Multiple thresholds and trajectories of microbial biodiversity predicted across browning gradients by neural networks and decision tree learning

Laurent Fontaine, Maryia Khomich, Tom Andersen, Dag Hessen, Serena Rasconi, Marie Davey, Alexander Eiler

ISME

www.nature.com/ismecomms

ARTICLE OPEN

Multiple thresholds and trajectories of microbial biodiversity predicted across browning gradients by neural networks and decision tree learning

Laurent Fontaine[1,2,5], Maryia Khomich[1,3,5], Tom Andersen[1,2], Dag O. Hessen[1,2], Serena Rasconi[4], Marie L. Davey[3] and Alexander Eiler[1,2]✉

Ecological association studies often assume monotonicity such as between biodiversity and environmental properties although there is growing evidence that nonmonotonic relations dominate in nature. Here, we apply machine-learning algorithms to reveal the nonmonotonic association between microbial diversity and an anthropogenic-induced large-scale change, the browning of freshwaters, along a longitudinal gradient covering 70 boreal lakes in Scandinavia. Measures of bacterial richness and evenness (alpha-diversity) showed nonmonotonic trends in relation to environmental gradients, peaking at intermediate levels of browning. Depending on the statistical methods, variables indicative for browning could explain 5% of the variance in bacterial community composition (beta-diversity) when applying standard methods assuming monotonic relations and up to 45% with machine-learning methods taking non-monotonicity into account. This non-monotonicity observed at the community level was explained by the complex interchangeable nature of individual taxa responses as shown by a high degree of nonmonotonic responses of individual bacterial sequence variants to browning. Furthermore, the nonmonotonic models provide the position of thresholds and predict alternative bacterial diversity trajectories in boreal freshwater as a result of ongoing climate and land-use changes, which in turn will affect entire ecosystem metabolism and likely greenhouse gas production.

ISME Communications 2211 ; https://doi.org/10.1038/s43705-021-00038-8

INTRODUCTION

For simplification, ecological associations such as between biodiversity and environmental properties are often assumed to be monotonic, i.e., either positive, negative, or neutral. But in nature, nonmonotonic interactions are commonly seen at the individual, population, community, and ecosystem levels. Most nonmonotonic relations reported in the ecological literature are periodic cycles in time (i.e., prey and predator relationship, ref. [1]) or humped-shaped curves when inferring for example relationships between productivity and biodiversity [2–4]. Nonmonotonicity has been suggested to represent an important driving force in ecological systems because environmental factors are highly variable in both space and time, and organisms do not interact with abiotic and biotic factors in a fixed way [5]. A common feature of nonmonotonic functions is that they define relationships with both increasing and decreasing sectors as well as different stable states where the nature of the response can change dramatically when an environmental factor (i.e., temperature) reaches a threshold (or ridge). Such thresholds are missed by monotonic (linear) models commonly used in ecological data interpretation and modeling. The assumption of monotonicity and resulting over-simplification of biological complexity has been criticized by many ecologists [6, 7].

Standard methods in ecology to perform classification and regression tasks over complex and noisy systems include distance-based regression (MRM), constrained ordinations (RDA and CCA), generalized linear and additive models (GLM, GAM). Decision tree-based machine-learning (regression trees, boosted regression trees, and random forests) and neural networks can fulfill the same objectives [8] and can perform better, especially in cases of non-monotonicity and high complexity [9]. Machine-learning models in microbial ecology literature can be divided into two broad categories: (1) predicting community composition from environmental variables [10] and (2) predicting environmental variables from community composition [11]. Decision tree algorithms and neural networks, however, have limitations like predicting multiple variables at once for the former or evaluating the importance of predictors for the latter. Thus, these methods require careful consideration of how to feed biotic and abiotic data to a model if one is to capture accurately the complexity of microbial ecosystems. Decision trees are well-suited for identifying thresholds in biological systems [12] while allowing one to examine individual trees to understand how each variable in a model contributes to the whole [13]. While neural networks do not easily allow one to examine the contributions of predictors to the model, they are not limited in their ability to capture continuous functions from the data.

[1]Section for Aquatic Biology and Toxicology, Department of Biosciences, University of Oslo, Oslo, Norway. [2]Center for Biogeochemistry in the Anthropocene, Department of Biosciences, University of Oslo, Oslo, Norway. [3]Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, Oslo, Norway. [4]Université Savoie Mont Blanc, INRAE, CARRTEL, Thonon-les-Bains, France. [5]These authors contributed equally: Laurent Fontaine, Maryia Khomich. ✉email: alexander.eiler@ibv.uio.no

Received: 6 April 2021 Revised: 27 July 2021 Accepted: 30 July 2021
Published online: 16 August 2021

An intensively studied relationship in microbial ecology is the link between microbial diversity and natural organic matter (NOM) which represents a major energy source for heterotrophic bacteria [14]. By far the largest NOM pool in aquatic environments is dissolved organic matter (DOM) which is of a complex and heterogeneous nature [15]. Subsets of the diverse DOM pool can have a strong influence on light attenuation, metal speciation, and bioavailability, while also acting as a pH buffer [16]. In recent decades, an increase of DOM loadings to boreal surface waters has been observed [17, 18]. This increase has been linked to a 30% increase in precipitation due to climate change and a projected 15–20% increase in runoff [19]. Exacerbated by land-use change, the increased supply of DOM to lakes and rivers [20] has direct and indirect effects on the microbial loop with implications for phenological events such as the timing of the spring phytoplankton bloom [21] and fish spawning time. Also, increased levels of chromophoric DOM will suppress primary production due to light limitation [22], while providing substratum for heterotrophic bacteria [14, 23], thereby promoting reduced production to respiration ratios. Thus, overall changes to carbon processing by heterotrophic bacterial communities can affect emissions of CO_2 and CH_4 from the boreal landscape and local water quality [24–26].

Complex interactions between heterotrophic bacteria and DOM have been suggested to shape the apparent composition of both of these key ecosystem components [27–31]. This coupling is corroborated by incubation experiments under controlled laboratory conditions where it has been shown that the availability and composition of organic substrates favor specific bacterial groups, and in this way shape bacterial community composition (BCC) and community metabolism [32–36]. Moreover, bacteria do not only consume and degrade DOM but also produce and release an array of autochthonous organic compounds during cell growth, division, and death [37], thereby influencing the availability, composition, and biogeochemical cycling of C in the biosphere [38, 39]. While community adaptation (i.e., composition shifts) has been found to precede bacterial degradation of specific carbon substrates [40], the contribution of bacterial community shifts and key bacterial players to the production and degradation of DOM is unclear [5]. As a result of these multiple levels of interactions and feedbacks, relationships between DOM and bacterial diversity are expected to be nonmonotonic.

Our study is based on samples from 70 large and relatively deep boreal lakes along a 750-km longitudinal gradient across southern Scandinavia. The Scandinavian diversity gradient is complex and not fully resolved as it coincides both with the main postglacial dispersal routes for freshwater biota, as well as with major changes in soil depth, altitude, and landscape productivity [41]. Previous molecular [41, 42] and non-molecular [43] studies have described the diversity and community composition of pelagic protists, aquatic fungi, zooplankton, and fish along a longitudinal gradient in these lakes. Generally, there is a strong decline in diversity across functional and taxonomic groups from east to west. The survey covers a wide longitudinal range and broad gradients in DOM quality and quantity as well as the nutrient status of the systems allowing us to parse out the spatial vs. local environmental effects on bacterial biodiversity.

Here, we aim to capture nonmonotonic features by using modern statistical tools such as generalized additive-models, maximal-information-based nonparametric-exploration (MINE), marginal-(maximum)-likelihood-model-fitting, eXtreme-Gradient-Boosting (XGBoost), and feed-forward-neural networks (FFNN). We tested the hypothesis that threshold responses and alternative trajectories exist in biodiversity responses across browning gradients in freshwater lakes. Taking into account co-varying factors such as nutrient status and other environmental abiotic gradients, we trained XGBoost and FFNNs to predict the interactions between DOM and bacterial community composition in the studied systems so as to identify thresholds in community composition along the studied DOM gradient. Ultimately, we intend to interpolate our findings in light of ongoing environmental change.

MATERIALS AND METHODS

Site description and sampling

Lakes were selected from the "Rebecca" [44] and "Nordic lake survey 1995" [45] datasets on Norwegian and Swedish lakes to create a subset fulfilling the following criteria: longitude 5–18 °E, latitude 58–62 °N, altitude <600 m, surface area >1 km^2, total phosphorus (TP) $<30\ \mu g\ L^{-1}$, total organic carbon (TOC) $<30\ mg\ L^{-1}$, and pH >5. Acidic, eutrophic, and highly dystrophic lakes were omitted. The final subset represents similarly sized boreal lakes within a narrow latitudinal and altitudinal range, with the best possible coverage and tentative orthogonality with respect to gradients of TP, TOC, and longitudinal position. In particular, longitude reflects the regional diversity gradient described in ref. [46], while TP and TOC represent two major and contradictory effects on aquatic productivity [22]. Water temperature, pH, and conductivity were measured in situ, and samples for nutrient analysis were collected as described in ref. [22]. There is a strong relationship between snap-shot temperature measured with the CTD and climatic average mean July air temperature, suggesting that the longitudinal temperature gradient is not confounded by the sampling scheme starting the survey in the west and moving eastward across the gradient. At each site, a water sample was collected from the lake epilimnion (0–5 m) in the central part of each lake during the daytime using an integrating water sampler (Hydro-BIOS, Germany). For DNA extraction, up to 100 mL of water was pre-screened in situ on 100-µm mesh to remove large non-microbial cells and then filtered through 0.2-µm pore size polycarbonate filters (25 mm diameter; Poretics, Spectrum Chemical Corp., NJ, USA) taken in three replicates. The filters were frozen in liquid nitrogen in situ and subsequently stored at −20 °C in cryovials until DNA extraction. The detailed sampling strategy and analytical methods have been previously described [22, 41, 42].

Carbon characterization

TOC was measured by infrared CO_2 detection after catalytic high-temperature combustion (using either a Shimadzu TOC-VWP analyzer or Phoenix 8000 TOC-TC analyzer). Particulate organic carbon (POC) was measured on an elemental analyzer (Flash EA 1112 NC, Thermo Fisher Scientific, Waltham, Massachusetts, USA) through rapid combustion of a pre-combusted GF/C filter with particulates in pure oxygen, where carbon was detected as CO_2 by gas-chromatography. DOC was calculated as the difference between TOC and POC. Carbon quality was assessed via absorbance spectra. After lake water had been filtered through a Acrodisc 0.2-µm polyethersulfone membrane syringe filter (Pall Life Sciences), the optical density of the filtrate ($OD_{CDOM}(\lambda)$) was measured in a 50-mm glass cuvette from 400 to 750 nm in steps of 1 nm. Absorption coefficient spectra of chromophoric DOM ($a_{CDOM}(\lambda)$; m^{-1}) were calculated according to ref. [47].

The absorbance measured at 400 nm (a_{CDOM}) was used as a proxy for aromaticity of chromophoric DOM (CDOM) after dividing by TOC concentrations. Iron can bind to humic substances and form complexes that may increase absorbance [48]. To account for this, a correction factor developed for a_{CDOM} using concentrations of dissolved iron Fe^{3+} was applied.

Non-algal particulate matter (NAP) was assessed by the optical density ($OD_{NAP}(\lambda)$), as described in ref. [22]. Absorption coefficients (m^{-1}) of total particulate matter ($a_p(\lambda)$), and NAP ($a_{NAP}(\lambda)$), were calculated according to ref. [47]. We used the algorithm of Bricaud and Stramski [49] to estimate the path-length amplification factor (β). Finally, we calculated the absorption coefficient spectra of phytoplankton pigments ($a_{ph}(\lambda)$; m^{-1}) as the difference between the total particulate and the NAP absorption coefficient spectra.

DNA extraction, amplification, and Illumina HiSeq sequencing of the V4 SSU

Total DNA was extracted from the filters using the PowerSoil DNA isolation kit (MoBio Laboratories Inc., Carlsbad CA, USA) according to the manufacturer's instructions and quantified using Qubit 2.0 Fluorometer (Invitrogen). The extracted DNA was sent to GATC Biotech (Konstanz, Germany) for amplification and HTS amplicon sequencing (INVIEW Microbiome Profiling 2.0 package). A set of universal primers was used to amplify the hypervariable regions V3–V5

(~569 bp) of the 16S rRNA gene. Amplicon sequencing was done on an Illumina HiSeq Rapid Run instrument using a paired-end 300 bp sequence run. The raw reads with corresponding mapping files were deposited in SRA under accession number PRJNA637765.

Bioinformatics

Raw sequence data were processed with CUTADAPT [50] to remove primers and then analyzed using DADA2 [51]. Forward and reverse reads were trimmed at 200 and 160 bp, respectively. Reads were denoised using the DADA2 machine-learning algorithm. Since trimming resulted in no overlap of the read pairs, forward and reverse reads were concatenated. Quality filtering removed any paired reads with missing primers or ambiguous base pairs as well as a Phred score below 20 somewhere in the paired reads. Taxonomic annotation was performed against the SILVA 132 database [52] using the Naive Bayesian classifier [53].

Statistics

All downstream statistical analyses were performed in R version 3.6.0 [54] using vegan [55], PHYLOSEQ [56], and MASS [57] for multivariate and species richness analyses unless otherwise noted. Missing values in the metadata were approximated using multiple imputation with fully conditional specification (FCS) implemented by the MICE algorithm as described in ref. [58]. CDOM variables used in this study included absorption coefficients at 400 nm (a_{CDOM}) and absorption spectral data between 400 and 750 nm. The entire absorption spectral data were scaled, and principal component analysis (PCA) was performed resulting in a PCA model with principal component 1 (PC1) explaining over 88% of the variance. As such PC1 scores can be used as an index to characterize the CDOM variability among the samples. Partial least-square modeling was performed with packages mdatools (function *randtest*) and plsdepot (functions *plsreg1* and *plsreg2* with cross-validation) using the first six principal components of the PCA from absorption data (*Y* variables) and scaled environmental data (*X* variables).

The two technical replicates were excluded from further downstream analyses as within replicate sequence variants were significantly more similar than between-sample comparisons (data not shown). To calculate diversity measures, the sequence variant table was rarefied to a common sampling depth of 392,082 reads/sample, based on the sample with the least number of reads. Species accumulation curves (SAC; calculated using the analytical version of the *specaccum* function) were applied to assess sampling effort in each lake. Rarefaction curves were constructed for each lake using the *rarecurve* function in vegan. Alpha-diversity indices (observed richness, Shannon, and Simpson diversity) were calculated for each lake using the function *diversity* (R package vegan) and in addition Faith's phylogenetic diversity was calculated. Associations between alpha-diversity indices and DOM descriptors were explored with generalized additive models (GAMs) using R package mgcv.

Non-metric multidimensional scaling (NMDS) ordinations [59] from multiple starting points (*metaMDS* function in vegan, *try* = 1000) were used to describe patterns in bacterial community composition (based on Hellinger-transformation and Bray–Curtis or on unweighted unifrac distance measures) between lakes. Permutation-based significance tests ($n = 999$) with the *envfit* function were used to fit spatial and environmental gradient variables to the NMDS ordination. The local environment was defined by the concentrations of total, particulate and dissolved CNP, and other parameters (see Supplementary Table S1 for a complete list of variables), while the spatial factors were represented by longitude, latitude, and altitude. In addition, a redundancy analysis (RDA) was performed on bacterial community composition (Hellinger-transformed) using scaled environmental data.

To determine the relative role of DOM descriptors (a_{CDOM}, PC1-CDOM and TOC), local (all other environmental variables), and regional (spatial factors) predictors on the distribution of bacterial communities along the biodiversity gradient, variance partitioning analysis was used. Variation partitioning by RDA (function *varpart* in vegan) [55] on Hellinger-transformed, normalized abundance data were used to estimate the fractions of bacterial community composition variation that could be explained independently by the local environment divided into DOM and other parameters, spatial gradients (latitude and longitude), or shared between them. Marginal (maximum) likelihood model fitting was used to fit a smooth response surface of TOC and a_{CDOM} values over the limits of the biplot of the bacterial community composition using the *ordisurf* function.

Machine-learning algorithms were used to identify beta-diversity patterns along the CDOM and TOC concentration gradients. Regression was performed with a_{CDOM} and TOC values as inputs and BCC Bray–Curtis distances as outputs using scikit-learn's implementation of XGBoost and random forest regressor (of the PyPi packages) as well as TensorFlow for the FFNNs using backpropagation [60]. In short, data were split into training and test sets comprising 80 and 20% of observations, respectively. For the FFNN, the weights of hidden layers were initialized using Xavier´s initialization [61], with ReLu activation and mean-squared error being used as a cost function. For visualization of the models, the original meshgrid of a_{CDOM} and TOC values spanning the minimum–maximum range of said gradient with a step size equal to the smallest pairwise a_{CDOM} and TOC differences was used. XGBoost stands for Extreme-Gradient Boosting and represents a specific implementation of the Gradient Boosting method and uses more accurate approximations to find the best (decision-)tree model. In prediction problems involving unstructured data (images, text, etc.), neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data as in our case, decision tree-based algorithms are considered to be better suited. XGBoost is exceptionally successful, particularly with structured data since it computes second-order gradients, i.e., second partial derivatives of the loss function (similar to Newton's method), which provides more information about the direction of gradients and how to identify minima in the loss function. XGBoost uses the 2nd order derivative as an approximation and advanced regularization, which improves model generalization. The accuracy of the methods was compared using mean-squared error (MSE) while the variance of raw data explained by the model was computed with R^2.

In addition, we performed ordinary least squares (OLS) regression by singular value decomposition (SVD) using polynomials of a_{CDOM} and TOC values as inputs. An appropriate polynomial degree was chosen in light of the bias-variance trade-off, where the error was minimal while bias and variance curves intersected.

In the maximal information-based nonparametric exploration (MINE, ref. [62]) analysis run with default settings, relationships with P values of <0.05 were recorded with a false discovery rate, as determined by Hochberg, of <0.05 (q values). The chosen P value set the maximal-information coefficient (MIC) cutoff to 0.3. The MIC is a statistical measure, similar to R^2 in general linear models, describing the goodness of fit between two variables [62]. Various statistics can be used to characterize the relationships identified by MIC, including measures of monotonicity, non-linearity, closeness to be a function, and complexity of relationships. The Maximum Asymmetry Score (MAS) measures the deviation from monotonicity. We plotted the variability in MIC and MAS between amplicon sequence variants (ASVs) and a_{CDOM} or TOC with q values <0.05 and linked them with the sign of the correlation coefficient (Spearman R).

RESULTS

The lake gradient through DOM quantity and quality

The sampled lakes spanned from the Norwegian coast of the North Sea to the Swedish East Coast of the Baltic Sea and represent summer conditions as samples were taken from July 20, 2011 to August 16, 2011 (Fig. 1A). Besides varying in latitude (58–62 °N) and temperature (9.9–21.4 °C), lakes varied in nutrient content with TOC in the lakes ranging from 0.3 to 12.9 mg l^{-1} (median 6.5 mg l^{-1}), TP from 0.5 to 27.5 µg l^{-1} (median 4.55 µg l^{-1}), total organic nitrogen (TON) from 87 to 1526 µg l^{-1} (median 298 µg l^{-1}), and chlorophyll *a* from 0.77 to 29.5 (median 2.7). Lake size varied from 1.09 to 140 km^2 with a median of 3.4 km^2 (Supplementary Table S1; ref. [22]).

The PCA revealed substantial differences in DOM quality along the sampled lakes as assessed by absorption spectra (Fig. 1B). First, the relative positioning of the sample scores was mainly a function of PC1 which explained 88% of the variability. This component was a function of TOC concentration and a_{CDOM} as revealed by Spearman rank-correlation analyses ($R = 0.75$; $P < 0.0001$ and $R = 0.8$; $P < 0.0001$, respectively). Other significantly correlated ($P < 0.0001$) environmental variables with CDOM (PC1) and a_{CDOM} included gas concentrations (O_2, CO_2, CH_4), chlorophyll a, total (TP), and particulate nutrient concentrations (PON, POC, and POP) (see also Supplementary Fig. S1).

Furthermore, partial least squares (PLS) were applied to predict CDOM (PC1-3) variability, (*Y* response variables) from lake water chemistry and climate variables (X predictor variables). Variables (*X*

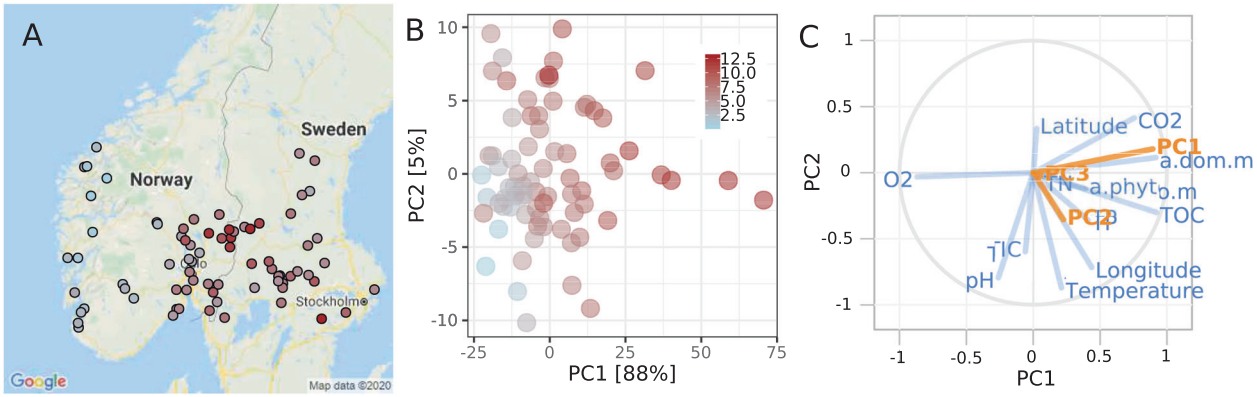

Fig. 1 Physico-chemical properties of study sites. Map of sampling locations (**A**) with total organic carbon concentrations (mg L^{-1}) in the lake system indicated by point color. Principal component analysis (PCA) (**B**) for the quality of dissolved organic matter (DOM) as assessed by absorbance spectra. Partial least-square (PLS) loading plot (**C**) revealing the covariation of the first three principal components for the quality of dissolved organic matter (CDOM), which were taken from analysis in panel B (*Y* variables in orange), and geographical, physical, and chemical lake characteristics as predictors (*X* variables in blue). The comparison of observed and model predictions of CDOM is shown in Supplementary Fig. S2 corroborating the high predictive power of the PLS model ($R^2 = 0.815$ and $Q^2 = 0.775$) when using environmental properties.

or *Y*) situated close together on the PLS plot such as CO_2, TOC, and a_{CDOM} can be interpreted as positively correlated with CDOM PC1, while variables opposite to CDOM PC1 as negatively correlated, such as O_2 (Fig. 1C). Next, we performed a PLS using only CDOM PC1 (Supplementary Fig. S2A) with internal cross-validation to test the repeatability of the analysis by removing a random subset of data (1/7th of the samples) to be used as the response dataset, while parallel models were run on the reduced calibration dataset. A comparison of predicted values from the calibration and response datasets allowed computation of the predictive residual sum of squares expressed as a Q^2Y. Overall, the PLS model performance was good with cumulative goodness of fit (R^2Y, explained variation) of 0.815, and the cumulative goodness of prediction (or Q^2Y, predicted variation) of 0.775 for the PLS model with two components. This was also corroborated by comparing original and predicted values (Supplementary Fig. S2B). As such the PLS model corroborates the results of the multiple correlation analysis (Supplementary Fig. S1) with CO_2, TOC, and O_2 concentrations representing environmental properties highly related to CDOM characteristics expressed by CDOM PC1 and a_{CDOM}.

Overall bacterial community features and diversity

A total of 15,120 unique sequence variants (including 864 archaeal and eukaryotic reads) were recovered from 25,574,631 high-quality reads across 72 lakes. After removing non-bacterial reads, an average of 764 (range = 354–1454, SD = 208) ASVs was detected per sample and the mean number of reads per lake was 352,572 (range = 234,930–502,704, SD = 57,780). A total of 174 ASVs were detected in more than 50 lakes with the mean number of total reads per lake ranging from 219 to 5754 reads and representing some of the most abundant sequence variants in our dataset. Rarefaction curves of ASV richness (Supplementary Fig. S3A) for each lake indicated that the total bacterial diversity was almost entirely recovered in all samples since the rarefaction curves approached asymptote and sampling saturation. Still, region-wide species accumulation curves based on the progressive or random addition of samples showed that the gamma diversity in the studied area has not been fully recovered (Supplementary Fig. S3B).

Various diversity indices were highly correlated in the present dataset. For example, bacterial diversities calculated using inverse Simpson, Shannon, Fisher, Faith's phylogenetic diversity, and ACE (abundance-based coverage estimators) diversity were highly correlated: $R > 0.46$, $P < 0.0005$. For example, ACE richness increased with TOC ($R = 0.23$, $P < 0.05$), CDOM PC1 ($R = 0.26$, $P < 0.03$), $a_{ph}(\lambda)$; m^{-1} ($R = 0.25$, $P < 0.05$) and a_{CDOM} ($R = 0.27$, $P < 0.025$), but not POC (for more details see Supplementary Fig. S1). Further assessment of the associations by GAMs revealed that including non-monotonicity improved the models between alpha-diversity (ACE richness and Shannon index), and organic matter descriptors considerably (i.e., as indicated by AIC, GCV, R^2 and chisq; Supplementary Table S2). Resulting GAMs revealed a peak in alpha-diversity (ACE richness and Shannon diversity) at intermediate browning, i.e., CDOM PC1 and a_{CDOM} (Supplementary Fig. S4). Such humped-shaped curves in association studies of alpha-diversity have been observed widely as for example when inferring relationships between productivity and biodiversity [2–4].

Spatial and environmental factors affecting bacterioplankton community composition

Bacterial community dissimilarity as estimated by Bray–Curtis distance increased significantly with geographic distance which, despite a pronounced scatter and low coefficient ($R^2 = 0.066$), exhibited significant distance-decay relationships ($P < 0.0001$). Similarly, variance partitioning analysis revealed that the fraction of the total community variation that could be explained solely by spatial factors (longitude and latitude 1.2%) was small. In comparison, the fraction that could be solely explained by local environment conditions was 11.2% combined for CO_2, TN, PO_4, and temperature, and 5.0% for TOC, CDOM, and a_{CDOM} while with shared effects of 20.4% and 13.1%, respectively (Fig. 2A). Approximately 72% of the community variance along the sampled lake gradient remained unexplained by the measured environmental and spatial gradient indicators assuming monotonic relationships.

The environmental properties showing a high co-variance with bacterial community composition (Bray–Curtis distance) were both spatial and environmental gradients including longitude ($R^2 = 0.247$, $P = 0.001$), latitude ($R^2 = 0.183$, $P = 0.001$), temperature ($R^2 = 0.341$, $P = 0.001$), and concentrations of total nitrogen ($R^2 = 0.242$, $P = 0.001$), total phosphorus ($R^2 = 0.235$, $P = 0.001$), CO_2 ($R^2 = 0.174$, $P = 0.004$) and PO_4 ($R^2 = 0.384$, $P = 0.001$). As revealed by RDA, the direction of maximal increase for the fitted vectors representing longitude, temperature, and CO_2 was similar but orthogonal to the vectors reflecting a nutrient state (i.e., TN, TP, and PO_4 concentrations) (Fig. 2B). This can be interpreted that there are two main directionalities driving bacterial community composition in lakes, corresponding to nutrient status and temperature.

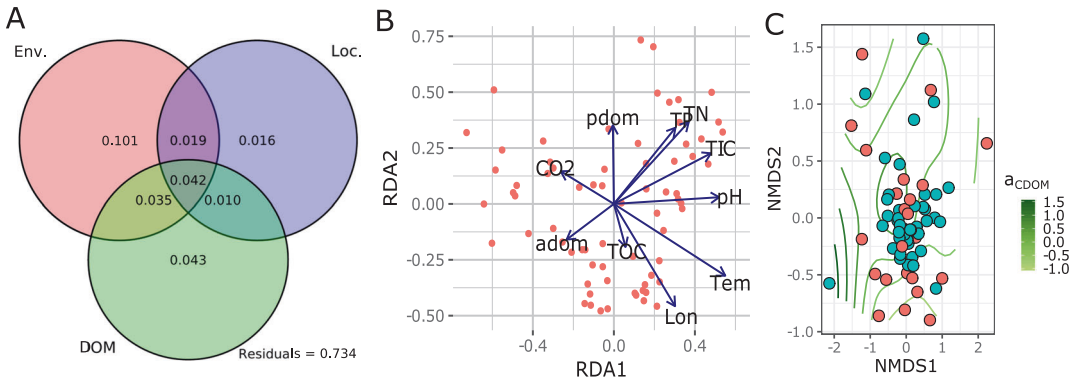

Fig. 2 Bacterial community composition along environmental gradients. Partitioning of the total variance in the bacterial community (given by Bray–Curtis distances) (**A**) with environmental (Env.), organic matter properties (DOM), and spatial (Loc.) descriptors. Results from an unconstrained redundancy analysis (**B**) showing the covariation between the composition of bacterial communities and environmental factors. Arrows represent fitted gradient vectors for spatial (Lon—longitude) and environmental (Tem—water temperature, pH, CO_2—carbon dioxide, TOC—total organic carbon, TP—total phosphorus, TN—total nitrogen, TIC—total inorganic carbon, adom—a_{CDOM} a proxy for aromaticity of CDOM and pdom—absorption coefficient spectra of phytoplankton pigments) variables. Ordisurf (**C**) with a_{CDOM} revealing the nonmonotonic relationship with bacterial community composition. In Supplementary Fig. S5, results using unifrac distances are presented for comparison.

Furthermore, when using phylogenetic distance (unifrac-based dissimilarity) the fraction of the total community variation that could be explained by spatial factors (longitude and latitude) was 11%, local environment conditions explained 21.9% (CO_2, TN, PO_4, and temperature), and TOC, CDOM, and a_{CDOM} 20% (Supplementary Fig. S5A). High co-variance with unifrac-based community composition were both spatial and environmental gradients including longitude ($R^2 = 0.185$, $P = 0.003$), latitude ($R^2 = 0.151$, $P = 0.005$), temperature ($R^2 = 0.344$, $P = 0.001$), and concentrations of total nitrogen ($R^2 = 0.128$, $P = 0.012$), total phosphorus ($R^2 = 0.175$, $P = 0.005$), CO_2 ($R^2 = 0.146$, $P = 0.007$), and PO_4 ($R^2 = 0.178$, $P = 0.003$). Furthermore, the Canonical Analysis of Principal Coordinates (CAP) on unifrac distances confirmed RDA results based on Bray–Curtis distances that there are two main directionalities driving bacterial community composition in lakes, corresponding to nutrient status and temperature (Supplementary Fig. S5B).

Monotonic functions as used in RDA revealed short vectors for TOC, a_{CDOM}, and CDOM PC1 which can be interpreted that organic matter is a poor predictor of bacterial community compositions in lakes. However, there is no reason to assume that TOC, a_{CDOM}, and CDOM PC1 vary in a monotonic fashion across the RDA's biplot (Fig. 2B), which is a prerequisite to identify relationships in unconstrained ordination. To reveal potential nonmonotonic relations, we fitted a smooth response surface of TOC and a_{CDOM} values over the limits of the biplot using *ordisurf* function (i.e., for a_{CDOM} see Fig. 2C and for TOC Supplementary Fig. S6; corresponding results from unifrac distances in Supplementary Fig. S5C). The fitted surfaces are far from monotonic and revealed that the relationships of a_{CDOM} and TOC with the bacterial community are significant ($P < 0.001$) and explained large parts of the variability (a_{CDOM}: adj. $R^2 = 0.3$; deviance explained 41.8%; and TOC: adj. $R^2 = 0.30$; deviance explained 36.5%) when performing smoothness selection via marginal (maximum) likelihood model fitting.

Similar beta-diversity patterns appeared along the a_{CDOM} gradient for both XGBoost, random forest and FFNN models (Fig. 3A–C). The mean value of the response surface (i.e., 0.916 in the XGBoost models for TOC and a_{CDOM}) can be treated as the baseline beta-diversity across all sites. Data points with values below the mean present higher similarities between sites; likewise, higher values represent lower similarity. Data points located on the diagonal are not presented as they are pairwise distances of a site to itself, thus assumed to be zero. To interpret the response surfaces, one may begin by looking at a point bordering the diagonal and then follow a line of points further up on the a_{CDOM} site 2 axis. A "ridge" indicates a a_{CDOM} value next to the diagonal to be a likely threshold from which the shift in bacterial community composition is greater than average. In the same manner, a "valley" indicates a a_{CDOM} value next to the diagonal is likely located on an interval of the a_{CDOM} gradient along which bacterial communities do not shift substantially. Following this interpretation, a_{CDOM} thresholds for high variation in BCC appear around 0.3, 0.5, and 1–1.5 absorbance units, while communities are more similar to others with higher a_{CDOM} around 0.4, 0.6, and 1.6–2.25 absorbance units. In comparison, the linear model captured the greater variation in beta-diversity pattern above 2 absorbance units on the a_{CDOM} gradient (Supplementary Fig. S7), but not the multiple ridges or valleys revealed by XGBoost, random forest regression, and FFNN (Fig. 3A–C).

Similarly, model results revealed "ridges" along the TOC gradient (Fig. 3D–F), indicative for thresholds at which shifts in BCC are greater than average. These TOC thresholds for high variation in BCC appear around 0.3, and 2–3 and 6.5 mgC L^{-1}. "Valleys" indicative for an interval of the TOC gradient where bacterial communities do not shift substantially were predicted to be around 1.5, 4–5, and 8 mgC L^{-1}. Overall, R^2 values of the XGBoost model predictions ($R^2_{TOC} = 0.446$; $R^2_{aCDOM} = 0.315$) showed smaller differences between the observed data and the fitted values, than the FFNN ($R^2_{TOC} = 0.068$; $R^2_{aCDOM} = 0.014$) and the random forest ($R^2_{TOC} = 0.414$; $R^2_{aCDOM} = 0.172$) models. Training the models with and without the duplicate samples did not affect the models.

Association between DOM and bacterial taxonomic groups

Altogether, 29 bacterial phyla were detected resembling results in line with the global synoptic meta-analysis of 16S rRNA gene sequences from lake epilimnia [63] and Zwart et al. [64], showing that four phyla (Proteobacteria, Actinobacteria, Bacteroidetes, and Cyanobacteria) were recovered commonly across the sampled freshwater ecosystems (Supplementary Fig. S8A). With regards to the number of ASVs, Proteobacteria was the most diverse phylum (4414 ASVs) followed by Bacteroidetes (1317 ASVs), while Cyanobacteria and Actinobacteria had similar richness (784 and 652 ASVs, respectively) (Supplementary Fig. S8B). By further resolving the taxonomy to the genus level, the most abundant identified groups were alfIV-A (LD12) (6.0%), *Aquincola* (4.9%), various acI (4.6%), *Synechococcus* (3.1%), *Niveitalea* (2.1%), and *Methyloferula* (1.5%).

To explore relationships between ASVs and environmental properties, we used MINE [62]. While this nonparametric approach identifies

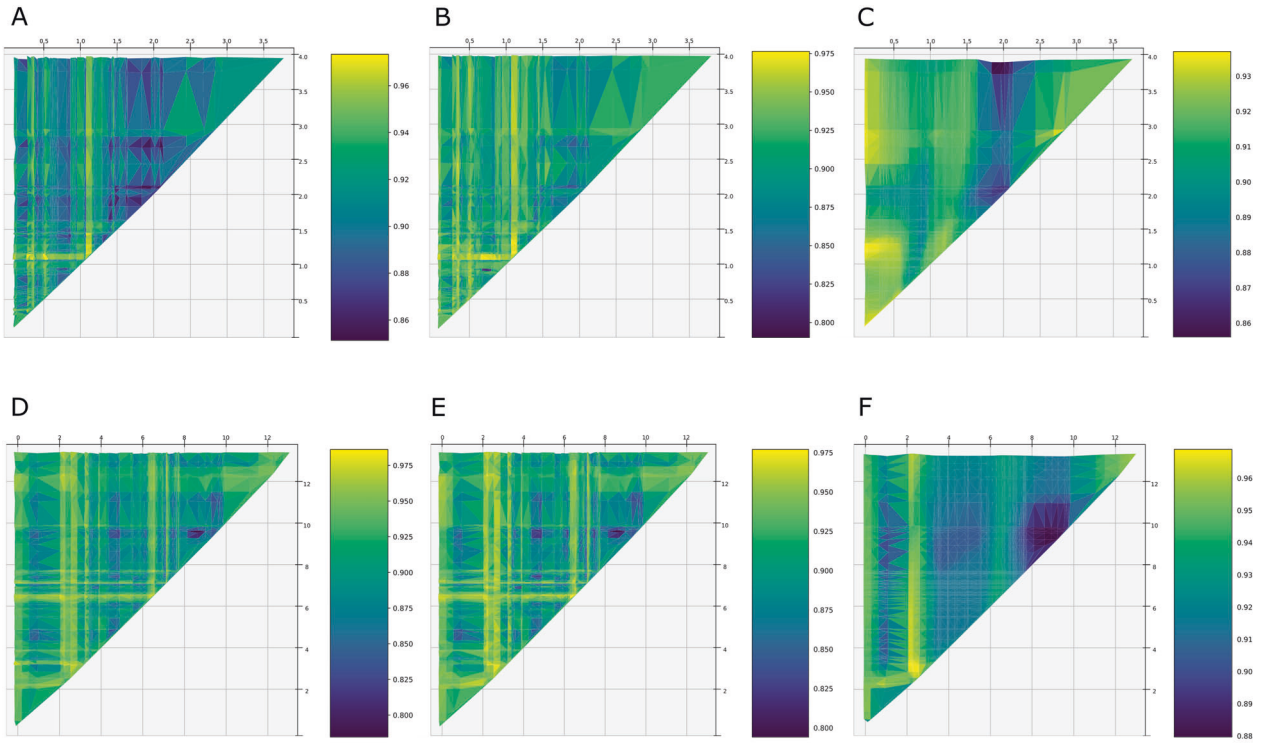

Fig. 3 Decision tree and neural network models for beta diversity along DOM gradients. Visualization of XGBoost (**A**, **D**), random forest (**B**, **E**), and feed-forward neural network (**C**, **F**) predictions of bacterial community compositional changes (Bray–Curtis distances). Compositional changes were predicted for a meshgrid of a_{CDOM} (**A**–**C**) and TOC (**D**–**F**) values spanning the minimum–maximum range of the gradient with a step size equal to the smallest pairwise a_{CDOM} and TOC differences. The mean value of the response surface can be treated as the baseline beta-diversity across all sites. Data points with values below the mean represent higher similarity between sites; likewise, higher values represent lower similarity. To interpret the response surfaces, one may begin by looking at a point bordering the diagonal and then follow a line of points further up on the a_{CDOM} or TOC site 2 axis. Here, a "ridge" indicates a a_{CDOM} or TOC value next to the diagonal to be a likely threshold from which the shift in bacterial community composition is greater than average. In the same manner, a "valley" indicates a a_{CDOM} or TOC value next to the diagonal which is likely located on an interval of the a_{CDOM} or TOC gradient along which bacterial communities do not shift substantially.

relationships of ASVs with all measured environmental variables, we will focus on the results from the analyses with a_{CDOM} and TOC. MINE identified 108 significant relationships (q value <0.05) with a MIC of 0.316–1 between single ASVs and TOC while 92 ASVs were identified with significant relationships with a_{CDOM} with a MIC of 0.316–1. The maximum asymmetry score (MAS) ranged from around 0.05–0.58 for a_{CDOM} and 0.05–0.67 for TOC. MAS values below 0.05 indicate a monotone relationship between ASVs and a_{CDOM} or TOC (Fig. 4). While purely monotone relationships were not detected, nonmonotonic responses dominated which are indicative of the existence of thresholds in the response of ASVs along the sampled CDOM and TOC gradients, similar to the model predictions of the entire bacterial community responses.

DISCUSSION

We show that freshwater microbial diversity is likely impacted by browning with implications for the functioning of lake ecosystems. Such anthropogenic-induced changes in microbial diversity have been reported in multiple studies [65, 66]. Here, the presence of thresholds within nonmonotonic relationships was revealed using machine-learning algorithms. Both alpha and beta-diversity were poorly predicted by monotonic functions, as the variation explained was scarcely exceeding 5% when using linear models, RDA, and variation partitioning. The variation explained increased with models taking deviations from monotonicity into account. For example, the fraction of variance explained in beta-diversity increased up to 45% when using XGBoost, 41% with random forest regressor, and 6.8% with FFNN while 30% with marginal likelihood models. In addition, we demonstrate that most relationships between bacterial taxa (ASVs), and TOC concentrations and chromophoric properties of the water were nonmonotonic.

A common feature of nonmonotonic functions is that they define relationships with both increasing and decreasing sectors as well as different stable states ("valleys") where the nature of the response can change suddenly when an environmental factor (i.e., browning) reaches a threshold ("ridge"). Results from the marginal likelihood model fitting can be interpreted along these lines since the model reveals distinct a_{CDOM} and TOC types coinciding with distinct environmental conditions and bacterial community composition profiles. Such non-monotonicity in response to DOM (a complex of substrates for microbial growth) can be predicted from kinetics studies emphasizing that growth may not be controlled by only a single compound but by two or more compounds simultaneously and that kinetic properties of a community might change due to adaptation of individual cells or community composition to ever-changing environmental conditions [67].

To capture further details and validate the findings of the marginal likelihood model fitting such as thresholds and non-monotonicity in bacterial community responses along the DOM gradient, we applied machine-learning methods, in particular FFNN, random forest, and XGBoost. A key finding revealed by the machine-learning methods is the apparent presence of multiple thresholds ("ridges") along the a_{CDOM} and TOC gradients where bacterial community composition shifts, corroborating the

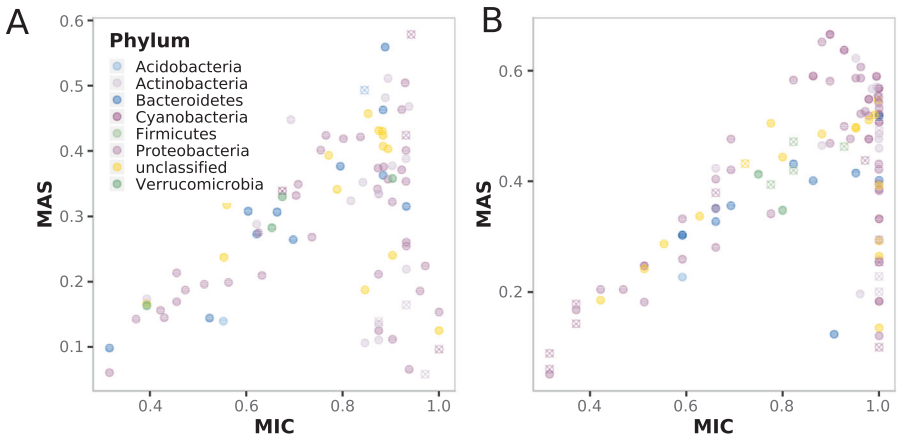

Fig. 4 MINE model for beta diversity along DOM gradients. Plots summarizing MINE statistics of the relationships between ASVs and a_{CDOM} (**A**), and ASVs and TOC (**B**). Depicted MINE statistics are MIC—coefficient, MAS—non-monotonicity. The color of the symbols indicates the taxonomic affiliation of the ASV at the phylum level.

predictions of the marginal likelihood model fitting. If bacterial community composition had been found to vary linearly along the TOC and a_{CDOM} gradients, they would have presented a pattern of isolines parallel to the diagonal in Fig. 3; and there would have been a linear relationship between BCC distances and a_{CDOM} or TOC differences between sites (Supplementary Fig. S7). The observed thresholds ("ridges") in TOC concentrations and a_{CDOM} values can be interpreted as "guardrails" of biodiversity along the browning gradient. These "guardrails" guard alternative trajectories at low browning which converge into a single almost monotonic (linear) trajectory when TOC concentrations (above 10 mgC L^{-1}) or chromophoric properties reach high levels (above 2.5 absorbance units). At low TOC and a_{CDOM}, community patterns seem to resemble alternative steady states persisting under equal environmental conditions [68]. Furthermore, our results point to alternative trajectories (dynamic regimes) in biodiversity separated by "guardrails" which start at 0.3, 0.5, and 1–1.5 absorbance units of a_{CDOM}, respectively. For TOC, the separating "guardrails" are predicted to start at around 0.5, and 2–3 and 6 mgC L^{-1}, interpretable as boundaries with enforced resilience keeping bacterial communities within different trajectories ("valleys"). This resembles Lyapunov function hills or ridges between attractor wells (the proverbial "marble in a cup" [68]) and emphasizes the usefulness of machine-learning models in predicting nonmonotonic biodiversity responses across environmental gradients when internal processes and external forcing mechanisms are unknown.

Predicting the entire ASV table prior to computing beta-diversity indices was avoided because random forest and XGBoost do not allow multi-target modeling. Ways around this are to use single-target modeling for each desired output variable, multi-regressor stacking, or regressor chains [69]. The issue with single-target modeling for a multi-target problem is that dependencies between targets are not taken into account. As for multi-regressor stacking and regressor chains, while they take dependencies between targets into account, the order of chaining matters and optimizing order by permutational tests quickly gets out of hand as the number of targets increases. These problems do not happen with neural networks as they allow multi-target outputs. In short, our approach of modeling beta-diversity as a single-target problem holds the advantages of predicting a single value in which the complexity of multi-target dependencies is contained within. This prevents a loss in model performance for random forest and XGBoost as well as eliminating the need for computationally unfeasible optimization. Free from the constraint of dataset size, it would be best to predict the whole microbial community from the whole environmental data using neural networks, but our dataset is too small to allow a satisfying performance. Directly predicting beta-diversity yielded satisfactory results with all three algorithms as is apparent in the similar beta-diversity patterns along the a_{CDOM} and TOC gradients. XGBoost reflected the raw data more closely (greater R^2) and was orders of magnitude faster than the FFNN. This corroborates the previous observation that decision tree-based algorithms such as XGBoost outperform neural networks when small-to-medium structured/tabular data is used, as in our case.

A potential explanation for the observed non-monotonicity between browning and microbial diversity could be attributed to other environmental parameters such as nutrients, temperature, and geography. These parameters will turn up as noise in the machine-learning models, and in addition the information of other measured variables was well contained in absorption spectra (CDOM) as indicated for instance by the PLS and correlation analysis. This emphasizes the problem of covariation and interdependence as in the case of TOC and CDOM with nutrients and other parameters, and as such machine-learning models cannot be used to attribute causality. Machine-learning methods are designed to optimize the ability to predict an outcome on an external dataset (i.e., biodiversity responses across browning gradients) using a training set to learn patterns associated with an outcome and a test set to determine the performance of the model.

A testable step in the causal chain to explain the apparent non-monotonicity in the relationship between browning and biodiversity is the non-interchangeable nature of individual taxa responses. Individual taxa responses can be direct and indirect with opposite and non-additive strategies based on changes in the environment. This is reflected by browning mostly leading to nonmonotonic relationships as shown by the high number of ASVs with high MAS (Fig. 4). The non-monotonicity in response to environmental stimuli can be explained by organisms' ability to adopt opposite strategies along the stimuli´s gradient. In the case of browning, terrestrially derived TOC provides a significant source of C for heterotrophic bacteria [14, 70] and where different fractions of this TOC are utilized with different efficiency [71]. The different fractions are also utilized by different taxa, which, as shown in our study, leads to different ASVs being present along the browning gradient. These opposing positive and negative effects on individual ASVs are only monotonic if they change in the same order or scale so that their net effect will be additive. However, if the positive and negative effects change in different orders or scales, which is common in nature, their net effect will not be additive, and the function will be nonmonotonic. This is reflected in the high number of nonmonotonic relationships in the co-occurrence patterns among ASVs (Supplementary Fig. S9). Additional potential explanations for the apparent nonmonotonic

responses of individual taxa are shifts in interaction behavior with examples such as the Prisoner's Dilemma [72] and opposing dual effects between organisms.

As shown by previous studies, seasonality, water mixing, as well as source and age of TOC, clearly offer different sources of energy that may select for different microbial community members and metabolic pathways at both short and long timescales. The nonmonotonic responses in community composition, as observed in our study, are likely also reflecting a trade-off between nutrients associated with CDOM and the increasing light attenuation caused by CDOM. Modest increases in TOC and CDOM have been shown to block out short-wave UV radiation [73] and to limit autochthonous production of TOC. Since browning is increasing by processes associated with climate change [17, 19] and the strong decline of atmospheric sulfur (S) deposition [18, 19], we predict, by translating our model results based on spatial data into a temporal context, that lake bacterioplankton diversity will develop along different trajectories ("valleys") guided by thresholds ("ridges" or "guardrails") at low browning (i.e., low TOC concentrations and low chromophoric properties). Lakes with high levels are predicted to follow a closely monotonic trajectory of biodiversity change over time. Considering that browning is an ongoing process, alternative trends of bacterial diversity in lakes currently experiencing low TOC and CDOM levels are expected while more uniform and monotonic trends are predicted in lakes with high levels of browning (above 10 mgC L^{-1}). As such our study provides some estimates on microbial biodiversity trends that can result from climate change, although our spatial gradient design centered on TOC needs to be complemented with long-term time series data for validation.

To conclude, our results highlight the need to explore nonmonotonic relationships common in biological systems which might provide part of the explanation of contrasting results among different studies, in addition to revealing the real complexity of associations between biodiversity and environmental properties. Most importantly, by using nonmonotonic functions and modeling the position of thresholds, alternative trajectories and guardrails can be revealed which are important for mitigation efforts and management decisions to counteract environmental changes [65] not only in freshwater microbiomes affected by browning.

REFERENCES

1. Vik JO, Brinch CN, Boutin S, Stenseth NC. Interlinking hare and lynx dynamics using a century's worth of annual data. Popul Ecol. 2008;50:267–74.
2. Luo G, Han Q, Zhou D, Li L, Chen X, Li Y, et al. Moderate grazing can promote aboveground primary production of grassland under water stress. Ecol Complex. 2012;11:126–36.
3. McNaughton S. Grazing as an optimization process: grass-ungulate relationships in the Serengeti. Am Nat. 1979;113:691–703.
4. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551:457–63.
5. Zhang Z, Yan C, Krebs CJ, Stenseth NC. Ecological non-monotonicity and its effects on complexity and stability of populations, communities and ecosystems. Ecol Modell. 2015;312:374–84.
6. Devin S, Giamberini L, Pain-Devin S. Variation in variance means more than mean variations: what does variability tell us about population health status? Environ Int. 2014;73:282–7.
7. Steudel B, Hector A, Friedl T, Löfke C, Lorenz M, Wesche M, et al. Biodiversity effects on ecosystem functioning change along environmental stress gradients. Ecol Lett. 2012;15:1397–405.
8. Christin S, Hervet É, Lecomte N. Applications for deep learning in ecology. Methods Ecol Evol. 2019;10:1632–44.
9. De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 2000;81:3178–92.
10. Larsen PE, Field D, Gilbert JA. Predicting bacterial community assemblages using an artificial neural network approach. Nat Methods. 2012;9:621–5.
11. Sperlea T, Kreuder N, Beisser D, Hattab G, Boenigk J, Heider D. Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework. Mol Ecol. 2021;30:2131–44.
12. Schiel DR, Lilley SA, South PM. Ecological tipping points for an invasive kelp in rocky reef algal communities. Mar Ecol Prog Ser. 2018;587:93–104.
13. Robinson B, Cohen JS, Herman JD. Detecting early warning signals of long-term water supply vulnerability using machine learning. Environ Model Softw. 2020;131:104781.
14. Hessen DO, Andersen T, Lyehe A. Carbon metabolism in a humic lake: pool sizes and cycling through zooplankton. Limnol Oceanogr. 1990;35:84–99.
15. Nebbioso A, Piccolo A. Molecular characterization of dissolved organic matter (DOM): a critical review. Anal Bioanal Chem. 2013;405:109–24.
16. Coble PG, Lead J, Baker A, Reynolds DM, Spencer RG (eds). Aquatic organic matter fluorescence. Cambridge: Cambridge University Press; 2014.
17. Finstad AG, Andersen T, Larsen S, Tominaga K, Blumentrath S, de Wit HA, et al. From greening to browning: catchment vegetation development and reduced S-deposition promote organic carbon load on decadal time scales in Nordic lakes. Sci Rep. 2016;6:1–8.
18. Monteith DT, Stoddard JL, Evans CD, de Wit HA, Forsius M, Høgåsen T, et al. Dissolved organic carbon trends resulting from changes in atmospheric deposition chemistry. Nature. 2007;450:537–40.
19. de Wit HA, Valinia S, Weyhenmeyer GA, Futter MN, Kortelainen P, Austnes K, et al. Current browning of surface waters will be further promoted by wetter climate. Environ Sci Technol Lett. 2016;3:430–5.
20. Meyer-Jacob C, Tolu J, Bigler C, Yang H, Bindler R. Early land use and centennial scale changes in lake-water organic carbon prior to contemporary monitoring. Proc Natl Acad Sci USA. 2015;112:6579–84.
21. Nelson DM, Smith W Jr. Sverdrup revisited: critical depths, maximum chlorophyll levels, and the control of Southern Ocean productivity by the irradiance-mixing regime. Limnol Oceanogr. 1991;36:1650–61.
22. Thrane J-E, Hessen DO, Andersen T. The absorption of light in lakes: negative impact of dissolved organic carbon on primary productivity. Ecosystems. 2014;17:1040–52.
23. Tranvik LJ. Allochthonous dissolved organic matter as an energy source for pelagic bacteria and the concept of the microbial loop. In: Salonen KKT, Jones RI, editors. Dissolved organic matter in lacustrine ecosystems. Vol. 73. Dordrecht: Springer; 1992. p. 107–14.
24. Bastviken D, Tranvik LJ, Downing JA, Crill PM, Enrich-Prast A. Freshwater methane emissions offset the continental carbon sink. Science. 2011;331:50.
25. Cole JJ, Caraco NF, Kling GW, Kratz TK. Carbon dioxide supersaturation in the surface waters of lakes. Science. 1994;265:1568–70.
26. Yang H, Andersen T, Dörsch P, Tominaga K, Thrane JE, Hessen DO. Greenhouse gas metabolism in Nordic boreal lakes. Biogeochemistry. 2015;126:211–25.
27. Cottrell MT, Kirchman DL. Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low-and high-molecular-weight dissolved organic matter. Appl Environ Microbiol. 2000;66:1692–7.
28. Crump BC, Kling GW, Bahr M, Hobbie JE. Bacterioplankton community shifts in an arctic lake correlate with seasonal changes in organic matter source. Appl Environ Microbiol. 2003;69:2253–68.
29. Jones SE, Newton RJ, McMahon KD. Evidence for structuring of bacterial community composition by organic carbon source in temperate lakes. Environ Microbiol. 2009;11:2463–72.
30. Kritzberg ES, Langenheder S, Lindström ES. Influence of dissolved organic matter source on lake bacterioplankton structure and function–implications for seasonal dynamics of community composition. FEMS Microbiol Ecol. 2006;56:406–17.
31. Lindström ES. Bacterioplankton community composition in five lakes differing in trophic status and humic content. Microb Ecol. 2000;40:104–13.
32. D'Andrilli J, Junker JR, Smith HJ, Scholl EA, Foreman CM. DOM composition alters ecosystem function during microbial processing of isolated sources. Biogeochemistry. 2019;142:281–98.
33. Eiler A, Langenheder S, Bertilsson S, Tranvik LJ. Heterotrophic bacterial growth efficiency and community structure at different natural organic carbon concentrations. Appl Environ Microbiol. 2003;69:3701–9.
34. Guillemette F, del Giorgio PA. Reconstructing the various facets of dissolved organic carbon bioavailability in freshwater ecosystems. Limnol Oceanogr. 2011;56:734–48.
35. Judd KE, Crump BC, Kling GW. Variation in dissolved organic matter controls bacterial production and community composition. Ecology. 2006;87:2068–79.
36. Romera-Castillo C, Sarmento H, Alvarez-Salgado XA, Gasol JM, Marrasé C. Net production and consumption of fluorescent colored dissolved organic matter by natural bacterial assemblages growing on marine phytoplankton exudates. Appl Environ Microbiol. 2011;77:7490–8.
37. Kawasaki N, Benner R. Bacterial release of dissolved organic matter during cell growth and decline: molecular origin and composition. Limnol Oceanogr. 2006;51:2170–80.
38. Battin TJ, Luyssaert S, Kaplan LA, Aufdenkampe AK, Richter A, Tranvik LJ. The boundless carbon cycle. Nat Geosci. 2009;2:598–600.
39. Osterholz H, Singer G, Wemheuer B, Daniel R, Simon M, Niggemann J, et al. Deciphering associations between dissolved organic molecules and bacterial communities in a pelagic marine system. ISME J. 2016;10:1717–30.

40. Cory RM, Kling GW. Interactions between sunlight and microorganisms influence dissolved organic matter degradation along the aquatic continuum. Limnol Oceanogr Lett. 2018;3:102–16.
41. Khomich M, Kauserud H, Logares R, Rasconi S, Andersen T. Planktonic protistan communities in lakes along a large-scale environmental gradient. FEMS Microbiol Ecol. 2017;93:fiw231.
42. Khomich M, Davey ML, Kauserud H, Rasconi S, Andersen T. Fungal communities in Scandinavian lakes along a longitudinal gradient. Fungal Ecol. 2017;27:36–46.
43. Andersen T, Hessen DO, Håll JP, Khomich M, Kyle M, Lindholm M, et al. Congruence, but no cascade-pelagic biodiversity across 3 trophic levels in Nordic lakes. Ecol Evol. 2020;10:8153–65.
44. Lyche Solheim A, Rekolainen S, Moe SJ, Carvalho L, Phillips G, Ptacnik R, et al. Ecological threshold responses in European lakes and their applicability for the Water Framework Directive (WFD) implementation: synthesis of lakes results from the REBECCA project. Aquatic Ecol. 2008;42:317–34.
45. Henriksen A, Skjelvåle BL, Mannio J, Wilander A, Harriman R, Curtis C, et al. Northern European lake survey, 1995: Finland, Norway, Sweden, Denmark, Russian Kola, Russian Karelia, Scotland and Wales. Ambio. 1998;27:80–91.
46. Ptacnik R, Andersen T, Brettum P, Lepistö L, Willén E. Regional species pools control community saturation in lake phytoplankton. Proc Royal Soc B. 2010;277:3755–64.
47. Mitchell BG, Kahru M, Wieland J, Stramska M. Determination of spectral absorption coefficients of particles, dissolved material and phytoplankton for discrete water samples. In: Mueller JL, Fargion GS, McClain CR, editors. Ocean optics protocols for satellite ocean color sensor validation, Revision IV. Vol. 4. Greenbelt, Maryland: Goddard Space Flight Center; 2003. p. 39–64.
48. Weishaar JL, Aiken GR, Bergamaschi BA, Fram MS, Fujii R, Mopper K. Evaluation of specific ultraviolet absorbance as an indicator of the chemical composition and reactivity of dissolved organic carbon. Environ Sci Technol. 2003;37:4702–8.
49. Bricaud A, Stramski D. Spectral absorption coefficients of living phytoplankton and nonalgal biogenous matter: a comparison between the Peru upwelling areaand the Sargasso Sea. Limnol Oceanogr. 1990;35:562–82.
50. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10–2.
51. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3.
52. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41:D590–D6.
53. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73:5261–7.
54. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. 2017.
55. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB et al. vegan: community ecology package. R package version 2.5-6. 2019. https://CRAN.R-project.org/package=vegan.
56. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE. 2013;8:e61217.
57. Venables W, Ripley B. Modern applied statistics with S. New York, NY: Springer; 2002.
58. van Buuren S. MICE: multivariate imputation by chained equations. R package version 2.22. 2015. https://mran.microsoft.com/snapshot/2014-11-17/web/packages/mice/mice.pdf. Accessed 12 Aug 2019.
59. Minchin PR. An evaluation of the relative robustness of techniques for ecological ordination. In: Prentice IC, van der Maarel E, editors. Theory and models in vegetation science. Dordrecht: Springer; 1987. p. 89–107.
60. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6.
61. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Malik J, Wang X, Yan S, Tang X, Jia J, editors. Proceedings of the IEEE international conference on computer vision. IEEE; 1730 Massachusetts Ave., NW Washington, DC, United States. 2015. p. 1026–34.
62. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011;334:1518–24.
63. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. Microbiol Mol Biol Rev. 2011;75:14–49.
64. Zwart G, Crump BC, Kamst-van Agterveld MP, Hagen F, Han S-K. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. Aquat Microb Ecol. 2002;28:141–55.
65. Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, et al. Scientists' warning to humanity: microorganisms and climate change. Nat Rev Microbiol. 2019;17:569–86.
66. Qiu Z, Coleman MA, Provost E, Campbell AH, Kelaher BP, Dalton SJ, et al. Future climate change is predicted to affect the microbiome and condition of habitat-forming kelp. Proc Royal Soc B. 2019;286:20181887.
67. Kovárová-Kovar K, Egli T. Growth kinetics of suspended microbial cells: from single-substrate-controlled growth to mixed-substrate kinetics. Microbiol Mol Biol Rev. 1998;62:646–66.
68. Scheffer M, Carpenter S, Foley JA, Folke C, Walker B. Catastrophic shifts in ecosystems. Nature. 2001;413:591–6.
69. Borchani H, Varando G, Bielza C, Larranaga P. A survey on multi-output regression. Wiley Interdiscip Rev Data Min Knowl Discov. 2015;5:216–33.
70. del Giorgio PA, Cole JJ. Bacterial growth efficiency in natural aquatic systems. Annu Rev Ecol Evol Syst. 1998;29:503–41.
71. Tranvik LJ. Bacterioplankton growth on fractions of dissolved organic carbon of different molecular weights from humic and clear waters. Appl Environ Microbiol. 1990;56:1672–7.
72. Hilbe C, Nowak MA, Sigmund K. Evolution of extortion in iterated prisoner's dilemma games. Proc Natl Acad Sci USA. 2013;110:6913–8.
73. Palen WJ, Schindler DE, Adams MJ, Pearl CA, Bury RB, Diamond SA. Optical characteristics of natural waters protect amphibians from UV-B in the US Pacific Northwest. Ecology. 2002;83:2951–7.

ACKNOWLEDGEMENTS

This study has been supported financially by the Department of Biosciences and the Centre of Biogeochemistry in the Anthropocene, University of Oslo, and by the Research Council of Norway (grant 'ECCO' 224779 to Dag O. Hessen and grant 'COMSAT' 196336 to Tom Andersen). We thank the COMSAT field sampling crew, especially Johnny Håll, Marcia Kyle, Robert Ptacnik, and Jan-Erik Thrane, for their efforts. Berit Kaasa and Sissel Brubak are acknowledged for laboratory support. The data analysis was performed on resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway.

AUTHOR CONTRIBUTIONS

This study was designed by TA and DOH. Sample and data collection were coordinated by TA. Molecular analyses were performed by MK under the supervision of SR and MD. Data were analyzed and visualized by MK, LF, MD, TA, and AE. The first version of the manuscript was drafted by MK but substantially modified after additional data analysis by LF and AE. All authors provided comments and were involved in writing the final version of the manuscript. Financial support for the project was acquired by TA, DOH, and AE.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43705-021-00038-8.

Correspondence and requests for materials should be addressed to A.E.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.