Accepted to Journal of Biogeography

Using traits to explain interspecific variation in diatom occupancy and abundance across lakes and streams

Running title: Explaining diatom occupancy and abundance

Annika Vilmi[1,2,*], Satu Maaria Karjalainen[2], Jianjun Wang[1,3], Jani Heino[2]

[1]State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences
[2]Freshwater Centre, Finnish Environment Institute
[3]The University of Chinese Academy of Sciences

*Corresponding author: annika.vilmi@gmail.com

Acknowledgements

Annika Vilmi was supported by the National Key Research and Development Program of China (2017YFA0605203), CAS Key Research Program of Frontier Sciences (QYZDB-SSW-DQC043), and the Chinese Academy of Sciences President's International Fellowship Initiative (2018PS0007). We thank all the people who participated in the collection of the diatom data. We would also like to acknowledge two anonymous reviewers, whose suggestions greatly improved the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interests.

Abstract

Aim: To discover how biological traits, ecological preferences and taxonomic relatedness are associated with occupancy and abundance of diatom species across lakes and streams.

Location: Finland.

Taxon: Diatoms.

Methods: We studied 288 diatom species from 492 stream sites and 230 diatom species from 290 lake sites. For each species, we calculated logit-transformed regional occupancy and log-transformed mean local abundance, and further determined biological traits, ecological preferences and taxonomic levels for each species. Boosted Regression Tree (BRT) analysis was used to reveal the linear and non-linear associations of biological, ecological and taxonomic predictors with occupancy or abundance of lake and stream diatoms.

Results: There were strong and positive interspecific occupancy-abundance relationships across both lakes and streams. The BRT models explained more deviances in variation in occupancy and abundance and their relationship for lakes than streams. Biological traits, especially cell size, but also life-form and guild, were the strongest predictors of diatom occupancy and abundance in lakes and streams when controlling for ecological preferences and taxonomic relatedness.

Main conclusions: In general, biological traits were the strongest predictors of occupancy and abundance in both freshwater systems. Species with similar biological traits thus tended to show similar occupancies and abundances. As indicated by lower explained deviances, occupancy and abundance in streams seemed to be more complexly structured than in lakes, suggesting that these two freshwater system types differ in the formation of biodiversity patterns. This difference may be related to the differences in hydrological connectedness between lakes and streams. Understanding how variations in species' occupancy and abundance are formed across various waterbodies is important for meaningful biodiversity conservation.

Keywords

Biodiversity, biological traits, cell size, boosted regression tree, species, taxonomy

Introduction

Species' abundances and distributions form the basis of all patterns in biodiversity across spatial and temporal scales, and thus they have been among the most studied features in biogeography and macroecology (Brown, 1984; Gaston et al., 2000). The main overall conclusion is that abundance and occupancy are positively associated with each other, with regionally widespread species showing high local densities and species occurring only in a limited number of places showing low local densities (Gaston & Blackburn, 2000; Gaston et al., 2000). During the past decades, researchers have asked which is the driver and which is the outcome: occupancy or abundance (Verberk, van der Velde & Esselink, 2010). In the field of metapopulation ecology, it has been suggested that abundant species are likely to disperse more widely than less abundant species, with abundant species simultaneously being less prone to local extinction (Gotelli, 1991; Nee, Gregory & May, 1991). Similarly, species that occupy many sites are more likely to find suitable habitats compared to species that occur at fewer sites, thus decreasing chances of extinction and showing increases in abundance (e.g. Hanski & Gyllenberg, 1993). Thus, the relationship between occupancy and abundance can, in this sense, be seen as a self-fulfilling pattern. However, metapopulation ideas rarely consider differences in species' habitat preferences (Verberk et al., 2010). For instance, when species differ in their environmental preferences, habitat partitioning and competition may exist. In this view, if local resource supply and regional distribution of resources are linked to each other, occupancy and abundance of a species are indirectly linked to each other (Gaston, Blackburn & Lawton, 1997; Verberk et al., 2010). Consequently, if the amount of resources and their regional distributions are positively associated with each other, so are the occupancies and abundances of species utilizing these resources (Verberk et al., 2010).

When trying to reveal mechanisms behind the positive occupancy-abundance relationship, researchers have studied if and how species characteristics (e.g. biological traits, ecological preferences and phylogeny) are associated with variation in abundances and distributions of species (White, Ernest, Kerkhoff & Enquist, 2007; Passy, 2012; Verberk et al., 2010; Verberk, van Noordwijk & Hildrew, 2013; Heino & Tolonen, 2018; Rocha et al., 2018). The connections among species characteristics, abundances and distributions have traditionally been addressed using sites as data points and observing how these variables vary in space, i.e., in a community ecology perspective (Verberk et al., 2013; Heino & Tolonen, 2018). In recent years, novel attempts have been made to disentangle the roles of species characteristics on interspecific variation in occupancy and abundance (Tales, Keith &

Oberdorff, 2004; Tonkin, Arimoro & Haase, 2016; Heino & Tolonen, 2018; Rocha et al., 2018). Instead of using sites as the foci, these approaches have used species as data points. Doing so allows one to study these patterns in a deconstructed way, in their simplest form, by focusing on single species. After all, although the general pattern described between occupancy and abundance is interspecific, abundances and occupancies of single species ultimately drive this interspecific relationship (Verberk et al., 2010). By using species as data points, it is possible to assess if and how much species characteristics account for interspecific variation in occupancy and abundance (Tales et al., 2004; Verberk et al., 2010; Heino & Tolonen, 2018).

Previous research has shown that species-specific niche preferences, illustrated frequently by niche breadth (Brown, 1984; Gregory & Gaston, 2000) and niche position (Hanski, Kouki & Halkka, 1993; Venier & Fahrig, 1996; Gregory & Gaston, 2000), are strong predictors of across-species variation in occupancy and abundance (Tales et al., 2004; Siqueira, Bini, Cianciaruso, Roque & Trivinho-Strixino, 2009; Slatyer, Hirst & Sexton, 2013). Typically, non-marginal species that occupy common habitat conditions show wider distributions than marginal species that occupy uncommon habitat conditions (Heino & Soininen, 2006; Passy, 2012). Niche parameters used in these kinds of studies are generally based on actual environmental measurements, thus reflecting species' realized niches. Based on existing knowledge on species' ecological preferences, one could also assume similar connections between species' ecological preferences and their regional distributions and local densities. Earlier research has also shown that biological traits, such as body size and dispersal mode, can be linked to abundances and distributions across space (De Bie et al., 2012). For instance, small species of microscopic passive dispersers tend to disperse more efficiently to areas further away, thus resulting in higher local abundances and wider regional distributions (Passy, 2012). Overall, the use of biological traits has increased because they help in understanding biological responses to ecological factors (Verberk et al., 2013).

Species characteristics are adaptations to enhance its survival and reproduction, have evolved through time, and are thus partly portrayed by phylogeny (Harvey, 1996). Some features have, however, appeared several times in evolutionary time, which can be seen as, for instance, similar traits between species that show low levels of phylogenetic relatedness (e.g. Rimet & Bouchez, 2012). In addition, it is important to acknowledge that species are in fact results of combinations of several traits (Verberk et al., 2013).

A deeper understanding of biological, ecological and phylogenetic effects on occupancy and abundance is highly important for meaningful biodiversity conservation (e.g.

Gaston et al. 2000). Information on predictors of occupancy and abundance, building blocks of biodiversity, is urgently needed given that human impacts on the abiotic environment and biota are so strong that we are currently living in a new geological era called the Anthropocene (Lewis and Maslin, 2015; Waters et al., 2016). One of the most vulnerable ecosystems to human impacts are the freshwaters (Heino, Virkkala & Toivonen, 2009; Vörösmarty et al., 2010; Vilmi et al., 2017a), which is why occupancy-abundance relationships and the drivers behind them deserve special attention in fresh waters. However, freshwater ecosystems are challenging to study, as they can greatly differ in their geomorphological and hydrological characteristics. Consequently, lakes and streams offer highly different types of habitats for freshwater organisms, with differing restrictions and possibilities for the movement of organisms. For example, lakes tend to be comparatively more isolated while streams are more connected via stream networks, and such levels of hydrological connectedness can thus affect biodiversity patterns in these systems (Lopes et al., 2014; Heino et al., 2015).

In aquatic ecosystems, one of the major biological groups are diatoms (Bacillariophyta). They are microscopic unicellular algae with outer cell structures formed of silica. Diatoms are plentiful in aquatic systems worldwide (Round, Crawford & Mann, 2007; see also Bar-On, Phillips & Milo, 2018), and they contribute strongly to carbon uptake through photosynthesis. Diatoms are also important primary producers in aquatic ecosystems, offering microhabitats and resources for other organismal groups. Ecology and taxonomy of freshwater diatoms is fairly well-known, as they are commonly utilized for bioassessment purposes (e.g. Hering et al., 2006). However, drivers of interspecific variation in occupancy and abundance of diatoms are still insufficiently recognized (Rocha et al., 2018; Vilmi, Tolonen, Karjalainen & Heino, 2019). To date, there is no sufficient knowledge on how variations in occupancy and abundance of diatoms are formed in freshwater systems with differing connectedness at a regional scale.

Here, we examined how diatom species characteristics are correlated with interspecific occupancy and abundance across lakes and streams. We utilized a large dataset of 230 species recorded from 290 lake sites and 288 species recorded from 492 stream sites. This dataset covered a comparatively large spatial extent (i.e. approximately 700 km in diameter), thus portraying occupancy and abundance patterns at a regional scale. Specifically, we aimed to answer the following questions: (1) What is the relationship between occupancy and abundance? (2) Do biological traits, ecological preferences and taxonomic relatedness explain variation in occupancy? (3) Do biological traits, ecological preferences and

taxonomic relatedness explain variation in abundance? (4) Are these findings similar for lakes and streams?

Material and methods

We collected diatom data from the Finnish Environment Institute's diatom database. The database contains monitoring and research data, so the data is comparatively uniform in quality, with similar methods used for sampling and identification. For this study, data on benthic diatom species from 492 stream and 290 lake sites across central and southern Finland were collected (Appendix 1 in Supporting Information). As the data are produced by many people, the data was carefully harmonized in order to ensure its suitability for the purposes of this study. Only samples with at least 400 identified valves were included in the final datasets. To rule out excessive seasonal effects, only samples taken in autumn were considered. A detailed description of the diatom database and data harmonization is presented in Vilmi, Karjalainen & Heino (2017b).

Collecting and processing data

From the harmonized diatom dataset other than species observations and also species that were present at one site only were removed. We then searched for species' environmental preferences on two most important environmental factors related to diatom distributions, i.e., trophic state and pH (Soininen, 2007; Gottschalk & Kahlert, 2012), and used Van Dam's (Van Dam, Mertens & Sinkeldam, 1994) diatom classifications retrieved from OMNIDIA software (Lecointe, Coste & Prygiel, 1993). Since the information was not available for all species, species without pH and trophic state preferences records were removed. In total, there were seven classes for trophic preferences (oligotraphentic, oligo-mesotraphentic, mesotraphentic, meso-eutraphentic, eutraphentic, hypereutraphentic and indifferent) and five classes for pH preferences (acidobiontic, acidophilous, circumneutral, alkaliphilous and alkalibiontic).

We further assigned the remaining species to guilds (low-profile, high-profile, motile and planktonic), and determined their sizes (i.e. biovolume) and life-forms (i.e. colonial or non-colonial). The guilds differ, for instance, in their ways to access resources (e.g. nutrients and light) and to cope with stressors (e.g. grazing or strong currents) (Passy, 2007; Rimet & Bouchez, 2012). Similarly, colonial and non-colonial taxa have different ways to compete for living space and resources (Rimet & Bouchez, 2012). We mainly followed Rimet and Bouchez (2012) when assigning species to the guilds, sizes and life-forms. For species not

included in their list, we used our expert opinion to determine their biological traits. The size classes were as follows: 0-99 μm^3 (class 1), 100-299 μm^3 (class 2), 300-599 μm^3 (class 3), 600-1499 μm^3 (class 4), and > 1500 μm^3 (class 5). The size classes were treated as an ordinal variable in the analyses.

To date, there is no true phylogenetic information for all diatom species included in our study. Thus, higher taxonomic levels were used to describe the taxonomic relationships between species. For each species, its genus, family, order, subphylum and phylum levels were determined. Recent literature (e.g. Lange-Bertalot, Hofmann, Werum & Cantonati, 2017) and AlgaeBase (www.algaebase.org) were used to provide robust taxonomical information. We acknowledge that our measures are just coarse proxies for phylogeny, and that they are silent to more fine-tuned variations between species' relatedness.

Finally, we had abundance data on 288 species from streams and 230 species from lakes, along with the information on their environmental preferences, biological traits and taxonomic relatedness. For each species, we calculated the proportion of sites occupied and made logit-transformations for the proportions. We also calculated species-specific mean abundances across lake or stream sites and transformed the values logarithmically. These transformations were done with the R package stats (R Core Team, 2018).

We then calculated trait distances separately for pH preferences, trophic preferences and biological traits using the Gower distance coefficient with the function gowdis available in the R package FD (Laliberté, Legendre & Shipley, 2014). Using principal coordinates analysis (PCoA; Legendre & Legendre, 2012), vectors were formed from the biological trait and ecological preference distances. This was done using the function pco in the R package labdsv (Roberts, 2016). Taxonomic distances were calculated using the function taxa2dist available in the R package vegan (Oksanen et al., 2018). Taxonomic vectors were formed from the taxonomic distances using the function pco from the R package labdsv (Roberts, 2016). For statistical analyses (see below), we selected the first four vectors describing trophic preferences (TT1–TT4), pH preferences (PT1–PT4), biological traits (BT1–BT4) and taxonomical relatedness (TAX1–TAX4), as the first four vectors were generally the strongest and diverged from the following ones. Thus, in total, we included 16 vectors as explanatory variables in our statistical analyses. We used a vector-based approach because, similar to true species (e.g. Verberk et al., 2013), vectors are combinations of several traits (e.g. Heino & Tolonen, 2018). Examples of interpreting vectors are presented in Appendix 2.

Statistical methods

The Boosted Regression Tree (BRT; Elith, Leathwick & Hastie, 2008) analysis was used to account for both linear and non-linear relationships between occupancy and abundance, and between each of these two variables and the biological, ecological and taxonomic vectors. The theory behind BRTs draws from both traditional statistics and machine learning. Decision tree-based regression models first relate the response to explanatory variables by consecutive, binary splits. The "boosting" means that several simple models are adaptively combined to enhance model accuracy. BRTs handle different types of explanatory variables and outliers well (i.e. there is no need to remove outliers), also acknowledging possible interactions between explanatory variables (Elith et al., 2008). BRTs produce predictor-specific results, so the relative effect of each predictor on the response variable can be easily detected. A partial dependency plot illustrates the relative influence of an explanatory variable on the response variable once the average effects of other explanatory variables have been acknowledged. The relative influence indicates the effect of each explanatory variable, being an extension of Friedman's (2001) relative influence to boosted models normalized to sum to 100 (Elith et al., 2008). The total explained deviance is the overall explanatory power of the final BRT model.

The BRT analyses were performed with the R package dismo (Hijmans, Phillips, Leathwick & Elith, 2017). The following parameters were used for the final BRTs: family = "gaussian", tree.complexity = 5, learning.rate = 0.001, and bag.fraction = 0.5. These parameter values were based on suggestions by Elith et al. (2008). The complexity of individual trees is determined by 'tree.complexity'. The weight applied to individual trees is set by 'learning.rate' (Hijmans et al., 2017). Decreasing the 'learning.rate' increases the number of trees required, shrinking the contribution of each tree and enhancing the reliability of the final model's estimation (Elith et al., 2008). The proportion of observations used in variable selection, controlling for stochasticity, is determined by 'bag.fraction' (Hijmans et al., 2017). As the results are affected by the determination of the used parameters, we also ran BRTs testing with different parameter values (e.g. tree.complexity = 3, tree.complexity = 7, learning.rate = 0.005, bag.fraction = 0.75). We subsequently found that the results based on the final BRTs were similar with the options tested. The total explained deviances by the final models were calculated as follows: (mean total deviance - mean residual deviance) / mean total deviance. The occupancy-abundance relationship was also studied using BRTs. As the interpretation of BRT results is easier if accompanied by visual inspections, boxplots and scatterplots were made with the package ggplot2 (Wickham 2016) in R. Stream and lake datasets were analyzed independently.

Results

Relationship between occupancy and abundance

The BRT analysis showed that occupancy and abundance were positively associated in both lakes and streams (Fig. 1). Scatterplots also similarly showed a rather positive relationship between occupancy and abundance in both lakes and streams (Fig. 2). The relationships between occupancy and abundance were generally linear, but saturated at some point. Abundance explained 42% and 36% of deviances for occupancy for lakes and streams, respectively.

Predictors of variation in occupancy

The BRT analysis explained approximately 24% and 18% of deviance in occupancy in lakes and streams, respectively. The BRT plots in general showed that not all illustrated relationships between predictor variables and occupancies were completely linear (Fig. 3).

For lakes, there were three predictor variables which all had over 10% relative influences on variation in occupancy (Fig. 3a). They were BT2, BT1 and PT1. BT2 was generally negatively and BT1 and PT1 non-linearly associated with the response variable. BT4 and all taxonomic vectors as well were slightly associated with variation in occupancy, with relative influences exceeding 5%. TT and PT vectors in general had weak, if any, relationships with occupancy (Appendix 3). Based on interpretations of the vectors (Appendix 2), species with similar sizes and life-forms tended to show comparatively similar occupancies (BT2 and BT1, respectively; see also Fig. 2).

For variation in occupancy in streams, there were two rather strong predictors, BT2 and BT1, which both had relative influences of over 15% (Fig. 3b). BT2 was negatively and BT1 was positively associated with variation in occupancy. BT4, BT3 and several taxonomic trait vectors along with PT1 had relative influences between 5 and 10% (Fig. 3b, Appendix 3). Based on interpretation of the vectors (Appendix 2), species with similar sizes tended to show similar occupancies (BT2). This was also likely the case with life-forms (BT1). Scatterplots in Fig. 2 partly showed similar patterns.

Predictors of variation in abundance

The BRT analysis explained approximately 41% and 24% of variation in mean local abundance in lakes and streams, respectively. The BRT plots again showed that not all

illustrated relationships between predictor variables and occupancies were completely linear (Fig. 3c and d).

For variation in abundance in lakes, there was one strong predictor, BT2, which had a relative influence of 41% (Fig. 3c). BT2 was mainly negatively associated with abundance. TAX1 also just exceeded the relative influence of over 10%, having a non-linear relationship with abundance. BT1, BT3, BT4 and TAX2 had relative influences between 5 and 10%. Earlier interpretations of vectors showed that BT2 is strongly related to cell sizes (Appendix 2). Thus, there was a strong pattern that species similar in sizes showed similar abundances. In scatterplots, this was most evident regarding the ends of the size-continuum (i.e. smallest vs. largest species; Fig. 2).

For streams, BT2 again had the largest relative influence (i.e. 33%; Fig. 3d). BT1 as well had a comparatively large relative influence, exceeding 15%. BT2 was negatively associated with abundance, while BT1 showed a linear and positive relationship with abundance. Also, all taxonomic vectors had relative influences of over 5% (Fig. 3d, Appendix 3). The importance of BT2 and BT1 indicate that biological trait similarity leads to similar levels of abundances (see also Fig. 2).

Discussion

We reported generally positive relationships between occupancy and abundance of diatoms in both lakes and streams. Although we found that the shape of the relationship was rather similar in both freshwater systems, occupancy and abundance were, interestingly, more closely linked with each other in lakes than in streams, as indicated by the amount of total explained deviances. In addition to the occupancy-abundance relationship, BRTs predicting occupancy and abundance in lakes similarly produced higher explained deviances than BRTs predicting occupancy and abundance in streams. Thus, it seems that interspecific variations in occupancy and abundance in lakes were structured in a more straight-forward manner, while the formation of occupancy and abundance in streams was a more complex process, or was related to traits not addressed in this study. Lakes are nevertheless comparatively isolated and generally more stable systems compared to streams (e.g. Dent, Cumming & Carpenter, 2002), indicating that there might not be so much noise in formation of occupancy and abundance patterns in lakes. Streams, on the other hand, tend to show higher hydrological connectedness, which may complicate the formation of clear patterns in occupancy and abundance (e.g. Shurin, Cottenie & Hillebrand, 2009). Earlier studies have shown that lakes and streams partly foster different diatom communities (Soininen & Weckström, 2009;

Kahlert & Gottschalk, 2014), and that they show differences in species richness and ecological uniqueness patterns (Vilmi et al., 2017b). Keeping these earlier observations in mind, it is possible that the differences we now found regarding occupancy and abundance in lakes and streams may be related to levels of connectedness. Hydrological connectivity has indeed been shown to affect biodiversity through dispersal (Lopes et al., 2014; Heino et al., 2015). Overall, the positive relationship between occupancy and abundance for both lakes and streams is in line with earlier observations on various freshwater organism groups (Verberk et al., 2010; Passy, 2012; Heino & Tolonen, 2018; Rocha et al., 2018).

When examining the roles of species' biological traits, ecological preferences and taxonomic relatedness for variation in occupancy and abundance in lakes and streams, we found that biological traits were the strongest predictors for both response variables and in both freshwater systems. Although the total explained deviances in the models were not very high, species with similar traits tended to show similar levels of occupancy and abundance, irrespective of the type of aquatic system. Earlier studies have also observed that biological traits can affect occupancy and/or abundance of freshwater species (Verberk et al., 2010; Passy, 2012; Heino & Grönroos, 2014; Rocha et al., 2018). First of all, we found rather clear indications that the size of the species is tied to its occupancy and abundance, with smaller species showing larger abundances and occupancies than larger species. The effect of cell size on diatom occupancy and abundance has been reported in earlier studies as well (Passy, 2012; Rocha et al., 2018), and the interpretation of the observed size-related pattern is the same in this study: for diatom species, the smaller the better, regarding prospects for local survival (cf. abundance) and regional dispersal (cf. occupancy). We also showed that the life-form, and also likely the guild of a species, is partly connected to its occupancy and abundance. For instance, colonial species tended to be more abundant and occupy more sites than non-colonial species.

Taxonomic relatedness was not a strong predictor of occupancy and abundance. There were, however, subtle indications that species with taxonomic similarity may have rather similar levels of occupancy and abundance. Similar small effects of taxonomic aspects on occupancy and abundance of freshwater macroinvertebrates were found by Heino and Tolonen (2018). As biological traits and phylogenetic relatedness are partly linked to each other (Harvey, 1996), it is also possible that the small taxonomic effects we found may in fact reflect effects of biological traits other than those addressed here. It is also advisable to keep in mind that taxonomy is only a coarse proxy for phylogeny, and the use of actual phylogenetic information might increase its impact on occupancy and abundance patterns.

Although two of the most important environmental aspects for structuring freshwater diatom communities, i.e., pH and nutrients (Soininen, 2007; Gottschalk & Kahlert, 2012), were indirectly considered here, no uniform evidence was found for the effects of species' environmental preferences on their occupancy and abundance. Species' pH preferences were variably associated with occupancy in lakes and streams, but in general, different environmental preferences of species were not clearly related to their occupancy and abundance. This finding is partly contradictory to earlier studies showing that species' 'true' niche parameters can greatly affect interspecific variation in occupancy and/or abundance of freshwater organisms (Tales et al., 2004; Verberk et al., 2010; Heino & de Mendoza, 2016; Heino & Tolonen, 2018; Rocha et al., 2018; Vilmi et al. 2019). The difference between these earlier investigations and our current study is that we used existing literature knowledge on species' pH and trophic preferences, while 'true' niche parameters are calculated based on actually measured environmental variables. In this regard, these two complementary ways do not measure the same aspects of the importance of environmental conditions, but both approaches nevertheless intend to shed light into how environmental conditions where species live (or prefer to live) affect their occupancies and abundances. As our current dataset was rather large, we could not have collected enough comparable information on local environmental variables to calculate reliable 'true' niche parameters. It is nevertheless possible that inclusion of niche parameters instead of environmental preferences in our models could have supported earlier findings that have reported clear effects of niche parameters on occupancy and abundance. Also, it is important to understand that reaching optimum environmental conditions is constrained by regional obstacles and processes, such as historical factors and dispersal limitation (e.g. Soininen, 2007). Thus, it may be that not all species we studied had reached their preferred environmental conditions, but instead were living in sub-optimal conditions. It is also possible that the environmental preferences that have been devised for the diatom species in other regions do not work in our focal study system. Ultimately, local adaptations to surrounding conditions may appear quite rapidly in diatom species (e.g. Sjöqvist, Godhe, Jonsson & Kremp, 2015).

Overall, the explained deviances by all BRTs were comparatively low (except for abundance in lakes). Occupancy was less well explained than abundance, and stream BRTs showed higher residual deviances than the respective lake models. Thus, it seems that abundance was better connected to our predictor variables, while formation of occupancy is either a more complex process and/or is related to traits other than those we actually investigated. Similarly, as discussed previously, occupancy and abundance variations in

streams were likely to be more strongly affected by factors and processes not investigated here compared to those of lakes. High residual variations may be a consequence of spatial patterns in environmental conditions, or stochastic colonization and extinction of different species (e.g. Verberk et al., 2010). These processes could not be evaluated in our interspecific analysis of occupancy and abundance, but they could probably be included in the study of single species' distribution modelling studies.

Figures

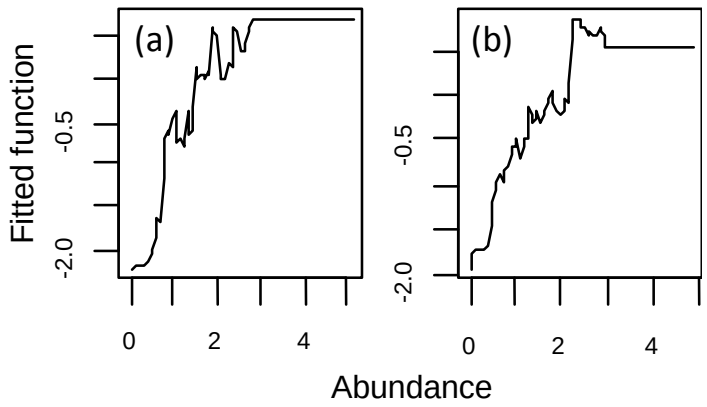

Fig. 1. Partial dependency plots of occupancy-abundance of the BRT analysis for lake (a) and stream (b) diatom species. The plots represent the fitted functions and the effects of log-transformed mean abundance on logit-transformed occupancy. For lakes, the BRT analysis explained 42.0% of deviances for occupancy. For streams, the BRT analysis explained 35.7% of deviances for occupancy.

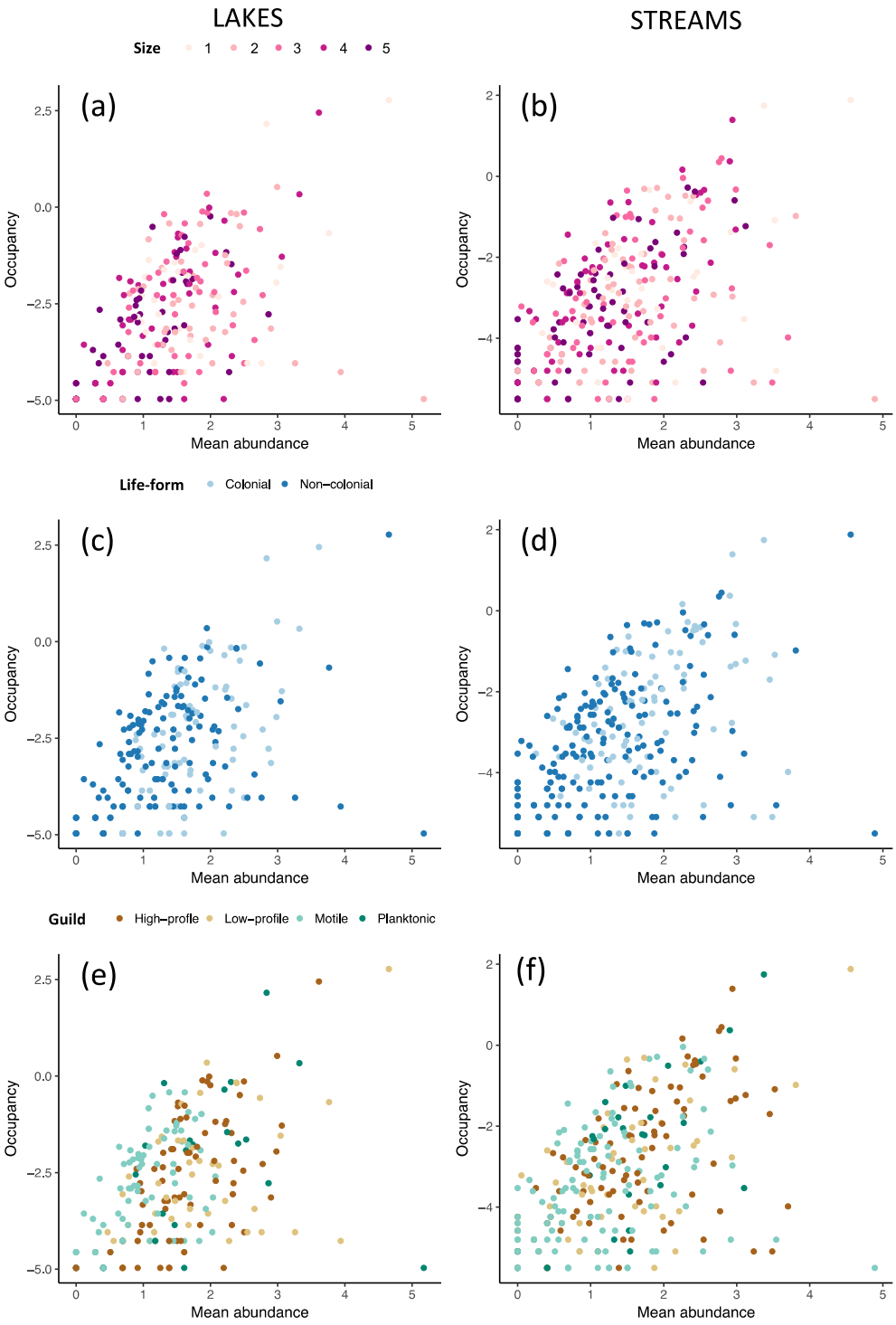

Fig. 2. Scatterplots illustrating the relationships of logit-transformed occupancy and log-transformed mean abundance for diatom species in lakes (left side) and streams (right side). Plots also illustrate the variation in occupancy and abundance according to size classes (a and b), life-forms (c and d) and guilds (e and f).

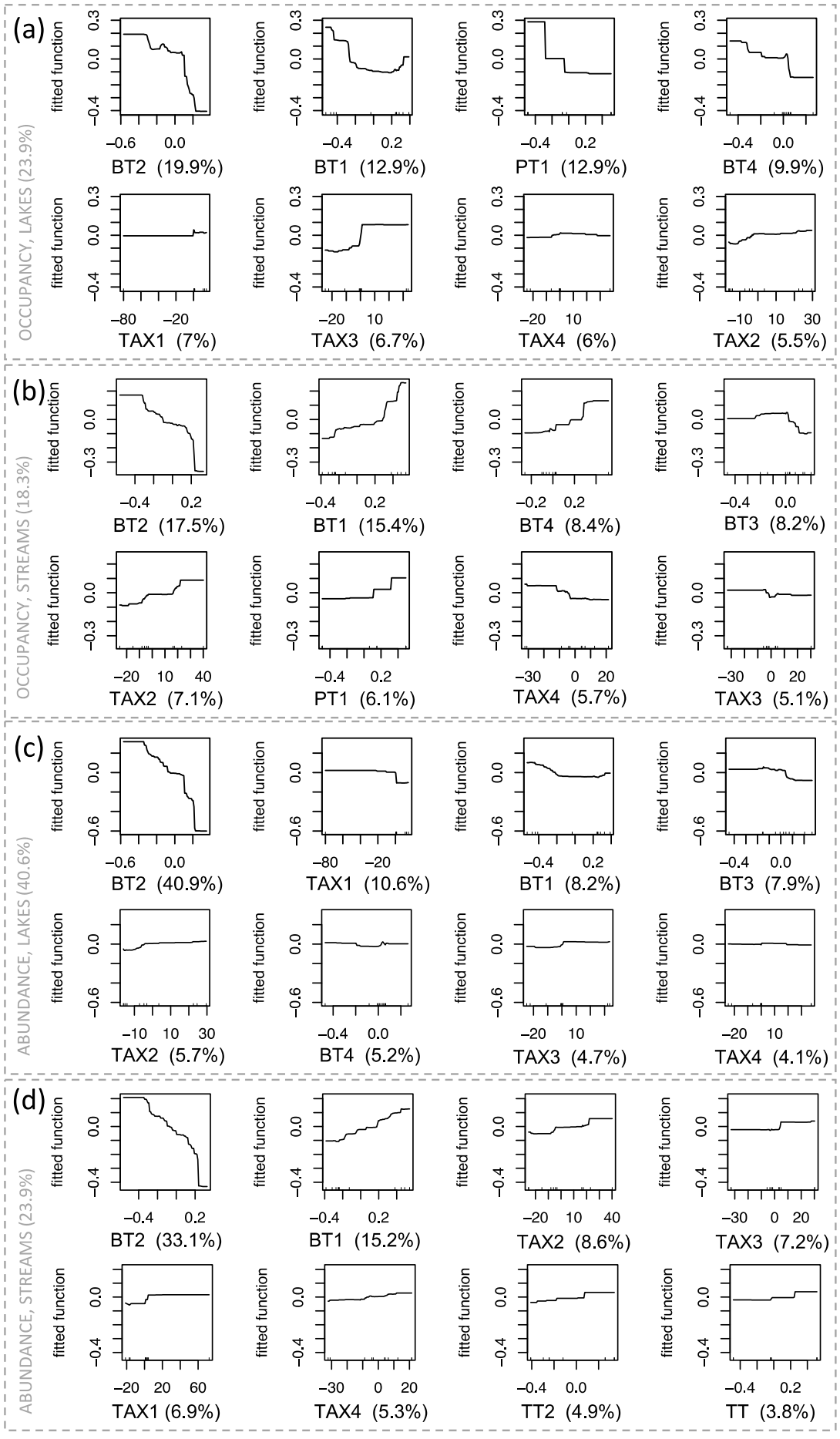
(a)
OCCUPANCY, LAKES (23.9%)
fitted function
BT2 (19.9%)
BT1 (12.9%)
PT1 (12.9%)
BT4 (9.9%)
TAX1 (7%)
TAX3 (6.7%)
TAX4 (6%)
TAX2 (5.5%)
(b)
OCCUPANCY, STREAMS (18.3%)
BT2 (17.5%)
BT1 (15.4%)
BT4 (8.4%)
BT3 (8.2%)
TAX2 (7.1%)
PT1 (6.1%)
TAX4 (5.7%)
TAX3 (5.1%)
(c)
ABUNDANCE, LAKES (40.6%)
BT2 (40.9%)
TAX1 (10.6%)
BT1 (8.2%)
BT3 (7.9%)
TAX2 (5.7%)
BT4 (5.2%)
TAX3 (4.7%)
TAX4 (4.1%)
(d)
ABUNDANCE, STREAMS (23.9%)
BT2 (33.1%)
BT1 (15.2%)
TAX2 (8.6%)
TAX3 (7.2%)
TAX1 (6.9%)
TAX4 (5.3%)
TT2 (4.9%)
TT (3.8%)

Fig. 3. Partial dependency plots of the eight most important predictor variables in the BRT analysis for diatom species occupancies and abundances in lakes (a and c) and streams (b and d). For occupancy, the whole BRT analyses explained 23.9% and 18.3% of deviances in lakes and streams, respectively. For abundance, the whole BRT analyses explained 40.6% and 23.9% of deviances in lakes and streams, respectively. The plots represent the fitted functions and relative influences of the predictor variables for logit-transformed occupancy and log-transformed abundance. BT = biological trait vector, PT = pH preference trait vector, TAX = taxonomic vector, TT = trophic preference vector.

References

Bar-On, Y.M., Phillips, R, & Milo, R. (2018) The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, doi: 10.1073/pnas.1711842115

Brown, J.H. (1984) On the Relationship between Abundance and Distribution of Species. *The American Naturalist*, **124**, 255–279.

De Bie, T., De Meester, L., Brendonck, L., Martens, K., Goddeeris, B., Ercken, D., Hampel,H., Denys, L., Vanhecke, L., Van der Gucht, K., Van Wichelen, J., Vyverman, W. & Declerck, S.A.J. (2012) Body size and dispersal mode as key traits determining metacommunity structure of aquatic organisms. *Ecology Letters*, **15**, 740–747.

Dent, C.L., Cumming, G.S. & Carpenter S.R. (2002) Multiple states in river and lake ecosystems. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **357**, 635–645.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Friedman, J.H. (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

Gaston, K.J. & Blackburn, T.M. (2000) *Pattern and Process in Macroecology*. Blackwell, Oxford.

Gaston, K.J., Blackburn, T.M., Greenwood, J.J.D., Gregory, R.D., Quinn, R.M. & Lawton, J.H. (2000) Abundance–occupancy relationships. *Journal of Applied Ecology*, **37**, 39–59.

Gaston, K.J., Blackburn, T.M. & Lawton, J.H. (1997) Interspecific abundance-range-size relationships: an appraisal of mechanisms. *Journal of Animal Ecology*, **66**, 579–601.

Gotelli, N.J. (1991) Metapopulation models: the propagule rain, the rescue effect, and the core-satellite hypothesis. *The American Naturalist*, **138**, 768–776.

Gottschalk, S., Kahlert, M. (2012) Shifts in taxonomical and guild composition of littoral diatom assemblages along environmental gradients. *Hydrobiologia*, **694**, 41–56.

Gregory, R.D. & Gaston, K.J. (2000) Explanations for commonness and rarity in British breeding birds: separating resource use and resource availability. *Oikos*, **88**, 515–526.

Hanski, I. & Gyllenberg, M. (1993) Two general metapopulation models and the core-satellite species hypothesis. *The American Naturalist*, **142**, 17–41.

Hanski, I., Kouki, K. & Halkka, A. (1993) Three explanations of the positive relationship between distribution and abundance of species. In R.E. Ricklefs & D. Schluter (Eds.),

Species diversity in ecological communities: historical and geographical perspectives (pp. 108–116). Chicago, IL: University of Chicago Press.

Harvey, P.H. (1996) Phylogenies for ecologists. *Journal of Animal Ecology*, **65**, 255–263.

Heino, J. & de Mendoza, G. (2016) Predictability of stream insect distributions is dependent on niche position, but not on biological traits or taxonomic relatedness of species. *Ecography*, **39**, 1216–1226.

Heino, J. & Grönroos, M. (2014) Untangling the relationships among regional occupancy, species traits and niche characteristics in stream invertebrates. *Ecology and Evolution*, **4**, 1931–1942.

Heino, J., Melo, A. S., Siqueira, T., Soininen, J., Valanko, S., & Bini, L. M. (2015). Metacommunity organisation, spatial extent and dispersal in aquatic systems: Patterns, processes and prospects. *Freshwater Biology*, **60**, 845–869.

Heino, J. & Soininen, J. (2006) Regional occupancy in unicellular eukaryotes: a reflection of niche breadth, habitat availability or size-related dispersal capacity? *Freshwater Biology*, **51**, 672–685.

Heino, J. & Tolonen, K.T. (2018) Ecological niche features override biological traits and taxonomic relatedness as predictors of occupancy and abundance in lake littoral macroinvertebrates. *Ecography*, **41**, doi: 10.1111/ecog.03968.

Heino, J., Virkkala, R., & Toivonen, H. (2009) Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions. *Biological Reviews*, **84**, 39–54.

Hering, D., Johnson, R.K., Kramm, S., Schmutz, S., Szoszkiewicz, K. & Verdonschot, P.F.M. (2006) Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. *Freshwater Biology*, **51**, 1757–1785

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2017) dismo: Species Distribution Modeling. R package version 1.1-4. URL https://CRAN.R-project.org/package=dismo

Kahlert, M., & Gottschalk, S. (2014) Differences in benthic diatom assemblages between streams and lakes in Sweden and implications for ecological assessment. *Freshwater Science*, **33**, 655–669.

Laliberté, E., Legendre, P. & Shipley, B. (2014) FD: measuring functional diversity (FD) from multiple traits, and other tools for functional ecology. R package version 1.0-12. URL https://CRAN.R-project.org/package=FD

Lange-Bertalot, H., Hofmann, G., Werum, M. & Cantonati, M. (2017) *Freshwater Benthic Diatoms of Central Europe: Over 800 Common Species Used in Ecological Assessment*. M. Cantonati, M.G. Kelly & H. Lange-Bertalot (Eds.). Koeltz Botanical Books.

Lecointe, C., Coste, M. & Prygiel, J. (1993) Omnidia: software for taxonomy, calcula-tion of diatom indices and inventories management. *Hydrobiologia*, **269/270**, 509–513.

Legendre, P. & Legendre, L. (2012) *Numerical Ecology* (3rd ed.). Amsterdam: Elsevier.

Lewis, S.L. & Maslin, M.A. (2015) Defining the Anthropocene. *Nature*, **519**, 171–180.

Lopes, P.M., Bini, L.M., Declerck, S.A.J., Farjalla, V.F., Vieira, L.C.G., Bonecker, C.C., Lansac-Toha F.A., Esteves F.A. & Bozelli, R.L. (2014) Correlates of Zooplankton Beta Diversity in Tropical Lake Systems. *PLoS ONE*, **9**, e109581. https://doi.org/10.1371/journal.pone.0109581.

Nee, S., Gregory, R.D. & May, R.M. (1991) Core and satellite species: theory and artefacts. *Oikos*, **62**, 83–87.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.M., Szoecs, E. & Wagner, H. (2018) vegan: Community Ecology Package. R package version 2.5-2. URL https://CRAN.R-project.org/package=vegan

Passy, S.I. (2007) Diatom ecological guilds play distinct and predictable behavior along nutrient and disturbance gradients in running waters. *Aquatic Botany*, **86**, 171–178.

Passy, S.I. (2012) A hierarchical theory of macroecology. *Ecology Letters*, **15**, 923–934.

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rimet, F. & Bouchez, A. (2012) Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. *Knowledge and Management of Aquatic Ecosystems*, **406**, 01.

Rocha, M.P., Bini, L.M., Siqueira, T., Hjort, J., Grönroos, M., Lindholm, M., Karjalainen, S.M. & Heino, J. (2018) Predicting occupancy and abundance by niche positions, niche breadth and body size in stream organisms. *Oecologia*, **186**, 205–216.

Roberts, D.W. (2016) labdsv: Ordination and Multivariate Analysis for Ecology. R package version 1.8-0. URL https://CRAN.R-project.org/package=labdsv

Round, F., Crawford, R. & Mann, D. (2007) *The Diatoms* (5th ed). Cambridge: Cambridge University Press.

Shurin, J.B., Cottenie, K. & Hillebrand, H. (2009) Spatial autocorrelation and dispersal limitation in freshwater organisms. *Oecologia*, **159**, 151–159.

Siqueira, T., Bini, L.M., Cianciaruso, M.V., Roque, F.O. & Trivinho-Strixino, S. (2009) The role of niche measures in explaining the abundance–distribution relationship in tropical lotic chironomids. *Hydrobiologia*, **636**, 163–172.

Sjöqvist, C., Godhe, A., Jonsson, P.R. & Kremp, A. (2015) Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea-Baltic Sea salinity gradient. *Molecular Ecology*, **24**, 2871-2885.

Slatyer, R.A., Hirst, M. & Sexton, J.P. (2013) Niche breadth predicts geographical range size: a general ecological pattern. *Ecology Letters*, **16**, 1104–1114.

Soininen, J. (2007) Environmental and spatial control of freshwater diatoms – a review. *Diatom Research*, **22**, 473–490.

Soininen, J., & Weckström, J. (2009) Diatom community structure along environmental and spatial gradients in lakes and streams. *Fundamental and Applied Limnology*, **174**, 205–213.

Tales, E., Keith, P. & Oberdorff, T. (2004) Density-range size relationships in French riverine fishes. *Oecologia*, **138**, 360–370.

Tonkin, J.D., Arimoro, F.O. & Haase, P. (2016) Exploring stream communities in a tropical biodiversity hotspot: biodiversity, regional occupancy, niche characteristics and environmental correlates. *Biodiversity and Conservation*, **25**, 975–993.

Van Dam, H., Mertens, A. & Sinkeldam, J. (1994) A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. *Netherlands Journal of Aquatic Ecology*, **28**, 117–133.

Venier, L.A. & Fahrig, L. (1996) Habitat availability causes the species abundance-distribution relationship. *Oikos*, **76**, 564–570.

Verberk, W.C.E.P., van Noordwijk, C.G.E. & Hildrew, A.G. (2013) Delivering on a promise: integrating species traits to transform descriptive community ecology into a predictive science. *Freshwater Science*, **32**, 531–547.

Verberk, W.C.E.P., van der Velde, G. & Esselink, H. (2010) Explaining abundance–occupancy relationships in specialists and generalists: a case study on aquatic macroinvertebrates in standing waters. *Journal of Animal Ecology*, **79**, 589–601.

Vilmi, A., Alahuhta, J., Hjort, J., Kärnä, O-M., Leinonen, K., Rocha, M.P., Tolonen, K.E., Tolonen, K.T. & Heino, J. (2017a) Geography of global change and species richness in the North. *Environmental Reviews*, **25**, 184–192.

Vilmi, A., Karjalainen, S.M. & Heino, J. (2017b) Ecological uniqueness of stream and lake diatom communities shows different macroecological patterns. *Diversity and Distributions*, **23**, 1042–1053.

Vilmi, A., Tolonen, K.T., Karjalainen, S.M. & Heino, J. (2019) Niche position drives interspecific variation in occupancy and abundance in a highly-connected aquatic system. *Ecological Indicators*, **99**, 159–166.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R. & Davies, P.M. (2010) Global threats to human water security and river biodiversity. *Nature*, **467**, 555–561.

Waters, C.N., Zalasiewicz, J., Summerhayes, C., Barnosky, A.D., Poirier, C., Galuszka, A., Cearreta, A., Edgeworth, M., Ellis, E.C., Ellis, M., Jeandel, C., Leinfelder, R., McNeill, J.R., Richter, D. de B., Steffen, W., Syvitski, J., Vidas, D., Wagreich, M., Williams, M., Zhisheng, A., Grinewald, J., Odada, E., Oreskes, N. & Wolfe, A.P. (2016) The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science*, **351**, aad2622.

White, E.P., Ernest, S.K.M., Kerkhoff, A.J. & Enquist, B.J. (2007) Relationships between body size and abundance in ecology. *TRENDS in Ecology and Evolution*, **22**, 323–330.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Biosketch

Annika Vilmi is a postdoctoral researcher at the Nanjing Institute of Geography and Limnology (Chinese Academy of Sciences). Her research focuses on freshwater biodiversity patterns. The research group works in the field of freshwater biodiversity research and development of bioassessment approaches.

Author contributions: A.V. and J.H. conceived the ideas; A.V. and S.M.K. collected the data; A.V. analysed the data; A.V. led the writing with contributions from S.M.K., J.W. and J.H.

Supporting Information

Appendix 1. Map of sites where the biological data was collected from.

Appendix 2. Interpretation of vectors.

Appendix 3. Detailed results of BRT analyses.