Supplementary of

Diagnosing Dissolved Organic Carbon Simulation of SWAT-C model by Coupling Machine Learning approaches

Zehong Huang et.al

Correspondence to: Yongshuo H. Fu (Email: yfu@bnu.edu.cn);

Introduction

This supplementary file provides additional technical details to support the main manuscript. It includes descriptions of the spatial characteristics, parameterization, and calibration scheme of the Yalong River Basin, as well as a comparison of the machine learning methods applied in the study. Further information is also provided on the ML model framework, algorithm principles, and hyperparameter optimization. In addition, a SHAP-based analysis is presented to illustrate the individual effects of the vegetation module and relative humidity on model outputs. These materials aim to enhance the clarity, transparency, and reproducibility of our modeling approach.

Text S1. Spatial Characteristics, Parameterization, and Calibration Scheme of the Yalong River Basin

In the SWAT-C modeling process, a 10% area threshold was applied to filter land use, soil, and slope combinations within each sub-basin. Consequently, the Yalong River Basin was delineated into 21 sub-basins comprising a total of 323 hydrological response units. As illustrated in Figure S1, the Yalong River flows from north to south, with an average basin slope of approximately 20°, and local maximum slopes approaching 90°.The basin exhibits considerable heterogeneity in soil distribution, predominantly featuring lithic pergelic soils (LPi), highly active leached soils (LVh), and cryic cambisols (CMi). In terms of land cover, grasslands and forests are the primary land use types, occupying approximately 58% and 36% of the total basin area, respectively.

The 2022 release of the SWAT-C model was employed in this study (Yang and Zhang, 2016; Zhang et al., 2013). Parameter sensitivity analysis and model calibration were conducted using SWAT-CUP. The final parameter values and sensitivity analysis results for the Yalong River basin are summarized in Table S3.

Text S2. Comparison of Machine Learning Methods

This study compared machine learning methods commonly applied in hydrological and carbon-related research, as summarized in the Table S1. Based on their performance in previous studies, eight well-performing machine learning algorithms were selected: Convolutional Neural

Network (CNN), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBOOST), Multi-Layer Perceptron (MLP), and four LSTM-based variants—Vanilla LSTM (V-LSTM), Stacked LSTM (St-LSTM), Bidirectional LSTM (Bi-LSTM), and Convolutional LSTM (Cv-LSTM).

Text S3. ML Model Framework, Algorithm Principles, and Hyperparameter Optimization

All machine learning (ML) models in this study were developed on the Python 3.9.20 platform. All LSTM-based models were implemented using PyTorch 2.10.1, while SVR, MLP, CNN, and XGBOOST models were constructed using Scikit-learn 1.5.2. Hyperparameter optimization for all ML models was conducted using Bayesian optimization via the Optuna framework, with a pre-training iteration count of 6000 and 200 Bayesian trials. Additionally, 20% of the training data was randomly selected for validation.

Through comparative analysis of SWAT-C-based coupled models, the optimal DOC simulation strategy was identified. The core principles of each algorithm are summarized as follows:

CNN: Inspired by biological vision mechanisms (Liu et al., 2021), CNN uses three key feature extraction mechanisms—local receptive fields, weight sharing, and subsampling-to efficiently learn spatial patterns. It is particularly suitable for recognizing spatial structures in hydrological and geographical data (LeCun et al., 1989).

SVR: As a representative of statistical learning theory (Cortes and Vapnik, 1995), SVR adopts the structural risk minimization principle to control model complexity. By applying kernel functions, it maps nonlinear problems from low-dimensional to high-dimensional spaces, exhibiting strong generalization in hydrological forecasting (Belayneh et al., 2014; Piri et al., 2023).

XGBOOST: This algorithm optimizes the loss function using second-order Taylor expansion and incorporates regularization terms to control model complexity. Its parallel computing capabilities significantly improve the efficiency of gradient boosting algorithms, making it advantageous for modeling non-stationary hydrological data.

MLP: As a typical architecture of artificial neural networks, MLP establishes input-output mappings through nonlinear transformations in hidden layers. Its fully connected structure has been widely applied in streamflow simulation (Singh et al., 2012; Wang and Lu, 2006).

V-LSTM: The vanilla LSTM uses gated mechanisms (input, forget, and output gates) to regulate information flow and address the vanishing gradient problem in traditional RNNs (Deulkar et al., 2025).

St-LSTM: The stacked LSTM enhances model expressiveness through layer-wise feature abstraction, making it suitable for complex hydrological time series modeling (Mirzaei et al., 2021).

Bi-LSTM: This model employs bidirectional information flow to capture both forward and backward dependencies in time series data, improving DOC dynamic simulation accuracy (Zhang et al., 2025).

Cv-LSTM: By integrating CNN's spatial feature extraction with LSTM's temporal modeling, Cv-LSTM enables joint learning of spatiotemporal features in hydrological processes(Yuan et al., 2022).

The hyperparameter optimization results for each algorithm are summarized as Table S2.

Text S4. SHAP Based Single-Feature Analysis of the Vegetation Module and Relative Humidity

SHAP single-feature analyses were conducted for LAI and RH, as illustrated in Figure S2. The results reveal that LAI exerts a negative effect on the coupled model, with relatively limited temporal variability. In contrast, RH also exhibits a negative influence on the model, but shows considerable variability over time.

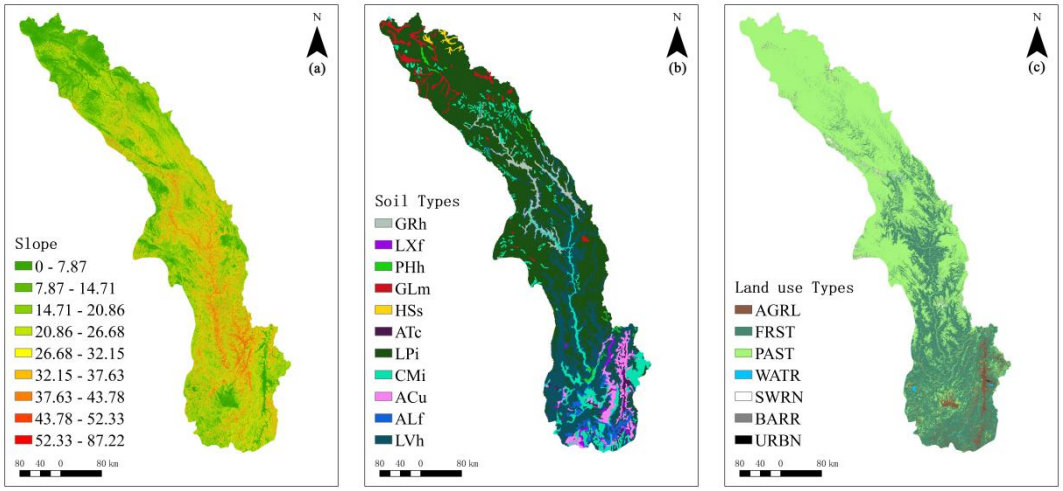

Figure S1. Spatial characteristics of the Yalong River Basin. (a) slope, (b) soil types, and (c) land use of the Yalong River Basin.

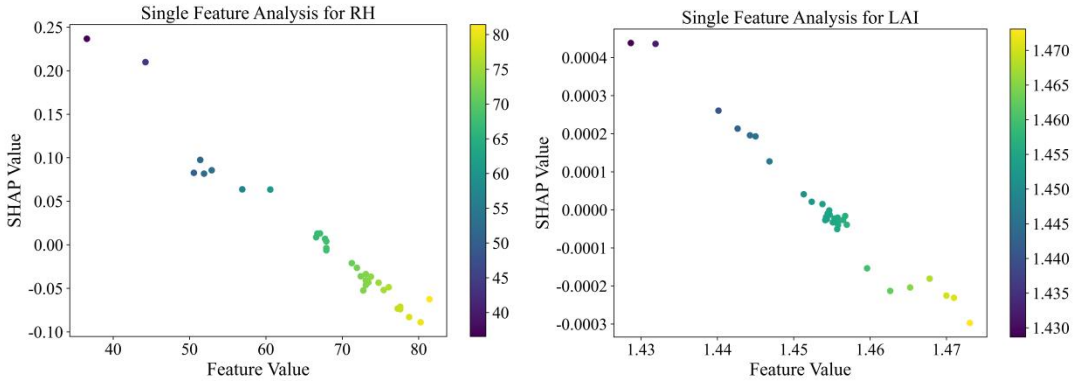

Figure S2. SHAP dependence plots for single-feature analysis of RH and LAI. These plots illustrate how variations in RH and LAI values influence their SHAP values, reflecting their individual impacts on the model's prediction of DOC. The color gradient represents the magnitude of feature values.

Table S1. Comparison of Machine Learning Applications in Hydrology.

Ref.	Modeling techniques compared	Input variables	Output variables	Best model
(Liu et al., 2021)	XGBOOST, SVM, ANN	POC	POC	XGBOOST
(Yi et al., 2023)	DTR, GBR, LR, RFR, SVR	DOM composition	DOM	SVR
(Liang et al., 2023)	SVR, XGBOOST, RF, LSTM	Streamflow	Streamflow	LSTM
(Khosravi et al., 2022)	CNN, RF, ANFIS, SVR, MLP	Precipitation	Streamflow	CNN
(Khairudin et al., 2020)	ANN, SVR, DT, RFA, LSTM	Rainfall,	Rainfall	LSTM
(Maußner et al., 2025)	LS, DT, KNN, SVR, RF, LSTM	Water demand, Calendar input, Weather parameters	Water demand	SVR, RF
(Rahimzad et al., 2021)	LR, MLP, SVM, LSTM	Precipitation, Streamflow	Streamflow	LSTM

Table S2. The hyperparameter of ML.

ML type	Hyperparameters	Value
CNN	n_filters	81
	kernel_size	2
	n_dense	57
	learning_rate	0.0022486618511447297
	look_back	5
SVR	C	36.40872092554611
	epsilon	0.0007796103050027225
	gamma	auto
	kernel	poly
	look_back	18
XGBOOST	n_estimators	324
	max_depth	5
	learning_rate	0.023006063654684675
	subsample	0.5386606459539173
	colsample_bytree	0.9433033312130754
	gamma	0.029017276570926617
	reg_alpha	0.00019354992590110747
	reg_lambda	0.05586083835565143

	look_back	16
	look_back	3
	n_layers	2
MLP	n_units	82
	dropout_rate	0.09661161287007619
	learning_rate	0.008830715997276343
	look_back	10
	lstm_units	67
V-LSTM	dense_units	62
	dropout_rate	0.11795608570064298
	learning_rate	0.0022638192384822464
	look_back	5
	n_layers	2
St-LSTM	dense_units	95
	dropout_rate	0.21328331790576532
	learning_rate	0.0030991349634583948
	look_back	3
	filters	42
	kernel_size	3
Cv-LSTM	lstm_units	61
	learning_rate	0.0022511362507718993
	dense_units	63
	look_back	1
	lstm_units	64
	dense_units	30
Bi-LSTM	dropout_rate	0.2047821669899484
	learning_rate	0.0032213437409123447
	epochs	21
	batch_size	60

Table S3. SWAT-C calibration results for YLR.

Parameter	Change	Calibration range	Calibrated value	Definition	T-Stat/ P-Value

v__CH_K1.sub	Replace	(0,300)	107.250	Effective hydraulic conductivity in tributary channel alluvium	15.72/0.04
v__CH_K2.rte	Replace	(0,500)	253.750	Effective hydraulic conductivity in main channel alluvium	16.27/0.04
v__CH_COV1.rte	Replace	(-0.05,0.6)	0.368	Channel erodibility factor	2.55/0.24
v__CH_COV2.rte	Replace	(-0.01,1)	0.513	Channel cover factor	-14.25/0.04
v__USLE_P.mgt	Replace	(-0.5,0.6)	0.548	USLE support practice factor	-17.91/0.04
r__SOL_K().sol	Relative	(-1,1)	-0.605	Saturated hydraulic conductivity	-15.96/0.04
v__CH_N2.rte	Replace	(-0.01,0.3)	0.047	Manning's "n" value for the main channel	-17.18/0.04
v__ALPHA_BNK.rte	Replace	(0,1)	0.483	Baseflow alpha factor for bank storage	17.54/0.04
v__GW_DELAY.gw	Replace	(0,500)	31.250	Groundwater delay (days)	-17.71/0.04
v__CANMX.hru	Replace	(0,100)	39.500	Maximum canopy storage	5.66/0.11
v__BIOMIX.mgt	Replace	(0,1)	0.385	Biological mixing efficient	-16.99/0.04
r__CN2.mgt	Relative	(-1,1)	0.415	SCS runoff curve number f	-8.33/0.08
v__SPEXP.bsn	Replace	(1,1.5)	1.146	Exponent parameter for calculating sediment re-entrained in channel sediment routing	16.53/0.04
v__ESCO.hru	Replace	(0,1)	0.175	Soil evaporation compensation factor	5.13/0.12
v__LAT_SED.hru	Replace	(0,5000)	883.333	Sediment concentration in lateral flow and groundwater flow	-14.59/0.04
v__SED_CON.hru	Replace	(0,5000)	783.333	Sediment concentration in runoff, after urban BMP is	-13.49/0.05

				applied	
v__CH_BNK_KD.rte	Replace	(0,3.75)	2.938	Erodibility of channel bank sediment by jet test (cm3/N-s)	14.85/0.04
v__CH_BED_KD.rte	Replace	(0,3.75)	1.688	Erodibility of channel bed sediment by jet test (cm3/N-s)	20.15/0.03
v__CH_BNK_TC.rte	Replace	(0,400)	156.000	Critical shear stress of channel bank (N/m2)	14.76/0.04
v__CH_BED_TC.rte	Replace	(0,400)	169.333	Critical shear stress of channel bed (N/m2)	12.14/0.05
v__CH_BNK_D50.rte	Replace	(1,10000)	7966.871	D50 Median particle size diameter of channel bank sediment	-9.66/0.07
v__CH_BED_D50.rte	Replace	(1,10000)	8033.530	D50 Median particle size diameter of channel bed sediment	-19.74/0.03
r__SPCON.bsn	Relative	(0,1)	0.994	Linear parameter for calculating the maximum amount of sediment that can be re-entrained during channel sediment routing	-10.81/0.06
r__EROS_SPL.bsn	Relative	(0.9,3.1)	1.074	Splash erosion coefficient for soil detachment by rainfall	-7.31/0.09
r__RILL_MULT.bsn	Relative	(0.5,2)	0.930	Multiplier to USLE_K for soil susceptible to rill erosion	16.03/0.04
r__SOL_BD().sol	Relative	(0,1)	0.462	Moist bulk density.	23.02/0.03
r__HRU_SLP.hru	Relative	(0,1)	0.828	Average slope steepness	-12.72/0.05
v__GWQMN.gw	Replace	(0,5000)	1012.5	Treshold depth of water in the shallow aquifer required for return flow to occur (mm)	-18.25/0.03

v_peroc_DOC_para.tes	Replace	(0,1)	0.011	DOC percolation coefficient	-19.31/0.03
v_part_DOC_para.tes	Replace	(4000, 10000)	5003.388	organic carbon partition coefficient	-20.26/0.03

Table S4. Specific locations of the SWAT-C modules

Data Name	Definition	SWAT-C Module	Locations
FLOW_OUT	Streamflow output	Runoff module	surface_change.f, surfst_h2o_change.f, wattable.f, etc.
SED_OUT	Sediment output	Sediment module	latsed.f, bacteria.f, cfactor_change.f, etc.
DOC_Simulate	Simulated total DOC	Carbon cycle module	orgncswat_change.f, gw_doc_new.f, substor_change.f, subbasin_change.f, etc.
BIOM	Biomass	Biomass module	plantmod_change.f, grow_change.f
ET	Evapotranspiration	Evapotranspiration module	etpot_change.f, etact_change.f
TOT_P	Total phosphorus output	Pollutant transport module	pestlch.f, pestw.f, pesty.f, nminrl.f, solp_change.f, etc.
LAI	Leaf area index	Vegetation growth module	plantmod_change.f, grow_change.f
SW	Soil water content	Soil moisture module	percmain.f
PCP, RH, TMAX, TMIN, SR, WIND	Precipitation, Relative humidity, Max/Min Temperature, Solar radiation, Wind speed	Meteorological forcing module	pmeas.f, hmeas.f, tmeas.f, smeas.f, wmeas.f,

References

Belayneh, A., Adamowski, J., Khalil, B., and Ozga-Zielinski, B.: Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models, Journal of Hydrology, 508, 418–429, https://doi.org/10.1016/j.jhydrol.2013.10.052, 2014.

Cortes, C. and Vapnik, V.: Support-vector networks, Mach. Learn., 20, 273–297, https://doi.org/10.1007/BF00994018, 1995.

Deulkar, A. M., Dixit, P. R., Londhe, S. N., and Jain, R. K.: Comparative assessment of artificial neural networks (ANNs), long short term memory network (LSTM) and hydrologic engineering centre-hydrologic modelling system (HEC-HMS) for runoff modelling, Water Resour. Manage., 39, 2049–2068, https://doi.org/10.1007/s11269-024-04055-9, 2025.

Khairudin, N. B. M., Mustapha, N. B., Aris, T. N. B. M., and Zolkepli, M. B.: Comparison of machine learning models for rainfall forecasting, in: 2020 International Conference on Computer Science and Its Application in Agriculture (ICOSICA), 2020 International Conference on Computer Science and Its Application in Agriculture (ICOSICA), Bogor, Indonesia, 1–5, https://doi.org/10.1109/ICOSICA49951.2020.9243275, 2020.

Khosravi, K., Golkarian, A., and Tiefenbacher, J. P.: Using optimized deep learning to predict daily streamflow: a comparison to common machine learning algorithms, Water Resour Manage, 36, 699–716, https://doi.org/10.1007/s11269-021-03051-7, 2022.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.: Backpropagation applied to handwritten zip code recognition, Neural Comput., 1, 541–551, https://doi.org/10.1162/neco.1989.1.4.541, 1989.

Liang, W., Chen, Y., Fang, G., and Kaldybayev, A.: Machine learning method is an alternative for the hydrological model in an alpine catchment in the tianshan region, central asia, J. Hydrol.: Reg. Stud., 49, 101492, https://doi.org/10.1016/j.ejrh.2023.101492, 2023.

Liu, H., Li, Q., Bai, Y., Yang, C., Wang, J., Zhou, Q., Hu, S., Shi, T., Liao, X., and Wu, G.: Improving satellite retrieval of oceanic particulate organic carbon concentrations using machine learning methods, Remote Sens. Environ., 256, 112316, https://doi.org/10.1016/j.rse.2021.112316, 2021.

Maußner, C., Oberascher, M., Autengruber, A., Kahl, A., and Sitzenfrei, R.: Explainable artificial intelligence for reliable water demand forecasting to increase trust in predictions, Water Research, 268, 122779, https://doi.org/10.1016/j.watres.2024.122779, 2025.

Mirzaei, M., Yu, H., Dehghani, A., Galavi, H., Shokri, V., Mohsenzadeh Karimi, S., and Sookhak, M.: A Novel Stacked Long Short-Term Memory Approach of Deep Learning for Streamflow Simulation, Sustainability, 13, 13384, https://doi.org/10.3390/su132313384, 2021.

Piri, J., Abdolahipour, M., and Keshtegar, B.: Advanced machine learning model for prediction of drought indices using hybrid SVR-RSM, Water Resour. Manage., 37, 683–712, https://doi.org/10.1007/s11269-022-03395-8, 2023.

Rahimzad, M., Moghaddam Nia, A., Zolfonoon, H., Soltani, J., Danandeh Mehr, A., and Kwon, H.-H.: Performance Comparison of an LSTM-based Deep Learning Model versus Conventional Machine Learning Algorithms for Streamflow Forecasting, Water Resour Manage, 35, 4167–4187, https://doi.org/10.1007/s11269-021-02937-w, 2021.

Singh, A., Imtiyaz, M., Isaac, R. K., and Denis, D. M.: Comparison of soil and water assessment tool (SWAT) and multilayer perceptron (MLP) artificial neural network for predicting sediment yield in the nagwa agricultural watershed in jharkhand, india, Agric. Water Manage., 104, 113–120, https://doi.org/10.1016/j.agwat.2011.12.005, 2012.

Wang, D. and Lu, W.-Z.: Forecasting of ozone level in time series using MLP model with a novel hybrid training algorithm, Atmospheric Environment, 40, 913–924, https://doi.org/10.1016/j.atmosenv.2005.10.042, 2006.

Yang, Q. and Zhang, X.: Improving SWAT for simulating water and carbon fluxes of forest ecosystems, Sci. Total Environ., 569–570, 1478–1488, https://doi.org/10.1016/j.scitotenv.2016.06.238, 2016.

Yi, Y., Liu, T., Merder, J., He, C., Bao, H., Li, P., Li, S., Shi, Q., and He, D.: Unraveling the linkages between molecular abundance and stable carbon isotope ratio in dissolved organic matter using machine learning, Environmental Science & Technology, https://doi.org/10.1021/acs.est.3c00221, 2023.

Yuan, X., Wang, J., He, D., Lu, Y., Sun, J., Li, Y., Guo, Z., Zhang, K., and Li, F.: Influence of cascade reservoir operation in the upper mekong river on the general hydrological regime: a

combined data-driven modeling approach, Journal of Environmental Management, 324, 116339, https://doi.org/10.1016/j.jenvman.2022.116339, 2022.

Zhang, X., Izaurralde, R. C., Arnold, J. G., Williams, J. R., and Srinivasan, R.: Modifying the Soil and Water Assessment Tool to simulate cropland carbon flux: Model development and initial evaluation, Sci. Total Environ., 463–464, 810–822, https://doi.org/10.1016/j.scitotenv.2013.06.056, 2013.

Zhang, X., Liu, J., Zhu, J., Cheng, W., and Zhang, Y.: Analysis of the spatiotemporal patterns of water conservation in the Yangtze River ecological barrier zone based on the InVEST model and SWAT-BiLSTM model using fractal theory: a case study of the minjiang river basin, Fractal Fract., 9, 116, https://doi.org/10.3390/fractalfract9020116, 2025.