Are drivers of microbial diatom distributions context dependent in human impacted and pristine environments?

Virpi Pajunen[1, 2], Jenny Jyrkänkallio-Mikkola[1], Miska Luoto[1] and Janne Soininen[1]

[1]Department of Geosciences and Geography, P.O. Box 64, FI-00014 University of Helsinki, virpi.pajunen@helsinki.fi, jenny.jyrkankalliomikkola@gmail.com, miska.luoto@helsinki.fi, janne.soininen@helsinki.fi.

[2]Correspondence: Virpi Pajunen, e-mail: virpi.pajunen@helsinki.fi, tel: +358 50 488 4437

Manuscript received 12 September 2018; revised 1 February 2019; accepted 16 April 2019.

Corresponding Editor: Stephen B. Baines

Abstract

Species occurrences are influenced by numerous factors of which effects may be context dependent. Thus, the magnitude of such effects and their relative importance on species distributions may vary among ecosystems due to anthropogenic stressors, for example. To investigate context dependency in factors governing microbial bioindicators, we developed

species distribution models (SDMs) for epilithic stream diatom species separately in human impacted and pristine sites. We performed SDMs using boosted regression trees for 110 stream diatom species, which were common to both data sets, separately in 164 human impacted and 164 pristine sites in Finland (c. 1000 km, 60˚ – 68˚ N). For each species and site group, two sets of models were conducted: climate model, comprising three climatic variables, and full model, comprising the climatic and six local environmental variables. No significant difference in model performance was found between the site groups. However, climatic variables had greater importance compared with local environmental variables in pristine sites, whereas local environmental variables had greater importance in human impacted sites as hypothesized. Water balance and conductivity were the key variables in human impacted sites. The relative importance of climatic and local environmental variables varied among individual species, but also between the site groups. We found a clear context dependency among the variables influencing stream diatom distributions as the most important factors varied both among species and between the site groups. In human impacted streams, species distributions were mainly governed by water chemistry, whereas in pristine streams by climate. We suggest that climatic models may be suitable in pristine ecosystems, whereas the full models comprising both climatic and local environmental variables should be used in human impacted ecosystems.

Key words: land use gradient; climate; local environment; stream diatoms; species distribution modelling; boosted regression trees

Introduction

Ecosystems are molded by a myriad of factors operating at multiple spatial scales (Cox et al. 2016). This is especially true in open systems, such as rivers and streams, characterized by unidirectional flow and supply of substances from terrestrial areas (Allan and Castillo 2007). Large scale factors, such as climate and catchment land use, strongly influence the local stream habitat and thus the species diversity therein (Allan 2004, Pajunen et al. 2017). Due to the ongoing anthropogenic environmental change, streams are subjected to multiple stressors including changes in land use and associated habitat degradation and changing climatic conditions. This increasing stress is resulting in biodiversity loss and homogenization of communities (Rahel 2002, Olden et al. 2004, Filipe et al. 2013, Dar and Reshi 2014). As a consequence of global warming, the stream water temperatures are predicted to rise correspondingly (Webb 1996, Morrill et al. 2005) and changes in precipitation will alter hydrological conditions. The complex interactions between climate, land use and water physicochemistry challenge the future predictions of stream conditions. Earlier studies have mostly concentrated in changes in community composition between different gradients of human impact (e.g., Pan et al. 2004, Soininen et al. 2004, Hering et al. 2006), yet the knowledge about the responses of individual species (based on their occurrences) to environmental and climatic factors in different environments are still limited. To fill this gap in knowledge, we investigated whether the effect of climatic variables on the distribution of diatom species is more pronounced in pristine sites than in human impacted sites.

The catchment land use affects stream physicochemistry (Foley et al. 2005) and even the past land use can have a long lasting imprint on stream conditions (Maloney et al. 2008, Walter and Merritts 2008, Maloney and Weller 2011). For example, land cover previously dominated by agriculture may sustain high nutrient loads from sediments to streams for a

long time period after a change in land use (Maloney and Weller 2011). Anthropogenic land use, comprising agriculture, development and urbanization, contributes to increased nutrient and ion concentrations (Taka et al. 2017), pollutants and turbidity (due to sediment load) in streams (Foley et al. 2005), and these effects cascade downstream (Levesque et al. 2017). Wang et al. (2008) found that nutrient loading and percent of urban land use were the most important drivers of deteriorating stream conditions. Climate also affects nutrient levels in streams as nutrient fluxes in a stream network are strongly driven by hydrology (Arvola et al. 2015). Furthermore, riparian vegetation regulates stream temperature and light conditions by shading, acts as organic matter input and as an agent in sediment retention. Its removal leads to increased water temperature, sediment load and nutrient leaching (Studinski et al. 2012, Sweeney and Newbold 2014), but also to increased periphyton biomass due to greater light intensities (Von Schiller et al. 2007). Such effects of land use are likely to increase with projected higher air temperatures and precipitation in the future (e.g., Holmberg et al. 2006, Piggott et al. 2015). The relationship between land use and microbial stream communities strengthens towards downstream because of the continuous accumulation of substances in a river continuum (Tudesque et al. 2014). Moreover, stream microbes may show in some circumstances stronger relationship with the changes in land use than with physicochemical gradients, e.g., pH, substrates and nutrients (Liu et al. 2016, Jyrkänkallio-Mikkola et al. 2017). This indicates that land use could provide a more robust measure of water chemistry variables – thus, reflecting stream chemistry at longer time scales than snapshot water samples. This implies that especially in the presence of human activities, microbial communities are strongly influenced by the relative proportions of local environmental factors (for example water chemistry) brought about by a certain type of land use.

Biological indicators, such as benthic diatoms, are widely used to assess the ecological status of freshwater ecosystems as they reflect water quality over a period of time (Sandin & Verdonschot, 2006). Many aquatic microbes, including diatoms, have species-specific responses towards water chemistry (Van Dam et al. 1994, Olapade and Leff 2005), but whether these responses stem from niche conservation or local adaption, is currently under debate (Finlay 2002, Wiens and Graham 2005). However, a clear evidence for niche conservation, i.e. the tendency of species to preserve inherited ecological characteristics, have been detected in lacustrine diatoms (Telford et al. 2006, Bennett et al. 2010). The relative importance of environmental variables (such as water chemistry and land use) affecting benthic diatoms can also vary between study regions and in different climatic zones (Charles et al. 2006, Jüttner et al. 2010), and are also influenced by the study scale (Verleyen et al. 2009, Heino et al. 2014). This suggests a certain context dependency among the most influential factors driving microbial distributions. Furthermore, previous studies have shown a strong influence of climatic factors on the distributions of stream micro-organisms, which can even exceed the effect of local environmental variables (Pajunen et al. 2016). Climate can be seen as a crucial factor that has a strong impact on water temperatures and terrestrial vegetation patterns (Cox et al. 2016), and thus also on variation in in-stream variables (Frissell et al. 1986, Stevenson 1997). The effect of climate is likely to be more apparent in pristine environments where, in the absence of human impact, natural processes are able to dictate the supply of substances and the disturbance regime in streams. The variation in water chemistry among pristine streams are expected to be smaller than in human impacted environments (Castillo et al. 2012), thus the relative role of climatic factors affecting stream communities may be stronger. In contrast, in human impacted environments, local environment sets a strong filter for species due to the large variation in local environmental factors among streams due to the varying forms and intensity of human influence.

Additionally, in boreal regions where productivity is generally low even minor human activities can have a notable impact on stream water chemistry (e.g. sewage discharges from dispersed settlement).

To investigate whether the distributions of commonly used microbial bioindicators are context dependent, i.e. species' responses vary between species and among sites with different magnitude of human impact, we developed species distribution models (SDMs) for stream diatoms separately in human impacted and pristine streams. We hypothesized that climatic variables affect the distribution of diatom species more in pristine sites than in human impacted sites. As a corollary, the effect of local environmental variables on species distributions is stronger in human impacted sites where the ranges of water chemistry are longer than in pristine sites. In human impacted sites, the addition of local variables to climate models would thus greatly enhance the model performance.

Methods

Data sampling and analysis

The data set comprised diatom (presence/absence), water chemistry and physical variable data collected from Finnish stream sites between 1986 and 2016 (328 sites in total) (Fig. 1). The samples were considered to be comparable as the sampling methods were identical and all sampling was performed during the base flow conditions in July to September. The sites were distributed relatively evenly across Finland and the measured environmental and climatic variables covered a wide range (Table 1). The most of the sampling sites (> 90 %) were located in headwater streams (orders 1 or 2), yet some samples were taken from larger rivers. More detailed information about the data set can be found in Eloranta (1995), Soininen et al. (2004) and Jyrkänkallio-Mikkola et al. (2017).

According to standard methods of benthic algae sampling in streams (Lowe and Pan 1996, Kelly et al.1998), each stream site was sampled for diatoms by collecting five to ten cobble sized stones along 10 m section comprising mainly riffle habitat. By collecting diatoms only from rock surfaces (i.e. epilithic diatoms), the potential effect of substrata on diatom communities could be ruled out in order to minimize the noise in the data (Lowe and Pan 1996). Biofilm was removed from the stones by brushing them with a toothbrush and combined as one composite sample at each site. Water samples were taken simultaneously with diatom samples, and were subsequently analyzed for total phosphorus (TP), pH, conductivity and water color using national standards. For minority of the sites (< 10 %), water chemistry data were taken from the national water quality database, using results from the nearest sampling occasion and location. Current velocity, canopy shading and stream width were measured at each site along the site perpendicular to the flow and covering the whole stream section. Samples were cleaned from organic material in the laboratory using wet combustion with acid (HNO_3:H_2SO_4; 2:1 or hydrogen peroxide [30%, H_2O_2]) and mounted in Naphrax or Dirax. A total of 250–500 diatom frustules per sample were identified to the lowest possible taxonomic level according to Krammer and Lange-Bertalot (1986–1991) and Lange-Bertalot and Metzeltin (1996), and counted using phase contrast light microscopy (magnification 1000×). A species was considered to be present at a site when at least one valve was observed.

Climatic and land use variables

We compiled a set of environmental variables presumed to affect diatom distributions. Climatic variables were chosen based on their use in previous SDMs conducted for Finnish diatom data (Pajunen et al. 2016). The variables were growing degree days adjusted to 5 ˚C (GDD), season precipitation sum from May to September (PRECS) and water balance

(WAB; calculated according to Skov and Svenning 2004). GDD represents the aerial temperature and the energy requirements of the species, while PRECS and WAB represent the moisture availability in the environment, connected to the extent of recharge and run-off. The climatic data set covered the years 1981–2010 and was obtained as a 10×10 km resolution grid from the Finnish Meteorological Institute (Venäläinen and Heikinheimo 2002). Using ArcGIS 10.3.1 software, site-specific catchment areas were created by calculating the patterns of flow direction and accumulation to each sampling point from digital elevation model (DEM; grid resolution 10×10 m, National Land Survey of Finland 2013). Classifications of land use were obtained from CORINE Land Cover data (20×20 m, Finnish Environment Institute 2013). Artificial and agricultural land use were merged to represent anthropogenic land use. The local and climatic variables were tested for covariance with nonparametric Spearman's rank correlation coefficient. All predictor variables had low collinearity ($r_s \leq |0.50|$, Appendix S1: Figs. S1–2).

Species distribution models

The data set (328 sites and in total 494 diatom species) was divided into two equal-sized groups: human impacted sites (n = 164, > 5% anthropogenic land use) and pristine sites (n = 164, < 5% anthropogenic land use). The 5% threshold for anthropogenic land use was chosen also to cover the impacts of point source pollution, which were also present in otherwise relatively pristine areas. These potential point sources were either observed in the field or estimated from the ground documents. In Finland, anthropogenic land use is typically of relatively low intensity. Therefore, even a small increase in human impact can have a notable effect on stream conditions. Diatom species that occurred in both groups and at least at 5% and maximum at 95% of the sites were included in the statistical analyses. These thresholds were chosen in order to perform robust species distribution models.

Two sets of diatom species distribution models were conducted for each of the 110 species separately for human impacted and pristine sites: climate models and full models. In the climate models, species distributions were modelled only by the three climatic variables: GDD, PRECS and WAB. In addition to the climatic variables, the full models had six environmental predictors: TP, conductivity, pH, water color (mainly reflecting the humic content of the water), canopy shading and current velocity. To account for the spatial concentration of the human impacted sites in the southern and western parts of Finland, we performed additional distribution models for subsampled data sets (n=100 sites and 104 species in both impacted and pristine data sets, Appendix S2: Table S1 and Fig. S1) using the same settings. Due to the exclusion of sites especially in the northern regions, the full and subsampled data sets differed by 16 species.

The SDMs were applied via the BIOMOD2 framework (Thuiller et al. 2016) fitted in R (version 3.3.3; R Development Core Team 2017) using boosted regression trees (BRT) as modelling algorithm. BRT is a machine learning technique, which has previously proven to be a robust method for creating SDMs for micro-organisms (Pajunen et al. 2016), as it is highly efficient at fitting nonparametric data, and can manage various types of predictor variables. It does not require prior data transformation and takes automatically into account the interaction effects between predictors (the principles of BRT in more detail: see Friedman 2001, De'ath 2007, Elith et al. 2008). BRTs were performed with a maximum number of 3000 trees, the interaction depth of 6 and the learning rate of 0.001.

The performance of each model was assessed with a cross validation (CV) approach, where the models were fitted four times by using a random sample of 70% of the data and subsequently evaluated against the remaining 30%. The predicted and observed occurrences

of species were compared at each CV run by calculating the area under the curve of a receiver operating characteristic plot (AUC) (Fielding and Bell 1997) and true skill statistics (TSS) (Allouche et al. 2006). The models have at least intermediate predictive performance if AUC values are > 0.7 (following Swets 1998) and TSS values are > 0.4 (following Landis and Koch 1977).

The importance of each predictor for a species in the models was assessed in BIOMOD2 by randomizing each variable individually and then projecting the model with the randomized variable while keeping the other variables unchanged. The model predictions containing the randomized variable were further correlated with those of the original models. Finally, the importance of the variable was calculated as one minus the correlation; higher values indicate predictors that are more important for the model (Thuiller et al. 2009). This analysis was repeated ten times. The differences in model performances and predictor relative importances between human impacted and pristine data sets were tested using paired t-test in R (version 3.3.3; R Development Core Team 2017). As a supplemental analysis, the variation in the species occurrence data among climatic and local environmental variables was decomposed using the variance partitioning approach based on redundancy analysis (RDA) applying the package VEGAN (Oksanen et al. 2015) in R. The analysis was performed separately for human impacted and pristine data sets.

Results

Both climatic and full models performed satisfactorily in human impacted and pristine sites (species distribution model (SDM) averages AUC > 0.70 and TSS > 0.40) and all sets of SDMs (i.e. climate and full models in both site groups) had similar patterns in predictive performances (Appendix S1: Fig. S3). The inclusion of local environmental variables into the

models did not improve the model performance compared to the climate models in pristine sites (climate model AUC 0.710 and TSS 0.447; full model AUC 0.707 and TSS 0.435), whereas in human impacted sites it slightly did (climate model AUC 0.708 and TSS 0.442; full model AUC 0.725 and TSS 0.464). However, no significant difference in model performance was found between human impacted and pristine sites (paired t-test, all $P > 0.05$).

In the full models, climatic variables had on average greater variable importance (the sum of median (md) importance: climatic 55% and local environmental variables 41%) compared with local environmental variables in pristine sites. In human impacted sites local environmental variables were more important (climatic 38% and local environmental variables 50%, respectively) (Fig. 2). Water balance (WAB) was the most important variable in both site groups (md importance: human impacted 11% and pristine sites 12%), whereas growing degree days (GDD) was as important in pristine sites (md = 12%) (Figs. 2 and 3). In human impacted sites, conductivity had the second greatest relative importance (md = 9%) on species distributions (Fig. 2). Precipitation (PRECS) had the second greatest importance (md = 8%) in pristine sites and the third greatest (md = 8%) in human impacted sites. Total phosphorus (TP) had the third greatest importance (md = 6%) on species distributions in pristine sites.

The relative importance of climatic and local environmental variables on individual species varied among species, but also between human impacted and pristine sites (Fig. 4, Appendix S1: Fig. S4). Overall variation between the two site groups was significant only for the relative importance of GDD (paired t-test; $P < 0.001$) being higher in pristine sites, whereas for other variables it was nonsignificant (paired t-test; $P > 0.05$). The between-site group

variation among the relative importance of variables on individual species was species-specific: some species responded similarly to climatic and/or local environmental variables in both site groups, yet the responses of some other species varied greatly (e.g. *Cocconeis placentula*, *Navicula rhynchocephala*) (Fig. 5).

The performances of the species distribution models conducted using the subsampled data sets were similar to the SDMs conducted for the full data sets (Appendix S2: Fig. S2). The models for the subsampled data sets showed that, on average, the local environmental variables were more important than climatic variables in impacted sites, whereas in pristine sites, the importance of climatic and local environmental variables did not differ significantly (P=0.141, Appendix S2: Fig. S3). Compared with the results of the full data sets, GDD and conductivity were less important for the species distributions in the subsampled pristine data set (Appendix S2, Fig. S4) as were WAB and pH in the human impacted data set, whereas, the importance of TP increased in both pristine and impacted sites. Overall variation between the two subsampled site groups was significant only for the relative importance of GDD, WAB and shading (paired t-test; $P < 0.05$) (Appendix S2, Fig. S5).

In RDA based variance partitioning, the climatic variables explained 5 % and local environmental variables 6 % of variation in both human impacted and pristine data sets (Appendix S1: Fig. S5). The joint effects explained 7 % of the variation in impacted sites and 4 % in pristine sites. The majority of the total variation was left undetermined in both data sets.

Discussion

Our study reveals a notable context dependency among the factors influencing the distributions of diatom species. The most important factors affecting diatom species distribution vary not only among species, but are also dependent on the degree of anthropogenic influence. Consistent with our hypothesis, the overall importance of climatic variables on species occurrence was greater in pristine streams than was the importance of local environmental variables. In contrast, the importance of local environmental variables was greatest in human impacted sites. We emphasize though, that water balance has a significant impact on stream diatom distributions in all stream environments, and its effect was stronger than the influence of any single local environmental variable not only in pristine locations, but also in human impacted sites. This corresponds to previous studies indicating that diatom species can be shaped by large-scale climatic (Weckström et al. 1997a, Pajunen et al. 2016) and historical factors (Vyverman et al. 2007).

Although climate is the ultimate factor influencing stream diatom distributions both directly (temperature) and indirectly (productivity and hydrology) (Pajunen et al. 2016, 2017), the importance of local physicochemical factors seem to be highlighted in the streams influenced by anthropogenic activities. This can be partly explained by the long ranges of water chemistry variables (here conductivity and TP; see Table 1) related to anthropogenic land use and by the strong species responses towards these variables (i.e. species filtering along environmental gradients). The tolerances of individual diatom species towards local environmental factors have been widely studied and many species have restricted tolerances and preferences towards certain environmental variables (for example, nutrients [Winter and Duthie 2000], conductivity [Potapova and Charles 2003] and pH [Andrén and Jarlman 2008]). For example, diatom communities in streams under human impact, such as

agriculture and point-source pollution, consist of species with a preference to high nutrient levels and tolerance towards pollutants (Lavoie et al. 2006, Moravcova et al. 2013).

Recently, growing evidence of context dependency in species responses toward environmental and spatial factors have been documented among stream organisms (for example, diatoms [Heino et al. 2012], bryophytes [Heino et al. 2012] and macroinvertebrates [Heino et al. 2012, Hawkins et al. 2015, Tonkin et al. 2016]). Stream diatoms are simultaneously affected by abiotic and biotic forcing which relative importance differ among sites and regions (Clements et al. 2015). For instance, the occurrence of *Frustulia rhomboides*, a species often classified as acidophilous i.e. occurring at pH <7 (Van Dam et al. 1994, Weckström et al. 1997b), was affected mainly by pH in pristine sites (relative importance = 32%). But in human impacted sites, it was mostly affected by conductivity (relative importance = 60%), while the relative importance of pH was negligible (1%, Fig. 5). Among-region variation in species-specific and community-level responses of diatoms to water chemistry has also been observed elsewhere (Charles et al. 2006, Jüttner et al. 2010, Chen et al. 2016). For example, Chen et al. (2016) found that diatom species indicating high nutrient conditions in U.S. streams occurred in low nutrient streams in China suggesting that diatom niches were not conserved. However, studies from different regions should be compared with caution as morphological taxonomic identification may contain locally adapted morphotypes of species (Rose and Cox 2014). Also, the spatial scale of the study may affect the relative importance of the most influential factors. The effect of climatic and dispersal-related factors operating at large spatial scales may become more important when the spatial scale is large (e.g., Soininen 2007, Vyverman et al. 2007, Benito et al. 2018). Thus, it can be envisaged that spatial signal in the communities and climatic effects on

species distributions need to be considered in diatom studies especially when operating at relatively large spatial scales.

The species' responses along anthropogenic gradients may also vary due to biotic interactions (such as competition), which intensity may vary along the shifts in community structure (Tilman 1977, Stelzer and Lamberti 2001). The species interactions and species occurrence on a site may be a result of metacommunity dynamics, such as species sorting and mass effects, and the processes structuring local communities may differ regionally (reviewed in Leibold et al. 2004). Context dependency may also be a sign of genotypic plasticity, i.e. species can be adapted to local conditions through rapid genetic evolution, enabled by the fast life-cycle of microbes (Birch 1960). Or, it may reflect phenotypic plasticity, i.e. the tolerances towards environmental factors vary among different morphotypes of individual species, which results in variable responses (Rose and Cox 2014). Context dependency may also depend on the spatial extent of the study area. In this study we performed additional SDMs using subsampled data sets. The results of these two parallel modelling efforts using slightly different study areas (i.e. full data sets vs. subsampled data sets) showed, for example, a lower importance of GDD in the subsampled pristine data set. This result is most likely due to the exclusion of northern sites and thus removing species that are strongly impacted by GDD. The number of possible processes causing context dependency highlights the need to study this topic further in the near future.

The moisture related factors, i.e. WAB and precipitation, were important both in human impacted and pristine sites. Precipitation and run-off are essential factors influencing aquatic biota, including stream diatoms, via weathering, transport of substances and flow regime (Stevenson et al. 1996, Leland and Porter 2000, Allan and Castillo 2007). In human impacted

streams, these climatic variables can enhance the effect of anthropogenic land use through run-off, which can consist of high amounts of allochthonous nutrients, organic matter, solids and pollutants (Pan et al. 2004, Death et al. 2015, Ponsati et al. 2016). The importance of climate is further emphasized by the fact that the impact of land use on water chemistry and further on diatom communities may weaken during summer base-flow conditions compared to wetter seasons (Pan et al. 2004). The hierarchical structure of environmental factors (for instance, climate influencing land cover which affects water physicochemistry) (Frissell et al. 1986, Stevenson 1997) may become more evident in the absence of human impact. For example, the relative importance of GDD was significantly greater in pristine than in human impacted sites suggesting that both the direct (temperature) and indirect (for example catchment and in-stream productivity) effects of GDD are more essential drivers in more pristine systems (Fig. 3). However, the relative importance of climatic variables in the human impacted data set may be influenced by the fact that the anthropogenic land use is mostly situated in the southern and western regions of Finland where the growing season is the longest. Thus, the human impacted data set contains less sites with cold and dry climatic conditions, more typical in the northern regions, compared to the pristine data set.

In conclusion, we found that the main drivers of epilithic stream diatom species distributions and also the species-specific responses to these drivers differed among human impacted and pristine environments. The effect of climate was important both in pristine and in human impacted streams in spite of wide gradients in local environmental variables and anthropogenic land use in the latter. However, the climatic influence was strongest in pristine streams, suggesting that climatic variables need to be considered in diatom models especially in regions where water chemistry gradients are only modest and stream physicochemistry is mainly dictated by natural landscape and the processes therein. The way that climatic and

environmental change will alter stream conditions in the future may be context dependent and differ among environments. Thus, it will be challenging to predict the distribution of micro-organisms under future climate scenarios.

Acknowledgments

This project was funded by Maj and Tor Nessling foundation, Nordenskiöld foundation and the Academy of Finland (grant 273560).

Literature cited

Allan, J. D. 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. *Annual Review of Ecology and Systematics* **35**:257–284.

Allan, J. D., and M. M. Castillo. 2007. *Stream Ecology: Structure and Function on Running Waters*. 2nd edn. Springer, Dordrecht, the Netherlands.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* **43**:1223–1232.

Andrén, C., and A. Jarlman. 2008. Benthic diatoms as indicators of acidity in streams. *Fundamental and Applied Limnology* **173**:237–253.

Arvola, L., E. Einola, and M. Järvinen. 2015. Landscape properties and precipitation as determinants for high summer nitrogen load from boreal catchments. *Landscape Ecology* **30**:429–442.

Benito, X., S. C. Fritz, M. Steinitz-Kannan, P. M. Tapia, M. A. Kelly, and T. V. Lowell, 2018. Geo-climatic factors drive diatom community distribution in tropical South American freshwaters. *Journal of Ecology* **106**: 1660–1672.

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

Birch, L. C. 1960. The genetic factor in population ecology. *The American Naturalist* **94**:5–24.

Castillo, M. M., H. Morales, E. Valencia, J. J. Morales, and J. J. Cruz-Motta. 2012. The effects of human land use on flow regime and water chemistry of headwater streams in the highlands of Chiapas. *Knowledge & Management of Aquatic Ecosystems* **407**: 9.

Charles, D. F., F. W. Acker, D. D. Hart, C. W. Reimer, and P. B. Cotter. 2006. Large-scale regional variation in diatom-water chemistry relationships: Rivers of the eastern United States. *Hydrobiologia* **561**:27–57.

Chen, X., W. Zhou, S. T. A. Pickett, W. Li, L. Han, and Y. Ren. 2016. Diatoms are better indicators of urban stream conditions: A case study in Beijing, China. *Ecological Indicators* **60**:265–274.

Clements, W. H., D. R. Kashian, P. M. Kiffney, and R. E. Zuellig. 2016. Perspectives on the context-dependency of stream community responses to contaminants. *Freshwater Biology* **61**:2162–2170.

Cox, C. B., P. D. Moore, and R. J. Ladle. 2016. *Biogeography: An Ecological and Evolutionary Approach*. John Wiley and Sons Ltd, Chichester, UK.

Dar, P. A., and Z. A. Reshi. 2014. Components, processes and consequences of biotic homogenization: A review. *Contemporary Problems of Ecology* **7**:123–136.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. *Ecology* **88**:243–251.

Death, R. G., I. C. Fuller, and M. G. Macklin. 2015. Resetting the river template: the potential for climate-related extreme floods to transform river geomorphology and ecology. *Freshwater Biology* **60**:2477–2496.

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

Eloranta, P. 1995. Type and quality of river waters in central Finland described using diatom indices. Pages 271–280 *in* D. Marino, and M. Montresor, editors. *Proceedings of the 13th International Diatom Symposium*. Biopress, Bristol, UK.

Fielding, A. H., and J. F. Bell. 1997. A review methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**:38–49.

Filipe, A. F., J. E. Lawrence, and N. Bonada. 2013. Vulnerability of stream biota to climate change in mediterranean climate regions: a synthesis of ecological responses and conservation challenges. *Hydrobiologia* **719**:331–351.

Finlay, B. 2002. Global dispersal of free-living microbial eukaryote species. *Science* **296**:1061–1063.

Finnish Environment Institute. 2013. *CORINE Land Cover 20 m*. https://avaa.tdata.fi/web/paituli. Accessed September 14, 2017.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P. K. Snyder. 2005. Global consequences of land use. *Science* **309**:570–574.

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

Frissell, C. A., W. J. Liss, C. E. Warren, and M. D. Hurley. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. *Environmental Management* **10**:199-214.

Hawkins, C. P., H. Mykrä, J. Oksanen, and J. J. Varder Laan. 2015. Environmental disturbance can increase beta diversity of stream macroinvertebrate assemblages. *Global Ecology and Biogeography* **24**:483–494.

Heino, J., M. Grönroos, J. Soininen, R. Virtanen, and T. Muotka. 2012. Context dependency and metacommunity structuring in boreal headwater streams. *Oikos* **121**:537–544.

Heino, J., M. Tolkkinen, A. M. Pirttilä, H. Aisala, and H. Mykrä. 2014. Microbial diversity and community-environment relationships in boreal streams. *Journal of Biogeography* **41**:2234–2244.

Hering, D., R. K. Johnson, S. Kramm, S. Schmutz, K. Szoszkiewicz, and P. F. M. Verdonschot. 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.

Holmberg, M., M. Forsius, M. Starr, and M. Huttunen. 2006. An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change. *Ecological Modelling* **195**:51–60.

Jüttner, I., P. D. J. Chimonides, S. J. Ormerod, and E. J. Cox. 2010. Ecology and biogeography of Himalayan diatoms: distribution along gradients of altitude, stream habitat and water chemistry. *Fundamental and Applied Limnology* **177**:293–311.

Jyrkänkallio-Mikkola, J., S. Meier, J. Heino, T. Laamanen, V. Pajunen, K. T. Tolonen, M. Tolkkinen, and J. Soininen. 2017. Disentangling multi-scale environmental effects on stream microbial communities. *Journal of Biogeography* **44**:1512–1523.

Kelly, M., A. Cazaubon, E. Coring et al. 1998. Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of Applied Phycology* **10:**215–224.

Krammer, K., and H. Lange-Bertalot. 1986–1991. *Bacillariophyceae. Süßwasserflora von Mitteleuropa 2 (1-4).* Gustav Fischer Verlag, Stuttgart, Germany.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33**:159–174.

Lange-Bertalot, H., and D. Metzeltin. 1996. *Iconographica diatomologica, Volume 2. Indicators of oligotrophy. 800 taxa representative of three ecologically distinct lake types: carbonate buffered, oligodystrophic, weakly buffered soft water*. Koeltz Scientific Books, Koenigstein, Germany.

Lavoie, I., S. Campeau, M. Grenier, and P. J. Dillon. 2006. A diatom-based index for the biological assessment of eastern Canadian rivers: an application of correspondence analysis (CA). *Canadian Journal of Fisheries and Aquatic Sciences* **63**:1793–1811.

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

Levesque, D., C. Hudon, P. M. A. James, and P. Legendre. 2017. Environmental factors structuring benthic primary producers at different spatial scales in the St. Lawrence River (Canada). *Aquatic Sciences* **79**:345–356.

Liu, S., G. Xie, L. Wang, K. Cottenie, D. Liu, and B. Wang. 2016. Different roles of environmental variables and spatial factors in structuring stream benthic diatom and macroinvertebrate in Yangtze River Delta, China. *Ecological Indicators* **61**:602–611.

Lowe, R. L., and Y. Pan. 1996. Benthic algal communities as biological monitors. In R. J. Stevenson, M. L. Bothwell, and R. L. Lowe (Eds.). Algal Ecology: Freshwater Benthic Ecosystems (pp. 705–739). Elsevier, San Diego.

Maloney, K. O., J. W. Feminella, R. M. Mitchell, S. A. Miller, P. J. Mulholland, and J. N. Houser. 2008. Landuse legacies and small streams: identifying relationships between historical land use and contemporary stream conditions. *Journal of the North American Benthological Society* **27**:280–294.

Maloney, K. O., and D. E. Weller. 2011. Anthropogenic disturbance and streams: land use and land-use change affect stream ecosystems via multiple pathways. *Freshwater Biology* **56**:611–626.

Moravcova, A., O. Rauch, J. Lukavsky, and L. Nedbalova. 2013. The responses of epilithic diatom assemblages to sewage pollution in mountain streams of the Czech Republic. *Plant Ecology and Evolution* **146**:153–166.

Morrill, J. C., R. C. Bales, and M. H. Conklin. 2005. Estimating stream temperature from air temperature: Implications for future water quality. *Journal of Environmental Engineering-asce* **131**:139–146.

National Land Survey of Finland. 2013. *Elevation model 10 m*. https://avaa.tdata.fi/web/paituli. Accessed September 14, 2017.

Olapade, O. A., and L. G. Leff. 2005. Seasonal response of stream biofilm communities to dissolved organic matter and nutrient enrichment. *Applied and Environmental Microbiology* **71**:2278–2287.

Olden, J. D., N. L. Poff, M. R. Douglas, M. E. Douglas, and K. D. Fausch. 2004. Ecological and evolutionary consequences of biotic homogenization. *Trends in Ecology & Evolution* **19**:18–24.

Pan, Y., A. Herlihy, P. Kaufmann, J. Wigington, J. van Sickle, and T. Moser. 2004. Linkages among land-use, water quality, physical habitat conditions and lotic diatom assemblages: A multi-spatial scale assessment. *Hydrobiologia* **515**:59–73.

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

Pajunen, V., M. Luoto, and J. Soininen. 2017. Unravelling direct and indirect effects of hierarchical factors driving microbial stream communities. *Journal of Biogeography* **44**:2376–2385.

Piggott, J. J., R. K. Salis, G. Lear, C. R. Townsend, and C. D. Matthaei. 2015. Climate warming and agricultural stressors interact to determine stream periphyton community composition. *Global Change Biology* **21**:206–222.

Ponsatí, L., N. Corcoll, M. Petrovic, Y. Picó, A. Ginebreda, E. Tornés, H. Guash, D. Barceló, and S. Sabater. 2016. Multiple-stressor effects on river biofilms under different hydrological conditions. *Freshwater Biology* **61**:2102–2115.

Potapova, M., and D. F. Charles. 2003. Distribution of benthic diatoms in U.S. rivers in relation to conductivity and ionic composition. *Freshwater Biology* **48**:1311–1328.

R Development Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org. Accessed September 14, 2017.

Rahel, F. J. 2002. Homogenization of freshwater faunas. *Annual Review of Ecology and Systematics* **33**:291–315.

Rose D. T., and E. J. Cox. 2014. What constitutes Gomphonema parvulum? Long-term culture studies show that some varieties of G. parvulum belong with other Gomphonema species. *Plant Ecology and Evolution* **147**:366–373.

Sandin, L., and P. F. M. Verdonschot. 2006. Stream and river typologies – major results and conclusions from the STAR project. *Hydrobiologia* **566**:33–37.

Skov, F., and J. Svenning. 2004. Potential impact of climatic change on the distribution of forest herbs in Europe. *Ecography* **27**:366–380.

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

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

Stelzer, R. S., and G. A. Lamberti. 2001. Effects of N : P ratio and total nutrient concentration on stream periphyton community structure, biomass, and elemental composition. *Limnology and Oceanography* **46**:356–367.

Stevenson, R. J., M. L. Bothwell, and R. L. Lowe. 1996. *Algal Ecology: Freshwater Benthic Ecosystems*. Elsevier, San Diego, USA.

Stevenson, R. J. 1997. Scale-dependent determinants and consequences of benthic algal heterogeneity. *Journal of North American Benthological Society* **16**:248–262.

Studinski, J. M., K. J. Hartman, J. M. Niles, and P. Keyser. 2012. The effects of riparian forest disturbance on stream temperature, sedimentation and morphology. *Hydrobiologia* **686**:107–117.

Sweeney, B. W., and J. D. Newbold. 2014. Streamside forest buffer width needed to protect stream water quality, habitat, and organisms: a literature review. *Journal of the American Water Resources Associations (JAWRA)* **50**:560–584.

Swets, K. 1988. Measuring the accuracy of diagnostic systems. *Science* **240**:1285–1293.

Taka, M., T. Kokkonen,, K. Kuoppamäki, T. Niemi, N. Sillanpää, M. Valtanen, L. Warsta, and H. Setälä. 2017. Spatio-temporal patterns of major ions in urban stormwater under cold climate. *Hydrological Processes* **31**:1564–1577.

Telford, R. J., V. Vandvik, and H. J. B. Birks, 2006. How many freshwater diatoms are pH specialists? A response to Pither & Aarssen (2005). *Ecology Letters* **9**: E1–E5.

Thuiller, W., D. Georges, R. Engler, and F. Breiner. 2016. *biomod2: Ensemble Platform for Species Distribution Modeling*. mran.microsoft.com/package/biomod2/biomod2.pdf. Accessed September 14, 2017.

Thuiller, W., B. Lafourcade, R. Engler, and M. B. Araújo. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* **32**:369–373.

Tilman, D. 1977. Resource competition between plankton algae: an experimental and theoretical approach. *Ecology* **58**:338–348.

Tonkin, J. D., J. Heino, A. Sundermann, P. Haase, and S. C. Jähnig. 2016. Context dependency in biodiversity patterns of central German stream metacommunities. *Freshwater Biology* **61**:607–620.

Tudesque, L., C. Tisseuil, and S. Lek. 2014. Scale-dependent effects of land cover on water physico-chemistry and diatom-based metrics in a major river system, the Adour-Garonne basin (South Westerns France). *Science of the Total Environment* **466**:47–55.

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

Venäläinen, A., and M. Heikinheimo. 2002. Meteorological data for agricultural applications. *Physics and Chemistry of the Earth* **27**:1045–1050.

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

Von Schiller, D., E. Marti, J. L. Riera, and F. Sabater. 2007. Effects of nutrients and light on periphyton biomass and nitrogen uptake in Mediterranean streams with contrasting land uses. *Freshwater Biology* **52**:891–906.

Vyverman, W., E. Verleyen, K. Sabbe, K. Vanhoutte, M. Sterken, D. A. Hodgson, D. G. Mann, S. Juggins, B. Van de Vijver, V. Jones, R. Flower, D. Roberts, V. A. Chepurnov, C. Kilroy, P. Vanormelingen, and A. De Wever. 2007. Historical processes constrain patterns in global diatom diversitymen. *Ecology* **88**:1924–1931.

Walter, R. C., and D. J. Merritts. 2008. Natural streams and the legacy of water-powered mills. *Science* **319**:299–304.

Wang, L., T. Brenden, P. Seelbach, A. Cooper, D. Allan, R. Clark Jr., and M. Wiley. 2008. Landscape based identification of human disturbance gradients and reference conditions for Michigan streams. *Environmental Monitoring and Assessment* **141**:1–17.

Webb, B. W. 1996. Trends in stream and river temperature. *Hydrological Processes* **10**:205–226.

Weckström, J., A. Korhola, and T. Blom. 1997a. The relationship between diatoms and water temperature in thirty subarctic Fennoscandian lakes. *Arctic and Alpine Research* **29**:75–92.

Weckström, J., A. Korhola, and T. Blom. 1997b. Diatoms as quantitative indicators of pH and water temperature in subarctic Fennoscandian lakes. *Hydrobiologia* **347**:171–184.

Wiens, J. J., and C. H. Graham. 2005. Niche condervatism: Integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics* **36**:519–539.

Winter, J., and H. Duthie. 2000. Epilithic diatoms as indicators of stream total N and total P concentration. *Journal of the North American Benthological Society* **19**:32–49.

Data availability

Data are available in Zenodo: http://doi.org/10.5281/zenodo.2644868

TABLE 1. The summary (minimum, maximum, range, median and standard deviation (Sd)) of the measured variables from 164 impacted (> 5 % anthropogenic land use) and 164 pristine (< 5 % anthropogenic land use) stream sites in Finland.

		Impacted sites					**Pristine sites**				
Variable	Unit	Min	Max	Range	Median	Sd	Min	Max	Range	Median	Sd
Growing degree days (GDD)		631.0	1465.9	834.8	1193.1	130.5	531.7	1450.6	919.0	953.8	220.5
Precipitation (PRECS)	mm	273.2	343.1	69.9	310.4	12.2	253.8	346.9	93.1	313.7	25.5
Water balance (WAB)	mm	269.9	573.4	303.5	364.0	63.5	272.8	624.1	351.3	332.0	95.7
Total phosphorus (TP)	$\mu g\ L^{-1}$	2.0	356.9	354.9	48.5	49.4	0.1	182.5	182.4	18.3	24.9
Conductivity	$\mu S\ cm^{-1}$	12.9	619.0	606.1	91.6	88.9	9.4	161.1	151.7	30	24.3
pH		5.7	8.2	2.5	7.1	0.5	4.5	8.1	3.6	6.7	0.6
Water color	$mg\ Pt\ L^{-1}$	5.0	375.0	370.0	90.0	74.9	2.5	625	622.5	100	87.0
Shading	%	0.0	100.0	100.0	39.8	26.4	0.0	100	100	35.5	26.0
Current velocity	$m\ s^{-1}$	0.0	1.7	1.7	0.3	0.3	0.0	1.7	1.7	0.3	0.3
Anthropogenic land use	%	5.0	65.7	60.8	18.5	15.0	0.0	4.9	4.9	1.3	1.5

Figure legends

FIG.1. Location of the sampling sites (n = 328) in Finland, Northern Europe, divided into two subgroups: human impacted sites (n = 164, > 5 % anthropogenic land use) and pristine sites (n = 164, < 5 % anthropogenic land use). The index map represents the location of Finland in the Northern Hemisphere.

FIG. 2. Relative influence (%) of climatic and local environmental variables and the sums of both variable groups for diatom species (n = 110) distributions separately in human impacted sites (n = 164, > 5 % anthropogenic land use) and pristine sites (n = 164, < 5 % anthropogenic land use). Models were conducted using boosted regression trees and the full set of predictors. The abbreviations stand for growing degree days (GDD), precipitation (PRECS), water balance (WAB) and total phosphorus (TP). Error bars represent standard errors.

FIG. 3. The most important variables for diatom species distributions in impacted and pristine sites. The y-axis represents the number of distribution models in which the variable was the most influential factor. The abbreviations stand for growing degree days (GDD), precipitation (PRECS), water balance (WAB) and total phosphorus (TP).

FIG.4. Relationships between the relative importance of six predictors for diatom species distribution in human impacted and pristine sites. The models were conducted using boosted regression trees as modelling method and the full set of predictors. The full models consist of three climatic predictors (growing degree days (GDD), precipitation (PRECS) and water balance (WAB)) and six local environmental predictors (conductivity, total phosphorus (TP),

pH, water color, shading by the canopy and current velocity). The differences between the site groups are compared in each plot using paired t-test. Dashed lines demonstrate the diagonal line (0, 1).

FIG. 5. Relationships between two predictor variables at the sites of occurrence of five diatom species in Finnish streams. Occurrences at impacted (> 5 % anthropogenic land use) sites are indicated by black dots and occurrences at pristine (< 5 % anthropogenic land use) sites by grey circles. The legends show the relative importance of each predictor on diatom species distributions separately at impacted and pristine sites. The abbreviations stand for growing degree days (GDD) and total phosphorus (TP).