 Trait reconstruction and estimation of functional diversity from ecological monitoring data

Alexey Ryabov*,†‡, Bernd Blasius*,†, Helmut Hillebrand*†‡, Irina Olenina§#, and Thilo Gross*†‡

*HIFMB, Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany,
†University of Oldenburg, Institute for Chemistry and Biology of the Marine Environment, Oldenburg, Germany
‡Alfred-Wegener Institute, Helmholtz Center for Marine and Polar Research, Bremerhaven, Germany
§Klaipedos universitetas, Marine Research Institute, Klaipeda, Lithuania
#Lithuanian environmental protection agency, Klaipeda, Lithuania

Abstract

The twin crises of climate change and biodiversity loss define a strong need for functional diversity monitoring. While the availability of high-quality ecological monitoring data is increasing, the quantification of functional diversity so far requires the identification of species traits, for which data is harder to obtain. However, the traits that are relevant for the ecological function of a species also shape its performance in the environment and hence should be reflected indirectly in its spatio-temporal distribution. Thus it may be possible to reconstruct these traits from a sufficiently extensive monitoring dataset. Here we use diffusion maps, a deterministic and de-facto parameter-free analysis method, to reconstruct a proxy representation of the species’ traits directly from monitoring data and use it to estimate functional diversity. We demonstrate this approach both with simulated data and real-world phytoplankton monitoring data from the Baltic sea. We anticipate that wider application of this approach to existing data could greatly advance the analysis of changes in functional biodiversity.

1 Introduction

Recent assessments have documented the ongoing precipitous loss of global biodiversity [1, 2, 3, 4]. More complex responses are observed on the regional scale, where stressors can lead to a transient increase in diversity [5, 6]. Meanwhile
our understanding of the complex dynamical interplay of dispersal, extinctions and speciation that has created earth’s biological diversity and drives current dynamics is still woefully incomplete [7]. Hence the scale of the unfolding crisis, the intricacy of the dynamics involved, but also the gaps in our understanding highlight the need for large-scale biodiversity monitoring.

Even quantifying biodiversity loss still poses challenges. It has been argued that, for simplicity, global policy goals should be phrased in terms of the number of extinctions [8]. Similarly to climate goals, quantified by temperature-increase, the number of extinctions has the benefit of being easily communicable, however unlike climate change, where many detrimental effects are directly triggered by rising temperatures, extinction numbers are a poor indicator of biodiversity loss, where a major concern is the loss of biological functions [9].

On a fundamental level biodiversity can be conceptualized as the genetic variation of forms. However, due to the complexity of biological life the genetic makeup is only a weak indicator of function [10]. Hence, for the assessment of ecosystem functioning, service provision, sustainability and quantification of responses to stressors robust measures of functional diversity are needed.

The need to understand functional diversity has been frequently highlighted [9, 11, 12, 13, 14]. Common measures, such as the Rao index [15, 16], compute functional diversity from pairwise functional distances between species. For defining such distances, researchers typically identify traits of the species under consideration and then compute functional diversity from distances in trait space, e.g. [17, 16, 18, 19, 20].

Trait-based approaches provide good estimates of functional diversity, but require the researcher to quantify the trait space of all organisms considered. The decision what constitutes a relevant trait is made based on the researcher’s experience and is dependent on the group of species and functions under consideration. Some traits may be difficult to measure and their values may be dependent on environmental conditions [21] and hence are context dependent. In practice, these constraints mean that trait-based quantification of diversity is presently constrained to comparatively small groups of similar well-studied organism and suffers from limited data availability.

In comparison to the manual determination of trait values, it is generally easier to quantify properties, such as species identity, biomass and/or abundance. Long-term ecological research programs have accumulated a treasure trove of monitoring data, recording this information with individual data sets spanning multiple decades and capturing dozens or hundreds of species. It is therefore attractive to try to infer functional traits from monitoring data. For example [22] used Bayesian model fitting to infer four researcher-identified traits from long-term timeseries.

In this paper we propose an approach that unlocks the potential of monitoring data as a tool to quantify species traits and functional diversity. We use diffusion maps [23, 24, 25], a manifold learning method, to reconstruct the trait space directly from monitoring datasets. In contrast to previous work [22] this approach does not require a model or a list of known traits and is not limited to time series data. Instead the diffusion map identifies both the trait values and
trait axes solely from species biomass in samples. The functional diversity can then be computed from the pairwise distances in the reconstructed trait space. We test this approach with a simulated data set from a mathematical model, before applying it to quantify functional diversity of phytoplankton communities in a monitoring data set from the Baltic Sea [26]. Our results show that the proposed method can be used to robustly estimate functional diversity from monitoring data. In the dataset analyzed here, it reveals an increase in functional diversity with time that is significantly more pronounced at the coastal stations.

Figure 1: Functional diversity estimation from a monitoring data set. A. We use data on the biomass of 516 phytoplankton species in 730 samples collected from 1993 till 2015 from 10 stations in the Baltic sea [26] (two species shown for illustration). B. We then compute the pairwise similarity between species from correlation between samples. C. From the similarities the trait space of the species (dots) is reconstructed using diffusion maps. In this trait space the pairwise functional dissimilarity is quantified by the diffusion distance $d_{ij}$. D. Once the distances between species have been quantified the diversity in a specific sample can be quantified by applying the Rao’s index to the species present (highlighted dots). In the sample shown most of the biomass is concentrated in a small area of trait space, leading to a comparatively low Rao index.
Trait space reconstruction from monitoring data

Quantifying differences between dissimilar objects poses a fundamental challenge: Whereas we may be able to compare two songs or two paintings, it is much harder to quantify how dissimilar a certain painting is from a certain song. The same challenge is encountered in assessments of functional diversity, where it is essential to quantify how dissimilar pairs of (potentially very different) species are. To circumvent this problem, the diffusion map [23, 24, 25] builds on the idea that the dissimilarity between pairs of objects can be robustly quantified if they are sufficiently similar. By finding all such short-distance comparisons that can be made in the dataset we obtain a set of “trusted” links between objects.

Here, we apply this approach to estimate functional diversity from monitoring data. Hence we need to quantify the similarities between species based solely on their biomass in samples. Our primarily notion of similarity is the Spearman correlation [27] between pairs of species across samples in the dataset, which provides an indicator of co-occurrence of species.

We follow [25] and consider a comparison between two species as a trusted link if it ranks in the top-10 most similar comparisons for either of the two species. The trusted similarities are stored in a similarity matrix $S$, while all others are set to zero. The result is a network of trusted links that spans the entire set of species while containing only relatively short-ranged and hence relatively accurate comparisons.

Once trusted links have been identified, we quantify the dissimilarity between species by their distance in the network of trusted links. Specifically, diffusion maps use the notion of diffusion distance [23], which takes all possible paths between network nodes into account. The diffusion distance $d_{ij}$ is related to the travel time between the respective species for a random walker traversing the network of trusted links. It can be computed efficiently from the eigenvectors $v_n$ and eigenvalues $\lambda_n$ of a Laplacian matrix describing the network (see Appendix for details). The result is a computationally efficient method (Fig. 1) that produces deterministic results, and where our choice of trusting ten neighbors is the only tuneable parameter.

The Laplacian eigenvectors that are identified in the process are also of interest for a different reason: The eigenvectors contain one element corresponding to each of the species. These numbers, normalized by their respective eigenvalue, $v_n/\lambda_n$ act as trait variables in a proxy trait space (Fig. 1C), whereas the corresponding eigenvalue indicates the importance of the respective trait dimension (see Appendix). We show below that the proxy trait spaces that are thus reconstructed align well with ecological intuition.

After the trait space has been reconstructed we consider individual samples that have been taken as part of the monitoring effort (Fig. 1D). Using the inter-species diffusion distances in the reconstructed trait space we quantify the functional diversity in the sample using Rao’s quadratic entropy [15]. Hence diffusion maps can be used to both quantify functional biodiversity and, to some extent, reconstruct biological trait spaces solely from monitoring data.
**Validation with model data**

To test the proposed method we generate synthetic data by simulating a meta-community model *in silico*. We consider a community of 200 primary producer species limited by three essential resources. Each species is characterized by a set of minimal resource requirements ($R^*$ values) reflecting the species ability to sequester the corresponding resource. We can thus envision the species as points in a three-dimensional trait space (Fig. 2A), where the $R^*$ values correspond to the traits. Specifically, we randomly draw the $R^*$ values such that they fill a triangular surface, modeling the existence trade-offs such that a greater ability in sequestering one resource is compensated by a lesser ability in sequestering others.

We adapt a model from [28] to simulate the population dynamics of species in 800 meta-community, each of which consists of a square lattice of 120 discrete patches arranged in a $10 \times 12$ grid. The individual patches are characterized by randomly drawn values of the supply of the three resources, mimicking real world spatial heterogeneity and facilitating coexistence of model species. The biomass density of each species in each patch changes dynamically according to an equation capturing local growth and mortality as well as dispersal to and from neighboring patches (see Appendix).

As a first test, we consider the distribution of species in the reconstructed trait space, spanned by the two most important eigenvectors, found by diffusion mapping the simulated biomass samples. While the reconstructed trait space is slightly deformed in comparison to the ground truth, it retains key characteristics (cf. Fig. 2A,B). In the reconstructed trait-space the species still form a triangular shape on a two-dimensional surface. This result gives us confidence that the diffusion map should also be capable of identifying meaningful trait-spaces from large real-world monitoring datasets.

The main purpose of the proposed method is the estimations of functional diversity. To test whether a diffusion map that has been constructed from one dataset can also be used to estimate the diversity in new samples we ran 100 additional simulations and estimated the functional diversities both for the ground truth, given by the known $R^*$ values, and for the reconstructed traits using the existing map. This was done both for the entire meta-community (mimicking regional diversity) and within each patch (local diversity). A comparison of the resulting Rao indices (Fig. 2C) shows a strong correlation ($R^2 = 0.92$) between the ground truth and the reconstructed values of regional diversity.

We also explored how limited data availability and different distance metrics impacts the accuracy of reconstruction (not shown, see Appendix). The power of the diffusion map hinges on our ability to construct a spanning network of trusted comparisons between the samples. If the underlying trait space is large or the number of species is small then we are forced to trust comparisons between comparatively dissimilar species, and the quality of the reconstruction degrades. By contrast, a larger number of observations reduces the noise in individual comparisons and improves the quality reconstruction.

In summary, the results from these numerical experiments show that, given
Figure 2: Numerical validation of the proposed method. (A) We numerically generate randomly distributed traits of 200 species (colored dots) that fill a triangle in a trait space spanned by three resource-requirement ($R^*$) parameters. Color indicates the resource ratio preferred by a species. (B) Reconstructed proxy trait space generated by diffusion mapping simulated biomass data from a meta-community model. The reconstruction identifies traits that span a space that is qualitatively similar to the ground-truth traits. Colors are the same as in A, illustrating that neighborhood relationships are mostly reconstructed correctly. (C) Rao’s functional diversity calculated from diffusion distances in the reconstructed trait space correlates strongly with the numerical ground truth (based on $R^*$ values). Indicated are local diversity (blue dots) in individual patches and regional diversity in meta-community (yellow dots). The $R^2$ value for regional diversity is 0.92, relative to a cubic regression (green line). These results show that the proposed method can be used to identify traits and robustly estimate functional diversity based on monitoring data.

Analysis of Baltic Sea phytoplankton species

We now turn back to the phytoplankton monitoring dataset (Fig. 1). The data was collected in the Lithuanian coastal area of the Baltic Sea and spans a period from May to November for 23 years (1993-2015). In total it contains 730 samples of the biomasses of 516 species measured at different times and stations (see Fig. 1A for examples).

We analyze the Baltic data using the same procedure that we applied to the simulation results. A projection of the reconstructed trait using the first reconstructed traits is shown in Fig. 1C. While the diffusion map guarantees that these traits are the most important variables that characterize the species, it does not provide an interpretation of the corresponding traits.

We now explore the meaning of the reconstructed trait variables, using some
Figure 3: Reconstructed traits from the monitoring dataset. Shown are species (dots) projected onto the space spanned by the most important reconstructed traits. Color-coded are environmental conditions under which the species were observed with high relative abundance (see text). The first reconstructed trait aligns very well with NO$_3^-$ concentration separating species by their nitrogen requirements. The water temperature (B) and day of the year (C) align with the second trait, separating the early from the late species. The PO$_4^-$ concentration is closely aligned with the third reconstructed trait (D).

additional data from the monitoring dataset. We consider four environmental variables that were recorded during sampling (day of year, water temperature, NO$_3^-$ concentration and PO$_4^-$). For each of the species we compute the mean environmental conditions at which it was observed. This is done by computing a weighted average of each environmental parameter where the biomass of the species under consideration is used as the statistical weight of the sample. Color coding the species in the reconstructed trait space (Fig. 3) shows that the first reconstructed trait aligns well with NO$_3^-$ concentrations (Spearman correlation, $r_S = 0.55$). We conclude that this trait represents adaptation to different levels of nutrient availability. The second reconstructed trait closely aligns with the temperature ($r_S = 0.50$) and day of year ($r_S = 0.43$), suggesting that this trait represents the growth strategy, separating early from late species. The third
trait correlates with the \( \text{PO}_4^– \) concentrations \( (r_S = 0.45) \).

The number of trait axes of the diffusion map space equals the dimensionality of the input data, i.e. the number of species. Although each reconstructed subsequent trait contains less information than the previous one, projections of the diffusion map on different trait axes give additional insights into the distribution of species ecotypes in higher dimensions of the reconstructed trait space (see Appendix).

### Diversity gains on the Lithuanian Coast

Once the diffusion map has been constructed it can be used to quantify the functional diversity. We use the reconstructed traits from the analysis of the whole dataset to compute the diffusion distances between all pairs of species. For each sample we then use the distances between the species to compute the Rao index.

The estimated day-to-day functional diversities are relatively noisy, likely due to intrinsic fluctuations in the system. However, when considered over the whole period, there is a significant biodiversity gain at all of the stations, Fig. 4. This gain is most pronounced at the coastal stations where the functional diversity is also the highest in the later years.

The local increase in functional diversity is consistent with previous findings and predictions. For example [5, 14, 29] observe comparable increases in species richness. A key mechanism in this context seems to be that species extinction events triggered by environmental change take longer to manifest than corresponding invasions [14]. Hence environmental disturbances are likely to trigger transient increases in biodiversity on a trajectory that eventually leads to diversity loss when either longer times or larger geographical scales are considered.

In the present case the difference between coastal and offshore stations provides strong evidence supporting the hypothesis of an invasion-triggered transient increase near the coast. The coastal stations are close to the freshwater communities of the Curonian Lagoon. An increased influence of the freshwater species due to changing environmental conditions could easily explain the observed diversity trends. We note that the only coastal station that did not experience any increase of functional diversity, is located directly at the exit of the Curonian Lagoon (circle in Fig. 4) and hence has always been strongly influenced by the freshwater communities (see Appendix).

The results on functional diversity are supported by the species composition dynamics. The low functional diversity in the spring samples (see Appendix) of early 1993 and 2000 coincides with the dominance of dinoflagellates, mainly *Peridiniella catenata*, whose numbers were over 50% of the total phytoplankton abundance, of up to 96% by biomass. During the period of increased functional diversity in spring samples from 1994 to 1999, *Peridiniella catenata* was found in small numbers and the community was dominated by 3-5 species constituting together more than 50% of the total abundance. During this time, the number of non-dominant species with relative abundance less than 10% also increased.
Figure 4: Phytoplankton diversity on the Lithuanian coast. We observe an increase in functional diversity over the measurement period at all of the 10 stations included in the dataset (circles). The station located at the exit of the Curonian Lagoon is marked with a black circle. The fastest increase (warmer colors) is found at some of the coastal stations. The coastal stations are also the most diverse in average (larger diameter).

Conclusions

In this paper we proposed a method by which functional trait axes and values can be reconstructed from monitoring data. This reconstruction enables a robust estimation of functional diversity within the system based solely on species biomasses in samples.

We demonstrated the method using simulated data and a phytoplankton monitoring dataset from the Baltic sea. The analysis of the real world data identified adaptation for early/late growth, high/low nitrogen levels and high/low phosphorus levels as the most important functional trait axes.

The analysis also revealed a local increase in functional diversity that is comparable to previous observations in other systems [5, 14, 29]. In the present analysis, the increase is most pronounced at coastal stations and can be linked to increasing influence from a nearby freshwater community. Hence our results provide additional evidence for the hypothesis that changing environmental parameters may lead to a transient increase in local diversity, that might ultimately lead to biodiversity loss on longer and larger scales [5, 6].

We expect that the proposed method will also be useful in the analysis of other datasets. The method is particularly attractive if we are interested in a large number and/or species for which the traits have not been quantified in detail. In such a context the proposed reconstruction from monitoring data avoids introducing biases in the construction of the trait space and can bridge knowledge gaps. Moreover the accuracy of the diffusion map is expected to
increase with the number of species in the dataset. The proposed method is thus very well suited for the analysis of diverse microscopic organisms, such as plankton, whereas the explicit construction of trait spaces will yield better results when smaller sets of better-studied macroscopic species are considered.

In principle the proposed method could also be applied to study bacterial diversity. However in bacteria the relevant traits are thought to be more closely related to their genetic makeup than in eucaryotes. Hence it is sensible to take the available genomic information into account when constructing diffusion maps of bacteria [30].

The accuracy of the estimated functional diversity should increase with increasing number of observations. This property makes it attractive to combine multiple datasets, and perhaps even all available high quality monitoring data to estimate species similarity. Different monitoring datasets should be relatively easy to fuse at this level, because only comparisons within samples from the same dataset need to be made, though different naming conventions may pose a problem. Nevertheless one could envision creating an aggregate similarity matrix for phytoplankton that takes a large number of data sets into account. Diffusion mapping such a matrix could effectively provide a functional diversity standard that can be used to quickly map the functional diversity of samples on a fixed scale.

Acknowledgements

This work was funded by the Ministry for Science and Culture of Lower Saxony (HIFMB project) and the Volkswagen Foundation (grant number ZN3285).

References

[1] Bradley J. Cardinale, J. Emmett Duffy, Andrew Gonzalez, David U. Hooper, Charles Perrings, Patrick Venail, Anita Narwani, Georgina M. Mace, David Tilman, David A. Wardle, Ann P. Kinzig, Gretchen C. Daily, Michel Loreau, James B. Grace, Anne Larigauderie, Diane S. Srivastava, and Shahid Naeem. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59–67, 2012. doi: 10.1038/nature11148.

[2] Georgina M. Mace, Mike Barrett, Neil D. Burgess, Sarah E. Cornell, Robin Freeman, Monique Grooten, and Andy Purvis. Aiming higher to bend the curve of biodiversity loss. *Nature Sustainability*, 1(9):448–451, 2018. doi: 10.1038/s41893-018-0130-0.

[3] Gerardo Ceballos, Paul R. Ehrlich, and Rodolfo Dirzo. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences*, 114 (30):E6089–E6096, 2017. doi: 10.1073/pnas.1704949114.
[4] E. S. Brondizio, J. Settele, S. Díaz, and H. T. Ngo. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES Secretariat, Bonn, Germany, 2019.

[5] Shane A. Blowes, Sarah R. Supp, Laura H. Antão, Amanda Bates, Helge Bruelheide, Jonathan M. Chase, Faye Moyes, Anne Magurran, Brian McGill, Isla H. Myers-Smith, Marten Winter, Anne D. Bjorkman, Diana E. Bowler, Jarrett E. K. Byrnes, Andrew Gonzalez, Jes Hines, Forest Isbell, Holly P. Jones, Laetitia M. Navarro, Patrick L. Thompson, Mark Vellend, Conor Waldock, and Maria Dornelas. The geography of biodiversity change in marine and terrestrial assemblages. Science, 366(6463):339–345, 2019. doi: 10.1126/science.aaw1620.

[6] Laura H. Antão, Amanda E. Bates, Shane A. Blowes, Conor Waldock, Sarah R. Supp, Anne E. Magurran, Maria Dornelas, and Aafke M. Schiper. Temperature-related biodiversity change across temperate marine and terrestrial systems. Nature Ecology & Evolution, 4(7):927–933, 2020. doi: 10.1038/s41559-020-1185-7.

[7] Thilo Gross, Korinna T. Allhoff, Bernd Blasius, Ulrich Brose, Barbara Drossel, Ashkaan K. Fahimipour, Christian Guill, Justin D. Yeakel, and Fanqi Zeng. Modern models of trophic meta-communities. Philosophical Transactions of the Royal Society B: Biological Sciences, 375(1814):20190455, 2020. doi: 10.1098/rstb.2019.0455.

[8] Mark D. A. Rounsevell, Mike Harfoot, Paula A. Harrison, Tim Newbold, Richard D. Gregory, and Georgina M. Mace. A biodiversity target based on species extinctions. Science, 368(6496):1193–1195, 2020. doi: 10.1126/science.aba6592.

[9] D. Tilman. Functional diversity. In S. A. Levin, editor, Encyclopedia of biodiversity, pages 109–120. Academic Press, San Diego, 2001.

[10] Guntram Weithoff, Marcia R. Rocha, and Ursula Gaedke. Comparing seasonal dynamics of functional and taxonomic diversity reveals the driving forces underlying phytoplankton community structure. Freshwater Biology, 60(4):758–767, 2014. doi: 10.1111/fwb.12527.

[11] O. L. Petchey and K. J. Gaston. Functional diversity (fd), species richness and community composition. Ecology Letters, 5:402–411, 2002. doi: 10.1046/j.1461-0248.2002.00339.x.

[12] Helmut Hillebrand, Danuta M. Bennett, and Marc W. Cadotte. Consequences of dominance. Ecology, 89(6):1510–1520, 2008. doi: 10.1890/07-1053.1.

[13] A. S. MacDougall, K. S. McCann, G. Gellner, and R. Turkington. Diversity loss with persistent human disturbance increases vulnerability to ecosystem collapse. Nature, 494(7435):86–89, 2013. doi: 10.1038/nature11869.
[14] Helmut Hillebrand, Bernd Blasius, Elizabeth T. Borer, Jonathan M. Chase, John A. Downing, Britas Klemens Eriksson, Christopher T. Filstrup, W. Stanley Harpole, Dorothee Hodapp, Heddi Seabloom, Dedmer B. Van de Waal, and Alexey B. Ryabov. Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring. *Journal of Applied Ecology*, 55(1):169–184, 2017. doi: 10.1111/1365-2664.12959.

[15] C. Radhakrishna Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982. doi: https://doi.org/10.1016/0040-5809(82)90004-1.

[16] C. Ricotta and M. Moretti. Rao’s quadratic diversity: a unified framework for functional ecology. *Oecologia*, 167:181–188, 2011. doi: 10.1007/s00442-011-1965-5.

[17] B. Walker, A. Kinzig, and J. Langridge. Plant attribute diversity, resilience, and ecosystem function: The nature and significance of dominant and minor species. *Oecologia*, 2:95–113, 1999. doi: 10.1007/s100219900062.

[18] Mike McWilliam, Morgan S. Pratchett, Mia O. Hoogenboom, and Terry P. Hughes. Deficits in functional trait diversity following recovery on coral reefs. *Proceedings of the Royal Society B: Biological Sciences*, 287(1918):20192628, 2020. doi: 10.1098/rspb.2019.2628.

[19] Clifton C. Nunnally, Mark C. Benfield, and Craig R. McClain. Trait-based diversity of deep-sea benthic megafauna communities near the deepwater horizon oil spill site. *Marine Ecology*, 41(5), 2020. doi: doi.org/10.1111/maec.12611.

[20] Pierre Ramond, Raffaele Siano, Sophie Schmitt, Colomban de Vargas, Louis Marié, Laurent Memery, and Marc Sourisseau. Phytoplankton taxonomic and functional diversity patterns across a coastal tidal front. *Scientific Reports*, 11(1):2682, 2021. doi: 10.1038/s41598-021-82071-0.

[21] Mridul K. Thomas, María Aranguren-Gassis, Colin T. Kremer, Marilyn R. Gould, Krista Anderson, Christopher A. Klausmeier, and Elena Litchman. Temperature–nutrient interactions exacerbate sensitivity to warming in phytoplankton. *Global Change Biology*, 23(8):3269–3280, 2017. doi: 10.1111/gcb.13641.

[22] CM Mutshinda, ZV Finkel, CE Widdicombe, and AJ Irwin. Phytoplankton traits from long-term oceanographic time-series. *Marine Ecology Progress Series*, 576:11–25, 2017. doi: 10.3354/meps12220.

[23] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005. doi: 10.1073/pnas.0500334102.
[24] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019. doi: 10.1038/s41587-019-0336-3.

[25] Edmund Barter and Thilo Gross. Manifold cities: social variables of urban areas in the uk. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2221):20180615, 2019. doi: 10.1098/rspa.2018.0615.

[26] Z.R. Gasiūnaitė, A.C. Cardoso, A.-S. Heiskanen, P. Henriksen, P. Kauppila, I. Olenina, R. Pilkaitytė, I. Purina, A. Razinkovas, S. Sagert, H. Schubert, and N. Wasmund. Seasonality of coastal phytoplankton in the baltic sea: Influence of salinity and eutrophication. *Estuarine, Coastal and Shelf Science*, 65(1-2):239–252, 2005. doi: 10.1016/j.ecss.2005.05.018.

[27] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72, 1904. doi: 10.2307/1412159.

[28] Dorothee Hodapp, Helmut Hillebrand, Bernd Blasius, and Alexey B. Ryabov. Environmental and trait variability constrain community structure and the biodiversity-productivity relationship. *Ecology*, 97(6):1463–1474, 2016. doi: 10.1890/15-0730.1.

[29] Helmut Hillebrand, Ute Jacob, and Heather M. Leslie. Integrative research perspectives on marine conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1814):20190444, 2020. doi: 10.1098/rstb.2019.0444.

[30] Ashkaan K. Fahimipour and Thilo Gross. Mapping the bacterial metabolic niche space. *Nature Communications*, 11(1):4887, 2020. doi: 10.1038/s41467-020-18695-z.

[31] David Tilman. Resources: a graphical-mechanistic approach to competition and predation. *Am. Nat.*, 116(3):362–393, 1980. ISSN 00030147. doi: 10.2307/2463311.

[32] Jesús Alberto León and Daniel B. Tumpson. Competition between two species for two complementary or substitutable resources. *Journal of Theoretical Biology*, 50(1):185–201, 1975. doi: 10.1016/0022-5193(75)90032-6.

[33] Jef Huisman and Franz J. Weissing. Fundamental unpredictability in multispecies competition. *The American Naturalist*, 157(5):488–494, 2001. doi: 10.1086/319929.

[34] Alex Ryabov. Simulation and analysis codes. [http://www.staff.uni-oldenburg.de/alexey.ryabov/download.html](http://www.staff.uni-oldenburg.de/alexey.ryabov/download.html) 2021.
A Computational Experiment

In this section we describe the detailed procedure for the simulation of the meta-community model, see [28] for the model analysis. We model spatial competition of \( n = 200 \) species for three limiting resources, heterogeneously distributed across a two-dimensional grid comprising of \( 10 \times 12 \) diffusively coupled cells. The resource availability in each cell \((x,y)\) is characterized by the equilibrium concentration of the three resources \((S_{x,y}^1, S_{x,y}^2, S_{x,y}^3)\) in the absence of consumers.

To obtain samples from environments with varying heterogeneity, we simulated 800 grids with different levels of resource variability. The regional variability of resource \( r \) in grid \( j \) is constrained by the parameters \( S_{jr}^{\min} < S_{jr}^{\max} \), drawn randomly from a uniform distribution in the range from 1 to 39 resource units. The local supply, \( S_{x,y}^r \), of resource \( r \) in cell \((x,y)\) of grid \( j \) is then drawn randomly from a uniform distribution in the range \([S_{jr}^{\min}, S_{jr}^{\max}]\) in the respective resource grid.

Following the mechanistic theory of resource competition, we use the \( R_{ir}^* \) values, i.e., the concentration of resource \( r \) at which species \( i \) can still persist, as basic species traits [31]. To define the trait of species \( i \), we draw the values \( R_{i1}^* \) and \( R_{i2}^* \) from a uniform distribution in the range \([0.5, 9.5]\) and then chose the third resource parameter \( R_{i3}^* \) such that the sum \( \sum_{r=1...3} R_{ir}^* = 19.5 \) is constant. We reject parameter combinations for which \( R_{i3}^* \) is greater than 9.5. The trait values that are thus selected are uniformly distributed in a triangle in the trait space, as shown in Fig. 2A in the main text.

We model the functional dependence of biomass growth using Monod kinetics of resource uptake and Liebig’s law of minimum [32, 33],

\[
g_i(R) = g_{\text{max}} \min_{r=1...3} \frac{R_r}{K_{ir} + R_r},
\]

where \( g_{\text{max}} \) is the maximum growth rate, \( R_r \) is the local concentration of resource \( r \), and \( K_{ir} \) is the half-saturation constant of growth of species \( i \) limited by
resource \( r \). We assume that the maximum growth rate \( g_{\text{max}} = 1 \) and mortality rate \( m = 0.25 \) are identical for all species, so that the dimensionality of the species trait space is defined by the \( R^* \) values only. Based on the condition \( g_i(R_{ir}^*) = m \), we can express the half-saturation constants for given values \( R_{ir}^* \) as

\[
K_{ir} = R_{ir}^* \frac{(g_{\text{max}} - m)}{m}.
\]

The system is initialized by assigning random biomass values in \([0, 0.1]\) to each species in each grid cell. These biomass values are small compared to the typical local equilibrium biomass of persisting species.

To describe the evolution of the system, we denote \( N_{i xy} \) as the local biomass of species \( i \), and \( R_{r xy} \) the local concentration of resource \( r \), in cell \((x, y)\). The biomass of species \( i \) then evolves according to

\[
\frac{d}{dt} N_{i xy} = \left[ g_i(R_{r xy}) - m \right] \cdot N_{i xy} + \Delta N_{i xy},
\]

where the first term captures local growth and loss processes and the second term captures dispersal in terms of the discrete Laplace operator

\[
\Delta N_{i xy} = \frac{N_{i,x+1,y} + N_{i,x-1,y} + N_{i,x,y+1} + N_{i,x,y-1} - 4N_{i xy}}{h^2},
\]

where \( h \) is the lattice constant. At the grid boundaries we assume no flux boundary conditions. The dynamics of resource \( r \) is given by the difference between local resource inflow and resource consumption (for simplicity, we neglect the diffusion of resources)

\[
\frac{d}{dt} R_{r xy} = D(S_{r xy} - R_{r xy}) - \sum_{i=1}^{n} c_{ir} N_{i xy} g_i(R_{r xy})
\]

where \( c_{ir} \) is the amount of resource \( r \) consumed for the production of one unit of biomass of species \( i \) and \( D \) is the dilution rate. We assume that \( c_{ir} \) is linearly related to \( R^* \), \( c_{ir} = 0.05 R_{ir}^* \). This parameterization provides an additional possibility for species coexistence not only due to species diffusion across the grid but also due to local resource partitioning.

We integrate the system numerically for 6000 simulation days. This simulation time is sufficient for species sorting and reducing the effect of initial conditions. If the local biomass of a species drops below 0.01 at the end of the simulation, we regard this species to be locally extinct and set its local biomass to zero (for comparison, the maximum local biomass of species is about 100). The local biomass of the 200 competitors at the end of the simulation are then analyzed using the diffusion map procedure described below.

Thus, the raw data includes \( 10 \cdot 12 \cdot 800 = 96,000 \) biomass samples of each of the 200 species. However, diffusion fluxes across cell boundaries lead to biomass correlations in neighboring cells, so that many samples obtained from the same grid are not independent. We estimate the minimal number of independent samples that we use below as the number of simulated grids.
Unless not otherwise specified above, parameter values are taken according to [28]. For numerical integration we use an explicit Runge-Kutta (4,5) algorithm (ode45 solver, MATLAB 2020). The source code for the simulation and data analysis is publicly available at [34].

B  Diffusion map procedure

We now explain the procedure that was used to identify reconstructed traits and to compute diffusion distances between species.

B.1  Notation

We consider \(m\) samples containing a subset of a total of \(n\) species. The biomass (or abundance) of species \(i\) in sample \(j\) is denoted by \(a_j^{(i)}\). We can then collect all observations of species \(i\) in the vector

\[
a^{(i)} = \begin{pmatrix} a_1^{(i)} \\ \vdots \\ a_m^{(i)} \end{pmatrix}
\]  

(6)

B.2  Similarity matrix

We compare different approaches to defining the similarity matrix (see below). We found that the most consistent results are obtained when similarity is determined through the Spearman correlation coefficient. In this case, the similarity of species \(i\) and \(j\) can be expressed as

\[
S_{ij} = r_s(a^{(i)}, a^{(j)}) = \frac{\text{cov}(rg_i, rg_j)}{\sigma_{rg_i} \sigma_{rg_j}}, \quad i \neq j
\]  

(7)

where \(rg_i\) and \(rg_j\) are vectors of rank variables for the biomasses of species \(i\) and \(j\) in different samples, \(\text{cov}(rg_i, rg_j)\) is the covariance between these variables, and \(\sigma_{rg_i}\) and \(\sigma_{rg_j}\) the standard deviations of the rank variables. We do not include the similarity of a species with itself and set diagonal elements to zero, \(S_{ii} = 0\).

An alternative approach (not used in the results in the main text) is based on defining species dissimilarity as the distance between vectors of either normalized biomass values \(b^{(i)} = a^{(i)}/\sigma_i\) or standardized biomass values, \(b^{(i)} = (a^{(i)} - \mu_i)/\sigma_i\), where \(\mu_i\) is the mean value and \(\sigma_i\) the standard deviation of the biomass of species \(i\) across all samples. We compute the dissimilarity \(d_{ij}\) between species as the Euclidean distance between these vectors

\[
d_{ij} = \sqrt{\sum_k \left( b_{(k)}^{(i)} - b_{(k)}^{(j)} \right)^2}
\]  

(8)
and then define the elements of the similarity matrix, \(S_{ij}\), as the inverse of the dissimilarity, such that
\[
S_{ij} = \frac{1}{d_{ij}}
\]
for \(i \neq j\) and \(S_{ij} = 0\) for \(i = j\).

### B.3 Thresholding

Long distance comparisons are unreliable and constitute a source of noise that can swamp the signal. Hence we want to eliminate all but the largest entries in the similarity matrix. We do this examining each row of \(S\) and setting all but the 10 largest entries to zero. Thresholding more aggressively has a stronger denoising effect, but risks splitting the network into disconnected components. The value of 10 is often a good middle ground that has been used in several previous publications [25, 35, 30]. We confirmed that for the present dataset the links described by the similarity matrix form a spanning component.

The thresholding procedure can leave us with an asymmetric matrix. For numerical reasons it is generally desirable to maintain the symmetry. We therefore resymmetrize the matrix by the operation
\[
S_{ij} \rightarrow \max(S_{ij}, S_{ji})
\]

### B.4 Laplacian

From \(S\) we construct the row normalized Laplacian \(L\) defined by
\[
L_{ij} = \begin{cases} 
-S_{ij}/\sum_j S_{ij} & i \neq j \\
1 & i = j
\end{cases}
\]
and solve the eigenvalue problem
\[
Lv_i = \lambda_i v_i, \quad i = 1 \ldots n.
\]

### B.5 Eigenvectors and eigenvalues

All further analysis builds on the eigenvalues and eigenvectors of \(L\). We compute these eigenvalues and eigenvectors numerically using the function \text{eig()} from MATLAB 2021, which solves eigenvalue problems using the QR algorithm.

The Laplacian always has at least one eigenvalue at zero. The multiplicity of this eigenvalue is identical to the number of components in the network. In our case, the network has only one component, hence the zero eigenvalue has multiplicity one. We denote this eigenvalue as \(\lambda_0\). The corresponding eigenvector \(v_0\) contains no further information. All other eigenvalues are positive and the corresponding eigenvectors contain trait information.

Since the Laplacian \(L\) is a symmetric matrix, all eigenvalues are real. The eigenvectors corresponding to small eigenvalues are respectively more important, i.e. explain more of the variation along the manifold. Hence the most important
eigenvector $v_1$ is the vector corresponding to the smallest non-zero eigenvalue. The eigenvector $v_2$ corresponding to the second smallest non-zero eigenvector corresponds to the second most important trait and so on.

Each eigenvector contains $n$ elements which assign a proxy trait value to each of the $n$ species. As the eigenvalues are inversely related to the importance of the trait we define the value of trait $k$ of species $i$ as $v_{k,i}/\lambda_k$.

### B.6 Diffusion distance

Once the trait values have been computed the dissimilarity can be quantified by the distance in trait space. To compare two species we compute the euclidean distance between species in the reconstructed trait space where the species traits are now given by the eigenvector elements corresponding to the species, rescaled by the respective eigenvalue. Hence the distance between two species $i,j$ is

$$d_{ij} = \sqrt{\sum_k \left( \frac{v_{k,i} - v_{k,j}}{\lambda_k} \right)^2}$$

### C Exploratory Analyses

In this section we explain the calculation of functional diversity and present some supporting information on the analysis of the simulated data. Moreover, we present some results from exploratory analysis that lead up to the selection of the Spearman correlation coefficient as our primary similarity measure.

#### C.1 Functional diversity

There are many ways to assess functional diversity. These methods have various advantages, and the choice of an appropriate index may depend on the specific problem [36]. In this article, we use the Rao’s quadratic entropy [37], because this index is sensitive not only to the specific species traits, but also to variation in species abundances. Functional diversity calculated as the Rao index for sample $k$ equals

$$FD_k = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} p_k^{(i)} p_k^{(j)},$$

where $p_k^{(i)} = a_k^{(i)}/\sum_j a_k^{(j)}$ is the relative biomass of species $i$ in this sample. This index represents a weighted average functional distance between all pairs of species $i$ and $j$, where the weighting factor $p_k^{(i)} p_k^{(j)}$ is the probability that one of two randomly selected individuals belongs to species $i$ and the other to species $j$. 

18
C.2 Effect of data availability

In the main text we report the results of trait reconstruction with simulated data from 800 grids. To explore how data availability affects the trait reconstruction we repeat the reconstruction using 100, 200, and 600 grids. For each of these trials we show the reconstructed trait space and the estimated functional diversity (shown for the full data set in the main text), as well as distances between traits measured in terms of $R^*$ values vs. reconstructed diffusion distances (Fig. 5). Using a small number of 100 grids the reconstruction is very noisy, yielding a square rather than a triangular shaped trait space. The results of the reconstruction improve steadily as we increase the number of grids used. With a number of 600 grids the quality of the reconstruction is comparable to that of the 800-grid reconstruction shown in the main text.

The diffusion map based on 100 grids allows us to reliably distinguish only species at small to intermediate functional distances in $R^*$ space, however reconstructed functional distance fluctuates around a constant level when the $R^*$ functional distance is large (Fig. 5D, G). The correlation between both functional distances and functional diversities measured in $R^*$ space and in reconstructed trait space increase with the number of samples included, and the relationships approach to monotonically increasing reversible functions (Fig. 5, middle and bottom row). This means that a sufficient input data volume provides a correct mapping for the entire range of functional distances.

C.3 Selection of a distance measure

When the original species traits are known, it is easy to estimate the amount of input data needed to construct a diffusion map with a targeted accuracy. For field data, however, we need an additional method to estimate the uncertainty for a given dataset and the possibility of reducing the uncertainty by increasing the amount of data. To be able to achieve this we suggest to use bootstrapping of samples, a common machine learning method for estimating the rate of uncertainty reduction with the number of training samples. This approach enables us to estimate the uncertainty of diffusion maps and allows the analysis of learning curves.

We denote a similarity matrix constructed by using $m$ samples as $S(m)$. To calculate this matrix we use bootstrapping, i.e., we select a random set of $m$ out of the $n$ samples with replacement. Thereby, on the basis of $n$ samples we obtain a reduced set of $m$ samples, which will probably include some samples several times, but have statistical properties close to the original set of samples. A bootstrap sample of $n$ out of $n$ samples will include, on average, 63% of the original data, while 37% will remain unused. This approach thus gives an estimate of the uncertainty of the similarity matrix generated from the $n$ samples.

To estimate the uncertainty, we quantify the 'difference' between a similarity matrix based on a subset of $m$ samples and a similarity matrix based the full set
Figure 5: Effects of data availability. Shown is the reconstructed trait space (top row), a scatter plot of ground truth ($R^*$) trait distances vs. reconstructed traits distances (center row), and a scatter plot of estimated functional diversity vs. ground truth functional diversity (bottom row) (cf. Fig. 2 from the main text). We compare different reconstructions using 100 (left column), 200 (center column), and 600 (right column) simulated grids to produce a diffusion map. The comparison and the indicated $R^2$ values in the second and third row illustrate how limited data availability degrades the reconstruction result.
of $n$ available samples by the correlation between the elements of these matrices

$$V(m) = \text{cor}(S(m), S(n))$$

(14)

Computing $V(m)$ over an increasing size $m$ of the subset yields a “learning curve” (Fig. 6). Constructing such curves using different common notions of distance between species (see Appendix) makes it possible to choose the metric that gives the most robust result. In this test the Spearman correlation performed very well, both for simulated and empirical data, yielding a good accuracy with limited data. Based on these results we choose the Spearman correlation as our primary notion of similarity between species.

The value $V(n)$ gives an estimate of the uncertainty in $S$ when all $n$ samples are used. Extrapolating these curves for $m > n$ provides a rough idea of the number of samples needed to calculate the similarity matrix with the desired accuracy. For the simulated data, when all samples are used, the correlations achieve values from 0.7 to 0.85 depending on the distance metric (Fig. 6A), and the correlation continues to grow monotonically. To obtain a substantial reduction of uncertainty in this case one need at least to double the number of samples. In the field data, compared to the model data, the number of independent variables (species) is larger and the number of observations is smaller. This increases the uncertainty of the similarity matrix, and the correlations obtained for the complete dataset achieve maximal values of $V(m) = 0.6$ (Fig. 6B). Thus, an increase of the monitoring area, frequency, or time-span should be able to significantly improve the accuracy of both the similarity matrix and reconstructed species traits.

C.4 Comparison of diffusion maps and PCA

Principal Component Analysis (PCA) is a very common data analysis procedure that identifies directions of large variation in a data cloud. Hence PCA can be used as a simple manifold learning method if the true manifolds in the data are close to linear. However, as we now demonstrate, using PCA for trait reconstruction results in reduced accuracy.

In the analysis of model data significant advantages of the diffusion map over PCA become apparent. We can highlight three main differences. First, the shape of trait distribution in the diffusion map, compared to PCA, is closer to the original distribution shape (compare Fig. 2A,B in the main text and Fig. 7). Second, the representation of data by diffusion map compared to PCA is much more condense in lower dimensions (Fig. 8). Finally the fractal dimension of the trait manifold reconstructed by diffusion map is closer to the dimension of the original manifold than that obtained by PCA (Fig. 9).

C.5 Dimensionality of the data space

To compare the representation of the original trait manifold by a minimal set of reconstructed traits obtained by PCA and diffusion map, it is convenient
Figure 6: **Learning curve of the similarity matrix for different distance metrics.** In both the simulated data (A) and Baltic Sea samples (B) the Spearman correlation provided the most robust indicator (higher values of $V(m)$) of species similarity over a wide range of values (see text for details).

Figure 7: **PCA reconstruction of modelling species traits.** The same as Fig. 2 in the main text but based on PCA of species similarity matrix calculated as Spearman correlation coefficients.
Figure 8: Comparison of transformed eigenvalues (loadings) calculated for diffusion maps (A) and PCA (B). The transformed eigenvalue shows the variability in data described by the given axes in the eigenvector space. As shown, the representation of data in diffusion map space is much more condensed in lower dimension of the eigenvector space, allowing us to use only a few first components to characterize the data.

to use the obtained eigenvalues. For diffusion maps, species traits are defined as $t_{j,i} = v_{j,i}/\lambda_j$, while PCA ordination of species uses the so-called scaling 2 representation where species traits are defined as $t_{j,i} = \sqrt{\lambda_j} v_{j,i}$. Since the absolute values of the eigenvectors $v_j$ equal one, the comparison of the ranges of species trait variation is reduced to a comparison of the transformed eigenvalues: $1/\lambda_j$ for the diffusion map and $\sqrt{\lambda_j}$ for PCA (Fig. 8). Note that $\sqrt{\lambda_j}$ is termed as loadings in PCA analysis.

The initial manifold of species traits is a two-dimensional object (Fig. 7A), and the values of first two inverted eigenvalues obtained by diffusion mapping significantly exceed the subsequent eigenvalues (Fig. 8A), while for PCA this difference is much smaller and even traits defined by eigenvectors with an index greater than 50 carry a significant part of information (Fig. 8B). Thus, diffusion maps compared to PCA are much better at concentrating information about species traits in a lower-dimensional space.

The reconstructed manifold is a multidimensional object, with the number of dimensions equal to the number of species. However, we expect that the resulting distribution of species traits should keep the original distribution dimensionality and be located on a two-dimensional surface. To estimate the dimensionality of the resulting distribution, we calculated its correlation fractal dimension in the space defined by the first 50 eigenvectors. The fractal dimension of the original distribution, $d = 1.9$, is less than 2, because a random distribution is never perfectly uniform and contains some halls. The diffusion map fractal dimension $d = 1.8$ is close to the original value, but the PCA distribution dimension exceeds it and equals 2.6 (Fig. 9). Implying that the
Figure 9: Estimation of the correlation fractal dimension $d$. The estimated dimension of the original trait distribution in space $R^*$ is $d = 1.8$, which is close to the actual dimension $d = 2$, since we put all $R^*$ on a plane. The dimensionality of the surface in the trait space reconstructed with diffusion map is $d = 1.9$. The distribution in PCA space has a larger dimensionality, $d = 2.6$, than the actual distribution, which implies that the data points in PCA space are likely located with random displacement around a two-dimensional surface, effectively creating an object with a fractional dimension greater than 2. The fractal dimension was estimated in the space of the first 50 eigenvectors.
distribution obtained with PCA is closer to a 3D object than a 2D surface.

D  Analysis of the Baltic Sea data

This section contains some additional notes on the Baltic Sea phytoplankton dataset and its analysis. We use data from at 10 stations the Lithuanian coastal area of Baltic Sea which were regularly visited for 23 years (1993-2015). The data includes 730 samples of biomass of 516 species and environmental data, such as temperature, NO$_3$, NO$_2$, PO$_4$, pH, and salinity.

D.1  Adaptation to environmental factors

Environmental data were measured at different depths for each sample. To obtain a single value for each sample, we interpolated these measurements over a range of depths from 1 to 10 m with resolution of 1 m and calculated the average value in this range.

We estimated species specific environmental condition as the average environmental condition weighted with species biomass

\[
\tilde{E}^{(r,i)} = \frac{\sum_{j=1}^{n} a_j^{(i)} E_j^{(r)}}{\sum_{j=1}^{n} a_j^{(i)}}
\]  (15)

where $a_j^{(i)}$ is the biovolume of species $i$ in sample $j$, $E_j^{(r)}$ is the environmental factor $r$ in this sample, and $n$ is the number of samples. In this way we obtain the species-specific day of year, temperature, nutrients, pH, etc., which are used to colorize the diffusion maps (Fig. 3 in the main text and Fig. 11), and to identify species trait-environment pairs (Fig. 10).

To find the best matching between traits and environmental factors shown in Fig. 3 in the main text, we performed a cross-correlation analysis of the relationships between the environmental factors and the reconstructed species traits. We calculated the Spearman correlation between the measured environmental factors and the first ten proxy traits, selected then the trait-environment pair with the highest correlation for each environmental factor, and sorted the results by correlation in descending order. Fig. 10 shows the top nine trait-environment pairs. The first trait correlates with NO$_3$ ($r_S = 0.55$) and NO$_2$ ($r_S = 0.49$); the second trait negatively correlates with temperature ($r_S = -0.5$), day of year ($r_S = -0.43$) and positively correlates with salinity ($r_S = 0.32$); the third trait correlates with PO$_4$ concentration ($r_S = 0.45$) and DIN ($r_S = 0.32$); the fifth trait correlates with NH$_4$ ($r_S = -0.27$), and the sixth trait with pH ($r_S = -0.34$). We found no strong correlation between the fourth trait and any of the environmental conditions, but this does not mean that this trait is not important, as it may reflect adaptation to factors missed in our data, such as zooplankton abundance, light radiation, or water turbidity.

By projecting the distribution of species traits from the multidimensional diffusion map space onto different species trait axes, we obtain more details
about the relationships between environmental factors and proxy traits. This is shown in Fig. 11 which is an extended version of Fig. 3 from the main text. As shown, adaptation to NO$_3$ correlates only with the first proxy trait, and adaptation to temperature correlates only with the second trait. At the same time, the day of the year is described by the second, third, and sixth trait (with decreasing correlation). This can be explained by the fact that this parameter strongly positively correlates with temperature, however species dominant in spring and autumn may have the same optimal temperature, but be characterized by different days of the year. Adaptation to PO$_4$ concentration is also related to several traits, but only the relationship with the third trait is monotonic, while the relationships with the second, fourth, and sixth trait is unimodal.

Fig. 13 shows the dynamics of functional diversity grouped by year (upper row) and additionally by season (lower rows). Functional diversity has remained high on average, with the exception of 1993, 1994, and 2000 (Fig. 13A). Spring samples showed an increase in functional diversity early in the observation period (1993-1999), followed by a decline in functional diversity in 2000, after which functional diversity remained at consistently high levels (Fig. 13B). Extremely large variation between stations, characterized by the size of boxes in Fig. 13 was observed in the spring samples of 1993, 1994, and 1996, in the summer samples of 1994, 1997, and 2001, and also in the fall sample of 2000.

D.2 The dynamics of functional diversity

For comparison of the dynamics of functional diversity at different stations (Fig. 4, the main text), we carried out a regression analysis of the dependence of functional diversity on the year at different stations (Fig. 12). Since we did not identify a statistically significant seasonal effect on functional diversity, the regression analysis was performed without grouping the data by seasons, but the data were grouped by seasons for clarity when plotting the figure. All coastal stations, with the exception of station 4, show a positive temporal trend in functional diversity, associated with the fact that samples at the beginning of the observations show a wide range of diversity, while samples closer to the end of the observations are characterized mostly by high functional diversity. Coastal station 4, located at the exit of the Curonian Lagoon, is the only coastal station that shows only a small statistically insignificant temporal trend, which is explained by the fact that the functional diversity at this station was at a high level throughout the entire observation period. In contrast to the coastal stations, the samples from off-shore stations (1B, 20, 64, 65) do not show a positive trend in functional diversity and are characterized by a reduced level of functional diversity throughout the entire observation period compared to the coastal stations.
Figure 10: Correlations between reconstructed species traits and species-specific environmental values. The top nine trait-environment pairs with the highest Spearman correlation coefficient are shown (see text for details).
Figure 11: Projection of diffusion map on different axes. Color shows the value of the species-specific environmental factor.
Figure 12: **Temporal trends of functional diversity at different stations.** Functional diversity at coastal stations 2, 3, 5, 6, 7 has increased over the time of observations. This increase was significant ($p < 0.05$) at stations 3, 5, and 7, marked with a star in the plot title. At coastal station 4 it was high during the entire monitoring period. In recent years, samples from all coastal stations exhibit only high functional diversity. By contrast, functional diversity at offshore stations 1B, 20, 64, and 65 has a weak positive trend and varied over a wide range throughout the observation period.
Figure 13: Temporal trends of functional diversity at different seasons. Boxplots showing the variation and median values of functional diversity in dependence of the year. The variation in diversity across observations and stations, shown both for the whole year (top row) and separately for each season. The boxplots show the median values (red line), the bottom and top of each box are the 25th and 75th percentiles of the sample, and the whisker length is \( \pm 2.7\sigma \) (corresponding to 99.3\% for normally distributed data).