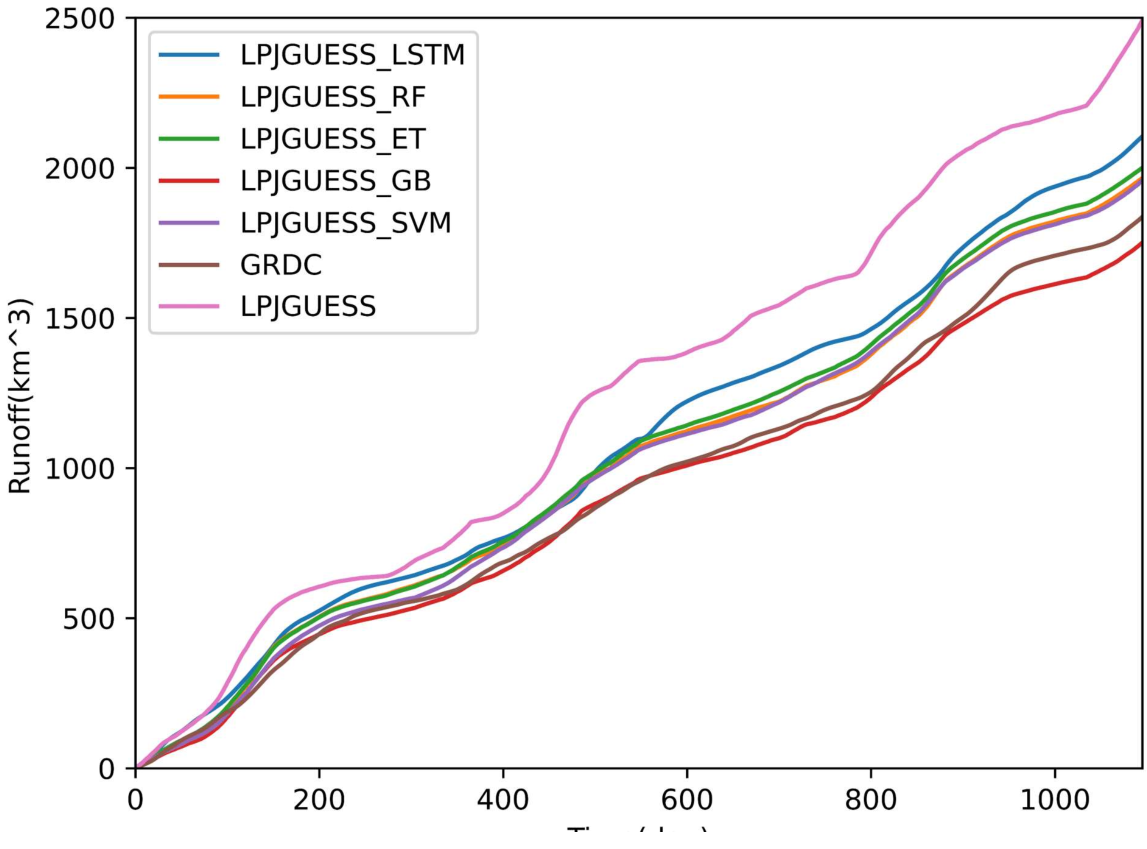
Monthly runoff



Improvement of advancing effect



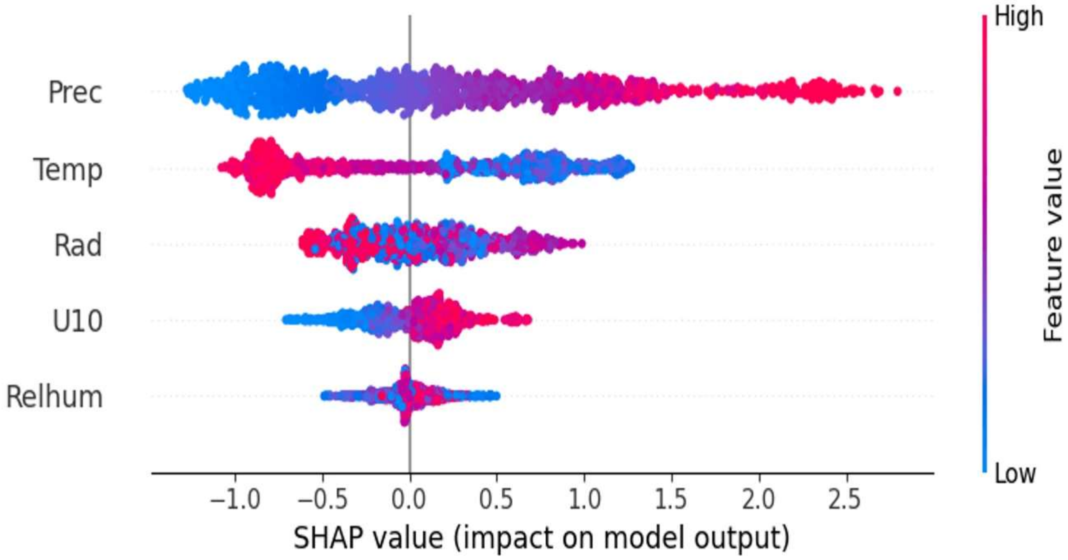
Cumulative daily runoff



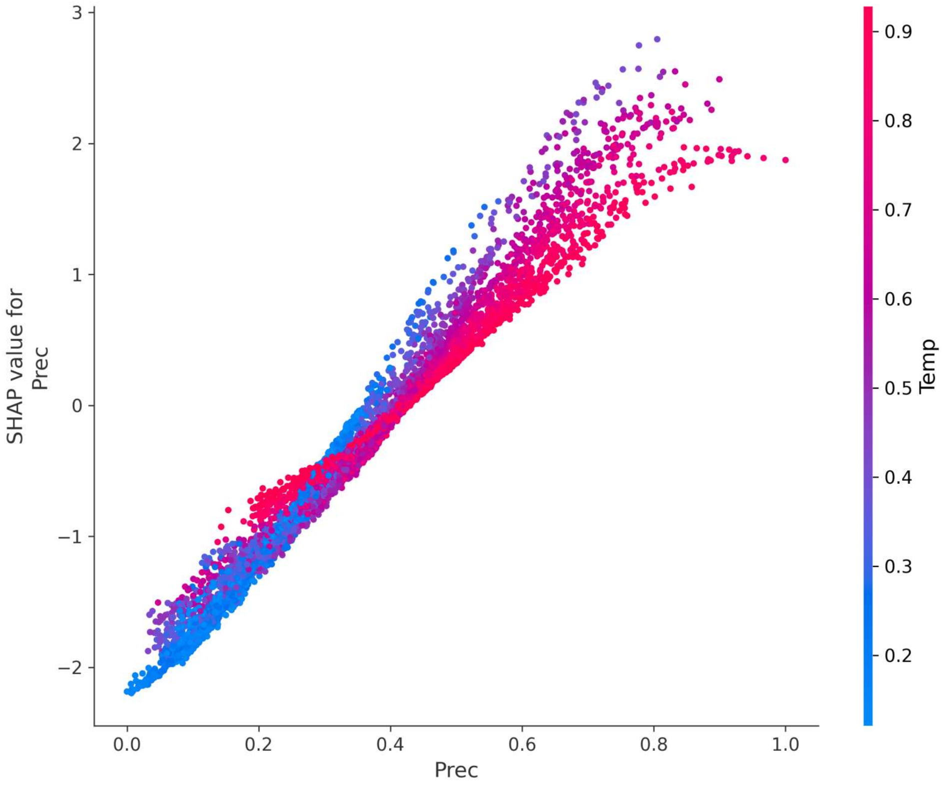
Metrics for sensitivity test

Training period	Predicting period	R	PBIAS	RMSE	MAE	PAE
2000-2012	2013-2015	0.77	6.62	15.68	11.72	0.75
2000-2009& 2013-2015	2010-2012	0.78	-15.10	19.97	13.92	0.72
2000-2005& 2010-2015	2007-2009	0.74	-14.97	18.12	13.15	0.77
2000-2002& 2006-2015	2003-2005	0.74	-14.63	17.38	12.54	0.78
2003-2015	2000-2002	0.71	19.75	16.11	13.48	0.59

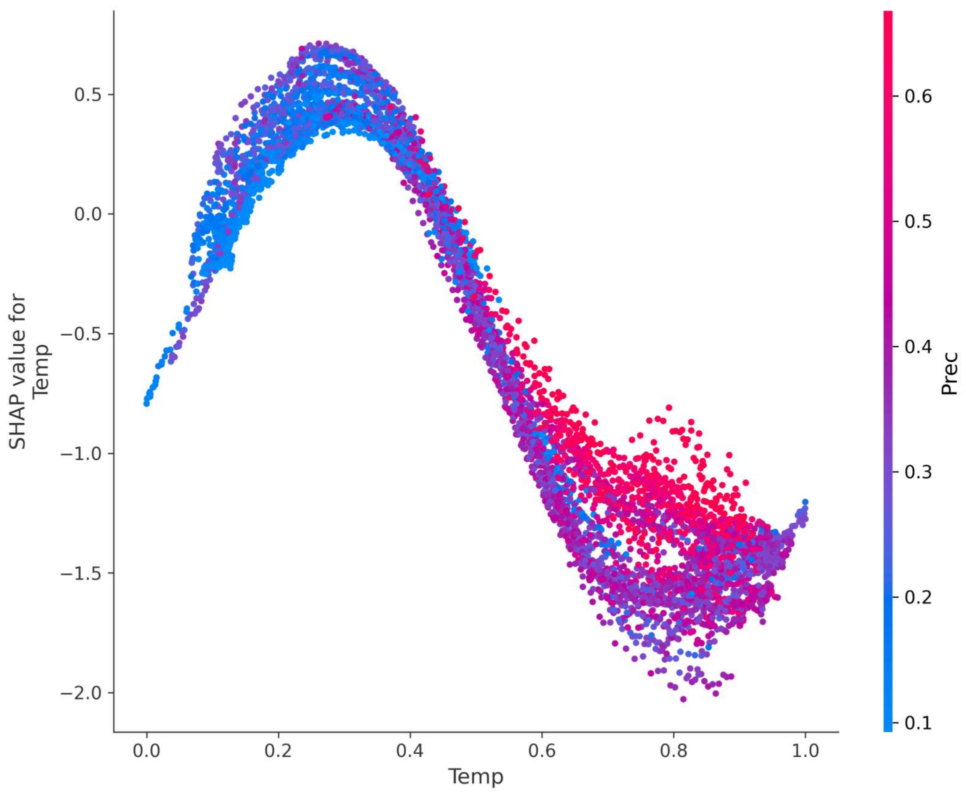
SHAP value



Interacted effect among variables



SHAP value for
Prec
Prec
Temp
3
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2



SHAP value for
Temp
Temp
Prec
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.5
0.4
0.3
0.2
0.1

CONCLUSION

•The performance of LPJ-GUESS integrated with ML models exhibited higher correlation coefficients ($R > 0.67$), PBIAS values within the range of -16.23 to 21.54, and lower Root Mean Square Error ($RMSE < 21.54$) compared to only LPJ-GUESS ($R = 0.48$, $PBIAS = 35.66$, $RMSE = 44.84$).

•The results shows the PBIAS values largely depended on the model types and training datasets while R and RMSE remain steady for different datasets.

•By the explainable machine learning SHAP, the importance for LPJ-GUESS runoff generation:Prec>Temp>Rad>U10>Relhum. The Prec and Temp has obvious interacted effect on modelled runoff.

AUTHOR INFORMATION

Hao Zhou, PhD-student

Email: hao.zhou@nateko.lu.se

Department of Physical Geography and Ecosystem Science

Lund University

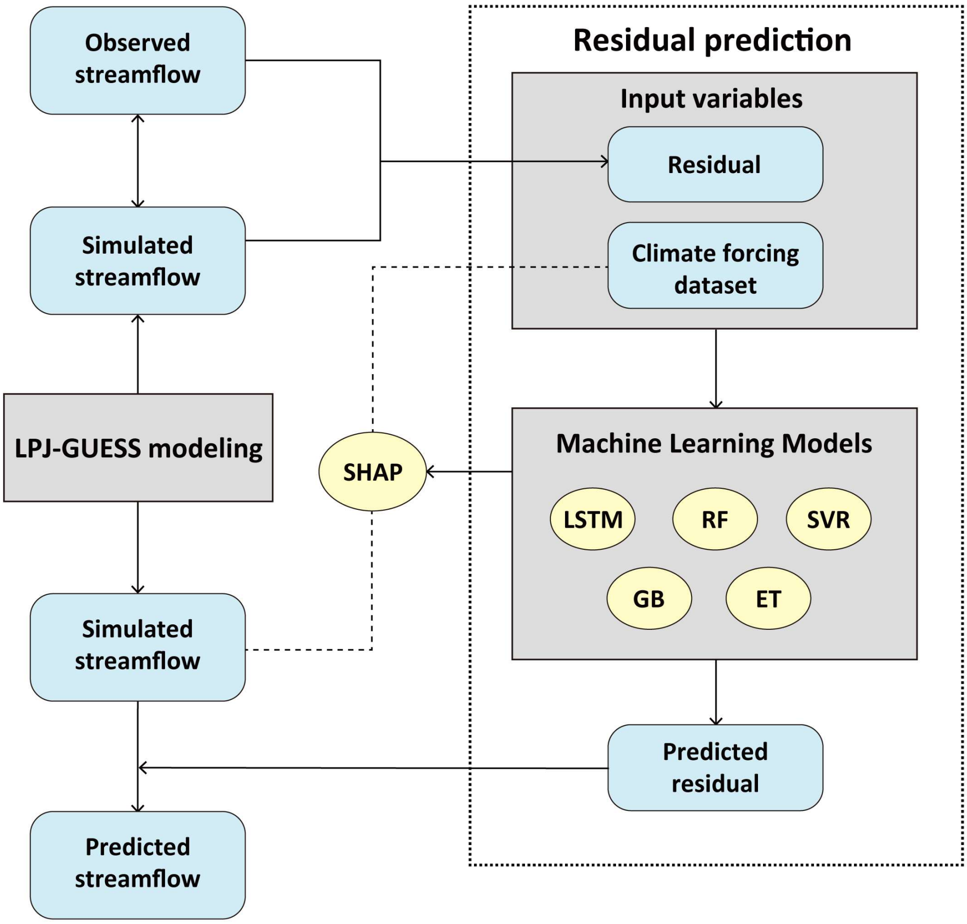
Sweden

TRANSCRIPT

ABSTRACT

Utilizing machine learning (ML) algorithms to help physically based models to predict streamflow has garnered significant attention as a potential improvement solution. However, most ML models have overfitting limitations and need large datasets. To address these shortcomings, we proposed an approach that integrates the dynamic global vegetation model LPJ-GUESS (Lund-Potsdam-Jena General Ecosystem Simulator) with multiple ML models to improve streamflow simulations. In this approach, ML models were employed to predict the residual errors of LPJ-GUESS using physically related model climate forcing variables such as temperature, precipitation, radiation, etc. We selected the Mississippi catchment in the United States as the study area for streamflow prediction, employing three different configurations: LPJ-GUESS integrated with ML models, only LPJ-GUESS, and only ML models. Evaluating with in-situ streamflow data of Vicksburg station from Global Runoff Data Centre (GRDC), the results demonstrate the superior performance of LPJ-GUESS integrated with ML models, exhibiting higher correlation coefficients ($R > 0.67$), PBIAS values within the range of -16.23 to 21.54, and lower Root Mean Square Error ($RMSE < 21.54$) compared to only LPJ-GUESS ($R = 0.48$, $PBIAS = 35.66$, $RMSE = 44.84$). In terms of the R and RMSE, LPJ-GUESS integrated with the ML models have an overall better value than the corresponding only ML models. Furthermore, a training data sensitivity experiment showes that the PBIAS values largely depends on the model types and training datasets while R and RMSE remain steady for different datasets. The analysis of LPJ-GUESS runoff generation driving factors by explainable machine learning SHAP shows the dominant importance of precipitation and temperature. Notably, LPJ-GUESS integrated with ML models successfully captures residual errors and effectively reduces inherent uncertainties, thus surpassing the performance of solely ML-based methods. Our study highlights the promising potential of integrating ML algorithms with LPJ-GUESS for streamflow prediction. This approach not only overcomes existing limitations but also offers a more robust representation of physical constraints, thereby fostering improved

accuracy and reliability in streamflow simulations.



(https://agu.confex.com/data/abstract/agu/fm23/2/2/Paper_1349022_abstract_1178201_0.jpg)