Unraveling the factors influencing spatiotemporal variations in riverine dissolved organic matter and iron through a machine learning approach

Dissolved organic matter (DOM) serves vital functions in aquatic ecosystems, such as regulating light availability in the water column and providing energy and nutrients to microorganisms, while excessive loading of light-absorbing colored DOM reduces light penetration, inhibiting photosynthesis by primary producers (e.g., phytoplankton, seaweeds). DOM also serves as an organic ligand for iron, an essential micronutrient for primary producers, and increases its availability. Based on the result of periodic water quality monitoring in the rivers of Ishigaki Island, a tropical island of Japan, the factors influencing spatiotemporal variations in the concentrations of riverine DOM and its components were identified by analyzing their relationships with catchment properties (e.g., land use, soil type) and seasonality (e.g., water temperature). Furthermore, the impact of the molecular composition of DOM on the concentration of dissolved iron (DFe) was assessed. The random forest (RF) machine learning algorithm was employed for the analyses because of its flexibility in handling non-parametric datasets and non-linear relationships, and its ability to measure the importance of predictor variables.

The RF models using catchment properties and water temperature as predictor variables accurately predicted the concentration of dissolved organic carbon (DOC) and the abundance of three humic-like components (C1 ~ C3) identified by fluorescence excitation-emission matrix coupled with parallel factor analysis (EEM-PARAFAC). Water temperature and areal share of poorly-drained lowland soil (Gleyic Fluvisols) were identified as the most important predictor variables for DOC and the humic-like components (Table 1) and positively influenced these DOM parameters (Fig. 1). This result indicates that the concentrations of DOC and humic-like components exhibit clear seasonal variations with their maxima in summer and that the poorly-drained lowland soil serves as the major source of riverine DOM (particularly humic-like components) in the studied catchments. The RF model for DFe using the abundance of EEM-PARAFAC components and other parameters relevant to iron solubility (e.g., water temperature, pH, concentrations of Ca^{2+} and Mg^{2+}) as predictor variables also explained a large portion of the variation in DFe concentration. A humic-like component derived from terrestrial material (C1) was the most important predictor variable and had a positive relation to DFe concentration (Fig. 2), emphasizing its significance as an organic ligand for iron.

The results obtained in this study improve our understanding of the spatiotemporal variability of terrestrial DOM and iron loadings and their impacts on tropical coastal ecosystems of high ecological and economic importance.

Authors: Kikuchi, T., Anzai, T. [JIRCAS]

Table 1. Importance of the predictor variables in the random forest (RF) models for dissolved organic carbon (DOC) and three humic-like components (C1 ~ C3)

	Variable	DOC	C1	C2	C3
	Water temperature	**16.1**	**20.8**	**21.3**	**17.5**
Land use	Upland fields	10.8	9.1	11.2	10.0
	Pastures	11.2	8.8	11.5	9.9
	Paddy fields	9.7	7.5	9.3	8.1
	Forests	14.1	11.1	13.2	11.9
	Livestock barns	7.0	8.0	7.2	8.2
Soil	Haplic Acrisols (Chromic)	11.9	11.1	11.3	10.9
	Haplic Cambisols	12.3	9.9	12.4	10.2
	Haplic Acrisols	12.7	8.9	8.3	8.6
	Haplic Cambisols (Eutric)	8.1	7.0	6.3	6.1
	Haplic Regosols (Calcaric)	7.7	6.7	6.2	5.8
	Gleyic Fluvisols	**16.1**	**22.2**	**16.8**	**21.4**
	Population density	11.7	9.6	8.9	8.6

(%) ~ 25 ~ 20 ~ 15 ~ 10

C1: Derived from terrestrial material by photochemical degradation
C2: Produced during microbial degradation of organic matter
C3: Produced during breakdown of lignin (e.g., syringaldehyde)

Importance was measured as the increase in mean squared error (MSE; in %) that occurred when the fitted model was run with the randomly permuted variable of interest. The greater the value, the more important the variable. Values greater than 15% are shown in white bold letters.

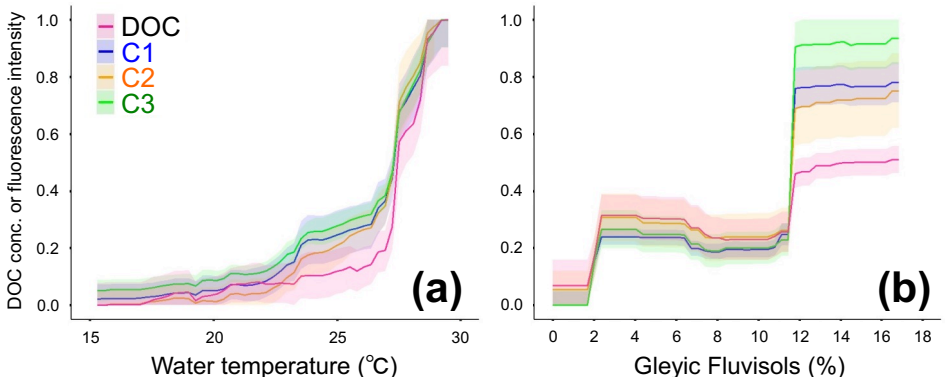

Fig. 1. Partial dependence plots (PDPs) of DOC and humic-like components on (a) water temperature and (b) areal share of Gleyic Fluvisols in the catchment

Solid lines and shaded areas represent the mean partial dependence and its standard deviation, respectively, for 15 RF models generated through three repetitions of five-fold cross-validation. The *y*-axes are scaled to a difference between the maximum and minimum values of each DOM parameter that is common in both panels.

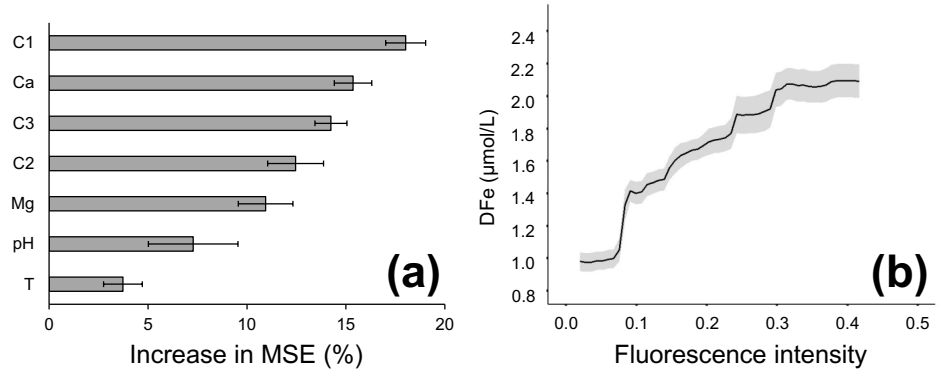

Fig. 2. (a) Importance of the predictor variables in the RF models for dissolved iron (DFe) and (b) PDP of DFe on a humic-like component (C1)

The bars and error bars in 'panel a' represent the mean value and standard deviation, respectively, for 15 RF models from three repetitions of five-fold cross-validation. T denotes water temperature.

Reference: Kikuchi, T., Anzai, T., Ouchi, T. (2023) Assessing spatiotemporal variability in the concentration and composition of dissolved organic matter and its impact on iron solubility in tropical freshwater systems through a machine learning approach. *Science of the Total Environment* 904: 166892.