Supporting information for

A Mechanism-Data Fusion Model for Algal Growth in Hydrologically Constrained River Segments

Ying Liu [1,&], Zhiwei Ren [1,&], Zelin Jing [1], Jihong Liu [1], Qingsong Chen [2,3], Yurou Wang [1], Liwenze He [4,*] , Yu Chen [1**]

[1] *Faculty of Environment Science and Engineering, Southwest Jiaotong University, Chengdu 611756, China*

[2] *Sichuan Academy of Eco-environmental Sciences, Chengdu 610000, China*

[3] *Sichuan Province Environmental Protection Technology & Engineering Co., Ltd, Chengdu 610000, China*

[4] *Department of Civil Engineering, Chengdu Technological University, Chengdu, 611730, People's Republic of China*

[&] *These authors contributed equally to this work and should be considered co-first authors*

[*] *Correspondence: hlwze1@cdtu.edu.cn*

[**] *Correspondence: chenyu1123@swjtu.edu.cn; Tel.: +86-1388-072-7070*

Text S1 Parameter Uncertainty and Model Limitations in the Jiuqu River Algal Growth Model

Parameter uncertainty analysis:

In this study, we developed a dynamic algal growth model for the Jiuqu River and identified key environmental factors influencing algal growth in the watershed using GBDT and SHAP analysis. The model includes eight biological constants (g_{max}, d_{max}, k_{CODMn}, k_{CODCr}, k_P, Topi, k_R, R_0), which are interrelated and closely linked to monitoring data such as total phosphorus (TP), chemical oxygen demand (COD_{Mn}, COD_{Cr}), runoff, and temperature (T) (Richter and Matuła, 2012; Solimeno et al., 2016).

The uncertainty of these biological constants mainly stems from three aspects:

(1) Measurement errors: Monitoring data may be influenced by instrument precision limitations, operational errors, and spatiotemporal sampling variations, all of which can affect the optimization of biological constants (Pianosi et al., 2016; Baker et al., 2023).

(2) Seasonal variations: Fluctuations in meteorological conditions such as temperature and sunlight duration, as well as seasonal changes in non-point source pollution, impact the algal growth environment and consequently affect parameter values.

(3) Model limitations: Discrepancies between the dynamic model and real-world conditions introduce additional uncertainty.

Since these eight biological constants not only reflect algal responses to environmental changes but are also influenced by variations in input variables, analyzing their sensitivity in isolation is insufficient to reveal their impact on model predictions. Therefore, this study does not conduct individual sensitivity analyses of these constants.

Given the strong correlation between biological constants and input variables, as well as their optimization based on real monitoring data, errors in the data and seasonal fluctuations may directly affect the optimization results. To evaluate the model's predictive capability under different datasets, we compared parameter optimization results using the first 50 data samples (from July to September, the peak algal growth period) with those using all 161 samples spanning summer, autumn, and winter. The results indicate that the model performs better when optimized using the first 50 samples. This may be because, during the algal growth season, key environmental drivers dominate algal growth, making it easier for the model to capture essential features, and the optimized biological constants exhibit higher representativeness. In contrast, when using the full dataset for optimization, the inclusion of multiple seasons increases data heterogeneity and complexity, potentially introducing additional noise and reducing optimization effectiveness.

Therefore, this study recommends considering different growth stages when optimizing biological constants using genetic algorithms. Specifically, separate optimizations based on data from the growth season and the full lifecycle could further enhance model performance and ensure its effective development.

Model limitations and future perspectives:

Although this study has attempted to construct an adaptable algal growth model by integrating biological mechanisms and data-driven methods under data-limited river segment conditions, the model still has certain limitations. First, the influence functions use simplified mathematical expressions (such as sine, cosine, logarithmic functions), which, while based on ecological process logic, still fall short in terms of biological mechanistic rigor compared to traditional mechanistic models. Second, to improve the model's generalizability, some biological processes have been approximated, which may overlook microscopic differences under specific conditions.

Moreover, the growth parameters in the model are optimized using a genetic algorithm, which, while improving prediction accuracy, makes the model structure relatively sensitive to parameter changes. The stability of the model under extreme conditions still needs further validation. Future research could consider incorporating more measured physiological data to optimize the biological interpretability of the parameters, while also integrating data assimilation and real-time monitoring techniques to enhance the model's adaptability in various types of water bodies and during sudden environmental events.

Text S2 Detailed Marginal Distribution Fitting Results

The goodness-of-fit of each distribution model was evaluated using the test statistics, and the best-fitting model that most accurately reflects the actual distribution characteristics of each environmental factor was selected. Fig.7 illustrates the distribution fitting results for the key environmental factors and chlorophyll a concentration. By comparing the fitting results and test statistics of different distribution models, the marginal distributions of the environmental factors and chlorophyll a concentration were clearly identified.

Tab. S4 presents the optimal distribution fitting results for different environmental factors and chlorophyll a, listing the best-fit distribution type for each variable. As shown in the table, CODCr and TP are best described by the Generalized Extreme Value (GEV) distribution, while CODMn fits well with the log-normal distribution. Both T and chlorophyll a exhibit a good fit to the exponential distribution, whereas R is best characterized by the Weibull distribution. These fitting results establish the optimal marginal distributions for each environmental factor, providing a robust foundation for subsequent Copula function modeling and the calculation of warning thresholds.

Text S3 Development of the Copula Model

After determining the optimal marginal distributions for each environmental factor, the study employed Copula functions to model the joint probability distribution between key environmental factors and chlorophyll a concentration. To evaluate the applicability of different Copula models, Gumbel Copula, Clayton Copula, Frank Copula, and t-Copula were compared.

Model parameters were optimized using the Maximum Likelihood Estimation (MLE) method. A comprehensive evaluation of the models was conducted by combining the Akaike Information Criterion (AIC) and cross-validation results. Based on these evaluations, the optimal Copula function for each pair of environmental factors and chlorophyll a concentration was selected (see Tab. S5). The chosen models provided a precise depiction of the nonlinear dependencies between critical environmental factors and chlorophyll *a*, facilitating accurate risk assessments.

Tab. S5 presents the optimal Copula model for each key environmental factor and chlorophyll *a* concentration. It reveals that the joint distributions of COD_{Cr}, COD_{Mn}, TP, T, and Runoff with chlorophyll *a* are all best described using the Gaussian Copula model, indicating relatively weak nonlinearity and strong linear characteristics in the relationships between these factors and chlorophyll *a* in this watershed.

Tab. S1 Ranking of key factor contribution values of GBDT and SHAP eigenvalue screening

Ranking	Selection Methods	
	GBDT	SHAP
1	COD_{Mn}	T
2	T	COD_{Mn}
3	COD_{Cr}	COD_{Cr}
4	Runoff	Runoff
5	TP	TP
6	ORP	Sunshine_Duration
7	Sunshine_Duration	ORP
8	NH_4^+-N	NH_4^+-N

Tab. S2 CCF analysis results of key environmental factors

Environmental factor	T	COD_{Cr}	COD_{Mn}	TP	Runoff
Delay unit	-7	-6	-6	0	-5

Tab. S3 Parameter optimization results of training different data sets

Data name	Parameter optimization range	50 sets of optimal parameters are trained	Optimum parameters for complete training
g_{max}	0.5-5	4.78	4.16
T_{opi}	10-25	19.81	22.83
d_{max}	0.5-2	0.51	0.96
k_P	0.05-0.5	0.48	0.3
k_{CODCr}	20-50	33.74	33.8
k_{CODMn}	2-15	10.47	14.26
k_R	0.5-1.5	0.9	0.53
R_0	0.1-0.5	0.42	0.49

Tab. S4 Test results of optimal distribution of key environmental factors and chlorophyll *a*

Data name	COD_{Cr}	COD_{Mn}	T	TP	Runoff	Chl.a
Fit the optimal distribution	Gev distribution	Lognormal distribution	Exponential distribution	Gev distribution	Weibull distribution	Exponential distribution

Tab. S5 Best Copula functions fitting different key environmental factors

Date name	COD_{Cr}	COD_{Mn}	TP	T	Runoff
Optimum Copula model	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian

Tab. S6 Risk thresholds of different key environmental factors

Date name	COD_{Cr}	COD_{Mn}	TP	T	Runoff
Risk threshold	18.78	5.2	0.09	12.21	0.04

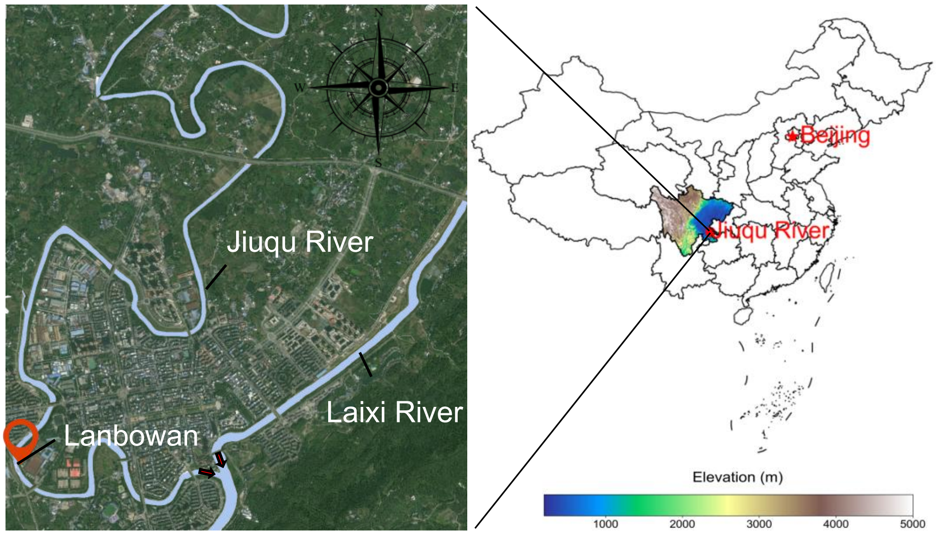

Fig. S7 Study Area and Sampling Site Layout (Red arrows show where the watershed flows)

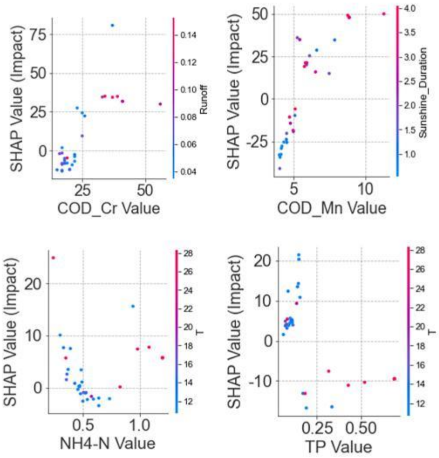

Fig. S8 Feature dependency graphs of chlorophyll *a* concentration on different environmental factors (COD_{Cr}, COD_{Mn}, NH_4^+-N, TP)

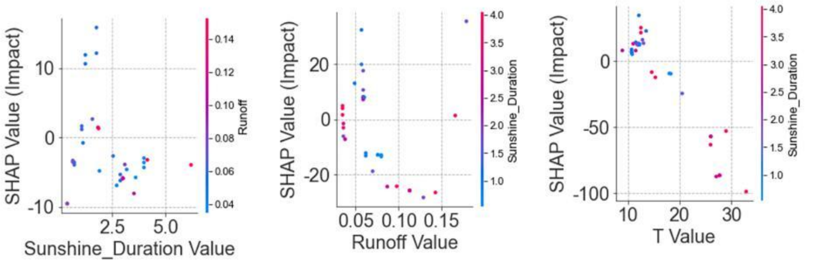

Fig. S9 Feature dependency graphs of chlorophyll *a* concentration on different meteorological and hydrological factors (T, Sunshine Duration, Runoff)

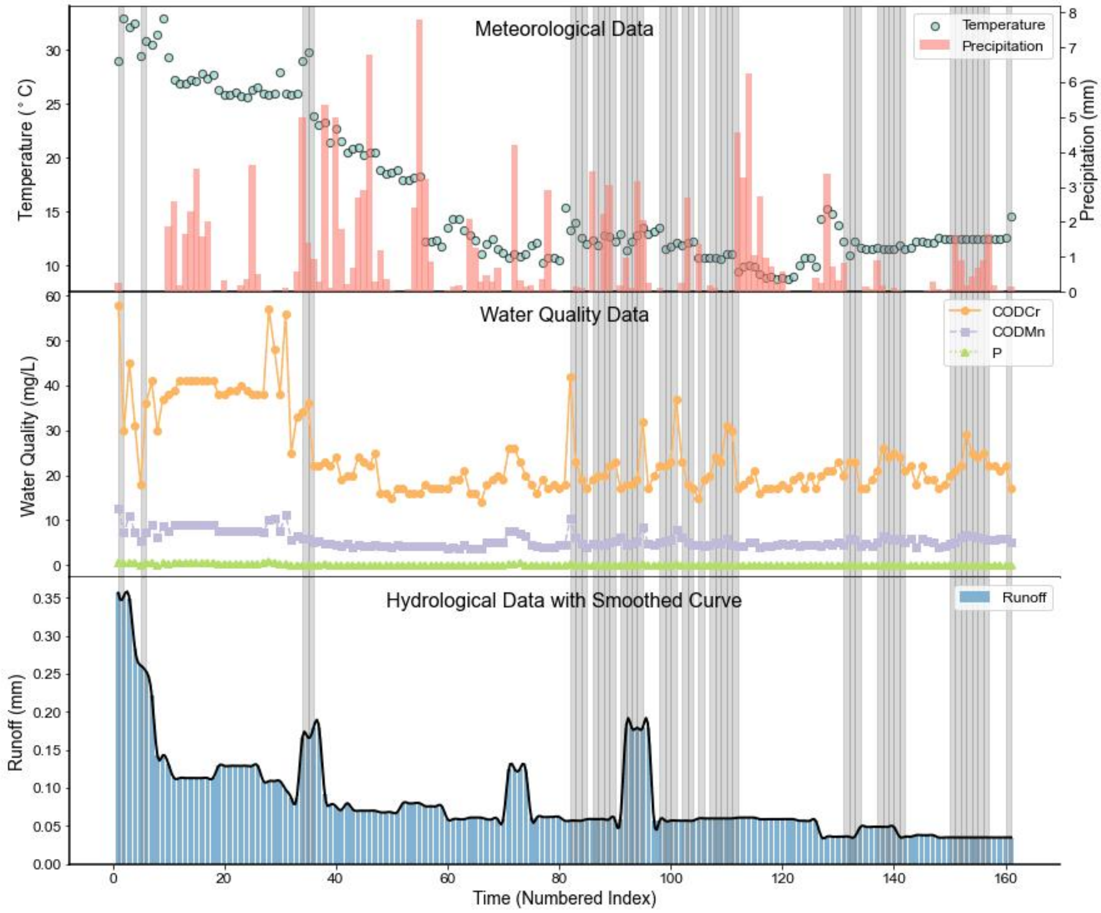

Fig. S10 Water quality, meteorological, hydrological data and rainfall data of Jiuqu River after screening of key factors

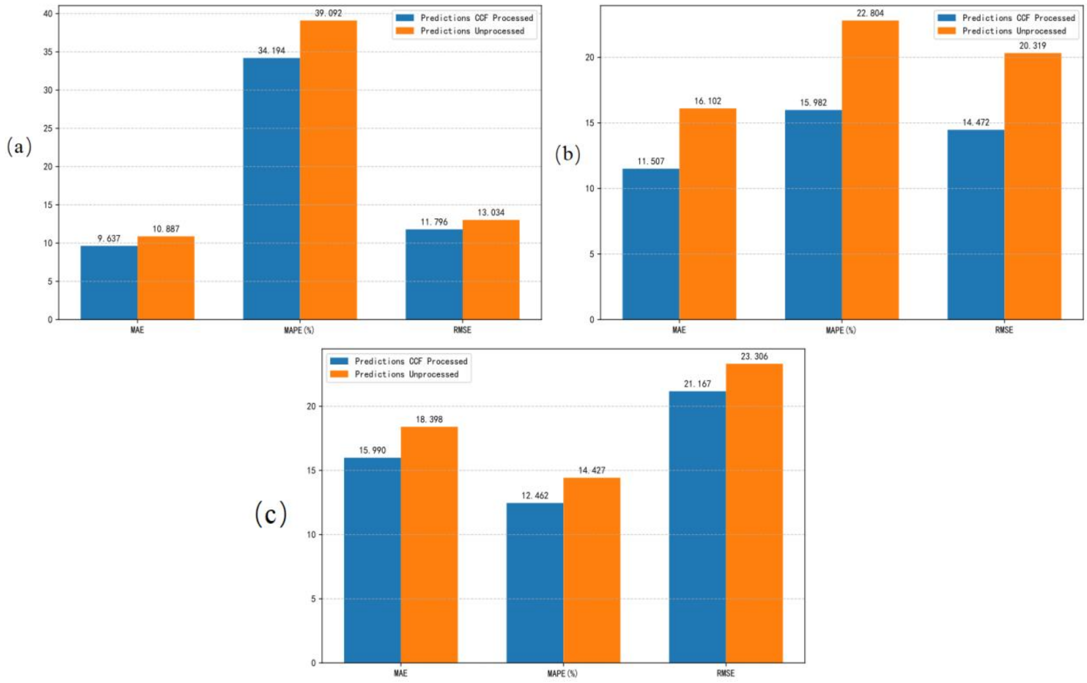

Fig. S11 The comparison of model evaluation metrics before and after delay analysis correction is

shown in the figure. (a) represents the range where chlorophyll *a* concentration is below 50 μg/L; (b) represents the range where chlorophyll *a* concentration is between 50 and 100 μg/L; and (c) represents the range where chlorophyll *a* concentration exceeds 100 μg/L

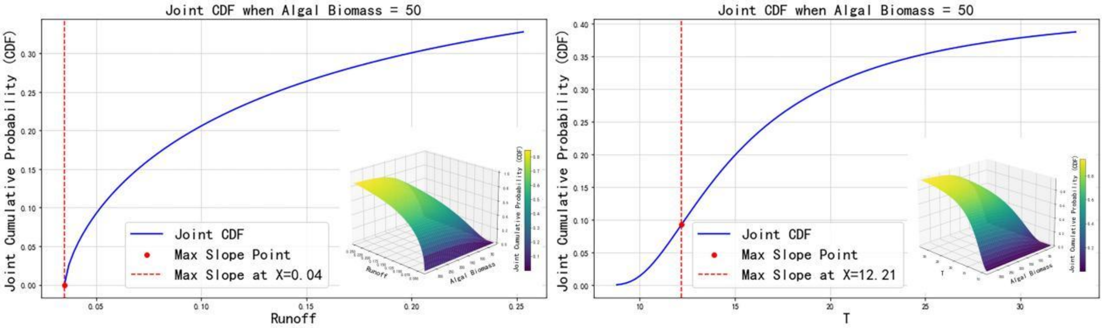

Fig. S12 Distribution fitting of meteorological and hydrological factors (T, Runoff)

References

Baker, E., Manenti, S., Reali, A., Sangalli, G., Tamellini, L., Todeschini, S. (2023). Combining noisy well data and expert knowledge in a Bayesian calibration of a flow model under uncertainties: an application to solute transport in the Ticino basin. *GEM - International Journal on Geomathematics,* 14(1). https://doi.org/10.1007/s13137-023-00219-8

Pianosi, F., Beven, K., Freer, J., Hall, J., Rougier, J., Stephenson, D., Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software,* 79, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008

Richter, D., Matuła, J. (2012). Response of cyanobacteria and algae community from small water bodies to physicochemical parameters. *Oceanological and Hydrobiological Studies,* 41(2), 18–28. https://doi.org/10.2478/s13545-012-0013-3

Solimeno, A., Samsó, R., García, J. (2016). Parameter sensitivity analysis of a mechanistic model to simulate microalgae growth. *Algal Research,* 15, 217–223. https://doi.org/10.1016/j.algal.2016.02.027