Identification of ecological thresholds from variations in phytoplankton communities among lakes: contribution to the definition of environmental standards

V. Roubeix, P.A. Danis, T. Feret, J.M. Baudoin

Author-produced version of the article published in : Environmental Monitoring and Assessment ,2016, p.188-246

Identification of ecological thresholds from variations in phytoplankton communities among lakes: contribution to the definition of environmental standards

Vincent Roubeix[1], Pierre-Alain Danis[2], Thibaut Feret[3] and Jean-Marc Baudoin[2]

[1]Irstea, UR RECOVER, Pôle Onema-Irstea hydroécologie plans d'eau, centre d'Aix-en-Provence, 3275 route Cézanne, F-13612 Le Tholonet Aix-en-Provence, France

[2]Onema, Pôle Onema-Irstea hydroécologie plans d'eau, F-13612 Le Tholonet Aix-en-Provence, France

[3]Irstea, UR EABX, centre de Bordeaux, F-33612 Gazinet Cestas, France

Corresponding author: vincent.roubeix@irstea.fr +33 4 42 66 79 35

Abstract

In aquatic ecosystems, the identification of ecological thresholds may be useful for managers as it can help to diagnose ecosystem health and to identify key levers to enable the success of preservation and restoration measures. A recent statistical method, gradient forest, based on random forests, was used to detect thresholds of phytoplankton community change in lakes along different environmental gradients. It performs exploratory analyses of multivariate biological and environmental data to estimate the location and importance of community thresholds along gradients. The method was applied to a dataset of 224 French lakes which were characterized by 29 environmental variables and the mean abundances of 196 phytoplankton species. Results showed the high importance of geographic variables for the prediction of species abundances at the scale of the study. A second analysis was performed on a subset of lakes defined by geographic thresholds and presenting a higher biological homogeneity. Community thresholds were identified for the most important physico-chemical variables including water transparency, total phosphorus, ammonia, nitrates and dissolved organic carbon. Gradient forest appeared as a powerful method at a first exploratory step, to detect ecological thresholds at large spatial scale. The thresholds that were identified here must be reinforced by the separate analysis of other aquatic communities and may be used then to set protective environmental standards after consideration of natural variability among lakes.

Keywords : community thresholds; environmental standards; gradient forest; lakes; phytoplankton

Introduction

The response of an ecosystem to a gradual change in environmental conditions may be smooth, in proportion to the change, or abrupt if a critical level is reached, with potentially an hysteresis effect when the change is reversed (Scheffer and Carpenter 2003). The hypothesis of multiple stable states of ecosystems and communities (May 1977) argues in favor of non-linear responses with ecological thresholds. This vision of ecosystems dynamics is supported by simple mathematical models and empirical observations of sudden shifts from one state to another. In the cases of shallow lakes, the existence of several states characterized by phytoplankton and macrophyte species has been demonstrated, with transitions possibly due to changes in nutrient concentrations (Scheffer et al. 1997; Scheffer et al. 2003). Following the idea of discontinuous variations triggered by controlling variables, numerous scientific studies have focused in the last 15 years on the research of thresholds from the response of aquatic communities to environmental parameters, such as water transparency, phosphorus and nitrogen concentrations or pH (Graham et al. 2004; Holt et al. 2003; Richardson et al. 2007). The identification of thresholds in the gradients of human-influenced variables is particularly interesting for environmental management, in order to anticipate dramatic ecological changes in case of degradation or to set targets for restoration (Chambers et al. 2012a; Soranno et al. 2008; Vollenweider 1975). In the European Water Framework Directive (WFD) (European Commission 2000) concerning all types of water masses including lakes, member states have to set environmental standards for general physico-chemical parameters which support the ecological assessment made from biological quality indices. These standards may be simply derived from the distribution of measured values by the use of quantiles or by the division of observed gradients (e.g. in Chambers et al. 2012a), without any link to biological elements. However, it is more relevant for lake management to propose standards based on ecological thresholds, through the analysis of community responses to water quality gradients (Poikane et al. 2014; Solheim et al. 2008; Penning et al. 2008; Free et al. 2006).

An ecological threshold may be generally defined as a critical point where moderate variations of an environmental parameter produce large responses in ecosystem state (Groffman et al. 2006). Thresholds can be *a posteriori* identified from time-series analysis (Andersen et al. 2009), but they can be also detected from the spatial comparisons of systems in environmental gradients (Catalan et al. 2009; Holt et al. 2003; Soranno et al. 2008). Amongst the response variables for lakes, those related to the composition of aquatic communities can be used to demonstrate the existence of thresholds.

Ecological thresholds imply non-linear relationships between environmental drivers and some biological variables. A large set of statistical methods has been proposed for threshold identification in environmental gradients (Brenden et al. 2008; Dodds et al. 2010; Qian 2014) with applications in various ecological contexts, mostly for streams (Black et al. 2011; Chambers et al. 2012b; Evans-White et al. 2009; Richardson et al. 2007; Smith and Tran 2010). The different approaches found in the literature can be classified according to their exploratory character and *a priori* knowledge of the relationships between environment and communities. Most of them focus on one environmental factor which is known to influence communities directly in water (e.g. phosphorus, pH) or from the watershed (e.g. urbanization, agriculture). Concerning the biological variables, some studies use the abundance of one sensitive species or a selection of species according to their indicator value (e.g. King et al. 2011; Richardson et al. 2007); in others the response variables are aggregate or synthetic metrics (e.g. Black et al. 2011; Evans-White et al. 2009). In some cases, the search for thresholds is carried out separately in different groups of sites, making the hypothesis that some typological features may influence the results (Utz et al. 2009).

In this study, an exploratory approach was followed to find thresholds through an analysis of multivariate biological and environmental data, minimizing preliminary variable selection. The phytoplankton communities of various types of lakes from different regions of France were considered together. Phytoplankton can be found in every lake and might be particularly sensitive to the commonly measured water physico-chemical parameters. It is therefore well-suited for an analysis of thresholds in environmental gradients on a large spatial scale. The abundances of all phytoplankton species were taken into account, assuming that every species can be an indicator of an environmental parameter. Thus, the thresholds that were sought out here, rather refer to the concept of community threshold that can be defined as a zone in an environmental gradient in which the rate of change in community structure is enhanced relative to the rest of the gradient, as a result of sharp increases or decreases in the abundances of several species (Baker and King 2010; Catalan et al. 2009).

With the method used in the present study, called gradient forest (Ellis et al. 2012; Pitcher et al. 2012), it is possible to include in the analysis a large number of potentially interacting environmental factors, without any assumption on their real influence on the biological communities. Phytoplankton data could therefore be simultaneously related either to human-influenced or natural factors. Gradient forest is an extension to the community level of random forest (Breiman 2001) which estimates variable importance for one species and allows to detect at which levels in an environmental gradient the main changes in abundance occur. By aggregating the results of random forests for all species reported in a survey, gradient forest orders environmental variables by

importance for communities and indicates along gradients the compositional turnover (cumulative rate of change for all species), whose potential peaks may indicate community thresholds.

The objectives of this work were to apply a recent statistical method, gradient forest, to a broad scale phytoplankton database, in order to identify potential thresholds in the gradients of the most important physico-chemical parameters, especially those related to eutrophication. The final goal is to give ecological justifications for the setting of environmental standards used in lake management.

Materials and methods

Database

Environmental description

The analysis was initially performed on 224 lakes located in France (Fig. 1). All types of lentic water masses were included (natural lakes, reservoirs, aquaculture ponds, gravel pits) provided that lake surface area was above 0.5 km². Data were provided by French water basin agencies that organize the monitoring of French lakes according to standard protocols and as required by the European Water Framework Directive (European Commission 2000). The 29 environmental variables which were taken into account in the analysis are presented in Table 1. This selection results from a trade-off between sufficient ecological description and the maximization of lake-sample size given the incompleteness of the national data set. The diversity of lakes was characterized by some basic geographic or physical variables such as latitude, longitude, altitude and lake maximal depth. Field physico-chemical measurements were done in each lake at the point of greatest depth. Water transparency was assessed using a Secchi disk (NF EN ISO 7027) and the depth of the euphotic zone (Zeu) was assumed to be 2.5 times Secchi depth. Average values of temperature, pH, conductivity and O_2 saturation were derived from the integration of vertical profiles over Zeu (1 m-measurement interval). Nutrients, dissolved organic carbon and alkalinity analyses were performed in an integrated water sample collected in the euphotic zone, following national and European standards (NF EN ISO 10304, NF EN ISO 6878, NF EN 1484, NF EN ISO 9963). Nutrient concentrations below the quantification limit (LQ) were given an arbitrary value of LQ/2 in order to keep the most oligotrophic lakes in the analysis.

Phytoplankton communities

Phytoplankton was analyzed from the integrated water sample used for nutrients (Laplace-Treyture et al. 2009). The composition of communities was determined at a specific level under inverted microscope according to Utermöhl's method (NF EN 15204). Cell counts of each species were

converted into biovolumes (in $mm^3.L^{-1}$) using standard specific cell volumes. The occurrence of more than 600 species was reported in the lakes. Rare species which were absent in more than 95 % of sites were excluded from the analysis, so that the final species number was 196.

Data temporal aggregation

Data were organized in campaigns along annual cycles between 2006 and 2012. There were 4 campaigns per annual cycle and 3 of them during the most productive months (from April to October). Physico-chemical measurements and phytoplankton samples were realized during the same campaigns. For 80 % of the lakes, data were available only for one annual cycle, whereas the rest of the lakes had data for 2 (14 %), 3 (4.5 %) or 4 (1 %) annual cycles. As the approach in this study is based on an inter-lake comparison, seasonal and inter-annual variations were not taken into account in the analysis. Thus, arithmetic means of phytoplankton biovolumes from all campaigns in each lake were considered. All environmental data except longitude, latitude, water temperature, pH and O_2 saturation level were log_{10}-transformed to facilitate graphical visualization of the results. Then, annual medians and maximal or minimal values (when relevant) of water physico-chemical variables were computed to describe abiotic conditions for phytoplankton in the lakes (Table 1). According to Ellis et al (2012), all variables were included in the GF analysis without selection, even if some of them were highly correlated. A sensitivity analysis with and without some correlated variables showed very little effect on the main results of gradient forest.

Data analysis

Method description

Gradient forest (Ellis et al. 2012) is a computer intensive method based on classification and regression tree analysis (De'ath and Fabricius 2000; Breiman et al. 1984). Regression trees repeatedly partitions the values of a single response variable (e.g. one species' abundance) into two mutually exclusive groups. These two groups correspond to the values of an explanatory variable which are below and above a split value. The explanatory variable splitting the data and the split value are determined so that the homogeneity of the groups is maximized as regards the response variable. Each split results in two branches and the recursive partitioning of sub-groups gives rise to a tree (e.g. in Fig. 2.2). At this step, it is important to notice that a split value can be interpreted as a threshold in an environmental gradient, from which the response variable changes substantially. The importance of the threshold can be measured by the fit improvement, i.e. the deviance reduction resulting from the split.

Regression trees can be used to predict the abundance of a species given the values of environmental variables. However, the results may depend on the observations and predictors

considered for the analysis. To gain more stability, it is preferable to repeat a high number of times the construction of the tree, with random selection of observations and predictors to be considered, and finally average the predictions of all trees (making a forest). In random forest (Prasad et al. 2006; Breiman 2001), each tree uses a random sample of the observations and each split is determined from a random subset of predictors. The observations which are not taken into account for the construction of a tree are used to cross-validate the performance of the tree. The predictive performance of the forest is the mean cross-validated performance of all trees (R^2). The importance of each predictor is estimated by the increase in mean square prediction error when the values of the predictor are randomly permuted.

Random forest can be applied to each species occurring in a survey using the same environmental variables. Then gradient forest combines the results of all random forests to derive information at the community level. Thus, the community level importance of an environmental variable is the average of all species level importances weighted by species R^2 (Fig. 2.4b). This overall importance in the survey can be fractionated along the variable gradient into quanta of community change, considering the splits due to the variable in all trees of all forests. The location of the quanta in the gradient is given by the split values and the magnitude, by the aggregation of the associated fit improvements. The aggregation of fit improvements takes into account the variable importance and the species R^2 in each random forest (see Ellis et al 2012 for details). The results can be represented in the form of a barplot of aggregated fit improvements in the environmental gradient (Fig. 2.4a). To help interpretation, the data can be smoothed with a density curve (density of splits). The density of splits, whose peaks may mark a community threshold, can be biased by a non-uniform distribution of data along the gradients of environmental variables. Therefore the ratio of split-over-data densities must be considered to better identify and characterize community thresholds (Fig. 2.4a).

Computer implementation

Gradient forest was performed using the two R packages ‘extendedForest’ and ‘gradientForest’ (R_Core_Team 2013; Ellis et al. 2012). A total of 500 trees were generated for each random forest. The split criterion was the sum of square deviations about the mean. In order to stabilize variance, biovolume data were log-transformed after the addition of the minimal strictly positive value for each species. For the estimation of variable importance, a conditional approach was followed in order to limit the importance of correlated variables (Strobl et al. 2008; Ellis et al. 2012). For each explanatory environmental variable, constrained permutations were carried out in each tree within partitions of correlated variables (>0.5 Pearson) obtained after up to 5 splits. For density estimation, the environmental gradients were divided into 201 bins. Local regression was applied to log-transformed biovolume data using R ‘lowess’ function. The probability of presence along a gradient

was calculated in the following way. The numbers of measurements and occurrences in each bin of the gradient were first determined. Then, the variations of these numbers were smoothed using R density function with a gaussian kernel and the same bandwidth as used for the density of data in the split density plot. Finally, the density of occurrences was divided by the density of measurements to get the probability of presence.

Results

National dataset

Gradient forest was first applied to the whole data set. Using biovolume data, 147 species had a positive R^2 and the mean value was 0.20. Indeed in gradient forest, species R^2 may be 0 or even negative (Ellis et al. 2012). Seven species had R^2 above 0.5 and were particularly well predicted by the 29 environmental variables considered (Fig. 3a). Surprisingly, latitude and longitude were clearly the most important variables for phytoplankton abundance (Fig. 3b). The next variables in order of decreasing importance, i.e Alk, PO4, NH4, Secchi, TP and DOC, had close values that were approximately 3-fold lower than latitude and longitude. The distribution of splits along the geographic gradients revealed two major thresholds in the dataset, a broad one around latitude 44°N and a more precise one at longitude 5°E (Fig. 4). The integration of the ratio of densities for each species involved in gradient forest (specific cumulative importance, Fig. 4) comfirms the existence of spatial community thresholds. The concentrations of vertical sections of the species curves just above 44°N and at 5°E shows that several species exhibited breakpoints in their relations to latitude and longitude.

Reduced dataset and physico-chemical thresholds

Given the importance of geographic variables and since the focus was rather on water physico-chemical characteristics, a second analysis was performed on a reduced data set with higher biological homogeneity. In order to include the highest number of lakes as possible, the lakes located between the two thresholds, with latitude higher than 44.5°N and longitude lower than 4.5°E, were selected (Fig. 1). These regionalisation of the study excluded the lakes of the east and south parts of the country and corresponds approximately to the Atlantic biogeographic area. Thus, 129 lakes remained in the analysis with 147 species associated. Considering water quality, the reduced data set was characterized by lower water transparency (median = 1.1 m) and alcalinity (0.74 $meq.L^{-1}$) and higher DOC (7.1 mg $C.L^{-1}$) and TP (44.7 $\mu g.L^{-1}$) concentrations. Following GF analysis, 77 species had a positive R^2 and the mean value was 0.15, slightly lower than that of the whole data set (Fig. 5a). The seven species with the highest R^2 (> 0.3) were *Phacotus lenticularis*, *Neodesmus danubialis*,

Monoraphidium arcuatum, *Pediastrum duplex*, *Nitzschia acicularis*, *Acutodesmus obliquus* and *Pediastrum tetras*. The order of variable importance was notably modified: although geographic variables remained important, several water physico-chemical parameters had the same levels of importance, such as Alk, TP, NH4 , NO3max and DOC. Secchi became the most important variable (Fig. 5b). Detailed results of this second analysis are given for the following variables which are amongst the most important and which are largely influenced by human activities, particularly in a context of eutrophicaton: Secchi, TP, NH4, NO3max and DOC. For each variable, in addition to global GF results concerning all species, the response of one species to the gradient was also provided as an example with a simple plot of its log-transformed biovolume. This species was selected according to two criteria: (1) the variable of interest was the most important for the species, otherwise it would be difficult to see a threshold from the plot of all biovolume data (without any preliminary split) and (2) the species contributed to the main thresholds detected by gradient forest.

In the Secchi gradient, splits are clearly restricted to the zone below 1.3 m (Fig. 6). It means that above this value, water transparency does not influence species abundance. The zone where the most important changes occur is around 0.4 m, where the ratio of densities peaks and several species exhibit a sudden rise in specific cumulative importance. The Chlorophyceae *Monoraphidium arcuatum* presents a threshold in its response to the Secchi gradient around 1 m. This can be seen through the raw abundance data of this species since Secchi is by far the most important variable determining its biovolume. The threshold can be identified on the graph by a decrease in the slope of the regression curve and by a more frequent absence of the species resulting in a drop in the probability of presence.

The density curves for TP reveal a threshold at 50 $\mu g.L^{-1}$ and a larger zone of change from 100 $\mu g.L^{-1}$ up to the end of the gradient (Fig. 6). The threshold results from the change in abundance of some species at the middle of the gradient, as illustrated by the specific cumulative importances. Among these species, the diatom *Aulacoseira granulata* shows the most important change. Its biovolume is mostly determined by TP. The regression curve shows a faster increase in biovolume at 50 $\mu g.L^{-1}$ and the species is present in almost all sites located further in the gradient.

Two main peaks of split density appear in the NH4 gradient at 50 and 125 $\mu g.L^{-1}$ (Fig. 7). Taking into account the density of data, the most important is rather the second one which corresponds to steep rises in specific cumulative importance for several species. From the beginning of the gradient, crossing this threshold implies for the Chlorophyceae *Monoraphidium tortile*, higher biovolumes and a sharp increase in the presence rate.

For NO3max, the most important splits are at the end of the gradient after 10 $mg.L^{-1}$ (Fig. 7), giving rise to a single large peak of the ratio of densities. The specific cumulative importance graph shows that most species respond only to large concentrations of NO3max. The diatom *Stephanodiscus hantzschii* exhibits one of the most important response between 20 and 30 $mg.L^{-1}$, characterized by higher abundance and a more frequent presence in the lakes.

In the DOC gradient, important splits only occur from 10 $mg.L^{-1}$ onwards (Fig. 8). Considering the density of data, it appears that most of the lakes are below the community thresholds which can be located at 11 and 18 $mg.L^{-1}$. The Chlorophyceae *Scenedesmus ecornis* shows a first breakpoint in its relation to DOC at 11 $mg.L^{-1}$, defined by marked increases in abundance and presence rate. A second step in specific cumulative importance corresponds to the second threshold and is due to even higher abundance and no absence.

Discussion

Statistical approach

The basic method used in gradient forest to find thresholds is the non-parametric deviance reduction (NDR) (Qian et al. 2003; Brenden et al. 2008) which is used in regression trees to determine splits. NDR has been used in many studies on ecological thresholds with univariate data (Chambers et al. 2012b; Evans-White et al. 2009; Holt et al. 2003; Smith and Tran 2010; Soranno et al. 2008). It is well suited for data following step function models but it is less effective in finding thresholds when biological response patterns are smoother with less abrupt thresholds (Brenden et al. 2008). Especially, the method may improperly detect thresholds in the case of linear relations (Fig. 2) (Daily et al. 2012). Nevertheless, when the abundance of individual species is considered, the data are often sparse and discontinuous with many zeros, and they generally exhibit stair-step patterns in environmental gradients rather than linear trends. Gradual responses are more often observed when species data are aggregated, e.g. into higher level taxonomic groups or functional metrics (King and Baker 2010; Utz et al. 2009). It is then important to examine graphically each relationship and to question the relevancy of a threshold. When response models differ largely from a step-function, the concept of threshold corresponds more generally to a change point (or change zone), rather to the classic definition given above, i.e. a small change in an explanatory variable giving rise to large variations in the response variable.

Another limit of the deviance reduction approach is its sensitivity to the distribution of data in the gradient. It was numerically demonstrated that such an analysis do not find the same thresholds whether sampling is uniform or not (Cuffney and Qian 2013; Daily et al. 2012). In gradient forest, the

use of the ratio of split-over-data densities adjusts for the bias created by data skewness and leads to a better identification of ecological thresholds (e.g. in Fig. 2).

Unlike most other threshold detection methods, gradient forest takes into account all species reported in a survey and assumes that each taxon can be an indicator of one of the drivers entering the analysis. In this point of view, it is similar to TITAN method (Baker and King 2010) which aims at finding a community threshold along a gradient; i.e. a zone in the gradient where there are 'synchronous' changes in the abundances of many species. Congruence in species' responses to a gradient may be unlikely given the diversity of ecological traits among taxa (Luck 2005). However, the existence of community thresholds may find a justification in an evolutionary point of view. Indeed, concordant declines of many species may be expected when an environmental variable (especially human-influenced) goes outside the range of variations in which these species have co-evolved (King and Baker 2010). In gradient forest, the absence of threshold in a gradient may mean either that all species respond randomly to the environmental variable which is not important (Fig. 2), or that the species have different thresholds along the gradient.

The changes in abundance due to limits in the biogeographic extension area of species, may interfere in the search for environmental thresholds (false zeros) when analysis is conducted at species level and at large spatial scale (Cuffney and Qian 2013; Utz et al. 2009). Aggregating species into metrics or higher taxonomic levels decreases spatial dependency but implies a loss of information. At large spatial scales, phytoplankton data are often analyzed at low taxonomic resolution level to cope with unharmonized species names and differences in taxonomic resolution among regional databases (Maileht et al. 2013). However, it may be always preferable to address the issue of thresholds at species level because species aggregation implies more difficult threshold identification due to the linearization of biological response.

The problem of spatial dependence of species abundance can be extended to any environmental variable interacting with the variable under study. For example, the identification of thresholds for zooplankton communities along acidity gradients in lakes can be confounded by morphometric factors such as lake depth or area (Holt et al. 2003). Regression tree analysis which is the basis of gradient forest, can deal with many explanatory variables and account for complex interactions (De'ath and Fabricius 2000). As a result of recursive partitioning, the effect of an environmental variable on biotic communities is assessed among sites which are the most homogenous as possible considering other more important variables. If natural variables are included (such as morphometric or geographic), a typology is implicitly made in the analysis. However, the community responses (thresholds) in the implicit types are aggregated in the outputs of gradient forest, so that it is not possible to analyze the interactions in details. Multivariate Regression Tree (De'Ath 2002) which also

deals with community change in a multivariate environment, is similar to gradient forest and more transparent as regards interactions. Nonetheless, gradient forest adds the performance of an ensemble method (random forest) and can indicate a rate of community change along environmental gradients, which is useful to identify thresholds.

If a natural factor is suspected by its importance, to modulate a community response to a variable of interest, the effect of the interaction may be investigated by analyzing separately different groups of sites, defined by the thresholds associated with this variable. In this study, the reduction of the data set according to longitude and latitude, assumed that there might be different thresholds in regions with distinct phytoplankton flora. The difference may be due to regional extirpation or adaptation of sensitive species (Utz et al. 2009). Even if gradient forest automatically accounts for regional effects on the abundance of each species, focusing on a more floristically homogeneous region limits the number of thresholds and facilitates the interpretation of the results. Lake typologies are commonly used in the context of the European Water Framework Directive for the application of biological indicators (European Commission 2000). The main natural criteria used to define lake types are altitude, depth and alkalinity. In the GF analyses, these variables had not a prominent importance compared to human-influenced variables. Thus, it did not appear necessary to carry out further typological data splits which would reduce the number of observations and decrease the robustness of the detected thresholds. Moreover, the biological phytoplankton index developed for French lakes does not consider any lake typology (Feret and Laplace-Treyture 2013). The thresholds found in this study are relevant for the concerned biogeographic region (NW France) and should not be used for the management of lakes in other regions without more investigations.

Drivers of phytoplankton communities

A striking result of the gradient forest analysis on phytoplankton species was the overwhelming importance of the geographic variables (latitude and longitude) (Fig. 3). This can be explained by a difference in spatial scales between species distribution areas and survey area. Generally, species distribution areas are determined by a combination of migration processes and a selection by environmental factors. As these distributions were not governed by a single dominant factor among those that were used in the analysis, geographic coordinates were the best predictors of species abundance. At the scale of the United States of America, Stomp et al. (2011) also demonstrated a large influence of geographic variables on phytoplankton diversity.

After exclusion of the eastern and southern regions, the relative importance of geographic variables was reduced in favor of structuring environmental variables, such as Secchi or TP (Fig. 5). Nevertheless, latitude remained an important variable since a secondary threshold was also detected

around 47°N (Fig. 4). Most of the important variables identified by gradient forest in the second analysis are related to lake eutrophication, as shown by the high correlation coefficients with the mean annual log-transformed chlorophyll-*a* concentration: -0.77, 0.68 and 0.59 for Secchi, TP and DOC respectively (Pearson correlation). A Secchi depth measures water transparency and determines the amount of light available for microalgal growth. Considering phytoplankton as a whole, Secchi does not constitute a limiting resource (as phosphorus does) but rather a consequence of algal biomass which reduces the penetration of light in the water column. However, at the species level, Secchi may select species according to their light requirements or their ability to adapt to low or high irradiance levels. Its prominent role in algal physiology and interspecific competition results in a leading position of Secchi among the other explanatory variables.

The importance of TP and dissolved inorganic nitrogen (DIN) is consistent with the well-known trophic control on phytoplankton production. An increase in nutrient concentrations generally stimulates phytoplankton biomass but its effect is not the same on all species or algal groups (Watson et al. 1997). Some species show preferences for low nutrient concentrations whereas others develop mostly in nutrient-rich waters. The classification of species according to their affinity for nutrients has been the basis for the development of many biological indices of eutrophication (Carvalho et al. 2013).

As lakes are generally limited by phosphorus, the importance of DIN might not be expected (Fig. 5). However, eutrophicated lakes with a high phosphorus concentration can become limited by nitrogen (Solheim et al. 2008; Donald et al. 2013). Changes in DIN concentration modify the N/P ratio and influence the composition of phytoplankton communities, particularly the abundance of nitrogen-fixing species (Schindler 1977). Several reasons may explain the higher importance of NH4 over NO3. Most ammonia comes from the decomposition of organic matter. Its concentration is thus closely related to lake productivity and eutrophication, as shown by the correlation between NH4 and TP (ρ=0.66), whereas NO3 was independent of TP (ρ=–0.01). High concentration of ammonia can also be related to organic pollution from sewage effluents (Beklioglu et al. 1999; García-Ferrer et al. 2003) which may differentially impact phytoplankton species (Katsiapi et al. 2013; Villena and Romo 2003). Finally, NH_4^+ can be toxic to phytoplankton when it turns into its unionized form (NH_3) as pH and temperature increase (Camargo and Alonso 2006; Konig et al. 1987). Unlike TP or NH4, the effect of nitrates was better expressed when the annual maximum was considered instead of the annual median. In average, the maximum concentration of nitrates was measured at the year's first campaign at the end of winter.

Another interesting result was the relatively high importance of DOC to predict phytoplankton abundance. The role of DOC in lake ecosystem functioning has been neglected in the last decades, whereas research in limnology has focused on the nutrient-productivity relationship (Williamson et al. 1999; Carpenter et al. 1998). DOC can affect phytoplankton communities through the attenuation of solar radiation and interactions with nutrients and contaminants (Jansson et al. 2000; Wall and Briand 1979). High DOC concentrations might result either from input of colored organic matter from the watershed, limiting light penetration and inhibiting primary production, or from the production of aquatic plants in a eutrophicated system (transparent, labile DOC) (Bade et al. 2007). A better assessment of DOC effect on phytoplankton would require a distinction between the allochthonous and autochthonous forms.

Due to its interactions with pH and nutrients (CO_2 concentration and phosphate bioavailability) and its link with water conductivity, alkalinity is a general factor influencing the distribution of microalgal species. In another study at broad spatial scale, it was shown that alkalinity could explain a large part of variations in phytoplankton communities (Maileht et al. 2013) but maximal depth was the most determining factor. Here, the morphological variable Zmax was considerably less important than Secchi and TP. This could be explained by the conditional approach used here to estimate variable importance, since lake depth is correlated with factors which might influence more directly phytoplankton communities, like TP or Secchi ($\rho<-0.7$). Indeed, when lake depth decreases, resuspension of sediment is more likely and inputs from the watershed (phosphorus) become concentrated in a lower lake water volume.

Thresholds and their use for lake management

The community thresholds that were identified in this study can be used to derive ecologically sound environmental standards. Unlike 'chemical' thresholds like those derived from percentile analysis (Chambers et al. 2011; Smith and Tran 2010), community thresholds take into account the response of biological compartments to environmental stressors and are in line with the preservation of biodiversity and ecosystem functioning (Brenden et al. 2008). The necessity to link environmental standards with the response of communities has been strengthened by the European Water Framework Directive (European Commission 2000), which states that physico-chemical quality elements must support the achievement of good status for the biological quality elements, such as phytoplankton, fish, macrophytes and macro-invertebrates. Recently, Poikane et al. (2014) have proposed ecological boundaries for chl-*a* on the basis of the response of lake ecosystems to eutrophication, considering phytoplankton (Cyanobacteria) and macrophytes.

The gradient forest analysis provided interesting thresholds for several environmental variables. Geographic thresholds isolated the Southern and especially the Eastern part of the country, which is characterized by several chains of mountains (mainly the Alps). These limits form a biogeographic typology which can be used to address the issue of thresholds separately for each region. Focusing in this study on the largest identified group without Mediterranean or alpine influence, interesting values were detected by gradient forest for the human-influenced variables (Fig. 6, 7 and 8). For TP and Secchi, the thresholds identified here can be compared to the trophic class limits proposed by OECD (1982): 35 µg $TP.L^{-1}$ and 100 µg $TP.L^{-1}$ or 1 m Secchi depth for the mesotrophic/eutrophic and eutrophic/hypereutrophic limits, respectively.

Ecological threshold values reported in the literature concern mostly total phosphorus and have the same order of magnitude as the threshold at 50 µg.L^{-1} found in the present study. Concerning lakes, TP thresholds have been reported at 18 µg.L^{-1} (Soranno et al. 2008), 10, 25 and 70 µg.L^{-1} (Free et al. 2006), 20 and 50 µg.L^{-1} (Penning et al. 2008) using phytoplankton or macrophyte data. Most other references of phosphorus thresholds come from periphyton or macro-invertebrates in rivers. The values are the following: 30-60 (Dodds et al. 2002), 12-15 (Richardson et al. 2007) , 60-90 (Evans-White et al. 2009), 9-70 (Smith and Tran 2010), 30-280 (Black et al. 2011) and 21-63 µg $TP.L^{-1}$ (Chambers et al. 2012b). Thresholds can be more or less broad in a gradient depending on their sharpness, the variability in response among species or biological compartments and the uncertainty associated with detection methods. A broad threshold, as observed here for TP, may also reflect hysteresis in the response of a biological compartment to the gradient, especially if some lakes under restoration are included in the analysis.

Since the variables presented in Fig. 6, 7 and 8 (Secchi, TP, NH4, NO3max, DOC) may respond to human disturbance, they can be used as physico-chemical indicators to evaluate the lake alteration level and to protect ecosystems. All of them might be sensitive to eutrophication or organic pollution, but also to hydromorphological alterations. Thresholds identified in this study can be translated into environmental standards which should not be exceeded to prevent ecosystems from important changes in their communities. However, variations in these physico-chemical parameters among lakes may be partly natural and independent of anthropogenic pressures. For example, a lake can present high values of colored DOC due to a large proportion of wetlands or coniferous forests in its watershed, resulting in a low Secchi depth (humic lakes). Similarly, naturally eutrophic lakes may exhibit high phosphorus concentrations which are not linked to any pollution (Borics et al. 2013), even if most European lakes in Europe have lower TP concentrations than the threshold of 50 µg.L^{-1} (Cardoso et al. 2007). Therefore, reference values must be defined in the evaluation process in order to take into account non-anthropogenic variations. Soranno et al. (2008) proposed a framework to

derive site-specific nutrient criteria from the knowledge of biological thresholds, reference values, and measured values. The lowest biological threshold that is above the modeled reference value, may be retained as an operational standard. When several thresholds are identified above a reference value, they may account for the increasing levels of alteration, delimiting the different classes of quality (Solheim et al. 2008; Free et al. 2006). According to the European Water Framework Directive (European Commission 2000), there must be four boundaries delimiting five quality classes for each physico-chemical quality element. The contribution of this study consists in the detection of some boundaries in some gradients. The research of thresholds should be extended to other lake communities (macrophytes, fish, macro-invertebrates, phytobenthos), which could reveal new critical zones because of their different sensitivities to pollution or on the contrary, reinforce the position of some thresholds affecting several biological groups (Richardson et al. 2007). Nevertheless, one critical question remains for lake management: which limit between quality classes should be attributed to a given ecological threshold? The main threshold following the reference conditions may be the good-moderate boundary, but if this threshold appears too distant from the reference, it may be assigned a lower quality limit. The choice can be guided by the levels of boundaries set in other regions through other methods (Claussen et al. 2012) or by the observed ranges of parameters in groups of lakes of each quality class, as determined by intercalibrated biological indices (Phillips et al. 2013). Missing limits can eventually be derived from simple interval divisions.

Conclusion

Gradient forest is an exploratory method which deals with multivariate biological and environmental data. It is well suited for the research of community thresholds on a large spatial scale. The main drivers of phytoplankton variations among the French lakes were identified and community thresholds were detected for spatial variables (longitude and latitude) and human-influenced variables in a chosen biogeographic region. These thresholds can contribute to the definition of environmental standards for lake management. Further investigations may include the definition of appropriate thresholds for the other biogeographic regions derived from the first analysis and the questions of criteria for other pressures on lakes such as acidification or climate change (Cardoso et al. 2009). Thresholds obtained from other lake communities (e.g. macrophytes or fish) could also be compared to those inferred here from phytoplankton. Finally, models that include hydromorphological explanatory variables should be developed to (1) identify the main causes of variations among lakes of the most important physico-chemical variables identified in this study, and (2) to predict natural reference conditions which are necessary for the evaluation of human alteration of ecosystems.

Acknowledgments

This research was funded by The French National Agency for Water and Aquatic Environments (ONEMA). The authors are grateful to the French Water Basin Agencies and their partners who contributed to the lake data acquisition and to the Onema-Irstea consortium for Lake Hydroecology (Pôle Onema-Irstea d'études et de recherche "hydroécologie des plans d'eau", Aix-en-Provence, France), who currently maintains the national biological and physico-chemical database for French lakes; special thanks to Nathalie Reynaud and Thierry Point (database) and Milena Borissova (text edition).

References

Andersen, T., Carstensen, J., Hernández-García, E., & Duarte, C. M. (2009). Ecological thresholds and regime shifts: approaches to identification. ***Trends in Ecology & Evolution, 24*****(1), 49-57, doi:http://dx.doi.org/10.1016/j.tree.2008.07.014.**

Bade, D. L., Carpenter, S. R., Cole, J. J., Pace, M. L., Kritzberg, E., Van de Bogert, M. C., et al. (2007). Sources and fates of dissolved organic carbon in lakes as determined by whole-lake carbon isotope additions. ***Biogeochemistry, 84*****(2), 115-129, doi:DOI 10.1007/s10533-006-9013-y.**

Baker, M. E., & King, R. S. (2010). A new method for detecting and interpreting biodiversity and ecological community thresholds. ***Methods in Ecology and Evolution, 1*****(1), 25-37, doi:10.1111/j.2041-210X.2009.00007.x.**

Beklioglu, M., Carvalho, L., & Moss, B. (1999). Rapid recovery of a shallow hypertrophic lake following sewage effluent diversion: Lack of chemical resilience. ***Hydrobiologia, 412*****, 5-15.**

Black, R. W., Moran, P. W., & Frankforter, J. D. (2011). Response of algal metrics to nutrients and physical factors and identification of nutrient thresholds in agricultural streams. ***Environmental Monitoring and Assessment, 175*****(1-4), 397-417, doi:10.1007/s10661-010-1539-8.**

Borics, G., Nagy, L., Miron, S., Grigorszky, I., László-Nagy, Z., Lukács, B. A., et al. (2013). Which factors affect phytoplankton biomass in shallow eutrophic lakes? ***Hydrobiologia, 714*****(1), 93-104.**

Breiman, L. (2001). Random forests. ***Machine Learning, 45*****(1), 5-32, doi:10.1023/A:1010933404324.**

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). ***Classification and Regression Trees*****: Chapman and Hall/CRC**

Brenden, T. O., Wang, L., & Su, Z. (2008). Quantitative Identification of Disturbance Thresholds in Support of Aquatic Resource Management. ***Environmental Management, 42*****(5), 821-832, doi:10.1007/s00267-008-9150-2.**

Camargo, J. A., & Alonso, A. (2006). Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: A global assessment. ***Environment International, 32*****(6), 831-849, doi:DOI 10.1016/j.envint.2006.05.002.**

Cardoso, A. C., Free, G., Nõges, P., Kaste, Ø., Poikane, S., & Solheim, A. L. (2009). Lake Management, Criteria. In G. E. Likens (Ed.), ***Encyclopedia of Inland Waters*** **(Vol. 1, pp. 310-331). Oxford.**

Cardoso, A. C., Solimini, A., Premazzi, G., Carvalho, L., Lyche, A., & Rekolainen, S. (2007). Phosphorus reference concentrations in European lakes. ***Hydrobiologia, 584*****, 3-12, doi:10.1007/s10750-007-0584-y.**

Carpenter, S. R., Cole, J. J., Kitchell, J. F., & Pace, M. L. (1998). Impact of dissolved organic carbon, phosphorus, and grazing on phytoplankton biomass and production in experimental lakes. ***Limnology and Oceanography, 43*****(1), 73-80.**

Carvalho, L., Poikane, S., Lyche Solheim, A., Phillips, G., Borics, G., Catalan, J., et al. (2013). Strength and uncertainty of phytoplankton metrics for assessing eutrophication impacts in lakes. *Hydrobiologia, 704*(1), 127-140, doi:10.1007/s10750-012-1344-1.

Catalan, J., Barbieri, M. G., Bartumeus, F., Bitusik, P., Botev, I., Brancelj, A., et al. (2009). Ecological thresholds in European alpine lakes. *Freshwater Biology, 54*(12), 2494-2517, doi:10.1111/j.1365-2427.2009.02286.x.

Chambers, P. A., Benoy, G. A., Brua, R. B., & Culp, J. M. (2011). Application of nitrogen and phosphorus criteria for streams in agricultural landscapes. *Water Science and Technology, 64*(11), 2185-2191, doi:Doi 10.2166/Wst.2011.760.

Chambers, P. A., Culp, J. M., Roberts, E. S., & Bowerman, M. (2012a). Development of Environmental Thresholds for Streams in Agricultural Watersheds. *Journal of Environmental Quality, 41*(1), 1-6, doi:10.2134/jeq2011.0338.

Chambers, P. A., McGoldrick, D. J., Brua, R. B., Vis, C., Culp, J. M., & Benoy, G. A. (2012b). Development of Environmental Thresholds for Nitrogen and Phosphorus in Streams. *Journal of Environmental Quality, 41*(1), 7-20, doi:10.2134/jeq2010.0273.

Claussen, U., Müller, P., & Arle, J. (2012). Comparison of Environmental Quality Objectives, Threshold Values or Water Quality Targets Set for the Demands of the European Water Framework Directive. *WFD CIS ECOSTAT WG A Report* (pp. 27). CIRCABC

Cuffney, T. F., & Qian, S. S. (2013). A critique of the use of indicator-species scores for identifying thresholds in species responses. *Freshwater Science, 32*(2), 471-488, doi:10.1899/12-056.1.

Daily, J. P., Hitt, N. P., Smith, D. R., & Snyder, C. D. (2012). Experimental and environmental factors affect spurious detection of ecological thresholds. *Ecology, 93*(1), 17-23.

De'Ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology, 83*(4), 1105-1117, doi:Doi 10.2307/3071917.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology, 81*(11), 3178-3192, doi:Doi 10.2307/177409.

Dodds, W. K., Clements, W. H., Gido, K., Hilderbrand, R. H., & King, R. S. (2010). Thresholds, breakpoints, and nonlinearity in freshwaters as related to management. *Journal of the North American Benthological Society, 29*(3), 988-997, doi:Doi 10.1899/09-148.1.

Dodds, W. K., Smith, V. H., & Lohman, K. (2002). Nitrogen and phosphorus relationships to benthic algal biomass in temperate streams. *Canadian Journal of Fisheries and Aquatic Sciences, 59*(5), 865-874, doi:Doi 10.1139/F02-063.

Donald, D. B., Bogard, M. J., Finlay, K., Bunting, L., & Leavitt, P. R. (2013). Phytoplankton-Specific Response to Enrichment of Phosphorus-Rich Surface Waters with Ammonium, Nitrate, and Urea. *Plos One, 8*(1), doi:DOI 10.1371/journal.pone.0053277.

Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: calculating importance gradients on physical predictors. *Ecology, 93*(1), 156-168.

European Commission (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a Framework for Community Action in the Field of Water Policy. (pp. 72): The European Parlament and Council.

Evans-White, M. A., Dodds, W. K., Huggins, D. G., & Baker, D. S. (2009). Thresholds in macroinvertebrate biodiversity and stoichiometry across water-quality gradients in Central Plains (USA) streams. *Journal of the North American Benthological Society, 28*(4), 855-868, doi:10.1899/08-113.1.

Feret, T., & Laplace-Treyture, C. (2013). IPLAC : l'indice Phytoplancton Lacustre : Méthode de développement, description et application nationale 2012. *Rapport convention Onema/Irstea 2012* (pp. 69). Bordeaux: Irstea, UR REBX.

Free, G., Little, R., Tierney, D., Donnelly, K., & Caroni, R. (2006). A reference based typology and ecological assessment system for Irish lakes. Prelimnary investigations. (pp. 266). Wexford, Ireland: Environmental Protection Agency.

García-Ferrer, I., Camacho, A., Armengol, X., Miracle, M. R., & Vicente, E. (2003). Seasonal and spatial heterogeneity in the water chemistry of two sewage-affected saline shallow lakes from central Spain. *Hydrobiologia, 506-509*, 101-110.

Graham, J. L., Jones, J. R., Jones, S. B., Downing, J. A., & Clevenger, T. E. (2004). Environmental factors influencing microcystin distribution and concentration in the Midwestern United States. *Water Research, 38*(20), 4395-4404, doi:DOI 10.1016/j.watres.2004.08.004.

Groffman, P., Baron, J., Blett, T., Gold, A., Goodman, I., Gunderson, L., et al. (2006). Ecological thresholds: The key to successful environmental management or an important concept with no practical application? *Ecosystems, 9*(1), 1-13, doi:DOI 10.1007/s10021-003-0142-z.

Holt, C. A., Yan, N. D., & Somers, K. M. (2003). pH 6 as the threshold to use in critical load modeling for zooplankton community change with acidification in lakes of south-central Ontario: accounting for morphometry and geography. *Canadian Journal of Fisheries and Aquatic Sciences, 60*(2), 151-158, doi:doi:10.1139/f03-008.

Jansson, M., Bergström, A.-K., Blomqvist, P., & Drakare, S. (2000). allochthonous organic carbon and phytoplankton/bacterioplankton production relationships in lakes. *Ecology, 81*(11), 3250-3255, doi:10.1890/0012-9658(2000)081[3250:AOCAPB]2.0.CO;2.

Katsiapi, M., Moustaka-Gouni, M., Vardaka, E., & Kormas, K. A. (2013). Different phytoplankton descriptors show asynchronous changes in a shallow urban lake (L. Kastoria, Greece) after sewage diversion. *Fundamental and Applied Limnology, 182*(3), 219-230.

King, R. S., & Baker, M. E. (2010). Considerations for analyzing ecological community thresholds in response to anthropogenic environmental gradients. *Journal of the North American Benthological Society, 29*(3), 998-1008, doi:10.1899/09-144.1.

King, R. S., Baker, M. E., Kazyak, P. F., & Weller, D. E. (2011). How novel is too novel? Stream community thresholds at exceptionally low levels of catchment urbanization. *Ecological Applications, 21*(5), 1659-1678, doi:10.1890/10-1357.1.

Konig, A., Pearson, H. W., & Silva, S. A. (1987). Ammonia Toxicity to Algal Growth in Waste Stabilization Ponds. *Water Science and Technology, 19*(12), 115-122.

Laplace-Treyture, C., Barbe, J., Dutartre, A., Druart, J.-C., Rimet, F., & Anneville, O. (2009). Standard protocol for sampling, conservation, observation and counting of lake phytoplankton for application of the WFD. Version 3.3.1. (pp. 42). Cestas, France: Cemagref UR REBX.

Luck, G. W. (2005). An introduction to ecological thresholds. *Biological Conservation, 124*(3), 299-300, doi:10.1016/j.biocon.2005.01.042.

Maileht, K., Nõges, T., Nõges, P., Ott, I., Mischke, U., Carvalho, L., et al. (2013). Water colour, phosphorus and alkalinity are the major determinants of the dominant phytoplankton species in European lakes. *Hydrobiologia, 704*(1), 115-126, doi:10.1007/s10750-012-1348-x.

May, R. M. (1977). Thresholds and Breakpoints in Ecosystems with a Multiplicity of Stable States. *Nature, 269*(5628), 471-477, doi:Doi 10.1038/269471a0.

OECD (1982). Eutrophication of waters -monitoring, assessment and control. Paris: Organisation for Economic Co-operation and Development.

Penning, W. E., Dudley, B., Mjelde, M., Hellsten, S., Hanganu, J., Kolada, A., et al. (2008). Using aquatic macrophyte community indices to define the ecological status of European lakes. *Aquatic Ecology, 42*(2), 253-264, doi:10.1007/s10452-008-9183-x.

Phillips, G., Lyche-Solheim, A., Skjelbred, B., Mischke, U., Drakare, S., Free, G., et al. (2013). A phytoplankton trophic index to assess the status of lakes for the Water Framework Directive. *Hydrobiologia, 704*(1), 75-95, doi:10.1007/s10750-012-1390-8.

Pitcher, C. R., Lawton, P., Ellis, N., Smith, S. J., Incze, L. S., Wei, C. L., et al. (2012). Exploring the role of environmental variables in shaping patterns of seabed biodiversity composition in regional-scale ecosystems. *Journal of Applied Ecology, 49*(3), 670-679, doi:DOI 10.1111/j.1365-2664.2012.02148.x.

Poikane, S., Portielje, R., van den Berg, M., Phillips, G., Brucet, S., Carvalho, L., et al. (2014). Defining ecologically relevant water quality targets for lakes in Europe. *Journal of Applied Ecology, 51*(3), 592-602, doi:10.1111/1365-2664.12228.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181-199, doi:DOI 10.1007/s10021-005-0054-1.

Qian, S. S. (2014). Ecological threshold and environmental management: A note on statistical methods for detecting thresholds. *Ecological Indicators, 38*, 192-197, doi:10.1016/j.ecolind.2013.11.008.

Qian, S. S., King, R. S., & Richardson, C. J. (2003). Two statistical methods for the detection of environmental thresholds. *Ecological Modelling, 166*(1–2), 87-97, doi:http://dx.doi.org/10.1016/S0304-3800(03)00097-8.

R_Core_Team (2013, R: A language and environment for statistical computing.

Richardson, C. J., King, R. S., Qian, S. S., Vaithiyanathan, P., Qualls, R. G., & Stow, C. A. (2007). Estimating ecological thresholds for phosphorus in the Everglades. *Environmental Science & Technology, 41*(23), 8084-8091, doi:10.1021/es062624w.

Scheffer, M., & Carpenter, S. R. (2003). Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends in Ecology & Evolution, 18*(12), 648-656, doi:10.1016/j.tree.2003.09.002.

Scheffer, M., Rinaldi, S., Gragnani, A., Mur, L. R., & vanNes, E. H. (1997). On the dominance of filamentous cyanobacteria in shallow, turbid lakes. *Ecology, 78*(1), 272-282.

Scheffer, M., Szabo, S., Gragnani, A., van Nes, E. H., Rinaldi, S., Kautsky, N., et al. (2003). Floating plant dominance as a stable state. *Proceedings of the National Academy of Sciences of the United States of America, 100*(7), 4040-4045, doi:10.1073/pnas.0737918100.

Schindler, D. W. (1977). Evolution of Phosphorus Limitation in Lakes. *Science, 195*(4275), 260-262, doi:DOI 10.1126/science.195.4275.260.

Smith, A. J., & Tran, C. P. (2010). A weight-of-evidence approach to define nutrient criteria protective of aquatic life in large rivers. *Journal of the North American Benthological Society, 29*(3), 875-891, doi:Doi 10.1899/09-076.1.

Solheim, A. L., Rekolainen, S., Moe, S. J., Carvalho, L., Phillips, G., Ptacnik, R., et al. (2008). Ecological threshold responses in European lakes and their applicability for the Water Framework Directive (WFD) implementation: synthesis of lakes results from the REBECCA project. *Aquatic Ecology, 42*(2), 317-334, doi:DOI 10.1007/s10452-008-9188-5.

Soranno, P. A., Cheruvelil, K. S., Stevenson, R. J., Rollins, S. L., Holden, S. W., Heaton, S., et al. (2008). A framework for developing ecosystem-specific nutrient criteria: Integrating biological thresholds with predictive modeling. *Limnology and Oceanography, 53*(2), 773-787, doi:10.4319/lo.2008.53.2.0773.

Stomp, M., Huisman, J., Mittelbach, G. G., Litchman, E., & Klausmeier, C. A. (2011). Large-scale biodiversity patterns in freshwater phytoplankton. *Ecology, 92*(11), 2096-2107.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *Bmc Bioinformatics, 9*, doi:Doi 10.1186/1471-2105-9-307.

Utz, R. M., Hilderbrand, R. H., & Boward, D. M. (2009). Identifying regional differences in threshold responses of aquatic invertebrates to land cover gradients. *Ecological Indicators, 9*(3), 556-567, doi:10.1016/j.ecolind.2008.08.008.

Villena, M. J., & Romo, S. (2003). Phytoplankton changes in a shallow Mediterranean lake (Albufera of Valencia, Spain) after sewage diversion. *Hydrobiologia, 506-509*, 281-287.

Vollenweider, R. A. (1975). Input-output models - With special reference to the phoshorus loading concept in limnology. *Source of the Document Schweizerische Zeitschrift für Hydrologie, 37*(1), 53-84.

Wall, D., & Briand, F. (1979). Response of lake phytoplankton communities to in situ manipulations of light intensity and colour. *Journal of Plankton Research, 1*(1), 103-112, doi:10.1093/plankt/1.1.103.

Watson, S. B., McCauley, E., & Downing, J. A. (1997). Patterns in phytoplankton taxonomic composition across temperate lakes of differing nutrient status. *Limnology and Oceanography, 42*(3), 487-495.

Williamson, C. E., Morris, D. P., Pace, M. L., & Olson, A. G. (1999). Dissolved organic carbon and nutrients as regulators of lake ecosystems: Resurrection of a more integrated paradigm. *Limnology and Oceanography, 44*(3), 795-803.

Table 1 List of environmental variables used in the analysis with basic distribution statistics (minimum, median and maximum) among lakes. The codes are used to denominate the variables in the following. For variables having annual variations (lines below Zmax), the first code corresponds to the annual median and the codes with 'min' and 'max' endings are the minimal and maximal annual values, respectively. CV is the mean annual coefficient of variation.

Parameter	code	min	med	max	CV (%)
Latitude (°N)	lat	41.47	46.42	50.87	-
Longitude (°E)	lon	-4.01	2.26	9.48	-
Altitude (m)	Alt	0	213	2082	-
Maximal depth (m)	Zmax	0.8	15	309.7	-
Water temperature (°C)	Temp	7.1	17	22.8	40
	Tempmin	0.3	7	14.5	-
	Tempmax	8.9	21.5	29.2	-
Secchi depth (m)	Secchi	0.1	1.6	18.7	38
	Secchimin	0.05	0.95	12	-
O_2 saturation (%)	SatO2	53.9	94.8	126.9	17
	SatO2min	15.2	78.3	114.3	-
Alkalinity (meq.L^{-1})	Alk	0.1	1.0	4.6	-
pH	pH	5.7	8.1	9.9	6
	pHmin	5.1	7.5	9.1	-
	pHmax	6.9	8.6	12.5	-
Conductivity (µS.cm^{-1})	Cond	14	239	1307	12
Dissolved organic carbon (mg.L^{-1})	DOC	0.3	4.9	36.7	19
	DOCmax	0.4	6.0	56	-
Nitrates (µg.L^{-1})	NO3	120	1220	39260	81
	NO3max	250	3950	72000	-
Ammonia (µg.L^{-1})	NH4	9	50	389	71
	NH4max	10	130	3300	-
Nitrites (µg.L^{-1})	NO2	3.5	18	229	67
	NO2max	6	50	690	-
Phosphates (µg.L^{-1})	PO4	6	16	972	72
	PO4max	6	40	2116	-
Total phosphorus (µg.L^{-1})	TP	6	30	599	49
	TPmax	7	46	1490	-
Dissolved silicon (mg SiO_2.L^{-1})	SiO2	0.18	3.3	20.78	54

1 **Table 2** List of taxa used in the gradient forest analysis restricted to the North-West part of France (*n*=number
2 of records).

Taxon	*n*	Taxon	*n*	Taxon	*n*
Achnanthidium minutissimum	8	Dinobryon sociale	17	Peridinium inconspicuum	8
Acanthoceras zachariasii	23	Dinobryon suecicum	16	Phacus tortus	8
Actinastrum hantzschii	28	Discostella pseudostelligera	26	Phacotus lenticularis	24
Acutodesmus acuminatus	16	Discostella stelligera	13	Planktothrix agardhii	73
Acutodesmus obliquus	55	Dolichospermum flos-aquae	8	Plagioselmis nannoplanctica	83
Ankistrodesmus falcatus	10	Elakatothrix gelatinosa	37	Planctonema lauterbornii	11
Ankyra ancora	16	Erkenia subaequiciliata	34	Pseudodidymocystis fina	33
Ankyra judayi	47	Fragilaria capucina	10	Pseudodidymocystis planctonica	7
Aphanocapsa delicatissima	15	Fragilaria crotonensis	43	Pseudanabaena catenata	9
Aphanocapsa elachista	17	Goniochloris mutica	14	Pseudanabaena limnetica	16
Aphanocapsa holsatica	7	Granulocystopsis coronata	8	Pseudanabaena mucicola	13
Aphanizomenon flos-aquae	15	Hariotina reticulata	18	Puncticulata radiosa	32
Aphanothece clathrata	7	Kephyrion rubri-claustri	7	Rhodomonas lacustris	65
Asterionella formosa	97	Kirchneriella obesa	7	Scenedesmus aculeolatus	15
Aulacoseira ambigua	19	Koliella longiseta	64	Scenedesmus bicaudatus	32
Aulacoseira distans	44	Koliella planctonica	21	Scenedesmus ecornis	37
Aulacoseira granulata	98	Lagerheimia ciliata	19	Scenedesmus ellipticus	31
Aulacoseira subarctica	17	Lagerheimia genevensis	59	Scenedesmus obtusus	14
Choricystis minor	12	Lanceola spatulifera	71	Scenedesmus pulloideus	13
Ceratium hirundinella	40	Limnothrix redekei	31	Scenedesmus quadricauda var. longispinus	17
Chroococcus minutus	11	Mallomonas akrokomos	30	Scenedesmus verrucosus	12
Chrysococcus rufescens	9	Melosira varians	24	Schroederia setigera	11
Closterium aciculare	11	Merismopedia tenuissima	31	Sphaerocystis schroeteri	31
Closterium acutum	16	Microcystis aeruginosa	17	Staurastrum cingulum	8
Closterium acutum var. variabile	18	Micractinium pusillum	10	Staurastrum pingue	17
Cocconeis placentula	22	Monoraphidium arcuatum	95	Stephanodiscus hantzschii	66
Coelastrum astroideum	10	Monoraphidium circinale	41	Stephanodiscus minutulus	7
Coelastrum microporum	31	Monoraphidium contortum	111	Stephanodiscus parvus	29
Crucigenia tetrapedia	93	Monoraphidium convolutum	10	Staurosira construens	11
Cryptomonas curvata	11	Monoraphidium griffithii	55	Stelexomonas dichotoma	45
Cryptomonas marssonii	27	Monoraphidium komarkovae	36	Spermatozopsis exsultans	15
Cryptomonas ovata	22	Monoraphidium minutum	50	Tabellaria flocculosa	12
Cyclotella meneghiniana	44	Monoraphidium nanum	8	Tetrachlorella alternans	30
Cyclotella ocellata	16	Monoraphidium tortile	37	Tetraedron caudatum	66
Cyclostephanos dubius	77	Monactinus simplex	50	Tetraedron incus	27
Cyclostephanos invisitatus	16	Mucidosphaerium pulchellum	20	Tetraedron minimum	104
Chrysolykos planctonicus	8	Navicula gregaria	13	Tetraedron triangulare	39
Desmodesmus abundans	21	Navicula lanceolata	18	Tetraëdriella regularis	19
Desmodesmus armatus	64	Nephrochlamys rostrata	13	Tetrastrum staurogeniaeforme	21
Desmodesmus communis	82	Neodesmus danubialis	11	Tetrastrum triangulare	7
Desmodesmus intermedius	45	Nitzschia acicularis	92	Trachelomonas volvocina	80
Desmodesmus opoliensis	61	Nitzschia palea	11	Trachelomonas volvocina var. punctata	19
Desmodesmus spinosus	21	Oocystis lacustris	17	Treubaria planctonica	33
Diatoma tenuis	17	Oocystis parva	11	Treubaria setigera	10
Dictyosphaerium subsolitarium	25	Pandorina morum	11	Ulnaria delicatissima var. angustissima	25
Dinobryon bavaricum	20	Pediastrum boryanum	52	Ulnaria ulna var. acus	42
Dinobryon crenulatum	11	Pediastrum duplex	83	Ulnaria ulna	70
Dinobryon divergens	40	Pediastrum boryanum var. longicorne	10	Urosolenia longiseta	77
Dinobryon sertularia	10	Pediastrum tetras	73	Woronichinia naegeliana	38

3

4

Figure captions

Fig. 1 Location of the 224 lakes composing the data set for analysis. The dashed lines indicate latitude +44.5°N and longitude +4.5°E.

Fig. 2 Gradient forest analysis applied to a simulated data set of 4 species' abundances and 5 environmental variables. (1) Specific simulated relationships for 5 species-variable pairs (following simple functions: logistic, Gaussian, step-wise and linear, with Poisson error), independence was assumed for the other pairs. (2) Example of a regression tree for one species; splitting factors and values are indicated at the nodes of the branches. Leaves give the mean values of species abundance in the corresponding groups. (3) For each species, *n* trees are constructed, making a random forest from which a species R^2 and a specific variable importance can be derived. (4) Gradient forest combines the results of the random forests to give a community level variable importance and to locate the splits and their importance in the environmental gradients (black curve for density of splits, dashed curve for density of data and thick grey curve for the ratio of densities; the horizontal dashed line indicates where the ratio is 1, each curve integrates to the variable importance). There is one community threshold in the gradient of A and two in the gradient of B. The variable E shows no threshold (independence). There are 2 equally important thresholds in the gradient of D after accounting for data distribution (grey curve), and variable C exhibits a spurious threshold (see discussion).

Fig. 3 Species R^2 derived from random forests (a) and overall variable importance as determined by gradient forest (b). The sum of variable importances equals the average specific R^2. Codes refer to Table 1.

Fig. 4 Density of splits (top) and specific cumulative importance (bottom) in the gradients of the most important variables identified by gradient forest : latitude (lat) and longitude (lon). The specific cumulative importance is the integration of the ratio of densities for each species.

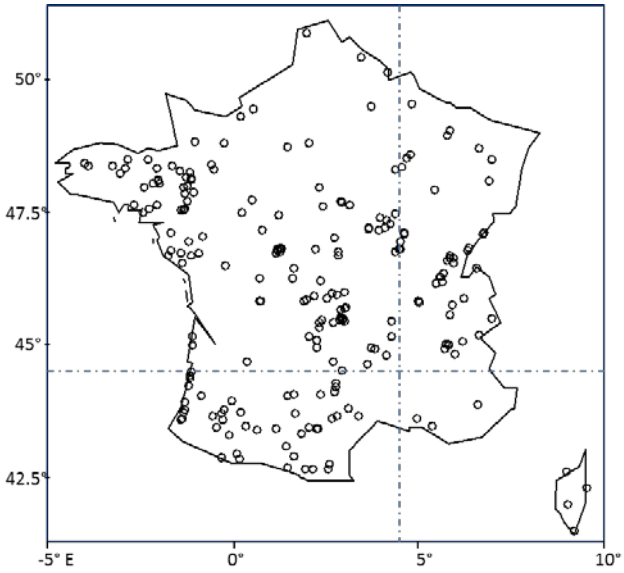
Fig. 5 Species R^2 derived from random forests with the reduced data set (a) and overall variable importance as determined by gradient forest with biovolume data (b). The sum of variable importances equals the average specific R^2. Codes refer to Table 1.

Fig. 6 Results of gradient forest for the variables Secchi and TP (one column per variable). (a) Graph of densities: the dashed curve indicates the density of data, the thin black curve, the density of splits, the thick grey curve, the ratio of densities and the horizontal dashed line shows where the ratio is one. Each curve integrates to the variable importance. (b) Curves of the specific cumulative importance which is the integration of the ratio of densities for each species. The black curve corresponds to the species shown as an example below. (c) Plots of the mean biovolume of *Monoraphidium arcuatum* (Chlorophyceae) and *Aulacoseira granulata* (Bacillariophyceae) in the gradients of Secchi and TP, respectively. The smoothing curve is continuous and the dashed curve indicates the probability of presence of the species among the sampled sites. (d) Barplot of the relative impurity importance of the variables for the species shown for example.

Fig. 7 Results of gradient forest for the variables NH4 and NO3max (one column per variable). (a) Graph of densities: the dashed curve indicates the density of data, the thin black curve, the density of splits, the thick grey curve, the ratio of densities and the horizontal dashed line shows where the ratio is one. Each curve integrates to the variable importance. (b) Curves of the specific cumulative importance which is the integration of the ratio of densities for each species. The black curve corresponds to the species shown as an example below. (c) Plots of the mean biovolume of *Monoraphidium tortile* (Chlorophyceae) and *Cyclostephanos invisitatus* (Bacillariophyceae) in the gradients of NH4 and NO3max, respectively. The smoothing curve is continuous and the dashed curve indicates the probability of presence of the species among the sampled sites. (d) Barplot of the relative impurity importance of the variables for the species shown for example.

Fig. 8 Results of gradient forest for the variable DOC. (a) Graph of densities: the dashed curve indicates the density of data, the thin black curve, the density of splits, the thick grey curve, the ratio of densities and the horizontal dashed line shows where the ratio is one. Each curve integrates to the variable importance. (b) Curves of the specific cumulative importance which is the integration of the ratio of densities for each species. The black curve corresponds to the species shown as an example below. (c) Plots of the mean biovolume of *Scenedesmus ecornis* (Chlorophyceae) in the gradients of DOC. The smoothing curve is continuous and the dashed curve indicates the probability of presence of the species among the sampled sites. (d) Barplot of the relative impurity importance of the variables for the species shown for example.

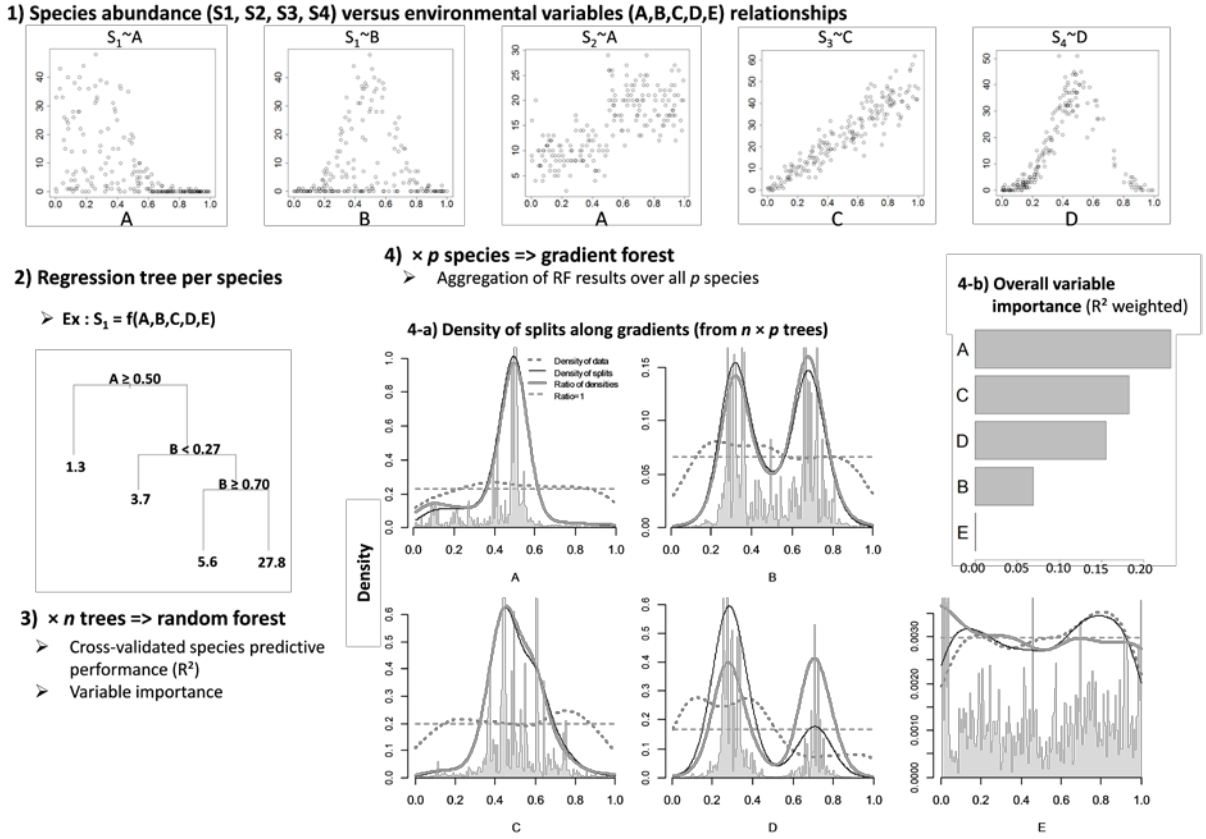
1 Fig 1



2

3

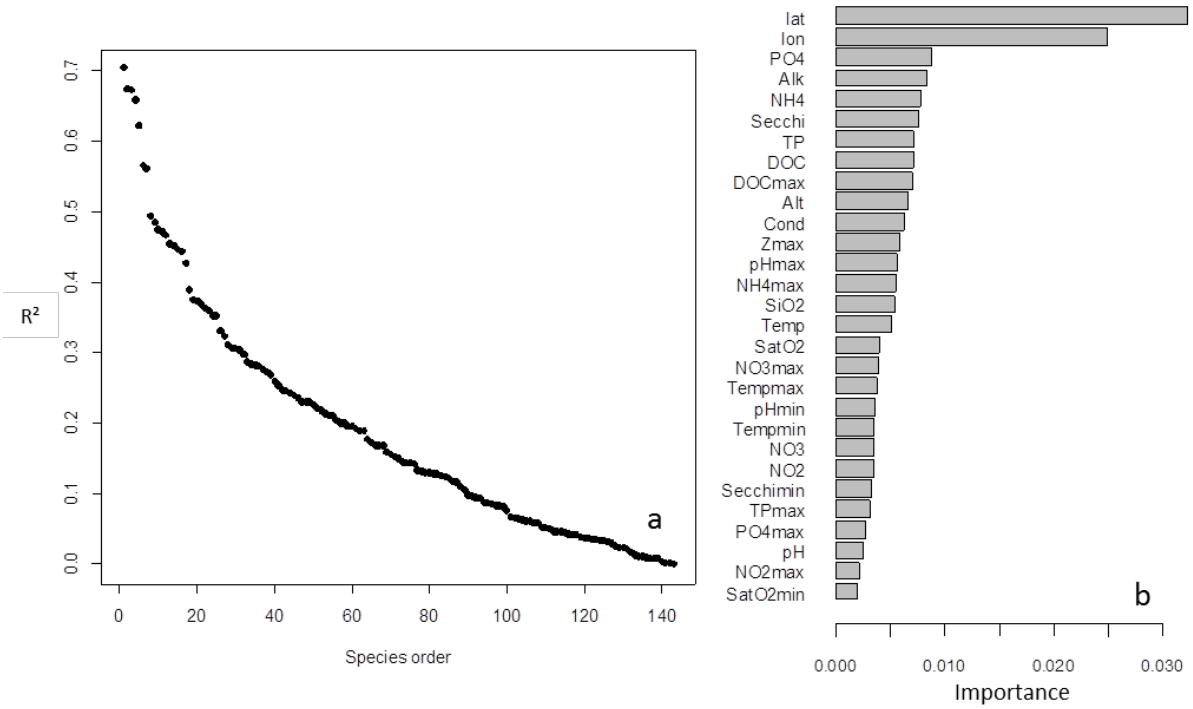
1 Fig 2


1) Species abundance (S1, S2, S3, S4) versus environmental variables (A,B,C,D,E) relationships
S1~A
S1~B
S2~A
S3~C
S4~D
A
B
A
C
D
2) Regression tree per species
Ex : S1 = f(A,B,C,D,E)
A ≥ 0.50
B < 0.27
B ≥ 0.70
1.3
3.7
5.6
27.8
3) × n trees => random forest
Cross-validated species predictive performance (R²)
Variable importance
4) × p species => gradient forest
Aggregation of RF results over all p species
4-a) Density of splits along gradients (from n × p trees)
Density of data
Density of splits
Ratio of densities
Ratio=1
Density
4-b) Overall variable importance (R² weighted)
A
C
D
B
E
0.00 0.05 0.10 0.15 0.20

2

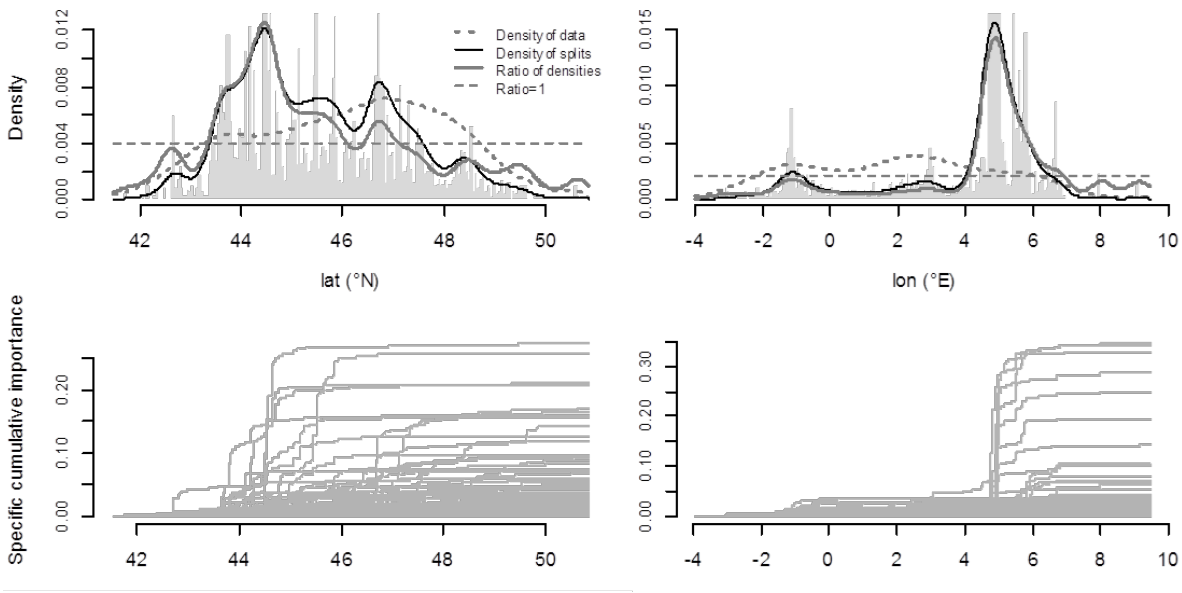
3

1 Fig 3


R²
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
20
40
60
80
100
120
140
Species order
a
lat
lon
PO4
Alk
NH4
Secchi
TP
DOC
DOCmax
Alt
Cond
Zmax
pHmax
NH4max
SiO2
Temp
SatO2
NO3max
Tempmax
pHmin
Tempmin
NO3
NO2
Secchimin
TPmax
PO4max
pH
NO2max
SatO2min
b
0.000
0.010
0.020
0.030
Importance

2

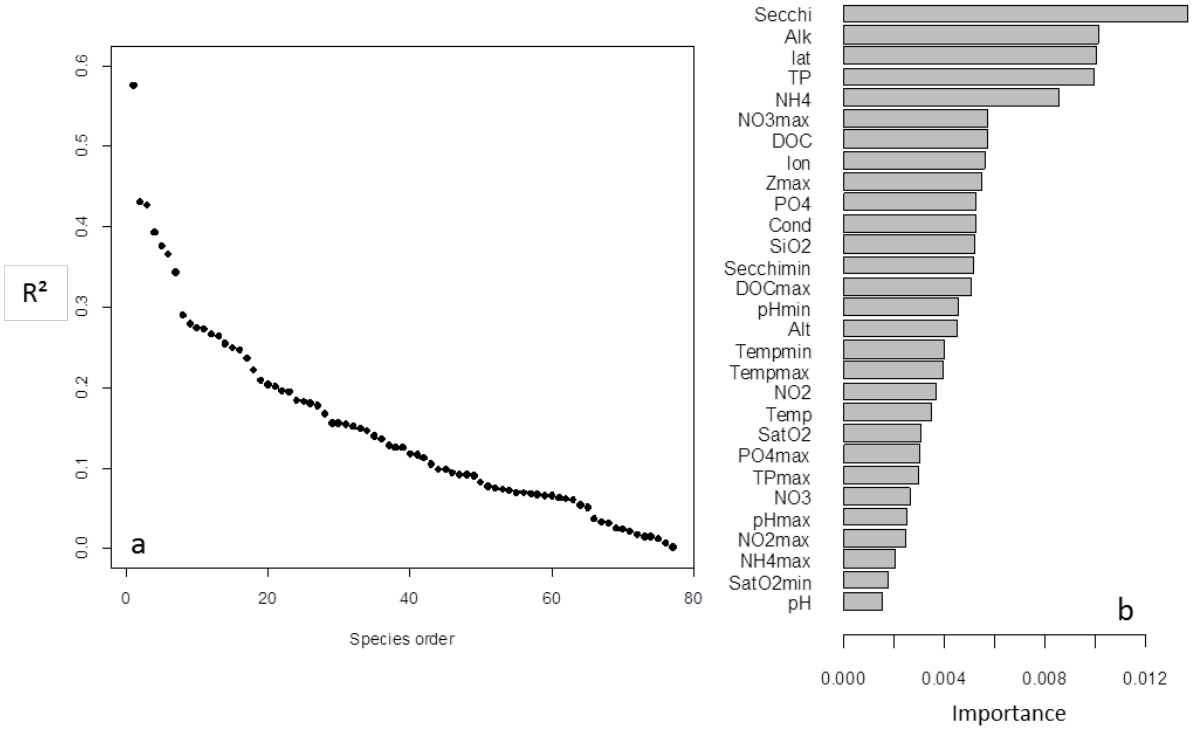
1 Fig 4


Density
Density of data
Density of splits
Ratio of densities
Ratio=1
0.000
0.004
0.008
0.012
42
44
46
48
50
lat (°N)
0.000
0.005
0.010
0.015
-4
-2
0
2
4
6
8
10
lon (°E)
Specific cumulative importance
0.00
0.10
0.20
0.30

2

3

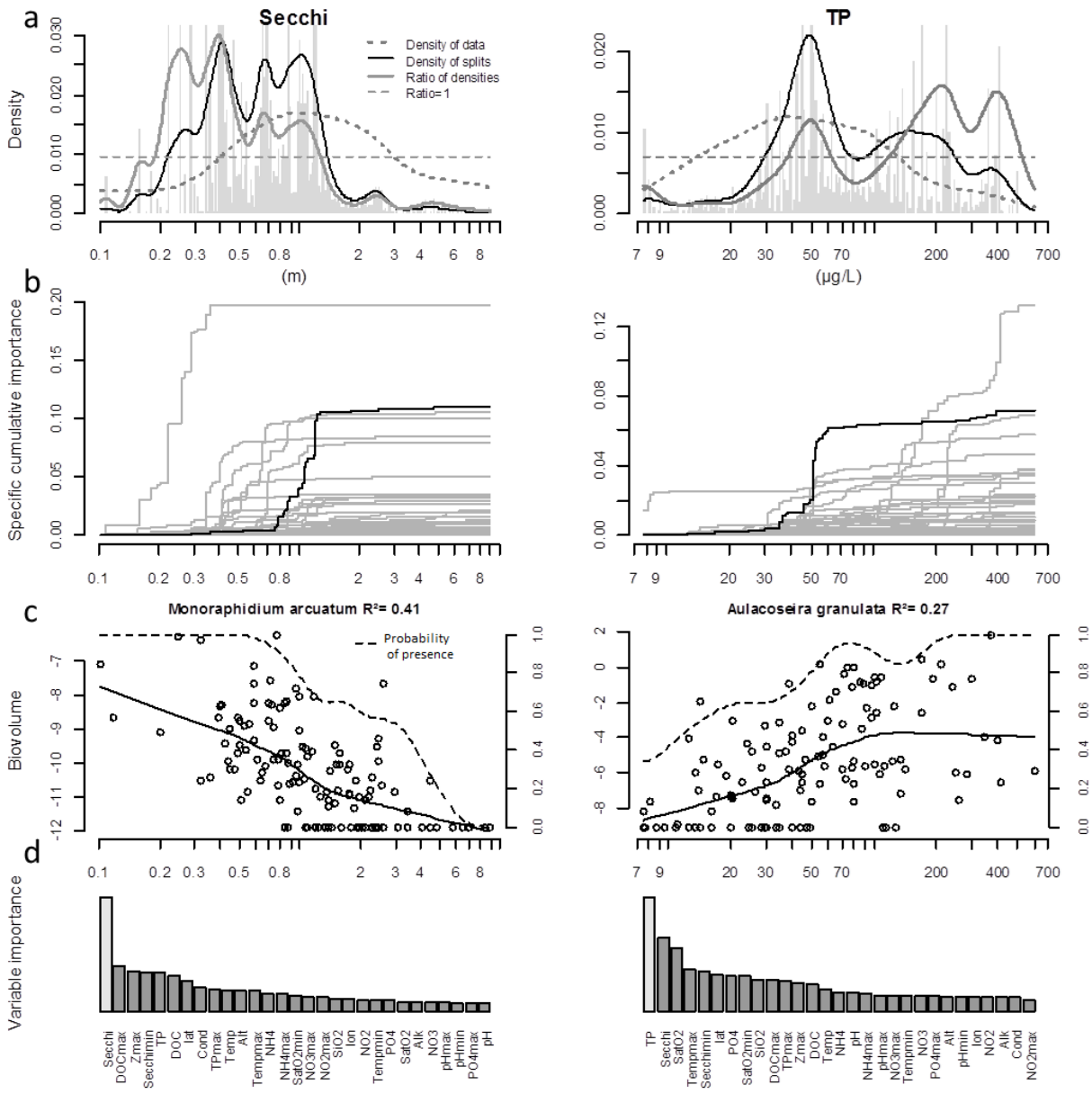
1 Fig 5


R²
0.0
0.1
0.2
0.3
0.4
0.5
0.6
a
0
20
40
60
80
Species order
Secchi
Alk
Iat
TP
NH4
NO3max
DOC
Ion
Zmax
PO4
Cond
SiO2
Secchimin
DOCmax
pHmin
Alt
Tempmin
Tempmax
NO2
Temp
SatO2
PO4max
TPmax
NO3
pHmax
NO2max
NH4max
SatO2min
pH
b
0.000
0.004
0.008
0.012
Importance

2

3

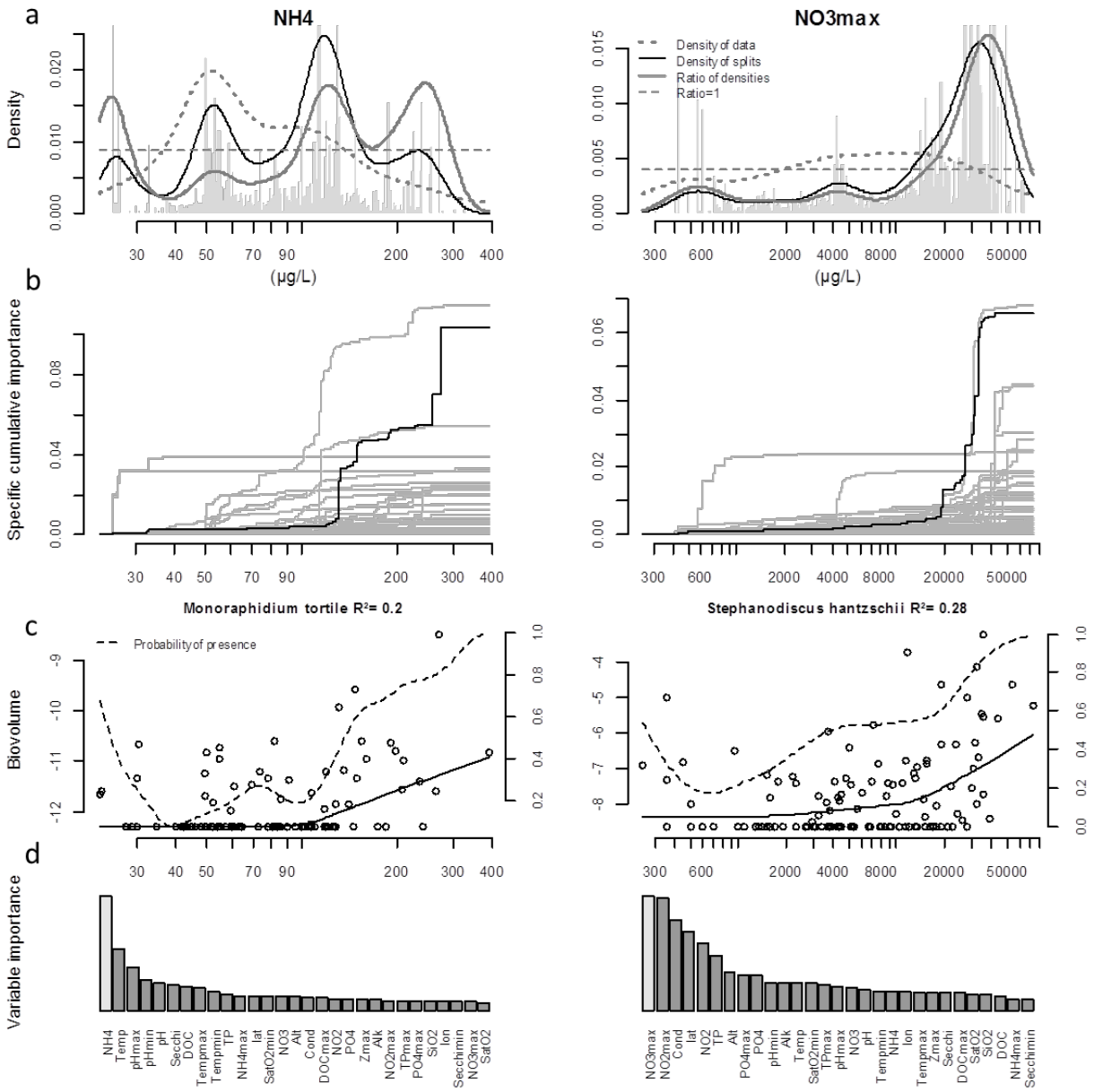
1 Fig 6


a
Secchi
TP
Density
Density of data
Density of splits
Ratio of densities
Ratio=1
(m)
(µg/L)
b
Specific cumulative importance
c
Monoraphidium arcuatum R²= 0.41
Aulacoseira granulata R²= 0.27
Biovolume
Probability of presence
d
Variable importance
Secchi DOCmax Zmax Secchimin TP DOC Iat Cond TPmax Temp Alt Tempmax NH4 NH4max SatO2min NO3max NO2max SiO2 Ion NO2 Tempmin PO4 SatO2 Alk NO3 pHmax pHmin PO4max pH
TP Secchi SatO2 Tempmax Secchimin Iat PO4 SatO2min SiO2 DOCmax TPmax Zmax DOC Temp NH4 pH NH4max pHmax NO3max Tempmin NO3 PO4max Alt pHmin Ion NO2 Alk Cond NO2max

2

3

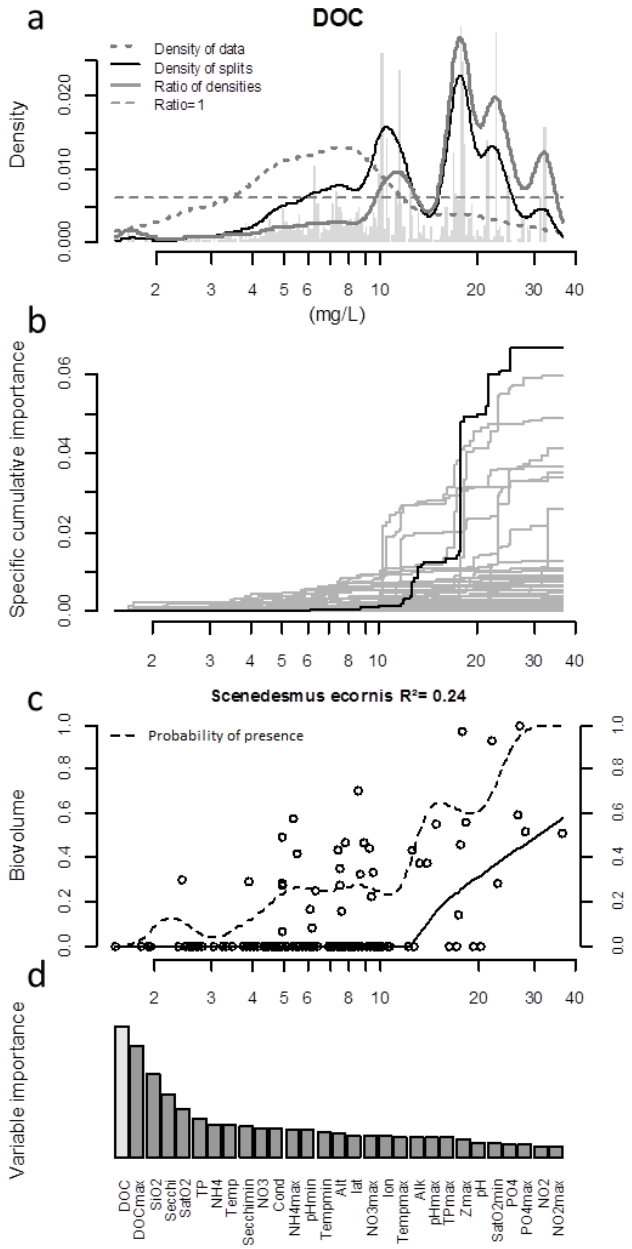
1 Fig 7


a
NH4
NO3max
Density
Density of data
Density of splits
Ratio of densities
Ratio=1
(µg/L)
b
Specific cumulative importance
Monoraphidium tortile R²= 0.2
Stephanodiscus hantzschii R²= 0.28
c
Biovolume
Probability of presence
d
Variable importance
NH4 Temp pHmax pHmin pH Secchi DOC Tempmax Tempmin TP NH4max Iat SatO2min NO3 Alt Cond DOCmax NO2 PO4 Zmax Alk NO2max TPmax PO4max SiO2 Ion Secchimin NO3max SatO2
NO3max NO2max Cond Iat NO2 TP Alt PO4max PO4 pHmin Alk Temp SatO2min TPmax pHmax NO3 pH Tempmin NH4 Ion Tempmax Zmax Secchi DOCmax SatO2 SiO2 DOC NH4max Secchimin

2

3

1 Fig 8


a
DOC
Density of data
Density of splits
Ratio of densities
Ratio=1
Density
0.000
0.010
0.020
2 3 4 5 6 8 10 20 30 40
(mg/L)
b
Specific cumulative importance
0.00
0.02
0.04
0.06
2 3 4 5 6 8 10 20 30 40
c
Scenedesmus ecornis R²= 0.24
Probability of presence
Biovolume
0.0 0.2 0.4 0.6 0.8 1.0
2 3 4 5 6 8 10 20 30 40
d
Variable importance
DOC
DOCmax
SiO2
Secchi
SatO2
TP
NH4
Temp
Secchimin
NO3
Cond
NH4max
pHmin
Tempmin
Alt
lat
NO3max
lon
Tempmax
Alk
pHmax
TPmax
Zmax
pH
SatO2min
PO4
PO4max
NO2
NO2max

2