Sun, Xiuming; Wu, Naicheng; Hörmann, Georg; Faber, Claas; Messyasz, Beata; Qu, Yueming; Fohrer, Nicola. 2022. **Using integrated models to analyze and predict the variance of diatom community composition in an agricultural area.**

The definitive version was published in *Science of the Total Environment*, 803, 149894. https://doi.org/10.1016/j.scitotenv.2021.149894

The definitive version is available at https://www.elsevier.com/

Using integrated models to analyze and predict the variance of diatom community composition in an agricultural area

Xiuming Sun[1*], Naicheng Wu[1*], Georg Hörmann[1], Claas Faber[1], Beata Messyasz[2], Yueming Qu[1] and Nicola Fohrer[1]

[1]: *Department of Hydrology and Water Resources Management, Institute for Natural Resource Conservation, Kiel University, 24118 Kiel, Germany*

[2]: *Department of Hydrobiology, Institute of Environmental Biology, Faculty of Biology, Adam Mickiewicz University in Poznan, Uniwersytetu Poznanskiego 6, 61-614 Poznan, Poland*

[*] *Corresponding authors:*

xsun@hydrology.uni-kiel.de (X. Sun)

nwu@hydrology.uni-kiel.de (N. Wu)

Key words: daily dataset, diatoms, lowland river, integrated modeling, prediction

Abstract

With the growing demand of assessing the ecological status, there is the need to fully understand the relationship between the planktic diversity and the environmental factors. Species richness and Shannon index have been widely used to describe the biodiversity of a community. Besides, we introduced the first ordination value from non-metric multidimensional scaling (NMDS) as a new index to represent the community similarity variance. In this study, we hypothesized that the variation of diatom community in rivers in an agricultural area were influenced by hydro-chemical variables. We collected daily mixed water samples using ISCO auto water samplers for diatoms and for water-chemistry analysis at the outlet of a lowland river for a consecutive year. An integrated modeling was adopted including random forest (RF) to decide the importance of the environmental factors influencing diatoms, generalized linear models (GLMs) combined with 10-folder cross validation to analyze and predict the diatom variation. The hierarchical analysis highlighted antecedent precipitation index (API) as the controlling hydrological variable and water temperature, Si^{2+} and PO_4-P as the main chemical controlling factors in our study area. The generalized linear models performed better prediction for Shannon index (R^2 = 0.44) and NMDS (R^2 = 0.51) than diatom abundance (R^2 = 0.25) and species richness (R^2 = 0.25). Our findings confirmed that Shannon index

and the NMDS as an index showed good performance in explaining the relationship between stream biota and its environmental factors and in predicting the diatom community development based on the hydro-chemical predictors. Our study shows and highlights the important hydro-chemical factors in the agricultural rivers, which could contribute to the further understanding of predicting diatom community development and could be implemented in the future water management protocol.

Introduction

Phytoplankton, the most important primary producers, contribute around 50% to the global primary production (Ptacnik et al., 2008) and to the global cycling of nutrients (Lomas et al., 2014). They play irreplaceable roles in aquatic ecosystems. Among the groups of phytoplankton, diatoms are the most widely spread in the world. Furthermore, based on the fast response to changes in water quality, diatoms have been widely used as bio-indicators to assess the ecological status of aquatic ecosystems (B-Béres et al., 2016; Hill et al., 2000, 2003; Stevenson et al., 1999; Wu et al., 2009, 2017; Zalack et al., 2010). Diatom abundance and diversity are the basic and traditional features to represent the variation of diatom communities. The variation in the phytoplankton community shows the recurring species composition, and biodiversity (Reynolds, 1988). The variation of the communities is caused by environmental variation and the response of each species in the community (Reynolds, 1988). Species richness is one of the important properties to describe the biodiversity of a community (Passy et al., 2017). The diversity index (e.g., Shannon index) is an efficient way to quantify the variation in community species composition (Kim et al., 2020; Zhou et al., 2019), which can show how each species contributes to the whole community (Weaver & Shannon, 1963). Another technique to represent the community composition variance is non-metric multidimensional scaling (NMDS), which shows the similarity between samples by calculating the Bray-Curtis distance (Bray & Curtis 1957). This technique is widely used to display the community pattern through ordination and clustering (Campos et al., 2021; Cotiyane-Pondo et al., 2020; Fukai et al., 2020). In this study, to couple with other models, we used the first ordination value to represent the similarity of each site (day). Species richness, biomass and Shannon indices have been much used in investigating the spatial community variance (Kafouris et al., 2019), seasonal variance (Woelfel et al., 2007), yet in the annual daily-based variance investigation is still rare (Sun et al., 2018; Wu et al., 2019).

The temporal variation of phytoplankton follows the seasonal variation of environmental variables (Lewis Jr 1978; Sommer et al., 1986; Winder & Hunter, 2008). Research has focused mostly on large scale interannual seasonality, normally on the differences between cold and warm seasons (Qu et al., 2019; Wang et al., 2015), dry and wet seasons (Zhou et al., 2019). However, the variability of meteorological conditions may cause smaller scale recurrent than seasonal periodic influences on the direct impact factors on stream biota, such as nutrients. Community composition and structure vary within different time periods. Compared to research with large sampling intervals (i.e., seasonal sampling), there have been fewer cases paying attention to the changes in phytoplankton on a short-term scale (Kim et al., 2020; Winder & Hunter, 2008). One recent research by Babitsch et al. (2021)

confirmed that low sampling frequencies of chemical pollutants and nutrients in rivers reduce the reliability of its performance in models. Research based on annual daily datasets is very rare, but they may provide the chance to understand the phytoplankton's features and variations. The understanding of these mechanisms could provide a solid base for predicting the future development of diatoms.

With the growing demand of assessing the ecological status, of investigating the relationships between biota and environmental influences (Rimet and Bouchez, 2011), there is the need to understand the relationship between the planktic diversity and the environmental factors (Cottenie, 2005; Franklin, 2009; Leibold et al., 2004; Laiolo et al., 2018; Santos et al., 2016; Soininen & Luoto, 2012). In an agricultural area, the biota in the rivers is mainly controlled by nutrients (Andrus et al., 2013; Cornejo et al., 2019). Fertilizers for enhancing crop growth have led to the enrichment of phosphorous (PO_4-P) and nitrogen (NO_3-N) in aquatic ecosystems (Guignard et al., 2017; Serediak et al., 2014). Nutrient enrichment (e.g., NO_3-N and PO_4-P) is one of the main forces to alter the abundance and diversity of diatoms (Wijewardene et al., 2021). Other than nutrients, hydrological effects also play important roles in shaping diatom communities in lentic aquatic ecosystem in agricultural areas (Sun et al., 2018; Qu et al., 2019). Compared to lakes and reservoirs, rivers show more hydrological dynamics resulting from precipitation and inflow from the upstream lakes.

However, the influence of hydro-chemical parameters on biota in rivers in agricultural areas has not yet been fully understood (Indermuehle et al., 2008; Schreiner et al., 2016; Wijewardene et al., 2021).

In this study, we aimed to address answers for the following questions: 1) what are the main annual variance of diatom community, 2) how the hydro-chemical parameters impact on diatom community regarding the abundance, diversity, and indices, 3) how can the selected hydro-chemical variables perform in predicting the diatom variance, and 4) whether the indices are adequate of representing the community composition variation? We conducted the research by using an annual daily dataset. An integrated modeling was adopted including random forest (RF) to decide the importance of the environmental factors influencing diatoms, generalized linear models (GLMs) combined with 10-folder cross validation to analyze and predict the diatom variation. The research questions were discussed based on the integrated modeling results.

Material and Methods

Study area

The Kielstau River is a lowland river with a length of ca. 17 km, a drainage area of ca. 50 km^2. It originates from the upper part of Lake Winderatt and it is a tributary of the Treene River (Fig. 1A), which runs into the Eider River. The dominant land use pattern of Kielstau catchment is agricultural land use, in which arable land is ~55% and pasture ~26% (Fohrer & Schmalz, 2012). Its annual precipitation is around 841 mm (station Satrup, 1961-1990) (DWD, 2010) and the mean annual temperature is 8.2 °C (station Flensburg, 1961-1990). There are six wastewater treatment plants in the Kielstau catchment (Point sources in Fig. 1, B). Discharge is measured at a gauging station (Fig. 1D) at the outlet of the catchment, which is part of the official gauging network of the Federal State Schleswig-Holstein. The catchment has been recognized as an UNESCO eco-hydrological demonstration site since 2010 (Fohrer & Schmalz, 2012).

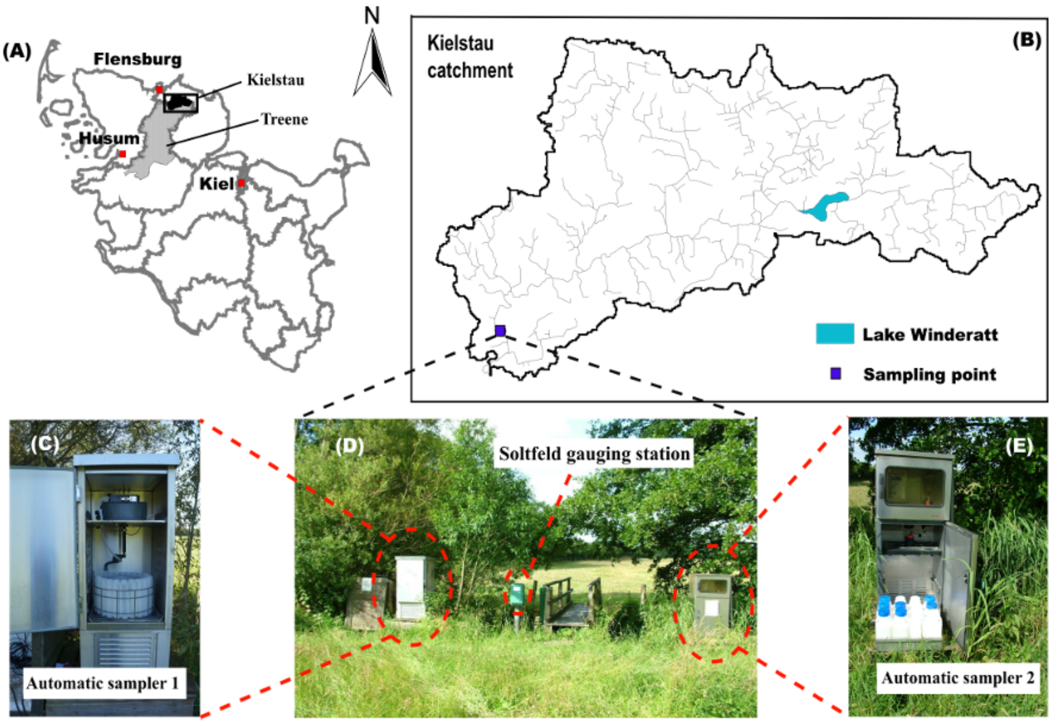

Fig. 1. Location of the Kielstau catchment (B) in Schleswig-Holstein state (A) and photos of the Soltfeld gauging station (D) and automatic water samplers for daily-mixed samples (C and E). (Photos by Sun, 2015)

<u>Sampling method</u>

Daily mixed water samples have been taken directly from the river by two auto-samplers (Fig. 1, C and E) close to the gauging station at the outlet, from April 29, 2013 to April 30, 2014. The physical variables pH, electric conductivity, water temperature and dissolved oxygen were measured weekly *in situ* with a portable instrument (WTW Multi 340i, Weilheim Germany). One of the auto-samplers (Fig. 1, C: ISCO 6712 Refrigerated Sampler Teledyne) kept the temperature at 4 °C and the water samples from it were used to determine the concentration of nutrients and metal ions in the laboratory of the Department of Hydrology and Water Resources Management of Kiel University according to the DIN standard methods. The water samples from the other auto-sampler (Fig. 1, E: Maxx Refrigerated Sampler SP 5 S) were used to prepare permanent diatom slides for further microscopic analysis.

Hydro-chemical analysis

The chemical variables analyzed included ammonium-nitrogen (NH_4-N), nitrate-nitrogen (NO_3-N), chloride (Cl^-), metal ions (K^+, Ca^+, Na^+, Mg^{2+} and Si^{2+}), orthophosphate-phosphorus (PO_4-P), sulphate (SO_4^{2-}), total phosphorus (TP) and total suspended solids (TSS). The concentration of metal ions was analyzed by inductively coupled plasma (IC) method (EN ISO 10304-1). Hydrological variables included daily discharge (Q), baseflow (BF), surface runoff (SR), precipitation (PREC), water depth (WD) and antecedent precipitation index (API). Surface runoff was calculated by end-member mixing analysis (EMMA) (Christophersen & Hooper, 1992). Baseflow end member was determined from stream water samples taken during dry periods with low discharge. These samples represented a state where the other end member did not contribute significantly. This state included slow catchment processes (which integrate catchment response over a long time) such as groundwater inflow and interflow from soil. In our catchment, upstream effects such as discharge from lake Winderatt and surrounding wetlands also contributed to the baseflow end member since these upstream sources were stable even during baseflow conditions.

Precipitation data was obtained from the nearby weather station of Moorau. API was an index to estimate the hydrological condition in the catchment and was calculated on a daily basis (Fedora & Beschta, 1989; Shaw, 1994).

$$API_t = (k*API_{t-1}) + P_{t-1} \quad (1)$$

where API_t = antecedent precipitation index (mm) at day t, P_{t-1} = precipitation (mm) at the day t−1, and k represents the potential loss of moisture, more details were given in (Wu et al., 2016).

Diatom preparation

The water samples from the second auto-sampler (Fig. 1, E) were transferred into 2.5 L separatory funnels and fixed in 5‰ non-acetic Lugol's iodine solution (Sabater et al., 2008). After a sedimentation period of 48 hours, the undisturbed water samples from the bottom of the separatory funnels were concentrated to 20 mL. The concentrated samples were used to prepare the diatom permanent slides (Fig. 2). We transferred 10 mL of the concentrated samples into centrifuge tubes and then siphoned off the supernatant. Afterwards, 5 mL of 30% hydrogen peroxide (H_2O_2) was added to eliminate the organic matter with heating in a water bath at 60°C for two days (flexibly adjust the time according to the oxidization process). Then we added 1 mL of 1 mol/L hydrochloric acid (HCl) to eliminate calcium carbonate with a reaction time for at least two hours. The supernatant was removed and refilled to 5 mL with distilled water after the samples staying in a centrifuge at a speed of 1200 rounds/min for 10 mins. This cleaning procedure has been repeated for three times or more until the pH value reached 7. The diatoms samples were determined to 0.5 mL with Ethanol, after which 0.1 mL of the well-mixed sample was put on and dried on a cover slip on a hotplate in a fume cupboard. The permanent diatom slides were afterwards mounted with Naphrax (Northern Biological supplies Ltd., UK, R1=1.74). These diatom slides were used to identify diatom species under a light microscope. A minimum of 300 individuals for each permanent slide was identified and counted with a Zeiss Axioskop microscope at 1000× under oil immersion. Diatoms were identified to the possible lowest taxonomic level (mostly species level) according to the key books by Bey and Ector, 2013; Lange-Bertalot, 2000a, 2000b, 2005, 2007; Round et al., 1990; and Simonsen, 1987.

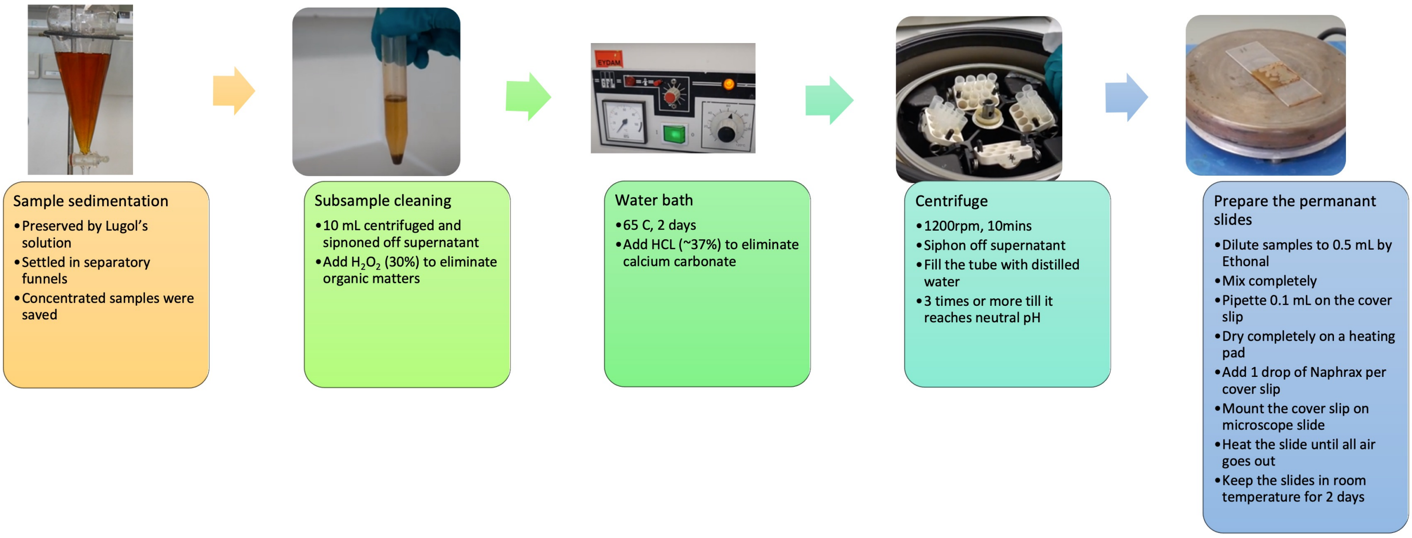

Fig. 2. Workflow diagram of the diatom permanent preparation.

The abundances were expressed as cells/L. The species richness was represented as the species number counted in the sample. The Shannon index was calculated according to Weaver & Shannon (1963) and the difference of diatom community composition was represented by the first dimension of non-metric multidimensional scaling (NMDS). The NMDS could be used to visualize differences in composition with Bray-Curtis similarity index (Stanish et al., 2012). In our study, we simplified it and chose the first ordination number to present the similarity of the community composition in temporal point of view.

Numerical analysis

First, we checked whether there were obvious linear relationships between the variables by applying the pairwise comparison analysis of all biotic and environmental variables. Second, we applied Pearson's correlation analysis to exclude the variables with high co-linearity. For the variable pairs with correlation coefficients greater than 0.6, we retained the variable which had a lower correlation with the other variables. In addition to that, we took empirical experiences into account to remain both variables if they are both very important for the growth of diatoms (i.e., temperature). We conducted the standard autocorrelation function (ACF) to analyze whether there is a lagged relationship of each variable. The maximum lag time was set to 100 days for ACF to make sure we won't oversee the lagged relationship. Besides the autocorrelation, we checked cross correlation, where one variable was correlated with lagged time series of a second. We applied cross correlation function (CCF) to explore whether there was a lagged relationship between diatom abundance and the other variables. The lag time for CCF was set to 100 days. Prior to the other analyses, we standardized (z-score normalization, *scale* function) all biotic and environmental datasets to avoid the effects from different measured units of the variables.

The machine learning models have been developing very fast recently and are now widely used in ecological research (Culter et al., 2007; Derot et al., 2020; Park et al., 2015). Random forest (RF) is a flexible and non-parametric regression tool which can not only be used to analyze non-linear relationships and complex interactions, but also to handle data sets with a large number of observations. RF generates the models by training two-thirds of the observations (“in the bag” data) and tests the models with the remaining ones (“out of bag” data). From the estimates of the out-of-bag error, RF can efficiently test the accuracy of the model by itself (Breiman, 2001). In addition, RF is a powerful statistical classifier to determine the variable importance and to model complex interactions between predictor variables. It is more flexible to deal with missing values (Cutler et al., 2007). RF also shows the importance of the predictors and thus provides the possibility of specifying the hierarchies of environmental factors influencing diatom assemblages. In addition to RF, generalized linear models (GLMs) can be used to quantitatively analyze and represent the variance of predictors by link functions. We conducted all the analysis in R (version 4.0.2; R Core Team, 2020). The RF was used to identify the hierarchy of variables with the package randomForestSRC (Ishwaran and Kogalur, 2021) and its *rfsrc* function. GLMs were applied after RF to analyze the interactions between diatom biotic indicators and environmental variables. Next, GLMs were used to get a regression model to detect the most significant environmental variables coupling with 10-folder cross validation (Kuhn 2020) and the best model was used to predict the diatom indices. The performance of the models was compared by root mean square error (RMSE), R squared, and mean absolute error (MAE). The analytical process was summarized as in Fig. 3.

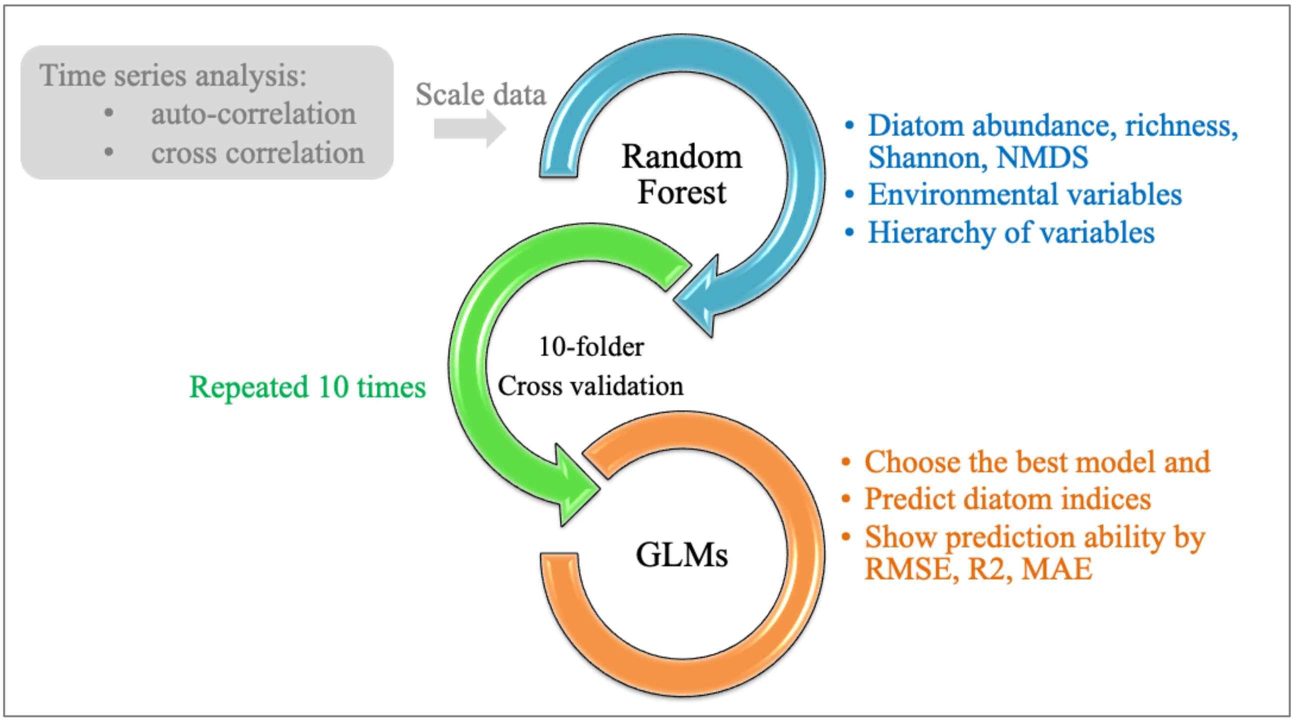

Fig. 3. Schematic workflow of the numerical analysis. NMDS: non-metric multidimensional scaling, GLMs: generalized linear models, RMSE: root mean square error, MAE: mean absolute error.

Results

Diatom succession

We recorded a total of 113 taxa from 45 genera of diatoms. The most dominant species (defined as relative abundance > 5%) were *Achnanthidium minutissimum* (Kütz.) Czarnecki (39.9%), *Navicula lanceolata* (Ag.) Ehr. (15.9%), and *Planothidium lanceolatum* (Bréb. ex Kütz.) Lange-Bertalot (6.4%). These three species dominated for almost the entire sampling year, especially *Achnanthidium minutissimum*. The highest diatom abundance was recorded as 5.96×10^6 cells/L in wintertime (November) and the lowest was 1.97×10^4 cells/L in spring (March). The averaged diatom abundance of the sampling period was 1.14×10^6 cells/L. The diatom abundance showed obvious seasonal variations throughout the year (Fig. 4). Diatom species richness showed relatively less variation, but the trend agreed with the variation of diatom abundance. The greatest diversity of the diatoms was 36 taxa per sample, the minimal was 7 and the mean value was 25 per sample. The Shannon index value ranged between 0.54 and 2.95. NMDS value ranged from -0.67 to 1.07 with an average of around 0 (scaled data). From a temporal point of view, the NMDS indicated a trend of more similarity for diatom community composition in spring and autumn than in summer and winter.

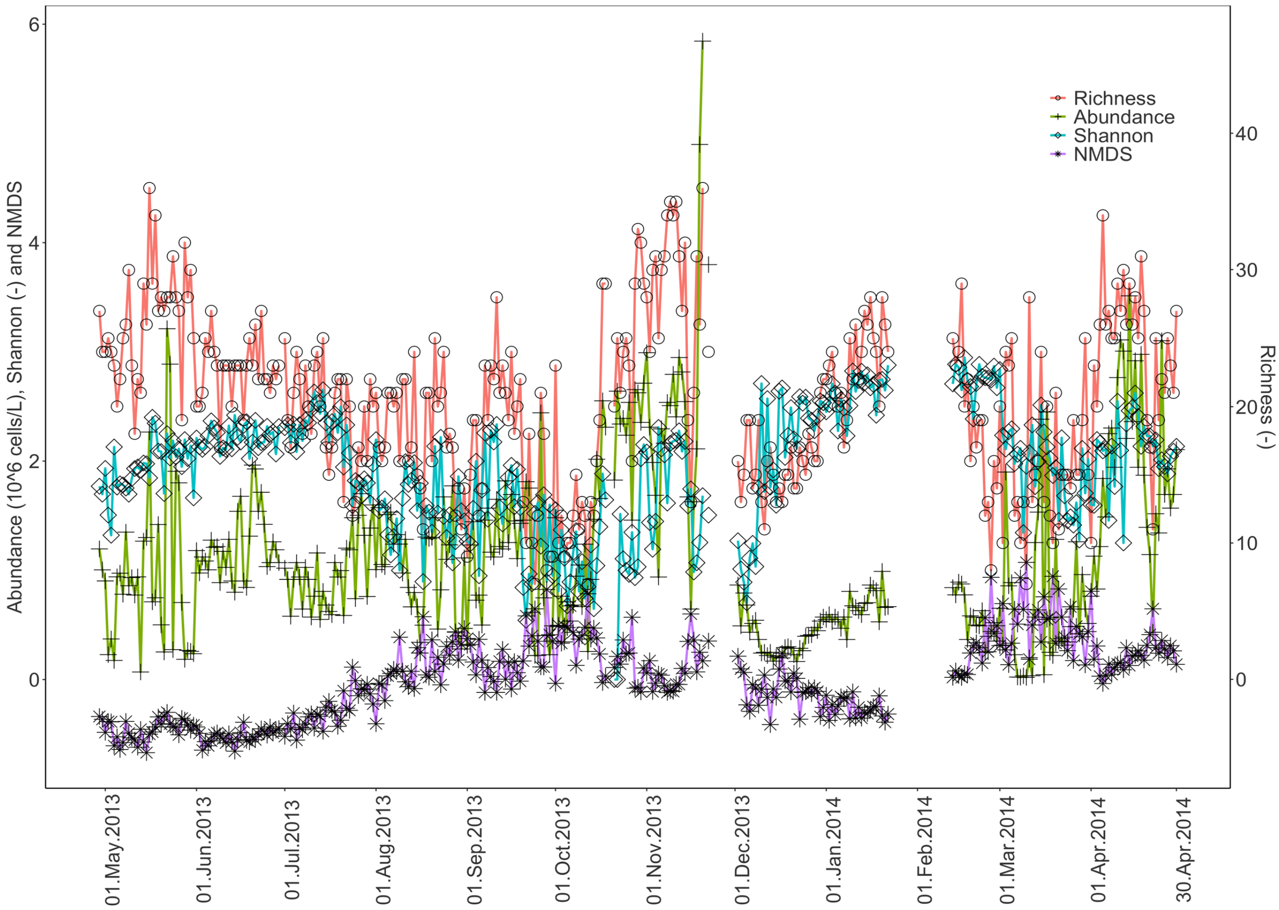

Fig. 4. Daily diatom indicators throughout the sampling period (April. 29, 2013- April.30, 2014): abundance (10^6 cells/L), Shannon index, non-metric multidimensional scaling (NMDS), and species richness; missing data are shown as blank.

Environmental variables

Eleven environmental variables were retained to run the following analyses. Their characteristics are shown in Table 2. Concentrations of NH_4-N, PO_4-P, pH, and Cl concentration remained stable. Except

Ca^{2+}, the metal ions of K^+ and Si^{2+} varied only in a narrow range. However, the hydrological parameters, specifically the API showed high variations throughout the sampling period.

Table 2. Summary of the selected environmental parameters used in the statistical analysis. WT: water temperature, API: antecedent precipitation index, BF: baseflow, PREC: precipitation.

Variables	**Unit**	**Minimum**	**Maximum**	**Median**	**Mean ± SD**
NH_4-N	mg/L	0.01	2.22	0.11	0.17 ± 0.20
PO_4-P	mg/L	0.02	0.50	0.10	0.12 ± 0.08
Cl^-	mg/L	16.26	43.72	29.49	29.88 ± 4.57
K^+	mg/L	3.24	7.66	4.87	4.95 ± 0.76
Ca^{2+}	mg/L	39.53	92.75	74.00	73.27 ± 7.67
Si^{2+}	mg/L	1.99	8.51	4.22	4.89 ± 1.68
WT	°C	1.20	16.70	10.60	10.35 ± 3.69
API	mm	0.88	122.54	18.91	30.23 ± 32.10
BF	m^3/s	0.01	0.37	0.11	0.12 ± 0.05
PREC	mm	0.00	40.30	0.10	1.74 ± 4.10
pH	-	7.20	8.20	7.60	7.60 ± 0.16

Time series analysis

We calculated the autocorrelation with a lag of 100 days but showed only the first 30 days in our figures (Fig. 5). As for diatom abundance, species richness, Shannon index and NMDS, the autocorrelation coefficients showed quite similar trend. There was a sharp decrease for all biotic parameters within 2 days. This indicated the quick changes in the abundance and richness of diatoms.

Shannon index and NMDS remained more stable than diatom abundance and species richness. Environmental variables showed larger differences. The AC of some variables like API and Si^{2+} remained high even after 15 days indicating they were quite stable; the sharp decline of the precipitation curve showed the random nature of precipitation.

Results of cross correlation showed the relationship between diatom indicators and environmental variables, respectively. Our results showed small cross correlation coefficients (Fig. 6) up to ± 0.4. The low coefficients indicated that there was no single linear relationship among them. However, the cross-correlation coefficients between Shannon index, NMDS and environmental parameters were slightly higher than diatom abundance and species richness.

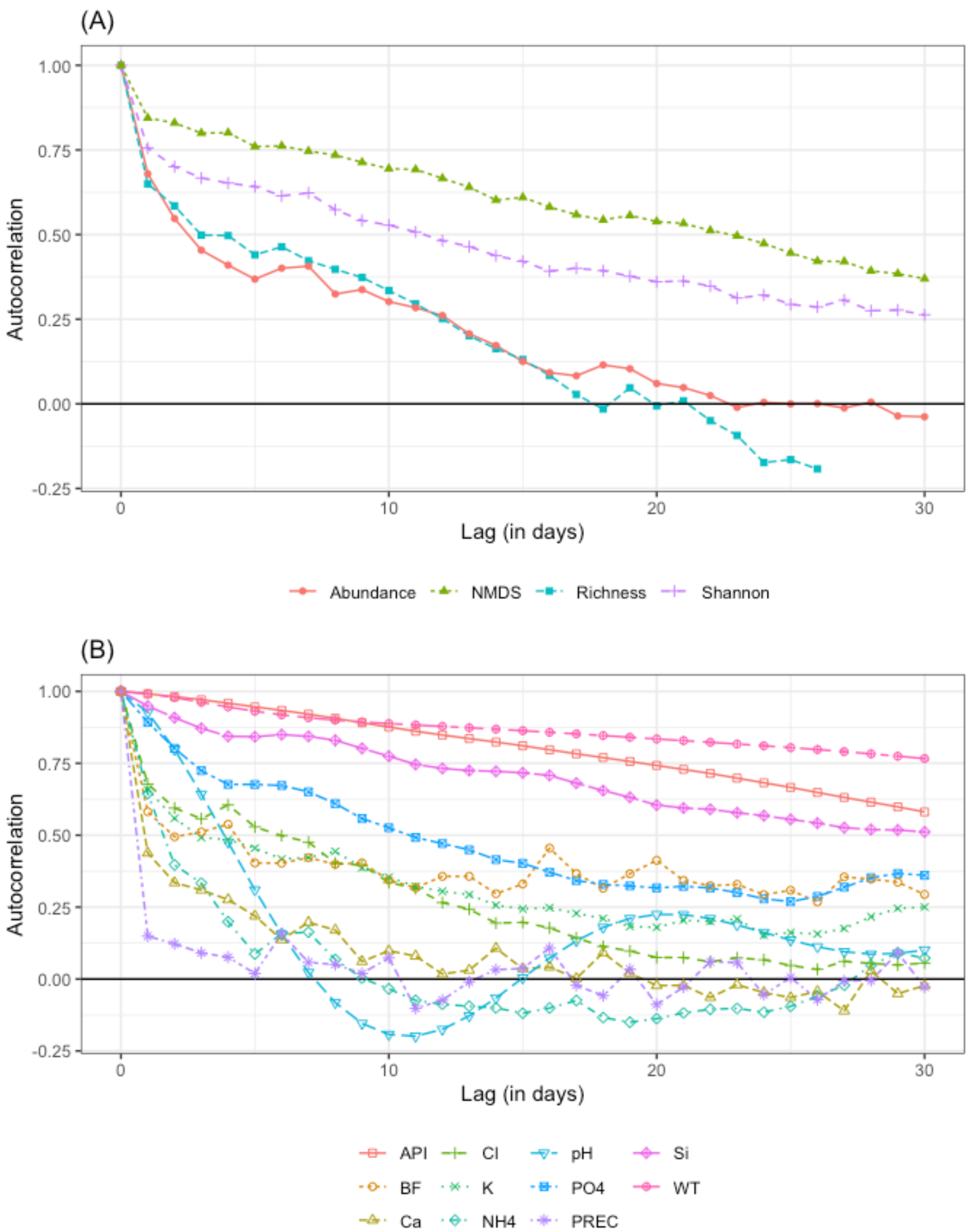

Fig. 5. (A) Autocorrelation of diatom abundance, species richness, Shannon index, and non-metric multidimensional scaling (NMDS) and (B) autocorrelation of selected environmental variables. WT: water temperature, API: antecedent precipitation index, BF: baseflow, PREC: precipitation.

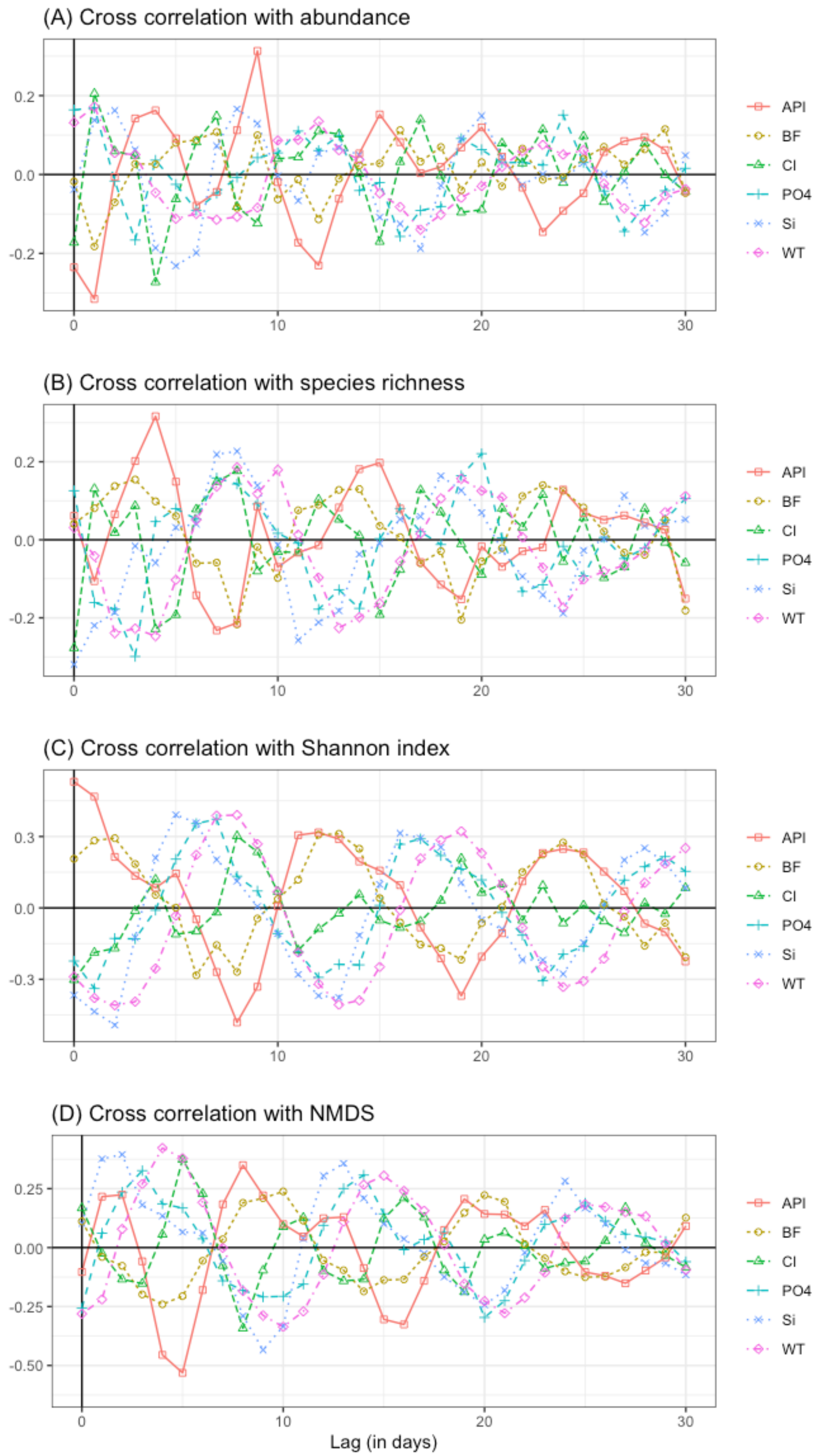
(A) Cross correlation with abundance
(B) Cross correlation with species richness
(C) Cross correlation with Shannon index
(D) Cross correlation with NMDS
API
BF
Cl
PO4
Si
WT
Lag (in days)

Fig. 6. Cross correlation coefficient between diatom (A) abundance, (B) species richness, (C) Shannon index, (D) non-metric multidimensional scaling (NMDS) and selected environmental variables, respectively. WT: water temperature, API: antecedent precipitation index, BF: baseflow.

Performance of models

The random forest (RF) models (Table 3) showed the variance of the biotic indicators (diatom abundance, species richness and Shannon index) explained by environmental variables and the variance of importance of the environmental variables. The results of the RF models showed satisfactory results. Namely 42% of the diatom abundance variance was explained by our selected environmental variables, while variances in diatom species richness, Shannon index, and NMDS were explained with 46%, 61%, and 69%, respectively. Shannon index and NMDS were better explained than traditional measurements. The out-of-bag error rates of the RF models were below 0.6%, thus indicated that the models were quite reliable.

In addition, the variable importance of the environmental variables was calculated (Fig. 7). Among all environmental variables, the antecedent precipitation index (API) showed the greatest importance. Additionally, the API, PO_4-P, Si^{2+} and K^+ were also important for diatom indicators. However, there were differences among the important variables related to different diatom indicators. For instance, the API and water temperature explained most of the variance of diatom abundance. The four most important environmental variables for species richness contributed less. Apart from other indicators, PO_4-P showed least importance of explaining variance of Shannon index. In contrast, PO_4-P was the second important variable for explaining the variance in NMDS. The statistically significant environmental variables were detected by the generalized linear models (Fig.7 shown by '*'). The performance of GLMs' prediction was showed in Table 3.

Table 3. Variance of biotic indicators explained by our selected environmental variables in random forest (first column), and the performance of the generalized linear models (GLMs) prediction in root mean square error (RMSE), R squared, and mean absolute error (MAE). NMDS: non-metric multidimensional scaling.

	Random Forest	GLMs		
	% Variance explained	Performance of model in prediction		
		RMSE	R^2	MAE
Diatom abundance	42	0.87	0.25	0.65
Species richness	46	0.88	0.25	0.71
Shannon index	61	0.75	0.44	0.60
NMDS	69	0.71	0.51	0.56

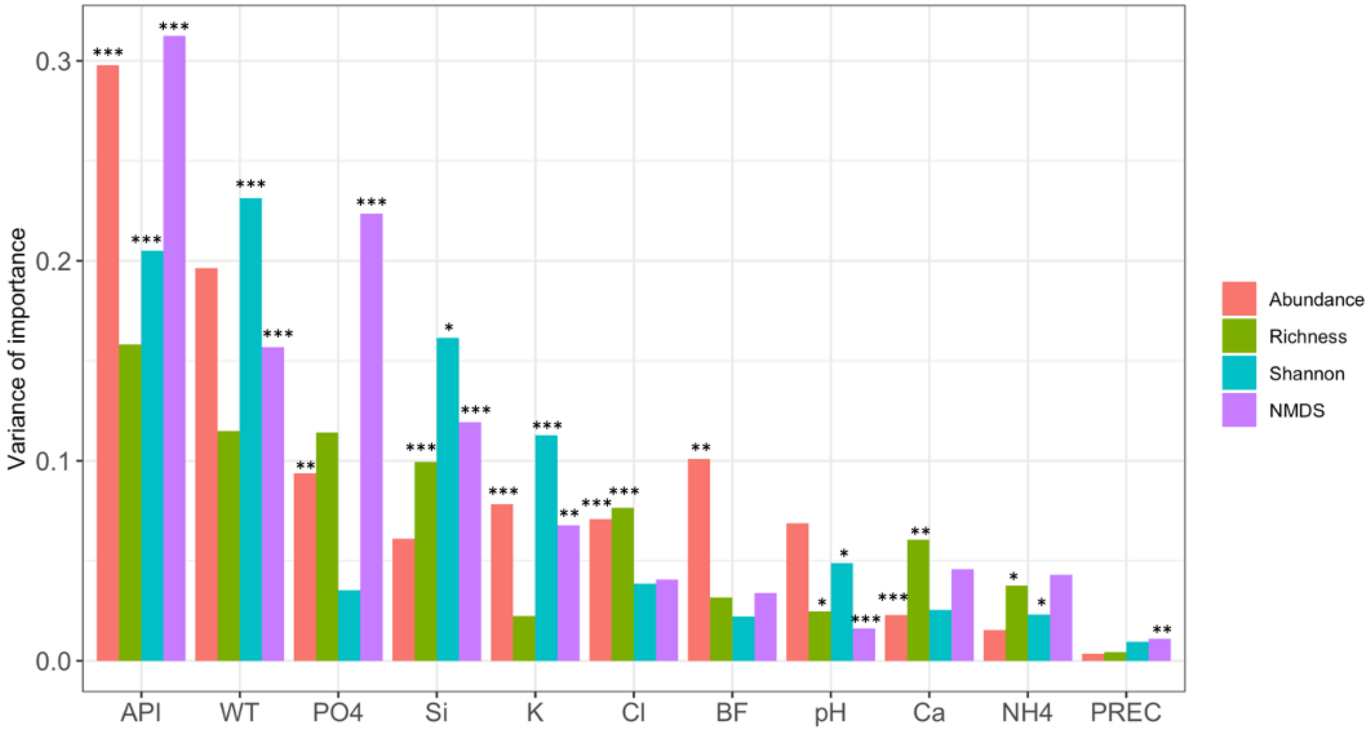

Fig. 7. The variable importance of environmental variables from random forest model for diatom indicators, and the significance of the environmental variables from the generalized linear models (*p* value): 0 < *** < 0.001 < ** < 0.01 < * < 0.05.

Discussion

Our random forest (RF) models showed good results in explaining variance of diatom indicators while our generalized linear models (GLMs) showed less satisfactory results in predicting the diatom indicators (Table 3). However, the performance of both explaining variance and predicting the Shannon index non-metric dimensional scaling (NMDS) was better than diatom abundance and species richness. We demonstrate the differences of the daily changing trend by time-series analysis, which shows the random nature of diatom abundance and species richness in the studied river (Fig. 4 (A)). This could be explained by the very short life spans of phytoplankton, where a single cell can exist less than one week and communities up to weeks (Morin et al., 2016). However, the Shannon index and NMDS showed higher autocorrelation, which indicates that the similarity of Shannon index and NMDS remain higher than the similarity of diatom abundance and species richness in a few consecutive days. This could be because the indices consider more of the biological and ecological characteristics of the species or so-called functional features (Mouchet et al., 2010; Passy 2007; Weithoff and Beisner, 2019; Wu et al., 2017), which generally can be shared by several species or even genera. This higher 'stability' also demonstrates better performance in the modeling prediction. In the predicting process of GLMs, the indices are better predicted than the abundance and richness.

Taking the critical need of professional knowledge and efforts for species taxonomy into account, the indices, especially trait-based indices (e.g., NMDS in our study) provide the fundament and more possibility of developing new taxonomy-free technologies (Arsenieff et al., 2020; Feio et al., 2020) to assess the status of aquatic systems.

The hierarchical analysis of RF highlighted the hydro-chemical parameters, e.g., API, WT, Si^{2+} and PO_4-P, in explaining the variance of the diatom indicators (Fig. 6). The autocorrelation analysis (Fig. 4 (B)) showed that the hydro-chemical variables remained more stable than the other variables in the study period, and also impacted the diatom community more than the others. The importance of API as a representative of hydrological variables in lowland rivers has been confirmed in previous studies (Wu et al., 2011a, 2011b; Sun et al., 2018). This leads to a higher focus on the hydrological effects on stream biota which have been previously neglected. In comparison with lakes, reservoirs and the other lentic aquatic habitats, a lowland river system is relatively more dynamic. That explains why hydrological conditions of lowland rivers play important roles in structuring biotic communities (Wu et al., 2017, Sun et al., 2018), higher wetness condition (higher API) increases diatom richness and diversity. This finding is in agreement with other research which reveals that drought is a strong negative stressor for diatom richness in lowland streams (B-Béres et al., 2019). A lowland river is a part of an open system, in which the impact from pre-riverine human activities cannot be neglected. API as an integrated proxy is based on the precipitation and gathers the surface runoff from the catchment to the stream and finally to the downstream catchment to the outlet. The present data was recorded at the outlet of the whole catchment, it includes all upstream impacts. Diatom diversity and composition are reported to be highly dependent on the enrichment of nitrogen (Kafouris et al., 2019). However, in the current study, nitrogen is not as important as phosphate. Diatom richness has been reported being controlled by water temperature and pH (Jyrkänkallio-Mikkola et al., 2018), and it could be higher in colder climates and lower water temperatures (Pajunen et al., 2016). The impact of water pH on diatom species richness was also revealed by global and continental scale studies (Soininen et al., 2016; Passy, 2010). Although the pH is one of the statistically significant variables in our study, the diatom variance explained by it is limited. This could be explained by the narrow range of pH (7.2 – 8.2) in the study. Water temperature is one of the most important controlling variables, however, it shows a conflicting effect regarding species richness, with the lowest richness being recorded in October (late autumn) but greatest richness in May, November, and April, compared with Jyrkänkallio-Mikkola et al. (2018).

In this study, we focused on the local environmental variables which can be seen as direct variables. Indirect effects from a larger spatial scale have not been included yet, for instance, land use patterns and global climate changes, that have effects on both hydro-morphology and physico-chemistry which leads to an effect on biological conditions (Villeneuve et al., 2018), and on the functional composition of phytoplankton communities (Qu et al., 2018, 2019).

Conclusion

Our findings confirm our hypothesis that diatom community variance is impacted by the hydro-chemical variables. The random forest modeling shows satisfactory results by explaining diatom indicators with a variance percentage ranged between 42% to 69%. The hierarchical analysis highlighted antecedent precipitation index (API) as the controlling hydrological variable, while water temperature, Si^{2+} and PO_4-P, as the main chemical controlling factors in our study area. Hydrological variables' effects on riverine phytoplankton should draw more attention in the future practical biomonitoring purposes. The generalized linear models performed a better prediction for Shannon index and non-metric multidimensional scaling than diatom abundance and species richness, which confirms that both indices perform adequately in explaining the relationship between stream biota and its environment. Our study shows and highlights the important hydro-chemical factors in the agricultural rivers, which could contribute to the further understanding of predicting diatom community development, and could be implemented in the future water management protocol.

Acknowledgements

We thank Mrs. Monika Westphal, Mrs. Bettina Hollmann and other colleagues for their supports during the field campaigns and lab processing.

Funding

This study was supported financially by DFG grants (FO 301/15-1, FO 301/15-2, WU 749/1-1, WU 749/1-2) and China Scholarship Council (CSC) (X. Sun).

References

Andrus, J. M., Winter, D., Scanlan, M., Sullivan, S., Bollman, W., Waggoner, J. B., ... & Brain, R. A. (2013). Seasonal synchronicity of algal assemblages in three Midwestern agricultural streams having varying concentrations of atrazine, nutrients, and sediment. *Science of the total environment*, *458*, 125-139.

Arsenieff, L., Le Gall, F., Rigaut-Jalabert, F., Mahé, F., Sarno, D., Gouhier, L., ... & Simon, N. (2020). Diversity and dynamics of relevant nanoplanktonic diatoms in the Western English Channel. *The ISME journal*, *14*(8), 1966-1981.

Babitsch, D., Berger, E., & Sundermann, A. (2021). Linking environmental with biological data: Low sampling frequencies of chemical pollutants and nutrients in rivers reduce the reliability of model results. *Science of The Total Environment*, *772*, 145498.

B-Béres, V., Lukács, Á., Török, P., Kókai, Z., Novák, Z., Enikő, T., ... & Bácsi, I. (2016). Combined eco-morphological functional groups are reliable indicators of colonisation processes of benthic diatom assemblages in a lowland stream. *Ecological indicators*, 64, 31-38.

B-Béres V., Tóthmérész B., Bácsi I., Borics G., Abonyi A., Tapolczai K., Rimet F., Bouchez A., Várbíró G., Török P. (2019): Autumn drought drives functional diversity of benthic diatom assemblages of continental streams. *Advances in Water Resources*, 126, 129-136.

Bey, M. Y., & Ector, L. (2013). Atlas des diatomees des cours d'eau de la region Rhone-Alpes (Tome 1–Tome 5). Caluire, France.

Bray J. R. & Curtis J. T. (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs.* 27, 325- 49.

Breiman, L. (2001). Random forests. *Machine Learning*. 45 (1), 5–32.

Campos, C. A., Kennard, M. J., & Júnior, J. F. G. (2021). Diatom and Macroinvertebrate assemblages to inform management of Brazilian savanna's watersheds. *Ecological Indicators*, *128*, 107834.

Christophersen, N. & Hooper, R.P. (1992). Multivariate analysis of stream water chemical data: the use of principal components analysis for the end-member mixing problem. *Water Resour*ces *Research* 28 (1), 99–107.

Cornejo, A., Tonin, A. M., Checa, B., Tuñon, A. R., Pérez, D., Coronado, E., ... & Boyero, L. (2019). Effects of multiple stressors associated with agriculture on stream macroinvertebrate communities in a tropical catchment. *PloS one*, *14*(8), e0220528.

Cotiyane-Pondo, P., Bornman, T. G., Dąbek, P., Witkowski, A., & Smit, A. J. (2020). Austral winter marine epilithic diatoms: Community composition and distribution on intertidal rocky substrate around the coast of South Africa. *Estuarine, Coastal and Shelf Science*, *242*, 106837.

Cottenie, K. (2005). Integrating environmental and spatial processes in ecological community dynamics. *Ecology Letters,* 8, 1175–1182.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawer, J.J. (2007). Random forest for classification in ecology. *Ecology*, *88 (11),* 2783–2792

Derot, J., Jamoneau, A., Teichert, N., Rosebery, J., Morin, S., & Laplace-Treyture, C. (2020). Response of phytoplankton traits to environmental variables in French lakes: New perspectives for bioindication. *Ecological Indicators*, *108*, 105659.

Di Battista, T., Fortuna, F., & Maturo, F. (2016). Environmental monitoring through functional biodiversity tools. *Ecological Indicators*, *60*, 237-247.

Evans, K. M., Wortley, A. H., Simpson, G. E., Chepurnov, V. A., & Mann, D. G. (2008). A molecular systematic approach to explore diversity within the Sellaphora pupula species complex (Bacillariophyta) 1. *Journal of Phycology*, *44*(1), 215-231.

Fedora, M.A. & Beschta, R.L. (1989). Storm runoff simulation using an antecedent precipitation index (API) model. *J. Hydrol. 112 (1–2),* 121–133.

Fohrer, N. & Schmalz, B. (2012). The UNESCO ecohydrology demonstration site Kielstau catchment – Sustainable water resources management and education in rural areas | Das UNESCO Ökohydrologie-Referenzprojekt Kielstau-Einzugsgebiet – Nachhaltiges Wasserressourcenmanagement und Ausbildung. *Hydrologie Und Wasserbewirtschaftung* 56 (4).

Franklin, J. (2009). Mapping Species Distribution. Spatial Inferences and Prediction, 3rd ed. Cambridge University Press, Cambridge.

Fukai, Y., Abe, Y., Matsuno, K., & Yamaguchi, A. (2020). Spatial changes in the summer diatom community of the northern Bering Sea in 2017 and 2018. *Deep Sea Research Part II: Topical Studies in Oceanography*, *181*, 104903.

Guignard, M. S., Leitch, A. R., Acquisti, C., Eizaguirre, C., Elser, J. J., Hessen, D. O., ... & Leitch, I. J. (2017). Impacts of nitrogen and phosphorus: from genomes to natural ecosystems and agriculture. *Frontiers in Ecology and Evolution*, *5*, 70.

Guo, K., Wu, N., Wang, C., Yang, D., He, Y., Luo, J., et al. (2019). Trait dependent roles of environmental factors, spatial processes and grazing pressure on lake phytoplankton metacommunity. *Ecological Indicators*. 103, 312–320.

Hallegraeff, G. M. "Harmful algal blooms: a global overview." *Manual on harmful marine microalgae* 33 (2003): 1-22.

Hill, B. H., Herlihy, A. T., Kaufmann, P. R., Stevenson, R. J., McCormick, F. H., & Johnson, C. B. (2000). Use of periphyton assemblage data as an index of biotic integrity. *Journal of the North American Benthological Society, 19(1),* 50-67.

Hill, B. H., Herlihy, A. T., Kaufmann, P. R., DeCelles, S. J., & Vander Borgh, M. A. (2003). Assessment of streams of the eastern United States using a periphyton index of biotic integrity. *Ecological Indicators*, 2(4), 325-338.

Indermuehle, N., Oertli, B., Biggs, J., Céréghino, R., Grillas, P., Hull, A., et al. (2008). Pond conservation in Europe: the European pond conservation network (EPCN). *SIL Proceedings*, 1922–2010 30, 446–448.

Interlandi, S. J., & Kilham, S. S. (2001). Limiting resources and the regulation of diversity in phytoplankton communities. *Ecology*, *82*(5), 1270-1282.

Ishwaran H. and Kogalur U.B. (2021). Fast Unified Random Forests for Survival, Regression,

and Classification (RF-SRC), R package version 2.11.0.

Jyrkänkallio-Mikkola, J., Siljander, M., Heikinheimo, V., Pellikka, P., & Soininen, J. (2018). Tropical stream diatom communities–The importance of headwater streams for regional diversity. *Ecological Indicators*, 95, 183-193.

Kafouris, S., Smeti, E., Spatharis, S., Tsirtsis, G., Economou-Amilli, A., & Danielidis, D. B. (2019).

Nitrogen as the main driver of benthic diatom composition and diversity in oligotrophic coastal systems. *Science of the Total Environment*, 694, 133773.

Kelly, M., Juggins, S., Guthrie, R., Pritchard, S., Jamieson, J., Rippey, B., et al. (2008). Assessment of ecological status in U.K. rivers using diatoms. *Freshwater Biology*. 53 (2), 403–422.

Kim, H. G., Hong, S., Kim, D. K., & Joo, G. J. (2020). Drivers shaping episodic and gradual changes in phytoplankton community succession: Taxonomic versus functional groups. *Science of The Total Environment*, 138940.

Kuhn, M. (2020). caret: Classification and Regression Training. R package version 6.0-86.

https://CRAN.R-project.org/package=caret

Laiolo, P., Pato, J., & Obeso, J. R. (2018). Ecological and evolutionary drivers of the elevational gradient of diversity. *Ecology letters, 21*(7), 1022-1032.

Lange-Bertalot, H., 2000a, 2000b, 2005, 2007. Iconographia Diatomologica. Annotated Diatom Micrographs, vol. 7, 9, 15, 18, Koenigstein, Germany: Koeltz Scientific Books.

Lange, K., Liess, A., Piggott, J.J., Townsend, C.R., Matthaei, C.D. (2011). Light, nutrients and grazing interact to determine stream diatom community composition and functional group structure. *Freshwater Biology*. 56 (2), 264–278.

Law, R.J., Elliott, J.A., Thackeray, S.J. (2014). Do functional or morphological classifications explain stream phytobenthic community assemblages? *Diatom Research*. 29 (4).

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., & Loreau, M., Gonzalez, A. (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecology letters, 7*:601–613.

Lewis Jr, W. M. (1978). Analysis of succession in a tropical phytoplankton community and a new measure of succession rate. *The American Naturalist*, *112*(984), 401-414.

Lomas, M. W., Bonachela, J. A., Levin, S. A., & Martiny, A. C. (2014). Impact of ocean phytoplankton diversity on phosphate uptake. *Proceedings of the National Academy of Sciences*, 111(49), 17540-17545.

Mann, D.G. (1999). The species concept in diatoms. *Phycologia* 38 (6), 437–495.

Mann, D.G., McDonald, S.M., Bayer, M.M., Droop, S.J.M., Chepurnov, V.A., Loke, R.E., et al. (2004). The Sellaphora pupula species complex (Bacillariophyceae): morpho-metric analysis, ultrastructure and mating data provide evidence for five new species. *Phycologia* 43 (4), 459–482.

Mouchet, M.A., Villéger, S., Mason, N.W., Mouillot, D. (2010). Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology*. 24, 867–876.

Morin, S., Gómez, N., Tornés, E., Licursi, M., & Rosebery, J. (2016). Benthic diatom monitoring and assessment of freshwater environments: Standard methods and future challenges. In: Romaní, A.M., Guasch, H., Balaguer, M.D. (Eds.), Aquatic Biofilms: Ecology, Water Quality and Water Treatment. Caister Academic Press, U.K, pp. 111–124.

Pajunen, V., Luoto, M., & Soininen, J. (2016). Stream diatom assemblages as predictors of climate. *Freshwater Biology*, 61(6), 876-886.

Park, Y., Cho, K. H., Park, J., Cha, S. M., & Kim, J. H. (2015). Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of the Total Environment*, *502*, 31-41.

Passy, S. I. (2007). Diatom ecological guilds display distinct and predictable behavior along nutrient and disturbance gradients in running waters. *Aquatic botany*, *86*(2), 171-178.

Passy, S. I. (2010). A distinct latitudinal gradient of diatom diversity is linked to resource supply. *Ecology, 91*,36–41.

Passy, S. I., Bottin, M., Soininen, J., & Hillebrand, H. (2017). Environmental filtering and taxonomic relatedness underlie the species richness–evenness relationship. *Hydrobiologia*, *787*(1), 243-253.

Ptacnik, R., Solimini, A. G., Andersen, T., Tamminen, T., Brettum, P., Lepistö, L., ... & Rekolainen, S. (2008). Diversity predicts stability and resource use efficiency in natural phytoplankton communities. *Proceedings of the National Academy of Sciences, 105*(13), 5134-5138.

Qu, Y., Wu, N., Makarevičiūtė, K., Guse, B., & Fohrer, N. (2018). Effects of land-use pattern and physiochemical conditions on phytoplankton communities in a German lowland catchment. *Fundamental and Applied Limnology/Archiv für Hydrobiologie*, 191(3): 175-187.

Qu, Y., Wu, N., Guse, B., Makareviciute, K., Sun, X., & Fohrer, N. (2019). Riverine phytoplankton functional groups response to multiple stressors variously depending on hydrological periods. *Ecological Indicators*, 101, 41-49.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rimet, F., Bouchez, A. (2011). Use of diatom life-forms and ecological guilds to assess pesticide contamination in rivers: lotic mesocosm approaches. *Ecological Indicators*. 11 (2), 489–499.

Reynolds, C. S. (1988). The concept of ecological succession applied to seasonal periodicity of freshwater phytoplankton: With 4 figures in the text. *Internationale Vereinigung für theoretische und angewandte Limnologie: Verhandlungen*, *23*(2), 683-691.

Round, F.E., Crawford, R.M., Mann, D.G., (1990). The Diatoms: Biology and Morphology of the genera. Cambridge University Press, Cambridge, UK.

Sabater, S., Artigas, J., Durán, C., Pardos, M., Romaní, A. M., Tornés, E., & Ylla, I. (2008). Longitudinal development of chlorophyll and phytoplankton assemblages in a regulated large river (the Ebro River). *Science of the total environment*, 404(1), 196-206.

Santos, A. M., Cianciaruso, M. V., & De Marco Jr, P. (2016). Global patterns of functional diversity and assemblage structure of island parasitoid faunas. *Global Ecology and Biogeography*, *25*(7), 869-879.

Serediak, N.A., Prepas, E.E., Putz, G.J. (2014). Eutrophication of freshwater systems. In: Holland, H.D., Turekian, K.K. (Eds.), Treatise on Geochemistry, Second edition Elsevier, Oxford, pp. 305–323.

Shaw, E.M. (1994). Hydrology in Practice, third ed. Taylor & Francis.

Simonsen, R. (1987). Atlas and Catalogue of the Diatom Types of Frederich Hustedt, 3 volumes. J. Cramer, Gebr. Borntraeger, Berlin, Stuttgart.

Soininen, J., & Luoto, M. (2012). Is catchment productivity a useful predictor of taxa richness in lake plankton communities? *Ecological Applications*, *22*(2), 624-633.

Soininen, J., Jamoneau, A., Rosebery, J., & Passy, S. I. (2016). Global patterns and drivers of species and trait composition in diatoms. *Global ecology and biogeography*, *25*(8), 940-950.

Sommer, U., Gliwicz, Z. M., Lampert, W., & Duncan, A. (1986). The PEG-model of seasonal succession of planktonic events in fresh waters. *Arch. Hydrobiol*, *106*(4), 433-471.

Sun, X., Wu, N., Faber, C., & Fohrer, N. (2018). Effects of hydrological variables on structuring morphological trait (cell size) of diatom community in a lowland river. *Ecological indicators*, *94*, 207-217.

Stanish, L. F., Kohler, T. J., Esposito, R. M., Simmons, B. L., Nielsen, U. N., Wall, D. H., ... & McKnight, D. M. (2012). Extreme streams: flow intermittency as a control on diatom communities in meltwater streams in the McMurdo Dry Valleys, Antarctica. *Canadian Journal of Fisheries and Aquatic Sciences*, 69(8), 1405-1419.

Stevenson, R. J., Pan, Y., & Van Dam, H. (1999). Assessing environmental conditions in rivers and streams with diatoms. *The diatoms: applications for the environmental and earth sciences, 1*(4).

Villeneuve, B., Piffady, J., Valette, L., Souchon, Y., & Usseglio-Polatera, P. (2018). Direct and indirect effects of multiple stressors on stream invertebrates across watershed, reach and site scales: A structural equation modelling better informing on hydromorphological impacts. *Science of the Total Environment*, *612*, 660-671.

Wang, Q., Yang, X., Anderson, N. J., & Ji, J. (2015). Diatom seasonality and sedimentation in a subtropical alpine lake (Lugu Hu, Yunnan-Sichuan, Southwest China). *Arctic, Antarctic, and Alpine Research*, *47*(3), 461-472.

Weaver, W., & Shannon, C. E. (1963). The Mathematical Theory of Communication. Illinois University Press, Urbana.

Weithoff, G., Beisner, B.E. (2019). Measures and approaches in trait-based phytoplankton community ecology–from freshwater to marine ecosystems. *Frontiers in Marine Science.* 6, 40.

Wijewardene, L., Wu, N., Qu, Y., Guo, K., Messyasz, B., Lorenz, S., ... & Fohrer, N. (2021). Influences of pesticides, nutrients, and local environmental variables on phytoplankton communities in lentic small water bodies in a German lowland agricultural area. *Science of The Total Environment*, 146481.

Winder, M., & Hunter, D. A. (2008). Temporal organization of phytoplankton communities linked to physical forcing. *Oecologia*, *156*(1), 179-192.

Woelfel, J., Schumann, R., Adler, S., Hübener, T., & Karsten, U. (2007). Diatoms inhabiting a wind flat of the Baltic Sea: species diversity and seasonal succession. *Estuarine, Coastal and Shelf Science*, *75*(3), 296-307.

Wu, N., Dong, X., Liu, Y., Wang, C., Baattrup-Pedersen, A., & Riis, T. (2017). Using river microalgae as indicators for freshwater biomonitoring: Review of published research and future directions. *Ecological Indicators*, *81*, 124-131.

Wu, N., Faber, C., Sun, X., Qu, Y., Wang, C., Ivetic, S., ... & Fohrer, N. (2016). Importance of sampling frequency when collecting diatoms. *Scientific reports*, *6*(1), 1-9.

Wu, N., Schmalz, B., & Fohrer, N. (2011a). A comparison of phytoplankton assemblages generated by two sampling protocols in a German lowland catchment. In *Annales de Limnologie-International Journal of Limnology* (Vol. 47, No. 4, pp. 313-323). EDP Sciences.

Wu, N., Schmalz, B., & Fohrer, N. (2011b). Distribution of phytoplankton in a German lowland river in relation to environmental factors. *Journal of Plankton Research*, *33*(5), 807-820.

Wu, N., Thodsen, H., Andersen, H.E., Tornbjerg, H., Baattrup-Pedersen, A., Riis, T., (2019). Flow regimes filter species traits of benthic diatom communities and modify the functional features of lowland streams-a nationwide scale study. *Science of The Total Environment*. 651, 357–366.

Wu, N., Tang, T., Zhou, S., Jia, X., Li, D., Liu, R., & Cai, Q. (2009). Changes in benthic algal communities following construction of a run-of-river dam. *Journal of the North American Benthological Society, 28*(1), 69-79.

Zalack, J. T., Smucker, N. J., & Vis, M. L. (2010). Development of a diatom index of biotic integrity for acid mine drainage impacted streams. *Ecological Indicators*, *10*(2), 287-295.

Zhou, Y., Zhang, Y., Liang, T., & Wang, L. (2019). Shifting of phytoplankton assemblages in a regulated Chinese river basin after streamflow and water quality changes. *Science of The Total Environment*, *654*, 948-959.

Zwart, J.A., Solomon, C.T., Jones, S.E. (2015). Phytoplankton traits predict ecosystem function in a global set of lakes. *Ecology*. 96, 2257–2264.