Citation for the published paper:
Keck F, Franc A, Kahlert M. (2018) Disentangling the processes driving the biogeography of freshwater diatoms: A multiscale approach. *Journal of Biogeography*. Volume: 45, Number: 7, pp 1582-1592.
http://dx.doi.org/10.1111/jbi.13239.

Disentangling processes driving freshwater diatoms biogeography: a multiscale approach

François Keck (francois.keck@slu.se)[1], Alain Franc (alain.franc@inra.fr)[2] and Maria Kahlert (maria.kahlert@slu.se)[1]

[1] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P. O. Box 7050, 750 07 Uppsala, Sweden

[2] UMR BIOGECO, Institut National de la Recherche Agronomique (INRA), F-33610 Cestas, France

Keywords: Biogeography, Diatoms, Dispersal limitation, Mass effect, Environmental response, Traits, Multiscale, Multivariate spatial analysis

Short running-title: Biogeographical processes in diatoms

Corresponding author: François Keck (francois.keck@gmail.com)

Number of words in the Abstract: 283

Number of words in main body of the paper, from the Introduction through the Biosketch: 4850

Number of references: 81

Abstract

Aim 1/ Disentangle the effects of local environmental conditions and space at multiple scales in order to uncover and explain biogeographical signals in microbial communities of freshwater diatoms. 2/ Assess the role of species traits related to dispersal capacities in diatoms biogeography.

Location Rivers and streams of Sweden sampled on a total area of 450000 km².

Methods We used partial redundancy analyses and variation partitioning coupled with spatial eigenfunctions analyses to separate the effects of environment and space at three different scales on diatoms community structure. As community analyses might give limited insight into species-specific mechanisms we also performed variation partitioning species by species and the amount of variance explained by each fraction was related to species cell length and attachment to the substrate.

Results We found a strong effect of environment and space at large scale suggesting environmental filtering and dispersal limitation over large distances. Additionally, species-based analyses uncovered an effect of space at fine scale suggesting possible species-specific mass-effect dynamic or species sorting driven by missing environmental factors varying at fine scale. Environmental and spatial effects varied strongly among species but this variation was not explained by cell length and attachment to the substrate.

Main conclusions Our results confirm the importance of both environment and dispersal-related processes in controlling diatoms community structures and bring new perspectives on the role of dispersal in shaping microbial communities at biogeographical scale. Dispersal being a species-specific mechanism, species association are not necessarily conserved from site to site and species-based approaches can detect different spatial patterns than traditional community analyses do. Contrary to what was reported in connected metacommunities at small-scale, interspecific traits variation seem to be of limited importance for long-distance overland dispersal in diatoms.

Introduction

Inferring ecological processes underlying species distribution and community structures in time and space is a major concern of ecology (Levin, 1992). It has long been recognised that the geographical patterns of biodiversity are the result of both species sorting, i.e. filtering of species by local abiotic and biotic factors, and dispersal (Hubbell *et al.*, 1999; Ricklefs, 2004). Spatial ecology has highlighted the importance of the interplay between those processes in explaining variation among communities at two main conceptual scales. First, at local scale, metacommunities which are usually defined as a set of communities that are linked by dispersal of multiple interacting species (Wilson, 1992; Leibold *et al.*, 2004; Holyoak *et al.*, 2005; Martiny *et al.*, 2006). Second, at larger scale, biogeography which often includes an evolutionary perspective (MacArthur, 1972; Ricklefs, 2004). Because they focus on the same types of processes, it can be difficult to draw a clear boundary between metacommunity ecology and biogeography (Gonçalves-Souza *et al.*, 2014). However, both approaches came to the conclusion that niche-based and dispersal-based processes are interacting closely in determining community structures. Additionally, they stressed that the relative roles of niche and dispersal-based processes vary among scales, ecosystems and organisms (Astorga *et al.*, 2012; De Bie *et al.*, 2012; Heino *et al.*, 2015, 2017).

Diatoms are unicellular eukaryotic algae encompassing a huge taxonomic diversity and contributing a significant share of marine and freshwater primary production. Because they are microscopic organisms (ca. 2–500µm), diatoms have traditionally been considered ubiquitously distributed. This idea is based on the Baas-Becking hypothesis which states that microbial species are not limited by dispersion and are cosmopolitan. This has been coined by the sentence "*everything is everywhere*, but *the environment selects*" (Baas Becking, 1934; Fenchel & Finlay, 2004) from which the second part has often been disregarded (De Wit & Bouvier, 2006). Because of this misinterpretation, the validity of the Baas-Becking hypothesis has been heavily discussed during the last hundred years, confounding being cosmopolitan and ubiquitous. However, a growing body of research is now suggesting that microbial biogeography is not the result of species sorting only, but that regional factors and historical dynamics (origin, dispersal and extinctions of species) are also important to determine the structure of local microbial communities (Martiny *et al.*, 2006; Telford *et al.*, 2006; Lindström & Langenheder, 2012).

Similar conclusions have been drawn from recent diatom studies where authors used variation partitioning (Borcard *et al.*, 1992) to separate the environmental and the spatial signals (Potapova & Charles, 2002; Soininen *et al.*, 2004; Vyverman *et al.*, 2007; Heino *et al.*, 2010). These studies showed that a significant part of the variation in diatom communities can be explained by a pure spatial effect (i.e. spatial variation without the environmental component). However, it can be challenging to clearly identify which ecological mechanisms generate a spatial effect (Fortin & Dale, 2005). Indeed, spatial patterns can be generated by dispersal limitation (by distance or physical barriers), by mass-effect (massive dispersal compensating for sub-optimal environmental condition), and environmental filtering (if important spatially structured variables are missing). Hence, spatial patterns must be interpreted cautiously, taking into account the geographical scale, the dispersal ability of the target organisms and the geographical variation of environmental conditions.

Diatoms are usually assumed to disperse passively (Kristiansen, 1996). Within a river system, individuals will naturally be transported by water flow, from upstream to downstream (Atkinson, 1988). This was confirmed by metacommunity studies taking into account flow directionality (Liu *et al.*, 2013; Dong *et al.*, 2016). Diatoms can also disperse from one water body to another and travel long overland distances through animal vectors like water birds (Proctor, 1959) and by airborne dispersal (Schlichting, 1969). Although these dispersal modes are documented and critically important for the colonization of new sites (Kristiansen, 1996), very little is known about their quantitative aspect, their geographical range, and how they impact communities composition dynamics across scales.

Furthermore, it is known that body size and dispersal mode are important traits that can modulate the exchange rates of individuals between sites and then affect the structure of communities across scales. So far, their role has been mainly investigated by comparing very large taxonomic groups (e.g. bacteria, macroinvertebrates, fishes; De Bie *et al.*, 2012) or generic dispersal types (e.g. active and passive dispersers; Shurin *et al.*, 2009; Soininen, 2016), but see Schwalb *et al.* (2015). In diatoms, the role of traits and guilds was studied in highly connected metacommunities (Liu *et al.*, 2013; Algarte *et al.*, 2014; Vilmi *et al.*, 2017). However, one may also expect that differences in dispersal capacities among diatoms species could have an effect on the structure of diatom communities at larger scales. Interspecific variation in dispersal-related traits like body size or cell attachment to the substrate could affect niche and dispersal related processes and partly explain diversity distribution patterns at different scales.

Multi-scale spatial analyses offer a powerful framework because patterns observed in communities and the ecological processes that generate them are not occurring at one single scale. Thus, multi-scale variation partitioning based on spatial eigenfunction analyses (Dray *et* *al.*, 2012) could potentially be an effective means of decomposing patterns and separating processes. However, unravelling the impact of environment and space from metacommunities to biogeographical regions require extensive datasets with good representativeness of each investigated spatial scale with precise and harmonized diatom identification, and a good assessment of environmental factors.

In this paper, we aim to investigate the role of environment and space in determining diatoms communities across multiple scales to uncover and explain biogeographical signals. We analysed a large dataset of 616 diatoms communities sampled across Sweden. This high-quality dataset combines harmonized fine-resolution taxonomy and wide geographic coverage and allows us to study diatoms spatial patterns at three different scales, possibly bridging metacommunities and biogeography (see Heino *et al.*, 2017). We used variation partitioning and Moran eigenvector maps (MEMs) to separate the effect of environment and space across scales. We performed variation partitioning using both multivariate community-based analyses (Verleyen *et al.*, 2009; Bennett *et al.*, 2010) and multiple univariate species-based (i.e. species-by-species) analyses (Ovaskainen *et al.*, 2017). We completed our study by relating the amount of explained variation of the different partitions to cell length and cell attachment to the substrate. Assuming these two traits have a significant impact on species dispersal capacity, we expected the amount of explained variation to be different between small and large cells and between attached and non-attached ones. These analyses will improve our understanding of ecological processes that drive diatom community structures from local to biogeographical scale.

Material and Methods

Sampling

The study area encompasses entire Sweden covering a surface of 450000 km². The climate of this area is subarctic north of 60°N and temperate in the southern part. We used a large dataset of 616 sites sampled in stream and rivers. Selected sites are separated at least by 5 km and none are located downstream of another. The location of each site is represented on the Figure 1. The sampling points are densely distributed allowing to map and study fine, medium and large scale patterns.

Diatom abundance data

Diatom samples were collected and prepared following standard protocols (European Committee for Standardization, 2003, 2004). Samples were collected during autumn, coinciding with the maximal development of diatom communities. At least 400 valves were counted and identified to the lowest taxonomic level possible, usually species, with the aid of literature in the Swedish standard (see Havs och Vattenmyndigheten, 2016). Species names were harmonized according to conventions adopted by the Nordic-Baltic Network for Benthic Algae in Freshwater (Kahlert & Albert, 2005). Unidentified valves and valves identified to genus level were not included in the analysis. To promote statistical robustness in the species-based analyses, species present in less than 15 sites have been excluded. The cleaning procedure resulted in a loss of 35 rare genera and ≈ 8% of valves/site. Counts were converted to relative abundances and were Hellinger-transformed (i.e. division of cell values by row sums and square root of the quotient; Legendre & Gallagher, 2001) prior to the analyses, as this transformation is recommended for variation partitioning based on linear ordination methods (Peres-Neto *et al.*, 2006). Correspondence analysis was performed on non-transformed relative abundance data.

Environmental descriptors

We extracted 26 parameters from the Swedish national database (http://miljodata.slu.se/) to describe the chemical and physical local conditions of the sites. Nonparametric random forest algorithm was used to complete the missing values (14.9% of the data). In order to avoid redundancies and high correlations (> 0.8) a subset of 10 parameters was selected, based on expert knowledge of their importance for diatoms. The final set included concentrations in total nitrogen, total phosphorus and total organic carbon, pH, alkalinity, water colour (measured as absorbance at 420 nm), turbidity, altitude, upstream catchment area and proportion of urbanized land cover in the upstream catchment area. Parameters were log transformed where necessary.

Spatial descriptors

We generated spatial descriptors in order to test the effect of space on diatom diversity across different scales using Moran Eigenvector Maps (MEMs; Dray *et al.*, 2006). This method consists on eigenvector decomposition of a spatial weighting matrix. The spatial weighting matrix (SWM) describes the spatial relationship among the different sites and is the product of a connectivity matrix (defining which sites are connected) and a weighting matrix (providing the intensity of the connections). Since there is no straightforward way to choose a SWM definition, Dray *et al.* (2006) recommend to use a data-driven approach to select the best SWM among different candidates.

We used the following procedure to generate 45 sets of MEMs which will be used later in the analyses as potential spatial descriptors. First, we generated 45 SWM based on 9 connectivity schemes combined with 5 weighting schemes. Connectivity matrices were created using the influence circle graph method (Legendre & Legendre, 2012) with 9 different radii ranging from 100 to 260 km. We weighted the connectivity matrices by a function based on Euclidean distance among sites: $f = 1-(d/dmax)^\wedge\alpha$ where d is the distance between sites and $dmax$ is the maximum distance between two connected sites. The five weighting schemes correspond to 5 different values for the coefficient α: 0, 1, 2, 3 and 4. Second, we computed the MEMs separately for each SWM. In each set, we kept only the MEMs associated with positive eigenvalues (i.e. exhibiting positive spatial autocorrelation). Third, we estimated the scale represented by each MEM. This is done for a given MEM by computing its spline-correlogramm and then finding the minimum distance for which a negative autocorrelation is estimated. Therefore, each MEM can be classified as fine scale (<30km), medium scale (30-100km) or large scale (>100km).

Community-based analyses

We selected the set of MEMs among the 45 candidates with the best fit to the community matrix using a multivariate AIC-like method (Godínez Domínguez & Freire, 2003) as recommended by Dray *et al.* (2006). To limit the number of variables included in the model and prevent type I error inflation, we selected a subset of spatial predictors within the selected set of MEMs using a forward selection procedure with double stopping criterion (Blanchet *et al.*, 2008). A similar forward selection approach was used to select the best environmental descriptors.

We used correspondence analysis and principal component analysis to examine the spatial variation of diatoms communities and environmental parameters, respectively. Partial redundancy analyses (pRDA; Borcard *et al.*, 1992) were used to examine the pure effects of space and environment on the community matrix. Multivariate analyses results were mapped using ordinary kriging to interpolate sites scores over a 5000 × 5000 cells grid covering the study area.

Next, we performed variation partitioning based on multiple pRDAs to quantify and test the variation in the community matrix that can be related to environmental predictors and spatial predictors (divided in fine, medium and large scale). Hierarchical partitioning of the shared fractions (Legendre *et al.*, 2012) was used to take into account the orthogonality of the spatial eigenfunction submodels. The variation explained by each fraction is reported using unbiased adjusted R^2 (Peres-Neto *et al.*, 2006). The individual fractions (i.e. the pure effects of environment and space at fine, medium and large scale) were tested by permutation.

Species-based analyses

Species-based variation partitioning has been carried out using the same procedure as for community-based variation partitioning. However, in this case, the procedure is applied independently, species by species. Hence, for each species, appropriate environmental and spatial descriptors were selected and used in a variation partitioning procedure based on multiple partial regressions. This approach generates a collection of N variation partitionings, with N being the number of species under study.

Performing one variation partitioning per species allowed us to relate the amount of variance explained by each fraction to the species traits. As the fractions of variance range from zero to one, we used beta regression models (Simas *et al.*, 2010; Soininen, 2014, 2016) to assess the effects of individual size and attachment to the substrate on the proportion of variance explained by pure spatial effects at fine, medium and large scale. We used the trait dataset from Rimet and Bouchez (2012) to test the effects of two traits: the cell length and the attachment to the substrate which is a binary categorical variable stating whether the species is attached to the substrate (by their valve face, their girdle view or by a pad or a stalk of mucilage) or non-attached (free moving or planktonic).

Software

We performed all the statistical analyses with R 3.3.1 software (R Development Core Team, 2016). MEMs were computed with the package adespatial (Dray *et al.*, 2016). Spline correlograms were estimated with the ncf package (Bjørnstad, 2016). Kriging was computed using gstat (Pebesma, 2004). We performed multivariate ordinations and variation partitioning with vegan (Oksanen *et al.*, 2016) and the additional code provided in Legendre *et al.* (2012). Finally, beta regression models were adjusted using the betareg package (Cribari-Neto & Zeileis, 2010).

Results

Community-based analysis

The variable selection procedure selected the 10 environmental parameters and 31 MEMs (19 at large scale, 12 at medium scale and none at fine scale). Both diatom communities and environmental parameters appeared to be spatially structured and correlated (Figure 2, see Appendix S1 for species loadings values) with an important contrast between the north and the south (axes 1 of the two analyses) and between the northern west mountains and the northern east coast (axes 2 of the two analyses). Northern species commonly found in boreal region of Sweden were *Eunotia bilunaris* var. *mucophila* Lange-Bertalot & Nörpel, *Peronia fibula* (Brébisson & Kützing) Ross, *Eunotia exigua* var. *exigua* (Brébisson & Kützing) Rabenhorst, *Eunotia incisa* var. *incisa* Smith & Gregory and *Eunotia rhomboidea* Hustedt, which all are known to prefer acid or very acid conditions. Species commonly found in agricultural regions in the south of Sweden were *Amphora pediculus* (Kützing) Grunow, *Rhoicosphenia abbreviate* (Agardh) Lange-Bertalot, *Navicula tripunctata* (O. Müller) Bory and *Eolimna subminuscula* Moser, Lange-Bertalot & Metzeltin, known for their preference of nutrient rich conditions, and

Achnanthidium lauenburgianum (Hustedt) Monnier, Lange-Bertalot & Ector, a species typical for very high alkalinity, such as occurring on Gotland, a large limestone island in the south of Sweden. Typical for the mountain area in the west of Sweden are *Gomphonema lateripunctatum* Reichardt & Lange-Bertalot, *Rossithidium nodosum* (Cleve) Aboal, *Sellaphora stroemii* Mann, *Encyonopsis cesatii* (Rabenhorst) Krammer and *Hannaea arcus* (Ehrenberg) Patrick, species that are commonly found in these mountains, but are rarer in Central Europe.

Mapping the two first axes of the pRDA of the community matrix explained by space after removing the effect of environment (Figure 3A) revealed large patches spread across the study area. On the other hand, the pRDA of the community matrix explained by environmental variables after removing the effect of space (Figure 3B) displayed a first axis where communities are primarily discriminated by altitude, the load of nutrients and alkalinity and a second axis mostly driven by the catchment area, pH and organic matter.

Results of variation partitioning are presented in Figure 4. The whole model including environment and space explained 24% of the total community matrix variation. Among the pure individual fractions, environment explained 6% of variance (p-value < 0.001). The effect of space was the strongest at large scale (6% of explained variation, p-value < 0.001) but was also detected at medium scale (1%, p-value < 0.001). No effect of space was found at fine scale, since no MEMs were selected. An important joint effect of environment and space (i.e. spatially-structured component of environmental factors) was found at large scale (10%).

Species-based analysis

The species-approach generated 278 variation partitioning analyses. The distribution of each fraction is presented in Figure 5. The whole model including environment and space explained, on average, 17.9% of the species variation. The effect of environment was found rather low compared to the community-based approach (mean explained variation = 2.4%). The effect of space was found to be the strongest at fine scale (mean explained variation = 5.5%), then at large scale (mean explained variation = 3.8%) and finally at medium scale (mean explained variation = 2.3%). Similarly to the community-based analysis, joint effect of space and environment was found at large scale (3.3%) while this effect was found below 0.5% at medium and fine scale.

Globally, the explained variation in each fractions varied strongly among species. However, the studied traits failed to explain this variation. Beta regressions pseudo-R^2 ranged from 0.001 for the medium scale pure spatial fraction model to 0.029 for the fine scale pure spatial fraction model. Cell length, attachment mode and their interaction were not significantly associated to any of the tested fractions (all p-values < 0.05, Appendix S2).

Discussion

Diatoms communities have been considered to be mainly structured by local environmental conditions for a long time. However recent studies have stressed the importance of spatial processes at different scales to understand species assemblage variation and pointed out the need for fine-grained multi-scales studies on large geographical area (Soininen, 2007; Vyverman *et al.*, 2007; Verleyen *et al.*, 2009; Heino *et al.*, 2010; Vilmi *et al.*, 2017). We found support for both environmental and spatial control of diatoms communities.

Both the community-based and the species-based analyses highlighted the control of environment on species distribution and communities' composition. The part of variation that can be related to environmental factors is traditionally interpreted as species sorting (Cottenie, 2005). These results are consistent with previous works which have clearly demonstrated the primary role of local environmental factors in determining benthic algae assemblage (Leland & Porter, 2000). The strong differences of ecological preferences among species and the direct effect of environment on diatom communities is also the reason why diatoms are widely recognized as efficient biological indicators of environmental conditions (Lowe & Pan, 1996; Stevenson & Smol, 2003).

In addition, pure spatial patterns were detected both in the community-based and the species-based analyses. Spatial patterns should be interpreted carefully as they can be the result of different dispersal-related processes: limitation or homogenization (Heino *et al.*, 2015); but can also reflect species sorting if important spatially structured variables are missing (Legendre & Legendre, 2012). In this study, we included as many environmental descriptors as possible, from dissolved nutrients to local land cover. The descriptors we used are known to be important drivers of diatom communities and we assume they give a general picture of the chemical and physical local conditions of the sites under study. However, it should be noted that some identified sources of variation were not covered in this paper. For example, local substrate composition and disturbance regime data are missing while they can have an important effect on diatom assemblages (Jüttner *et al.*, 1996; Passy, 2001; Cardinale *et al.*, 2006). Furthermore, biotic interactions are part of the species sorting perspective and Göthe *et al.* (2013) have shown that biotic predictors can capture a significant part of the variation in diatoms communities. Including the biotic component to diatoms biogeographical studies is a real challenge given the complexity of interactions networks but would probably improve our models and understanding of the processes underlying community structures both in time and space. At larger scale, climate (Pajunen *et al.*, 2016; Jyrkänkallio-Mikkola *et al.*, 2017) and geology (Lavoie *et al.*, 2010) are also important to take into account although it can be challenging to decouple their direct effects from pure spatial effects. Overall, the total amount of variance explained by our models may seem low (e.g. R^2_{adj} = 24% for the community-based analyses).

However, low amount of explained variation are frequently reported in empirical works using variation partitioning to study diatoms (e.g. Heino *et al.*, 2010; Algarte *et al.*, 2014, respectively 26% and 16.8%). Low explanatory power of models in microbial ecology are discussed by Vilmi *et al.* (2017) and can be explained by the complexity of communities with many rare species, the lack of important explanatory variables or the high frequency of stochastic events that cannot be integrated into models.

Community-based analysis revealed a significant pure spatial effect in diatoms communities. This result is in line with others research studying diatom biogeographical patterns at regional (e.g. Heino *et al.*, 2010; Bottin *et al.*, 2014) and at global scale (e.g. Vyverman *et al.*, 2007). In our study, most of the pure spatial effect is detected at large scale (>100km) suggesting that dispersal limitation plays an important role over large distances. Dispersal limitation is known to be a prominent factor at continental and inter-continental scales (Potapova & Charles, 2002; Vyverman *et al.*, 2007; Verleyen *et al.*, 2009; Bennett *et al.*, 2010) but it has also been suggested that diatoms dispersal limitation can be significant at smaller scale in boreal streams (Soininen *et al.*, 2004; Heino *et al.*, 2010). Our study is in line with these results and shows that dispersal limitation can have a noticeable impact from a distance of a few hundreds kilometres.

At finer scale, community-based analysis found no pure spatial structure. The absence of pure spatial structures at fine scale (<30km) can be interpreted as an evidence that dispersal-related processes are non-existent or obscured by other processes. Other studies have stressed that species sorting is likely to dominate at fine scale (e.g. Verleyen *et al.*, 2009; Bennett *et al.*, 2010) and have reported an effect of space that increase linearly and monotonously with distance (cf. Soininen, 2007). However, unlike the community approach, the species-based analysis detected pure spatial structures at fine scale. As mentioned above, pure spatial patterns can be the result of species sorting in a fragmented landscape as important environmental factors varying at fine scale may have been missed. Alternatively, a pure spatial effect at fine scale can be the signature of mass-effect (i.e. net emigration of individuals from source sites to sink sites where environmental conditions are sub-optimal). So far, mass-effect has only been considered in highly connected meta-community networks. However, given the microscopic nature of diatoms and the fundamental idea that microbial species have strong dispersal capacities, our results raise the question of the scale and context to which the mass-effect applies with diatoms. Additional work should be done to investigate further overland dispersal capacities of diatoms and the possible implications for community dynamics.

The differences found between the community-based approach and the species-based approach are obviously important. They can be explained by the fact that the species-based approach decompose the community matrix species-by-species while the community-based approach uses partial RDAs that focus on patterns of covariation between species. Large regions detected by community level analyses are large scale geographical units inferred from the combined distribution of taxa. However, the fact that diatoms species populations can disperse does not guarantee that species associations and combinations are conserved from site to site at fine scale. Hence, the use of multivariate community-based analyses that focus on the total community variation can be unsuitable to capture species-specific mechanisms. This is likely to be strengthened if the pool of species is large and several species are ecologically interchangeable (i.e. share the same niche). Thus, further to the fact that dispersal is a species-specific mechanism, the large taxonomic diversity of diatoms and the strong role of stochasticity and instability in shaping microbial communities assemblage (Stegen *et al.*, 2012) could explain the limits of multivariate analyses in detecting spatial effects at the finest scales. Finally, it should also be noted that the community-based approach is more likely to detect large scale patterns than the species-based approach for species having a restricted distribution range. In our study, this bias is limited by excluding species occurring in less than 15 sites, ensuring that the species included in our analyses have a sufficiently large distribution.

Overall, we found a high degree of joint effect of environment and space on diatom communities. Joint effect is the result of a strong covariation of the environmental parameters with space which is clearly depicted in Figure 2. Interestingly, we found this joint effect at large scale indicating that the main components of environment investigated in this study are autocorrelated over large distances. This result also suggests that environmental local heterogeneity has a lower effect at fine scale. Studies using variation partitioning frequently report a large part of variation explained by the joint effect of environment and space. This fraction is hard to interpret in terms of ecological mechanisms because the effects of space and environment are inextricable (Cottenie, 2005). The detected biological patterns delineate three large biogeographical regions: the south, the north and the mountainous region located in the north-eastern part of Sweden. Sites of these regions have contrasted environmental conditions (water chemistry, anthropogenic pressures, and climate) and are separated by potential physical barriers (long overland distances, mountains). Hence, it is likely that both species sorting and dispersal limitation act together at large biogeographical scale.

Species-based analysis highlighted that the pure effect of environment and space on diatom distribution can vary greatly from species to species. Previous studies have shown that body size and other dispersal related traits have important effect on metacommunity structure among different groups of organisms (Hillebrand *et al.*, 2001; Drakare *et al.*, 2006; De Bie *et al.*, 2012). More recently, Liu *et al.* (2013) and Dong *et al.* (2016) have shown that flow-mediated downward dispersal has a stronger impact on loosely attached diatoms. However, our results do not support the idea that dispersal-related traits could explain the species variation explained by pure environmental or spatial effects. This could be explained by the fact that we used different traits and we are investigating communities at larger scale and in a different context where dispersal is probably mainly mediated through animals and atmospheric deposition. Thus, it can be argued that the tested traits were not appropriate or sufficient to represent overland dispersal capacities and that other physiological, morphological or life history missing traits would be important to include. Moreover, it is also possible that at large scale diatoms dispersal capacities are comparable among species. Cell length and attachment may not vary enough to significantly discriminate diatom species in function of their dispersal capacities. Finally it should be noted that traits may impact both species sorting and dispersal-related mechanisms. For example, while we investigate the potential role of cell attachment on dispersal capacity, Heino and Soininen (2006) reported the niche breadth of non-attached diatoms species to be significantly larger than the niche breadth of attached species. They argued that it could be explained by the better capacity of attached species to persist in disturbed environment rather than a difference of dispersal capacity between attached and non-attached species. This example illustrate how intricate niche requirements and dispersal capacities can be in explaining habitat occupancy.

Our results have also practical implications since diatoms are commonly used as bioindicators. Dispersal limitation over large biogeographical regions can limit the power of biological indices and can motivate the development of regional indices (Soininen, 2004). Sweden, for example, is using diatom biomonitoring indices at a national scale (Naturvårdsverket, 2007). However, further investigations are required, in particular because anthropogenic effects have not been separated in our study but included as any other ecological factor. Recent studies have also investigated the potential of species distribution modelling techniques to test the role of biotic interactions (Bottin *et al.*, 2016) or climate (Pajunen *et al.*, 2016) on diatom communities structures. Including explicitly multiscale dispersal-related mechanisms into these models could improve their performances and reliability in estimating species realized niche (Guisan & Thuiller, 2005; Boulangeat *et al.*, 2012).

Acknowledgements

We thank Stefan Hellgren for his assistance with GIS and three anonymous reviewers for providing thoughtful comments on this manuscript.

Data Accessibility statement

Diatoms and environmental data were extracted from the Swedish national database (http://miljodata.slu.se/).

Biosketch

François Keck is a postdoctoral researcher at the Swedish University of Agricultural Sciences. His research interests include dynamics of freshwater communities and the development and implementation of computational methods for community ecology and ecological assessment.

Alain Franc is a senior scientist at the National Institute for Research in Agronomics (INRA) working in Bordeaux (France). His research interests include macroecology, ecological modeling, community assembly, focusing on connections between ecology and mathematical models.

Maria Kahlert is a senior researcher at the Swedish University of Agricultural Sciences. Her research interests are freshwater benthic algae: their diversity, regulation by environmental factors and their use for environmental assessment in lakes and streams.

References

Algarte, V.M., Rodrigues, L., Landeiro, V.L., Siqueira, T. & Bini, L.M. (2014) Variance partitioning of deconstructed periphyton communities: does the use of biological traits matter? *Hydrobiologia*, **722**, 279–290.

Astorga, A., Oksanen, J., Luoto, M., Soininen, J., Virtanen, R. & Muotka, T. (2012) Distance decay of similarity in freshwater communities: do macro- and microorganisms follow the same rules? *Global Ecology and Biogeography*, **21**, 365–375.

Atkinson, K.M. (1988) *The initial development of net phytoplankton in Cow Green Reservoir (Upper Teesdale), a new impoundment in Northern England*. *Algae and the aquatic environment* (ed. by F.E. Round), pp. 30–43. Biopress, Bristol, UK.

Baas Becking, L.G.M. (1934) *Geobiologie of inleiding tot de milieukunde*, W.P. Van Stockum & Zoon, Den Haag.

Bennett, J.R., Cumming, B.F., Ginn, B.K. & Smol, J.P. (2010) Broad-scale environmental response and niche conservatism in lacustrine diatom communities. *Global Ecology and Biogeography*, **19**, 724–732.

Bjørnstad, O.N. (2016) *ncf: Spatial Nonparametric Covariance Functions*,.

Blanchet, F.G., Legendre, P. & Borcard, D. (2008) Forward Selection of Explanatory Variables. *Ecology*, **89**, 2623–2632.

Borcard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the Spatial Component of Ecological Variation. *Ecology*, **73**, 1045–1055.

Bottin, M., Soininen, J., Alard, D. & Rosebery, J. (2016) Diatom Cooccurrence Shows Less Segregation than Predicted from Niche Modeling. *PLoS ONE*, **11**, e0154581.

Bottin, M., Soininen, J., Ferrol, M. & Tison-Rosebery, J. (2014) Do spatial patterns of benthic diatom assemblages vary across regions and years? *Freshwater Science*, **33**, 402–416.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, **15**, 584–593.

Cardinale, B.J., Hillebrand, H. & Charles, D.F. (2006) Geographic patterns of diversity in streams are predicted by a multivariate model of disturbance and productivity. *Journal of Ecology*, **94**, 609–618.

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

Cribari-Neto, F. & Zeileis, A. (2010) Beta Regression in R. *Journal of Statistical Software*, **34**.

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.

De Wit, R. & Bouvier, T. (2006) ‘Everything is everywhere, but, the environment selects’; what did Baas Becking and Beijerinck really say? *Environmental Microbiology*, **8**, 755–758.

Dong, X., Li, B., He, F., Gu, Y., Sun, M., Zhang, H., Tan, L., Xiao, W., Liu, S. & Cai, Q. (2016) Flow directionality, mountain barriers and functional traits determine diatom metacommunity structuring of high mountain streams. *Scientific Reports*, **6**.

Drakare, S., Lennon, J.J. & Hillebrand, H. (2006) The imprint of the geographical, evolutionary and ecological context on species–area relationships. *Ecology Letters*, **9**, 215–227.

Dray, S., Blanchet, G., Borcard, D., Guénard, G., Jombart, T., Larocque, G., Legendre, P. & Wagner, H.H. (2016) *adespatial: Multivariate Multiscale Spatial Analysis*,.

Dray, S., Legendre, P. & Peres-Neto, P.R. (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, **196**, 483–493.

Dray, S., Pélissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P.R., Bellier, E., Bivand, R., Blanchet, F.G., De Cáceres, M., Dufour, A.-B., Heegaard, E., Jombart, T., Munoz, F., Oksanen, J., Thioulouse, J. & Wagner, H.H. (2012) Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, **82**, 257–275.

European Committee for Standardization (2004) *Water quality - Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters.*, Brussels.

European Committee for Standardization (2003) *Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers*, Brussels.

Fenchel, T. & Finlay, B.J. (2004) The Ubiquity of Small Species: Patterns of Local and Global Diversity. *BioScience*, **54**, 777–784.

Fortin, M.-J. & Dale, M.R.T. (2005) *Spatial Analysis: A Guide for Ecologists*, Cambridge University Press.

Godínez Domínguez, E. & Freire, J. (2003) Information-theoretic approach for selection of spatial and temporal models of community organization. *Marine Ecology Progress Series*, **253**, 17–24.

Gonçalves-Souza, T., Romero, G.Q. & Cottenie, K. (2014) Metacommunity versus Biogeography: A Case Study of Two Groups of Neotropical Vegetation-Dwelling Arthropods. *PLoS ONE*, **9**, e115137.

Göthe, E., Angeler, D.G., Gottschalk, S., Löfgren, S. & Sandin, L. (2013) The Influence of Environmental, Biotic and Spatial Factors on Diatom Metacommunity Structure in Swedish Headwater Streams. *PLoS ONE*, **8**, e72237.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Havs och Vattenmyndigheten (2016) *Programområde: Sötvatten. Undersökningstyp: Påväxt i sjöar och vattendrag – kiselalgsanalys*,.

Heino, J., Bini, L.M., Karjalainen, S.M., Mykrä, H., Soininen, J., Vieira, L.C.G. & Diniz-Filho, J.A.F. (2010) Geographical patterns of micro-organismal community structure: are diatoms ubiquitously distributed across boreal streams? *Oikos*, **119**, 129–137.

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., Soininen, J., Alahuhta, J., Lappalainen, J. & Virtanen, R. (2017) Metacommunity ecology meets biogeography: effects of geographical region, spatial dynamics and environmental filtering on community structure in aquatic organisms. *Oecologia*, **183**, 121–137.

Hillebrand, H., Watermann, F., Karez, R. & Berninger, U.-G. (2001) Differences in species richness patterns between unicellular and multicellular organisms. *Oecologia*, **126**, 114–124.

Holyoak, M., Leibold, M.A. & Holt, R.D. (2005) *Metacommunities: spatial dynamics and ecological communities*, University of Chicago Press.

Hubbell, S.P., Foster, R.B., O'Brien, S.T., Harms, K.E., Condit, R., Wechsler, B., Wright, S.J. & Lao, S.L. de (1999) Light-Gap Disturbances, Recruitment Limitation, and Tree Diversity in a Neotropical Forest. *Science*, **283**, 554–557.

Jüttner, I., Rothfritz, H. & Eisenreich, S.J. (1996) Diatoms as indicators of river quality in the Nepalese Middle Hills with consideration of the effects of habitat-specific sampling. *Freshwater Biology*, **36**, 475–486.

Jyrkänkallio-Mikkola, J., Meier, S., Heino, J., Laamanen, T., Pajunen, V., Tolonen, K.T., Tolkkinen, M. & Soininen, J. (2017) Disentangling multi-scale environmental effects on stream microbial communities. *Journal of Biogeography*, n/a-n/a.

Kahlert, M. & Albert, R.-L. (2005) NorBAF – The Nordic-Baltic Network for Benthic Algae in Freshwater.

Kristiansen, J. (1996) Dispersal of freshwater algae — a review. *Hydrobiologia*, **336**, 151–157.

Lavoie, I., Grenier, M., Campeau, S. & Dillon, P.J. (2010) The Eastern Canadian Diatom Index (IDEC) version 2.0: including meaningful ecological classes and an expanded coverage area that encompasses additional geological characteristics. *Water Quality Research Journal of Canada*, **45**.

Legendre, P., Borcard, D. & Roberts, D.W. (2012) Variation partitioning involving orthogonal spatial eigenfunction submodels. *Ecology*, **93**, 1234–1240.

Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.

Legendre, P. & Legendre, L.F.J. (2012) *Numerical Ecology*, 3rd edn. Elsevier, Amsterdam.

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

Leland, H.V. & Porter, S.D. (2000) Distribution of benthic algae in the upper Illinois River basin in relation to geology and land use. *Freshwater Biology*, **44**, 279–301.

Levin, S.A. (1992) The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, **73**, 1943–1967.

Lindström, E.S. & Langenheder, S. (2012) Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, **4**, 1–9.

Liu, J., Soininen, J., Han, B.-P. & Declerck, S.A.J. (2013) Effects of connectivity, dispersal directionality and functional traits on the metacommunity structure of river benthic diatoms. *Journal of Biogeography*, **40**, 2238–2248.

Lowe, R.L. & Pan, Y. (1996) *Benthic algal communities as biological monitors*. *Algal ecology: Freshwater benthic ecosystems* (ed. by R.J. Stevenson), M.L. Bothwell), and R.L. Lowe), pp. 705–739. Academic Press, San Diego, California, USA.

MacArthur, R.H. (1972) *Geographical ecology: patterns in the distribution of species*, Princeton University Press.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-Devine, M.C., Kane, M., Krumins, J.A., Kuske, C.R., Morin, P.J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V.H. & Staley, J.T. (2006) Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, **4**, 102–112.

Naturvårdsverket (2007) *Bedömningsgrunder för sjöar och vattendrag*,.

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.H., Szoecs, E. & Wagner, H. (2016) *vegan: Community Ecology Package*,.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*.

Pajunen, V., Luoto, M. & Soininen, J. (2016) Climate is an important driver for stream diatom distributions. *Global Ecology and Biogeography*, **25**, 198–206.

Passy, S.I. (2001) Spatial Paradigms of Lotic Diatom Distribution: A Landscape Ecology Perspective. *Journal of Phycology*, **37**, 370–378.

Pebesma, E.J. (2004) Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**, 683–691.

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006) Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions. *Ecology*, **87**, 2614–2625.

Potapova, M.G. & Charles, D.F. (2002) Benthic diatoms in USA rivers: distributions along spatial and environmental gradients. *Journal of Biogeography*, **29**, 167–187.

Proctor, V.W. (1959) Dispersal of Fresh-Water Algae by Migratory Water Birds. *Science*, **130**, 623–624.

R Development Core Team (2016) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ricklefs, R.E. (2004) A comprehensive framework for global patterns in biodiversity. *Ecology Letters*, **7**, 1–15.

Rimet, F. & Bouchez, A. (2012) Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. *Knowledge and Management of Aquatic Ecosystems*, **406**, 1–14.

Schlichting, H.E. (1969) The Importance Of Airborne Algae and Protozoa. *Journal of the Air Pollution Control Association*, **19**, 946–951.

Schwalb, A.N., Morris, T.J. & Cottenie, K. (2015) Dispersal abilities of riverine freshwater mussels influence metacommunity structure. *Freshwater Biology*, **60**, 911–921.

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

Simas, A.B., Barreto-Souza, W. & Rocha, A.V. (2010) Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, **54**, 348–366.

Soininen, J. (2014) A quantitative analysis of species sorting across organisms and ecosystems. *Ecology*, **95**, 3284–3292.

Soininen, J. (2004) Determinants of Benthic Diatom Community Structurein Boreal Streams: the Role of Environmental and Spatial Factors at Different Scales. *International Review of Hydrobiology*, **89**, 139–150.

Soininen, J. (2007) Environmental and Spatial Control of Freshwater Diatoms—a Review. *Diatom Research*, **22**, 473–490.

Soininen, J. (2016) Spatial structure in ecological communities – a quantitative analysis. *Oikos*, **125**, 160–166.

Soininen, J., Paavola, R. & Muotka, T. (2004) Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. *Ecography*, **27**, 330–342.

Stegen, J.C., Lin, X., Konopka, A.E. & Fredrickson, J.K. (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *The ISME Journal*, **6**, 1653–1664.

Stevenson, R.J. & Smol, J.P. (2003) *Use of algae in environmental assessments*. *Freshwater Algae in North America: Ecology and Classification* (ed. by J.D. Wehr) and R.G. Sheath), pp. 775–804. Academic Press, San Diego, California, USA.

Telford, R.J., Vandvik, V. & Birks, H.J.B. (2006) Dispersal Limitations Matter for Microbial Morphospecies. *Science*, **312**, 1015–1015.

Verleyen, E., Vyverman, W., Sterken, M., Hodgson, D.A., De Wever, A., Juggins, S., Van de Vijver, B., Jones, V.J., Vanormelingen, P., Roberts, D., Flower, R., Kilroy, C., Souffreau, C. & Sabbe, K. (2009) The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. *Oikos*, **118**, 1239–1249.

Vilmi, A., Tolonen, K.T., Karjalainen, S.M. & Heino, J. (2017) Metacommunity structuring in a highly-connected aquatic system: effects of dispersal, abiotic environment and grazing pressure on microalgal guilds. *Hydrobiologia*, **790**, 125–140.

Vyverman, W., Verleyen, E., Sabbe, K., Vanhoutte, K., Sterken, M., Hodgson, D.A., Mann, D.G., Juggins, S., Vijver, B.V. de, Jones, V., Flower, R., Roberts, D., Chepurnov, V.A., Kilroy, C., Vanormelingen, P. & Wever, A.D. (2007) Historical Processes Constrain Patterns in Global Diatom Diversity. *Ecology*, **88**, 1924–1931.

Wilson, D.S. (1992) Complex Interactions in Metacommunities, with Implications for Biodiversity and Higher Levels of Selection. *Ecology*, **73**, 1984–2000.

Figures and Tables

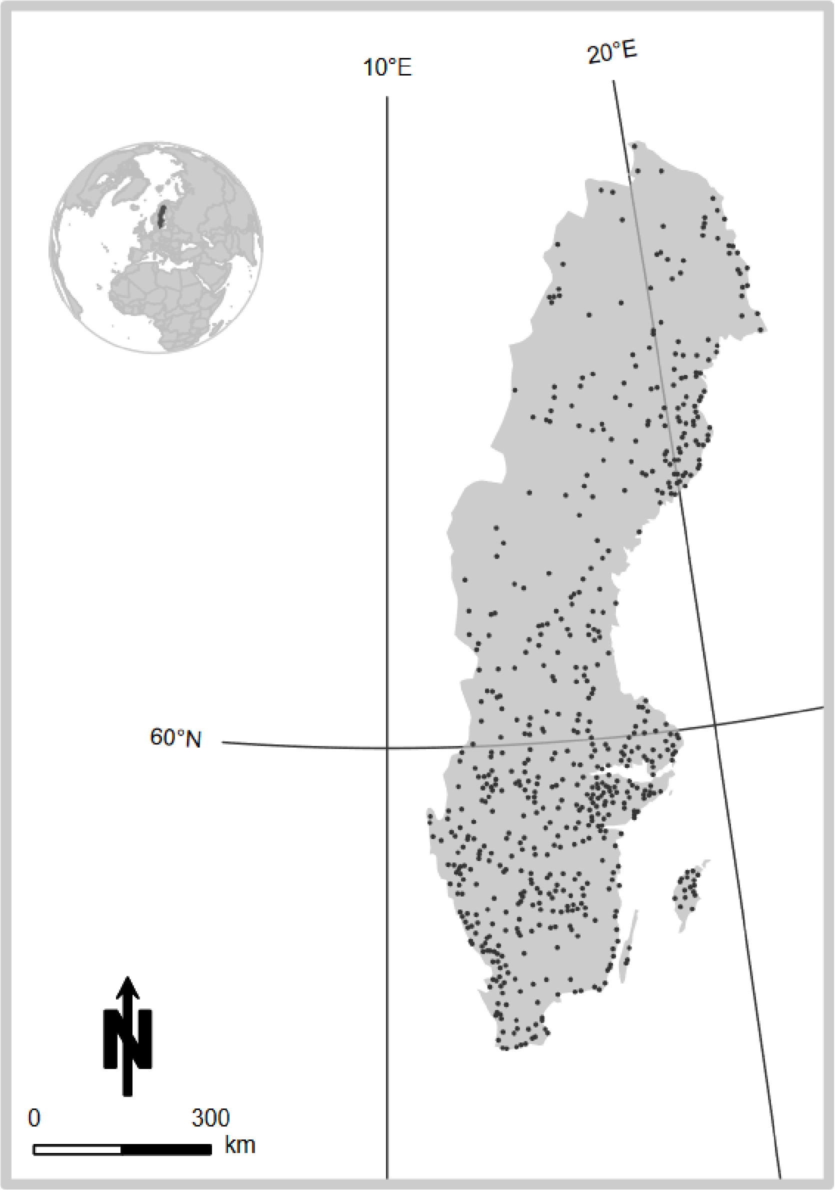

Figure 1. Map of Sweden showing the locations of the 616 sampling sites (black dots).

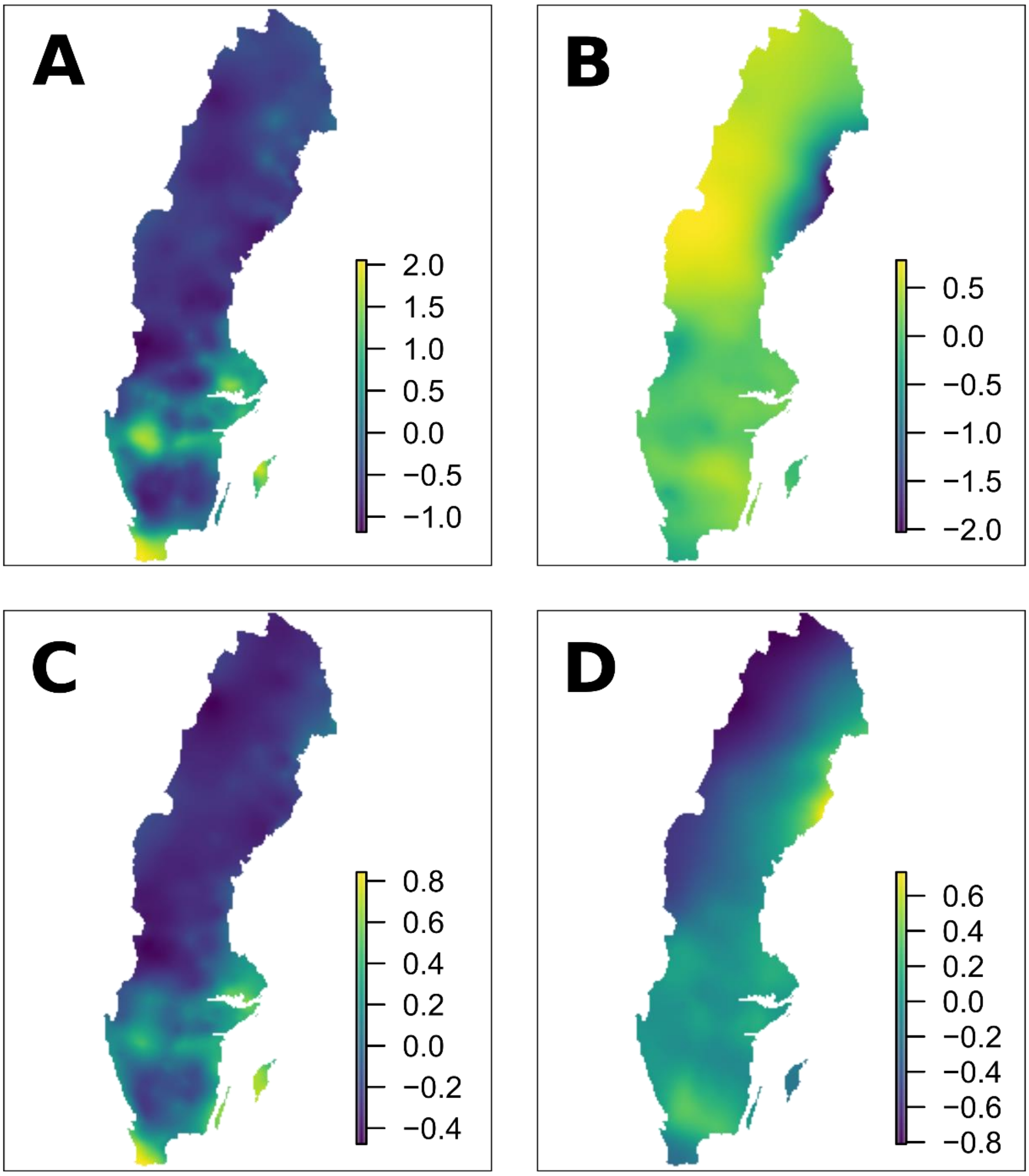

Figure 2. Maps of interpolated sites scores for (A-B) the correspondence analysis on the community matrix (Axis 1 and Axis 2 accounting for 4% and 3% of the variation respectively) and (C-D) for the principal component analysis on the environmental parameters matrix (Axis 1 and Axis 2 accounting for 33% and 23% of the variation respectively). For loadings values see Appendix S1.

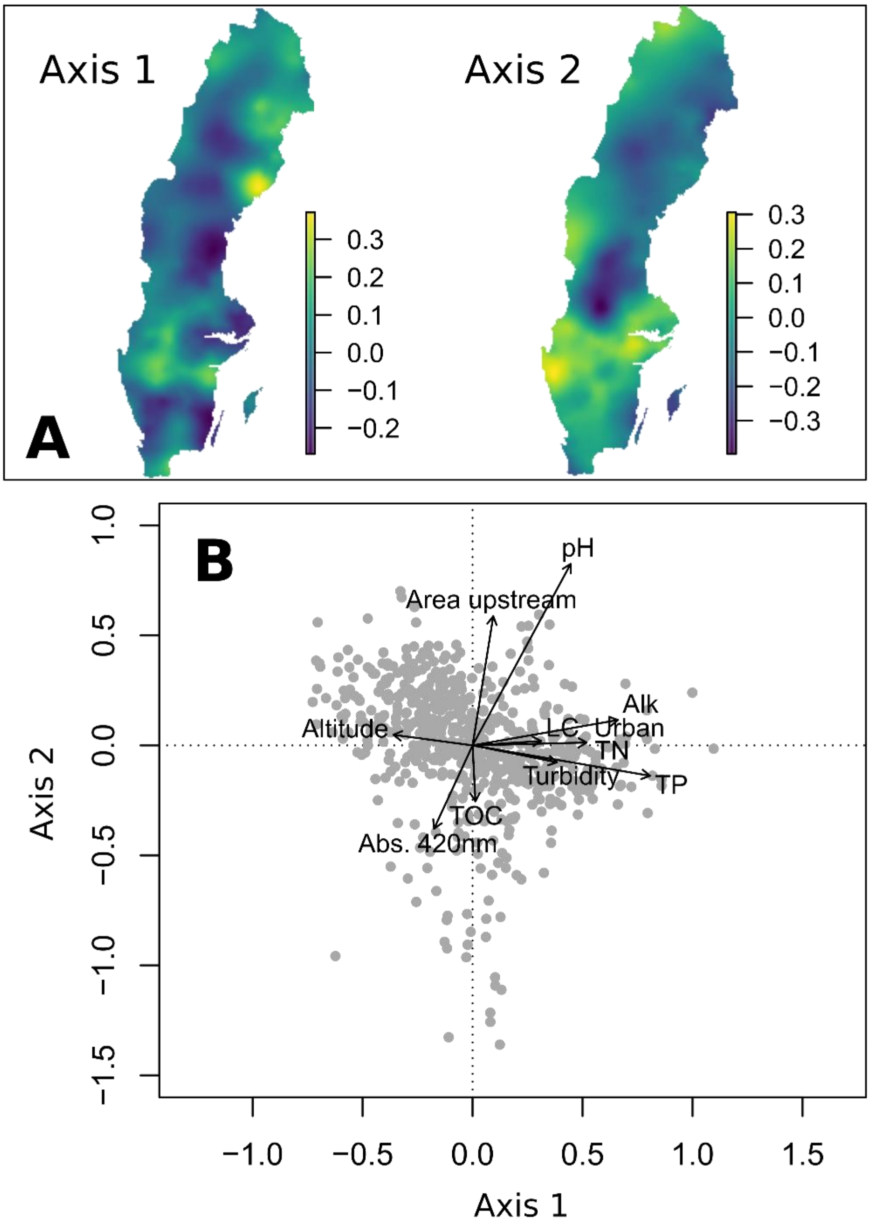

Figure 3. (A) Maps of interpolated sites scores for the partial redundancy analysis of the community matrix constrained by space after partialling out the effect of environment (Axis 1 and Axis 2 accounting for 2.3% and 1.5% of the variation respectively). (B) Ordination plot for the partial redundancy analysis of the community matrix constrained by the environment after partialling out the effect of space (Axis 1 and Axis 2 accounting for 3.1% and 1.9% of the variation respectively). For loadings values see Appendix S1.

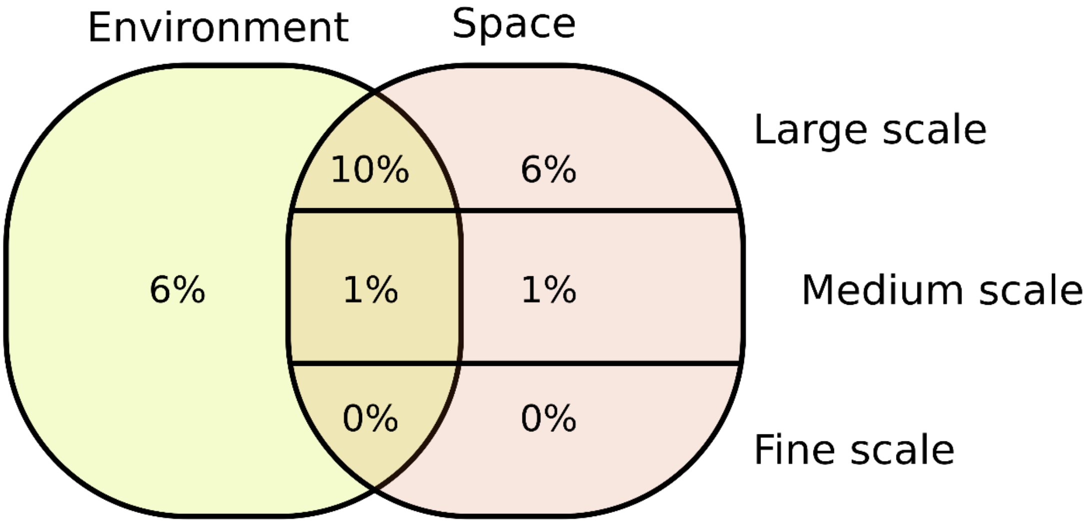

Figure 4. Venn diagram showing the results of the variation partitioning analysis on the community matrix (community-based approach). Adjusted R-squared is given for each fraction.

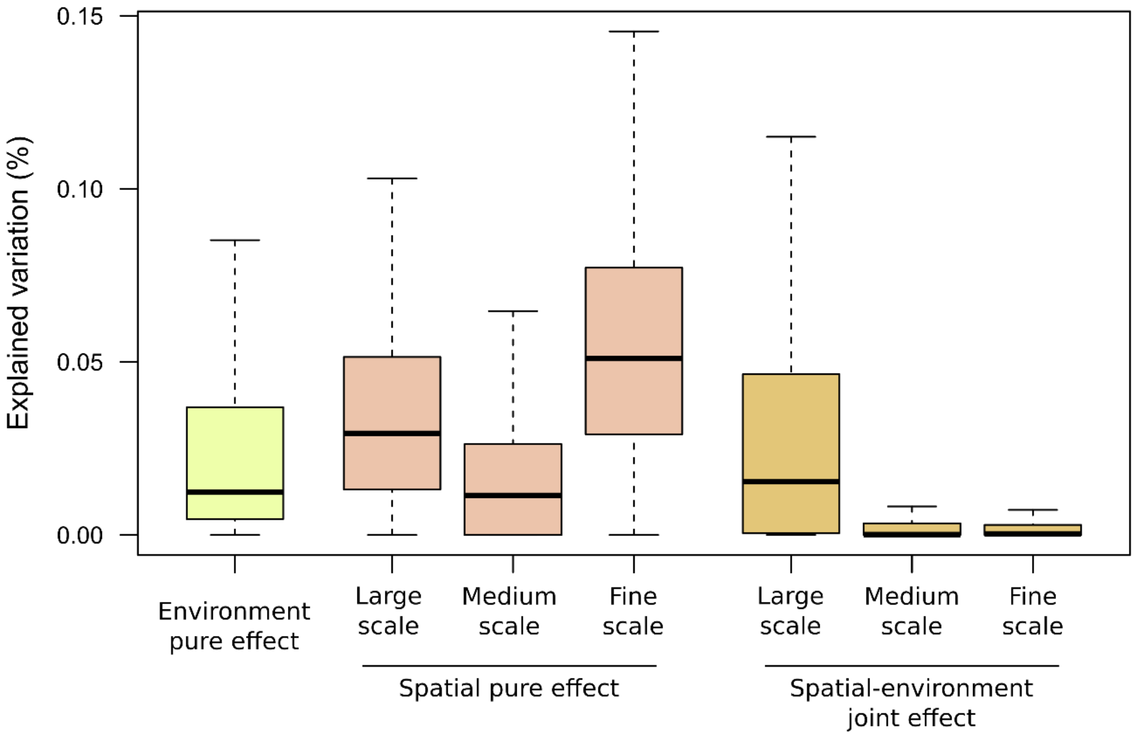

Figure 5. Boxplot showing the appended results of the variation partitioning analyses performed on each species (species-based approach).