

Figure S1. Comparison of satellite-derived parameters acquired by the Aqua and the Terra satellite.

Pearson's correlation coefficient was shown for each parameter pairs.

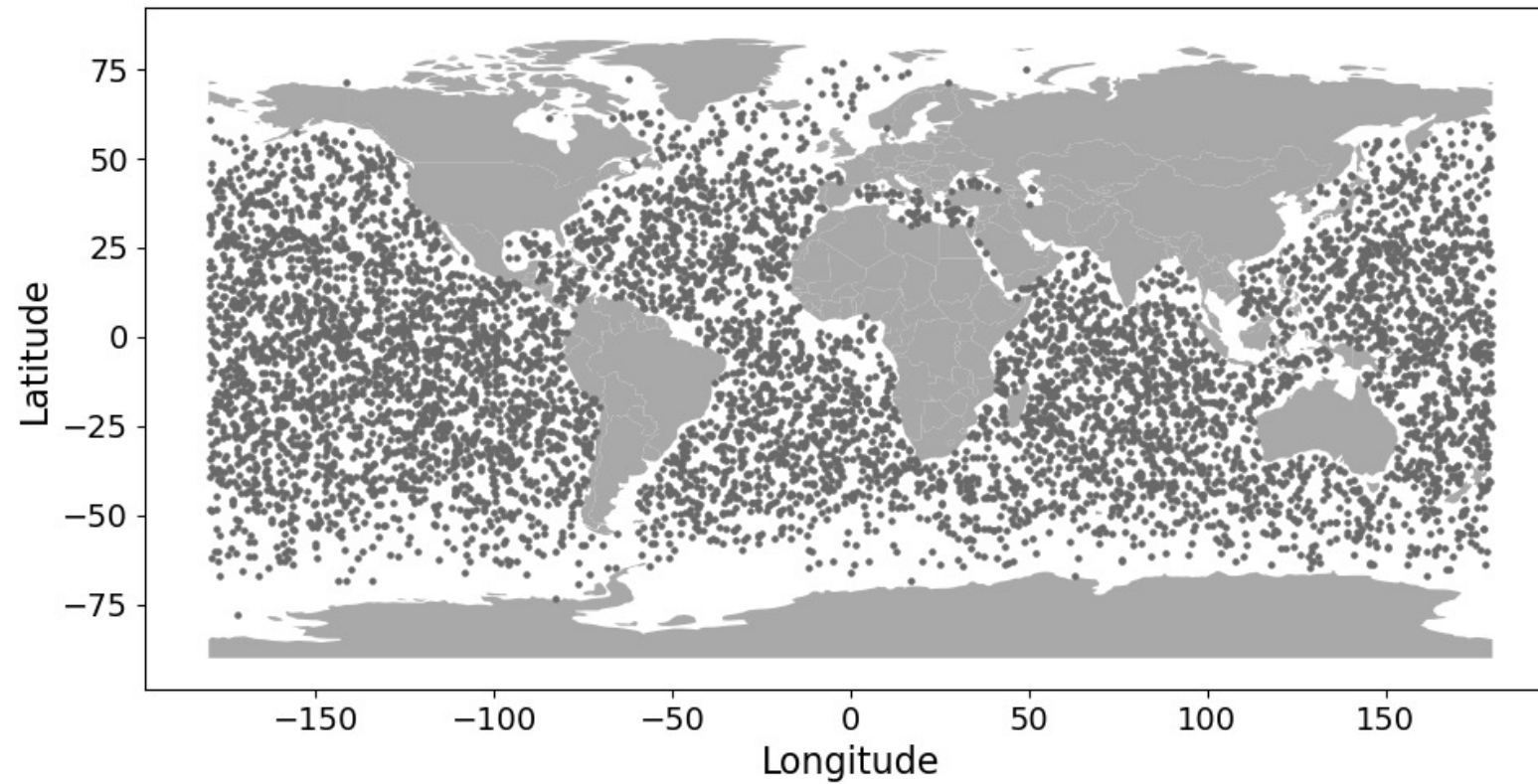


Figure S2. Randomly selected grid cells used to train a UMAP projection. Dark gray points are location of randomly selected grid cells. Sampling month was also randomly selected for each grid cell.

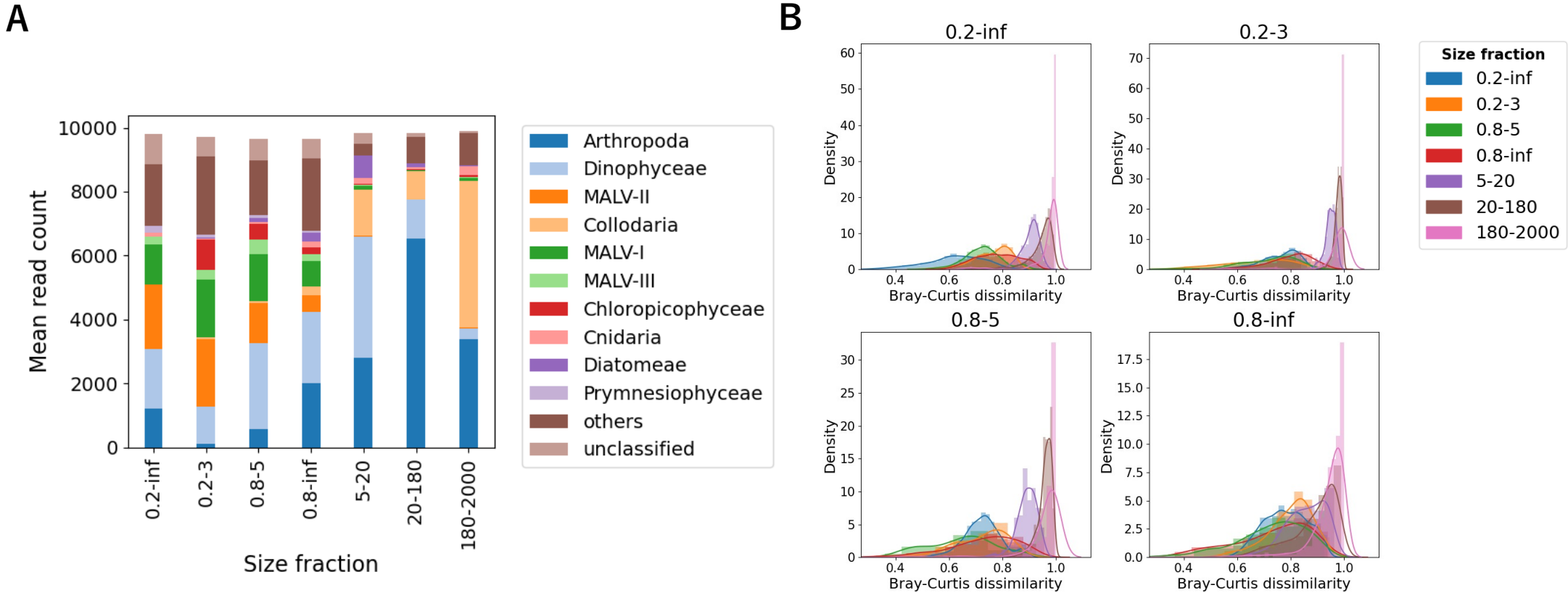


Figure S3. Comparison of size fractions in the South Pacific Subtropical Gyre. Taxonomic difference among size fractions was examined for samples from the South Pacific Subtropical Gyre, which included samples from all major size fractions. **A** Mean taxonomic composition of each size fraction. Taxonomic level is “taxogroup 2” in the EukRibo. **B** Taxonomic dissimilarity of intra- and inter-size-fraction sample pairs. Bray-Curtis dissimilarity was calculated based on OTU read counts.

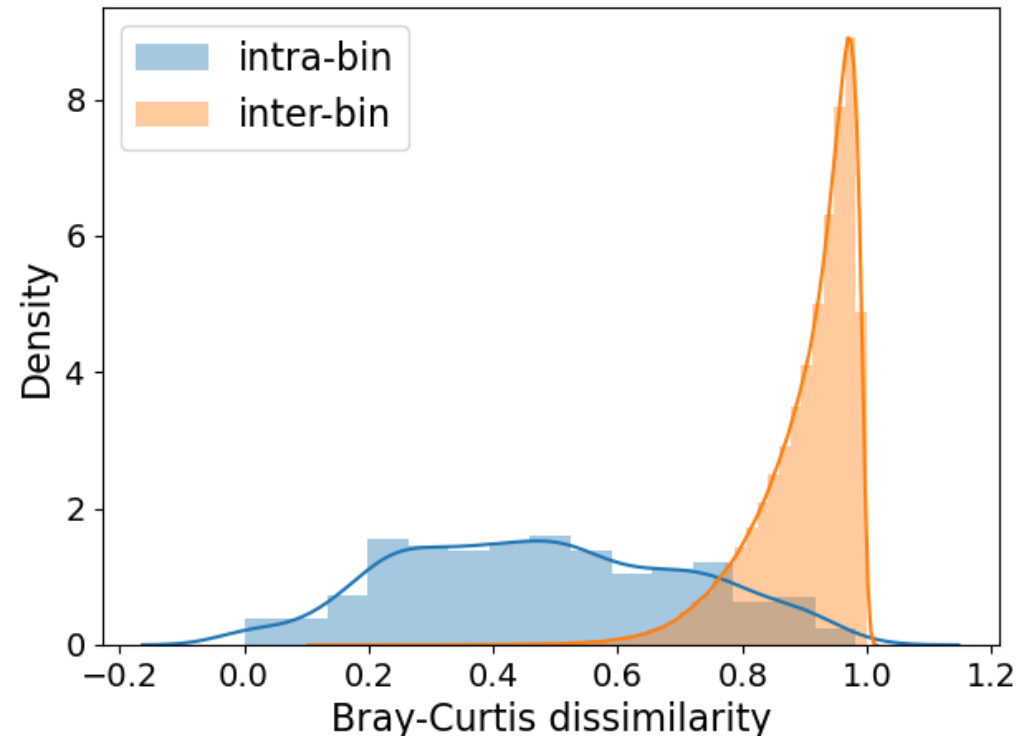


Figure S4. Comparison of dissimilarity of intra- and inter-bin sample pairs.

Bray-Curtis dissimilarity was calculated based on OTU read counts. Intra-bin sample dissimilarity was small enough compared to inter-bin sample dissimilarity in the most cases.

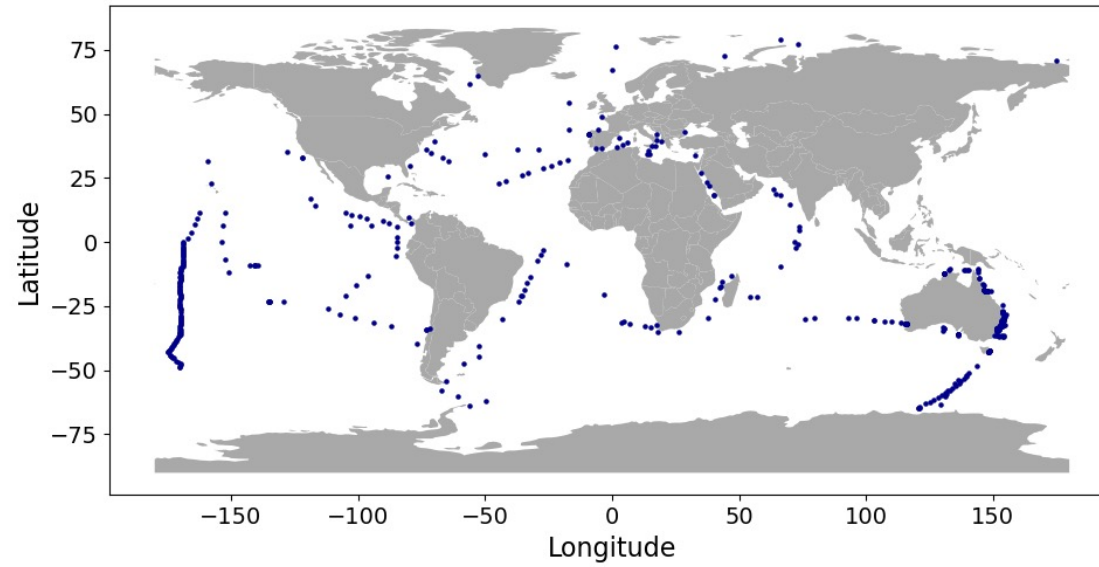
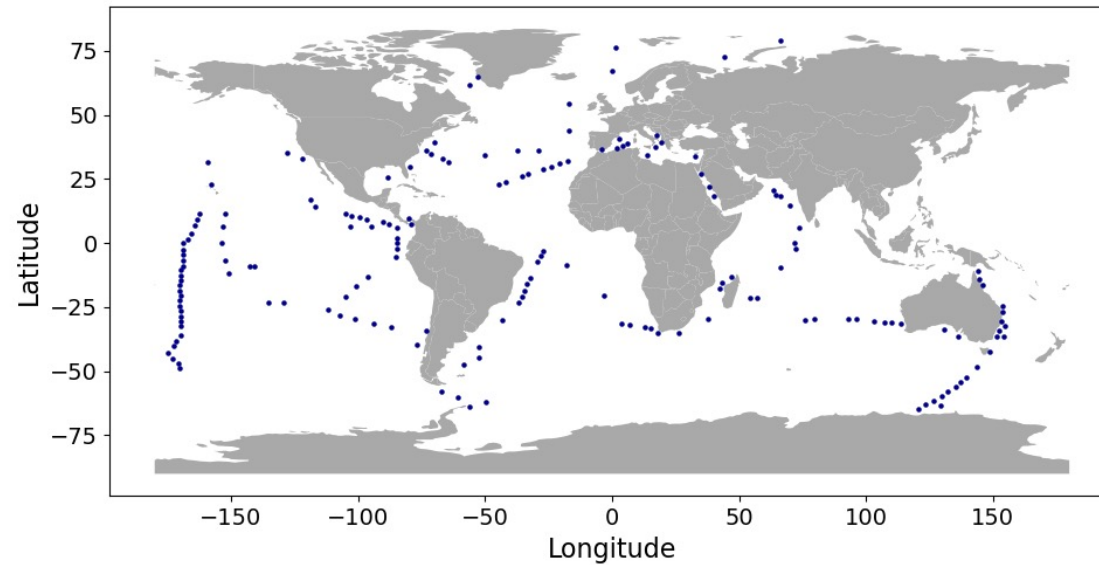
A**B**

Figure S5. Spatial resampling of metabarcoding data.

A Geographic location of 653 metabarcoding samples (bins) before spatial resampling. **B** 177 samples retained and used for the analysis.

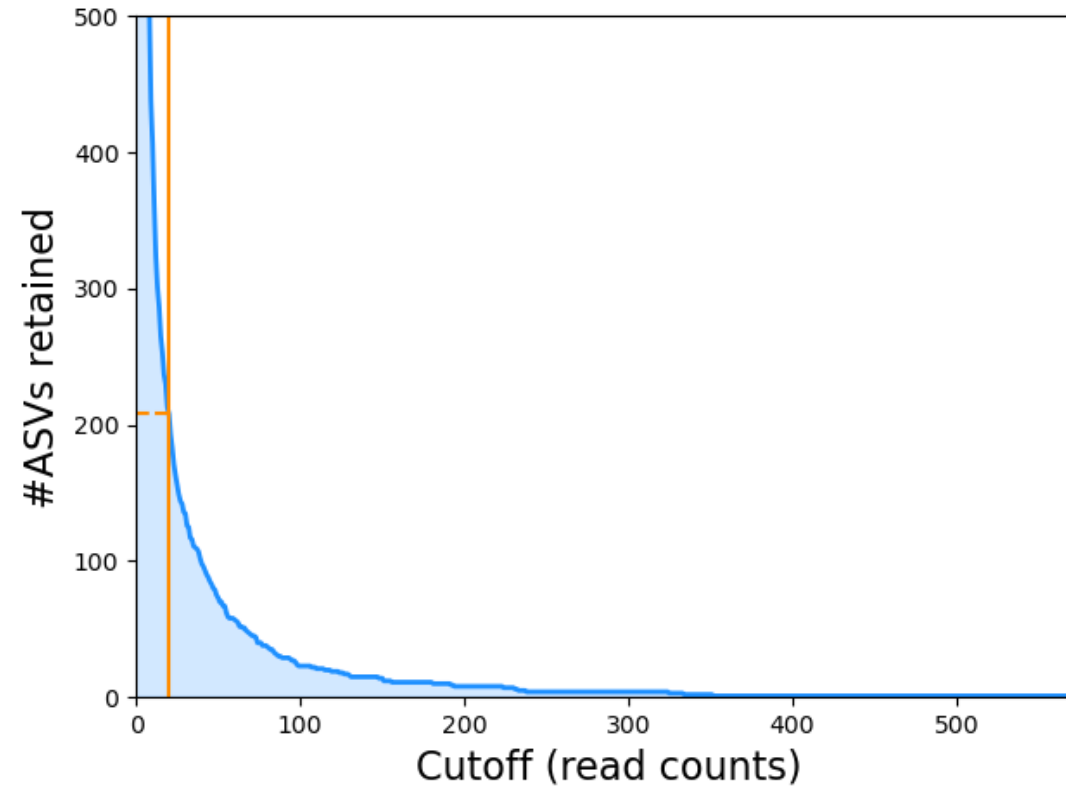


Figure S6. Number of OTUs retained with changing occurrence cutoff. Blue curve shows the number of OTUs retained with given cutoff used for selection. Orange line is the chosen cutoff (20 reads).

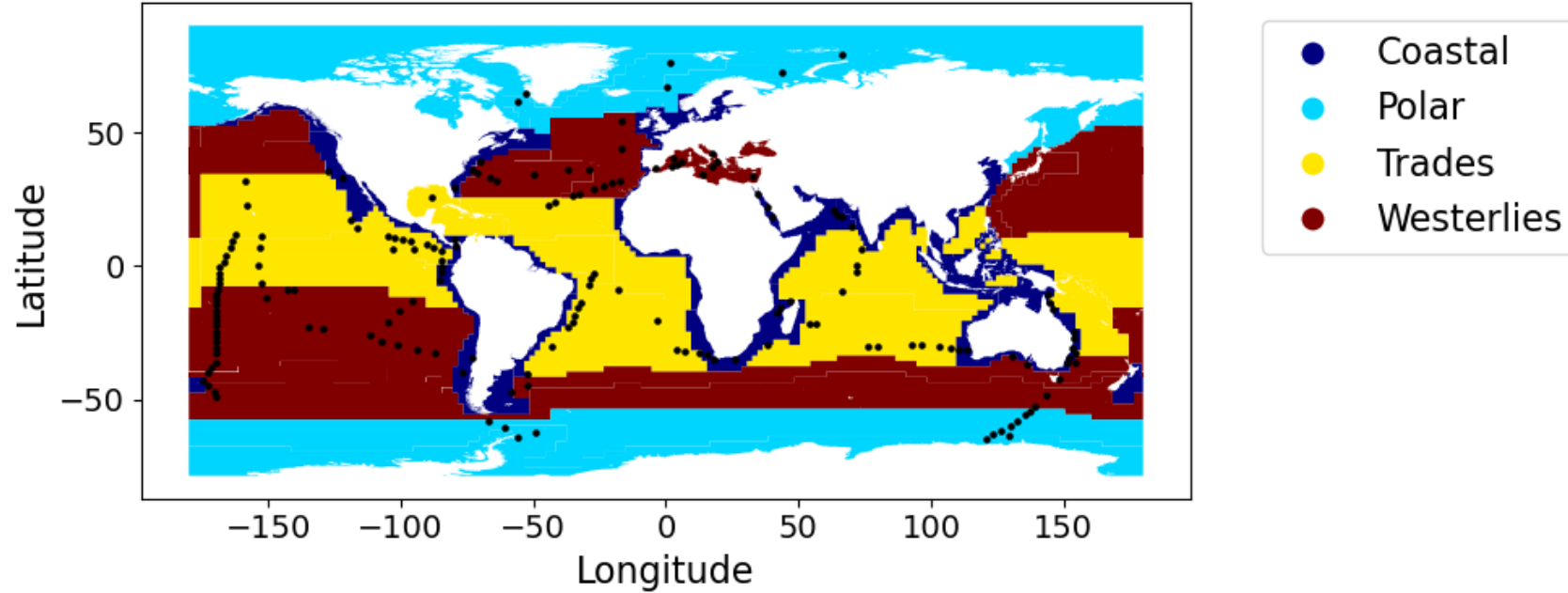
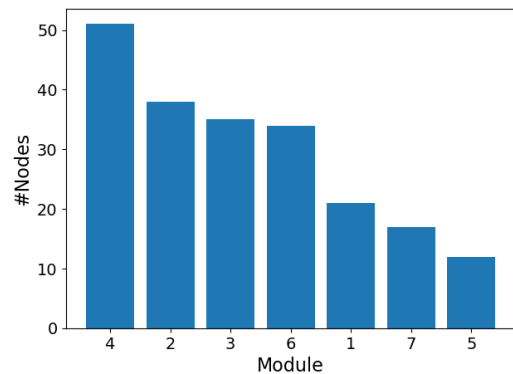


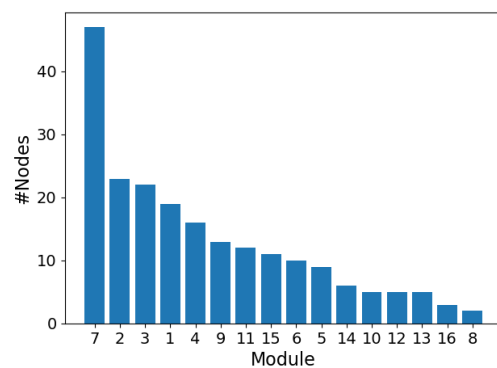
Figure S7. Map of the Longhurst biomes.

Black points are metabarcoding samples. The shape file of the Longhurst biomes was downloaded from Marine Regions (<https://www.marineregions.org>).

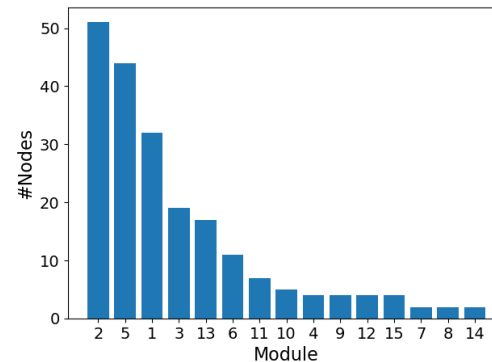
Fast Greedy
Modularity = 0.52



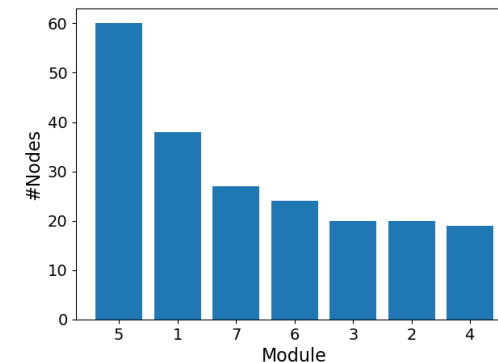
Infomap
Modularity = 0.51



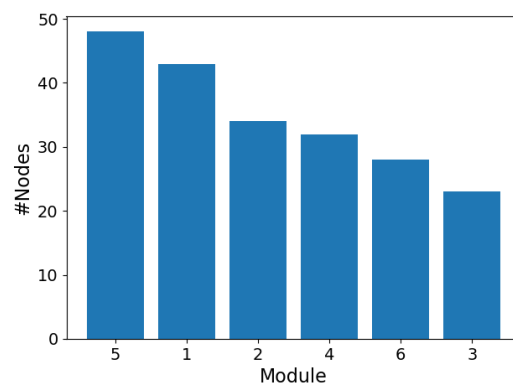
Label Propagation
Modularity = 0.49



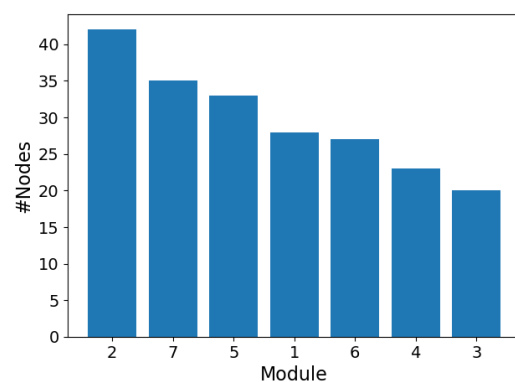
Leading Eigenvector
Modularity = 0.50



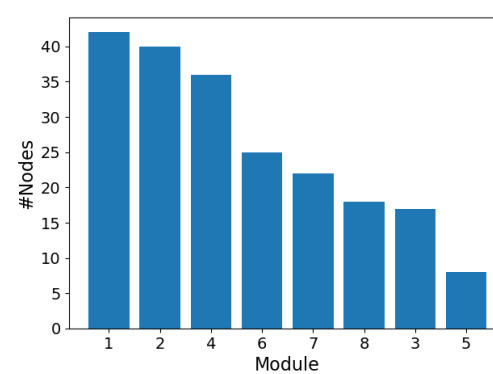
Leiden
Modularity = 0.55



Louvain
Modularity = 0.50



Spinglass
Modularity = 0.55



Walktrap
Modularity = 0.52

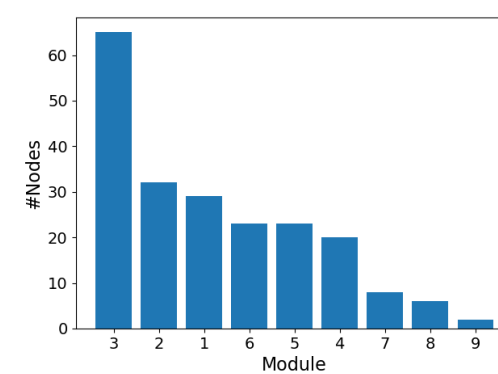


Figure S8. For each algorithm, modularity index and size of the modules.
Modularity index of the module division by each algorithm and size of detected modules by it.

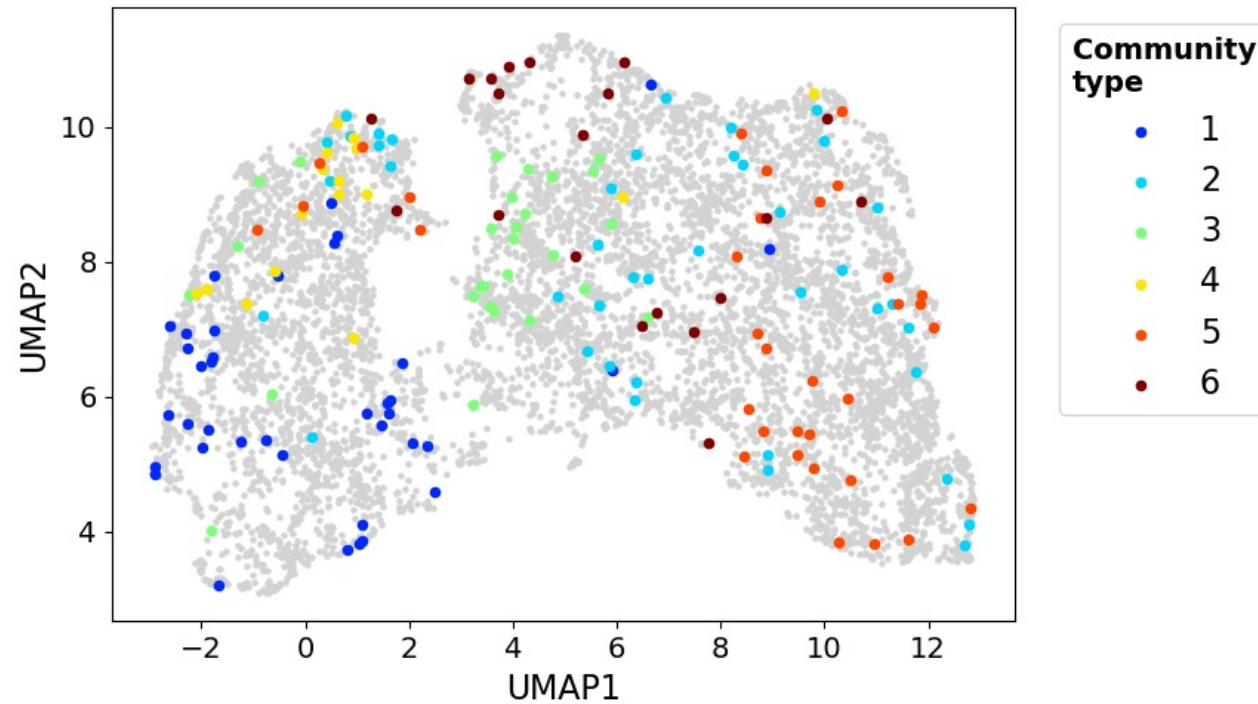
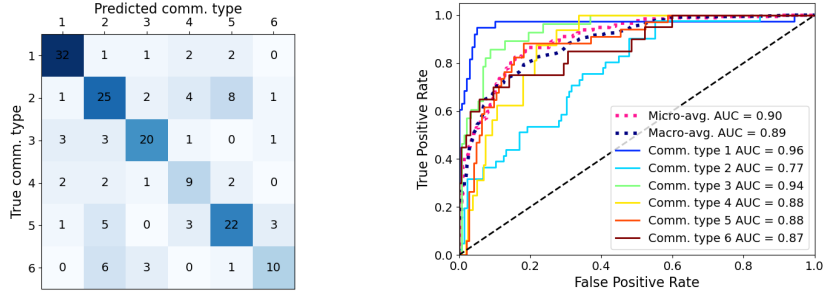


Figure S9. Community type distribution in the satellite-derived parameter space.

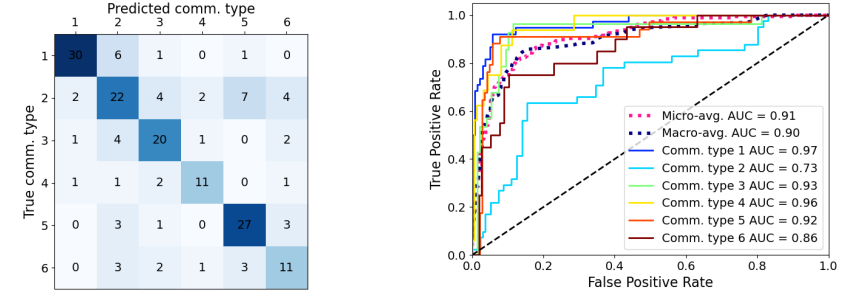
Metabarcoding samples projected on the 2-D map of the satellite-derived parameter space colored by community types. Small gray points are randomly selected grid cells used to train a UMAP projection.

Leave-one-out cross-validation

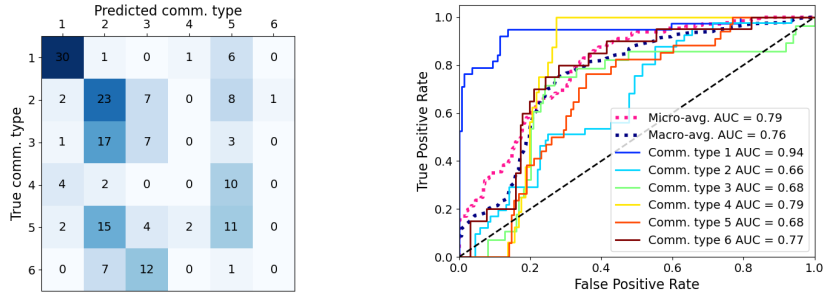
All 17 satellite-derived parameters (Acc = 0.67, AUC = 0.90)



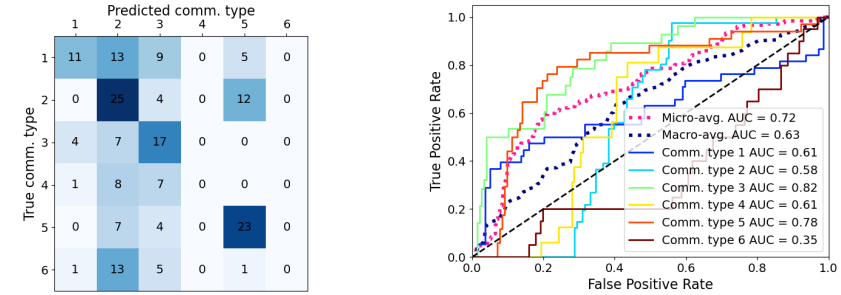
Latitude, Longitude (Acc = 0.68, AUC = 0.91)



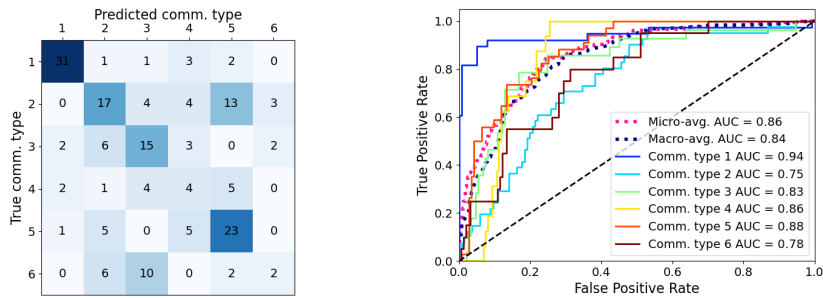
SST (Acc = 0.40, AUC = 0.79)



Chl *a* (Acc = 0.43, AUC = 0.72)



SST, Chl *a* (Acc = 0.52, AUC = 0.86)



All seven environmental parameters (Acc = 0.58, AUC = 0.88)

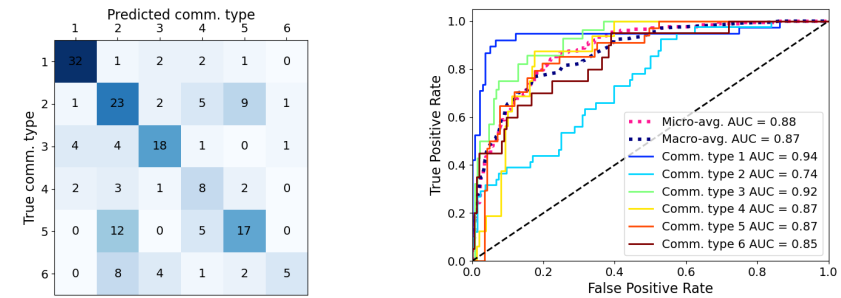
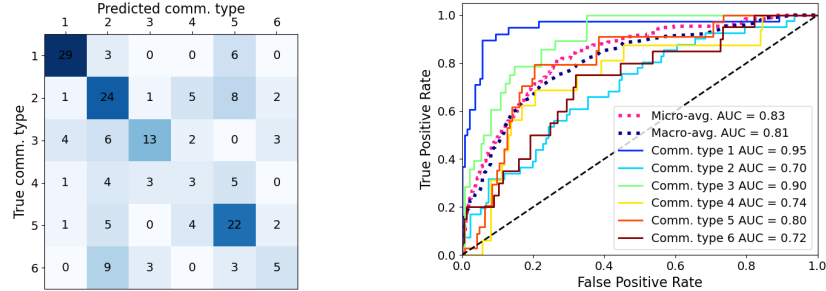


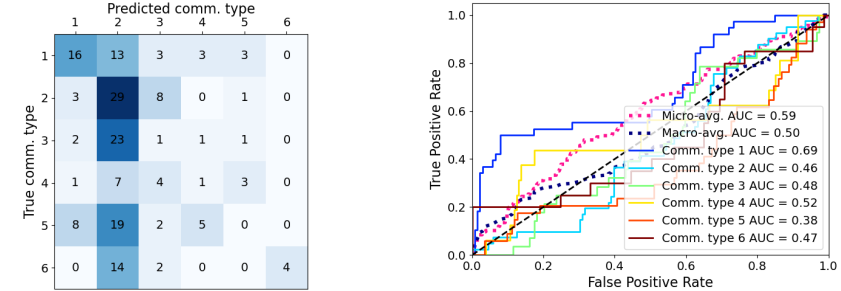
Figure S10. The leave-one-out cross-validation confusion matrix and ROC curve of SVM on community type prediction when different sets of parameters were used, related to Table 1.

Buffered cross-validation

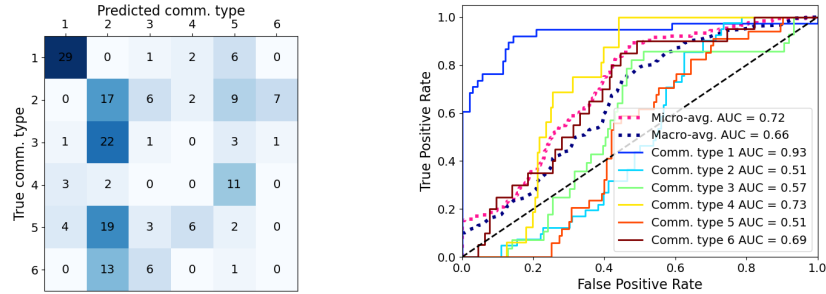
All 17 satellite-derived parameters (Acc = 0.54, AUC = 0.83)



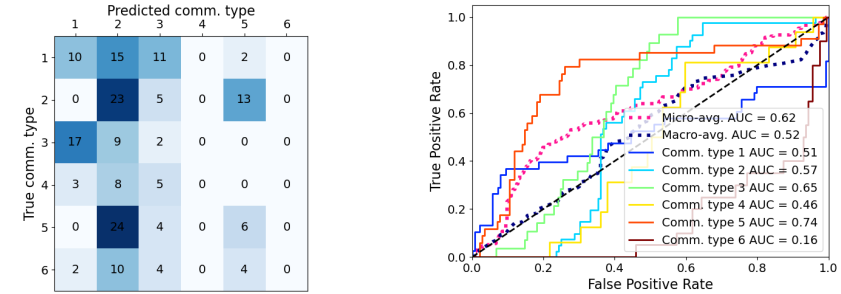
Latitude, Longitude (Acc = 0.29, AUC = 0.59)



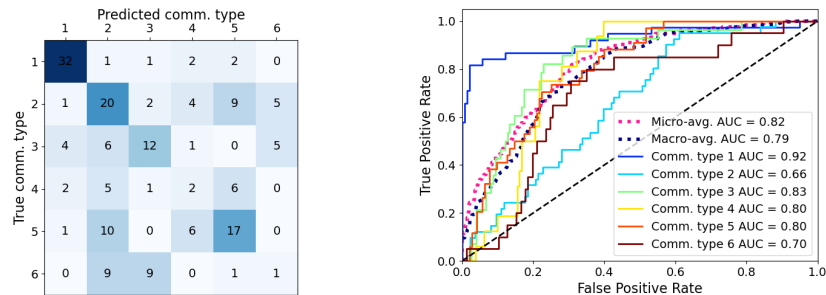
SST (Acc = 0.28, AUC = 0.72)



Chl *a* (Acc = 0.23, AUC = 0.62)



SST, Chl *a* (Acc = 0.47, AUC = 0.82)



All seven environmental parameters (Acc = 0.50, AUC = 0.83)

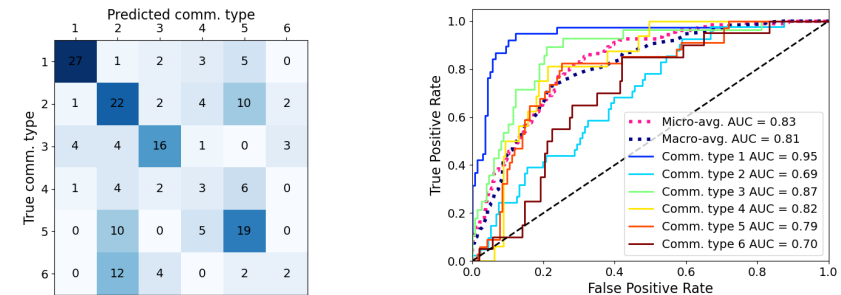


Figure S11. The buffered cross-validation confusion matrix and ROC curve of SVM on community type prediction when different sets of parameters were used, related to Table 1.

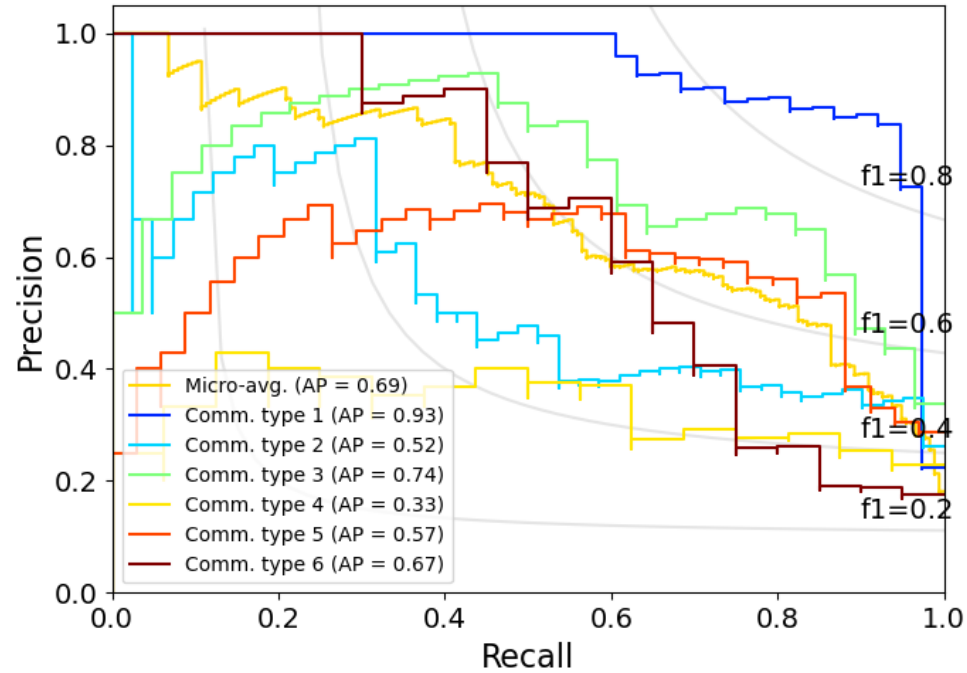
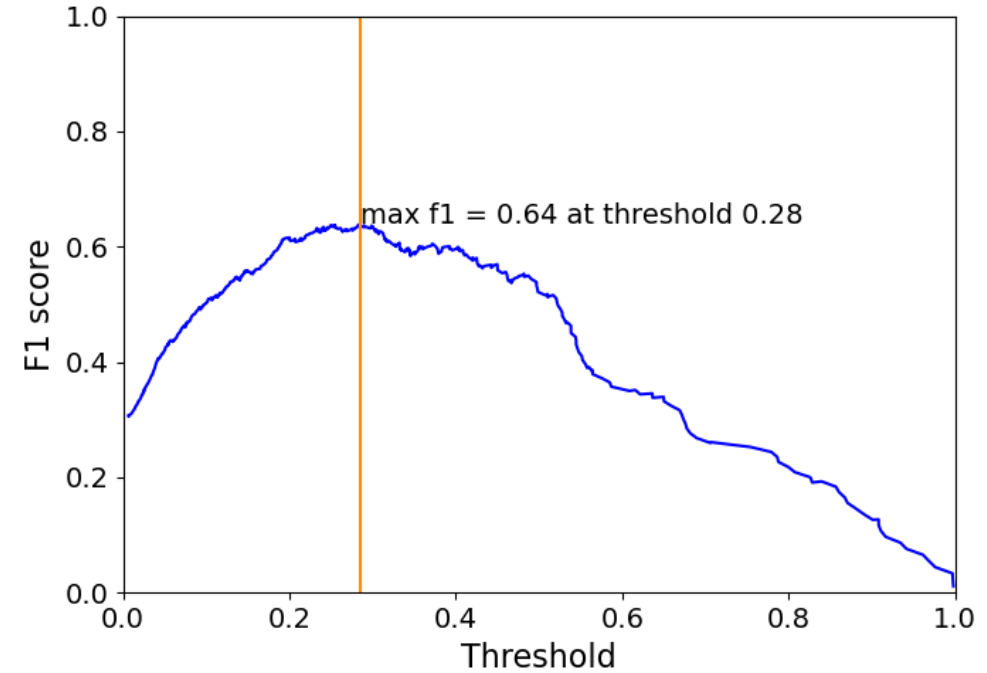
A**B**

Figure S12. Precision, recall, and F1 score of SVM on community type prediction using all 17 satellite-derived parameters.

A The precision-recall curve in the condition of leave-one-out cross-validation same as Figure 4B. **B** F1 score versus threshold of probabilistic output of SVM. Orange line shows the threshold making highest F1 score.

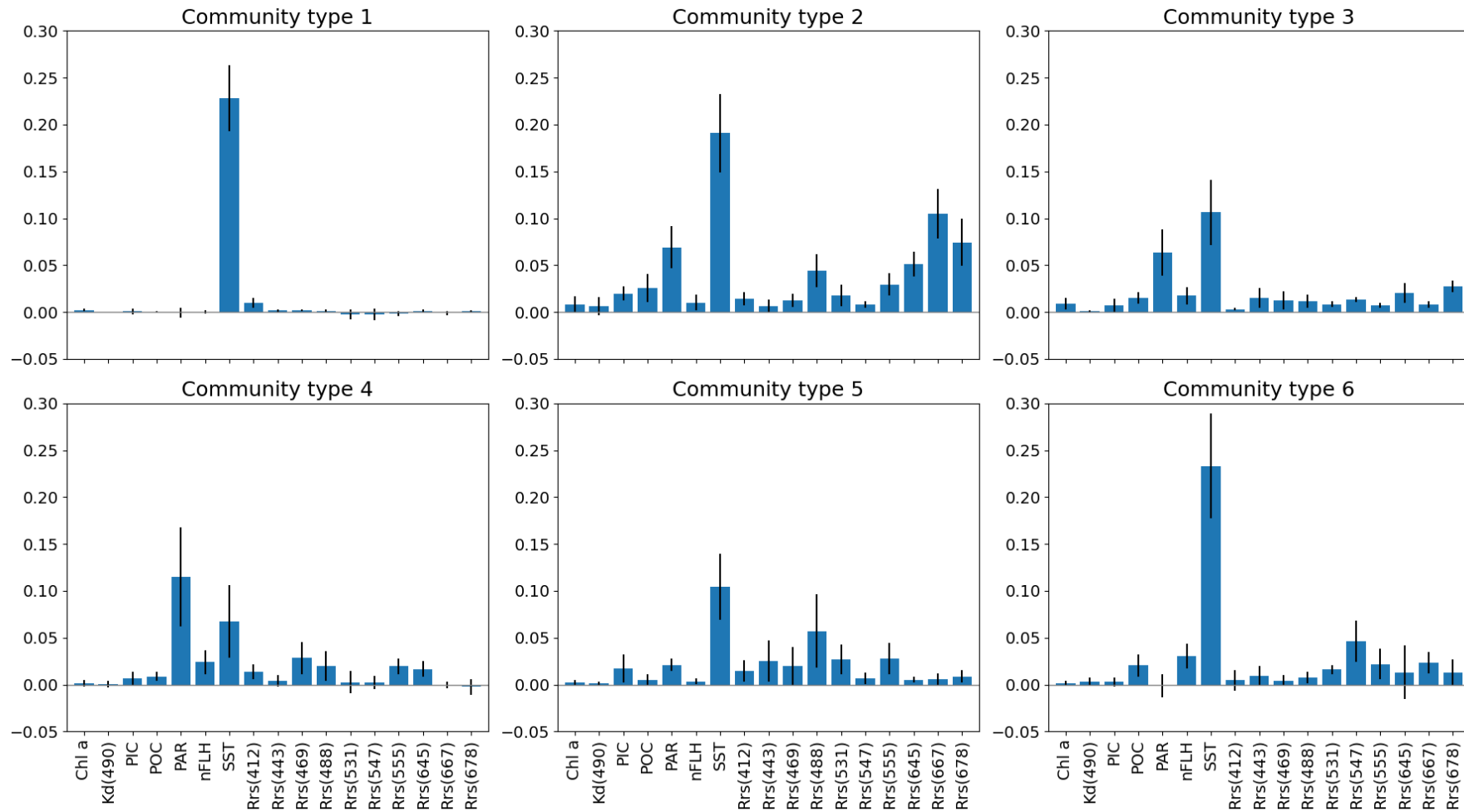


Figure S13. Permutation importance of each parameter for individual community type prediction.

Blue bars show mean values of parameter importance over 5 times repeats (error bars: standard deviation). ROC-AUC was used for scoring.

Polar

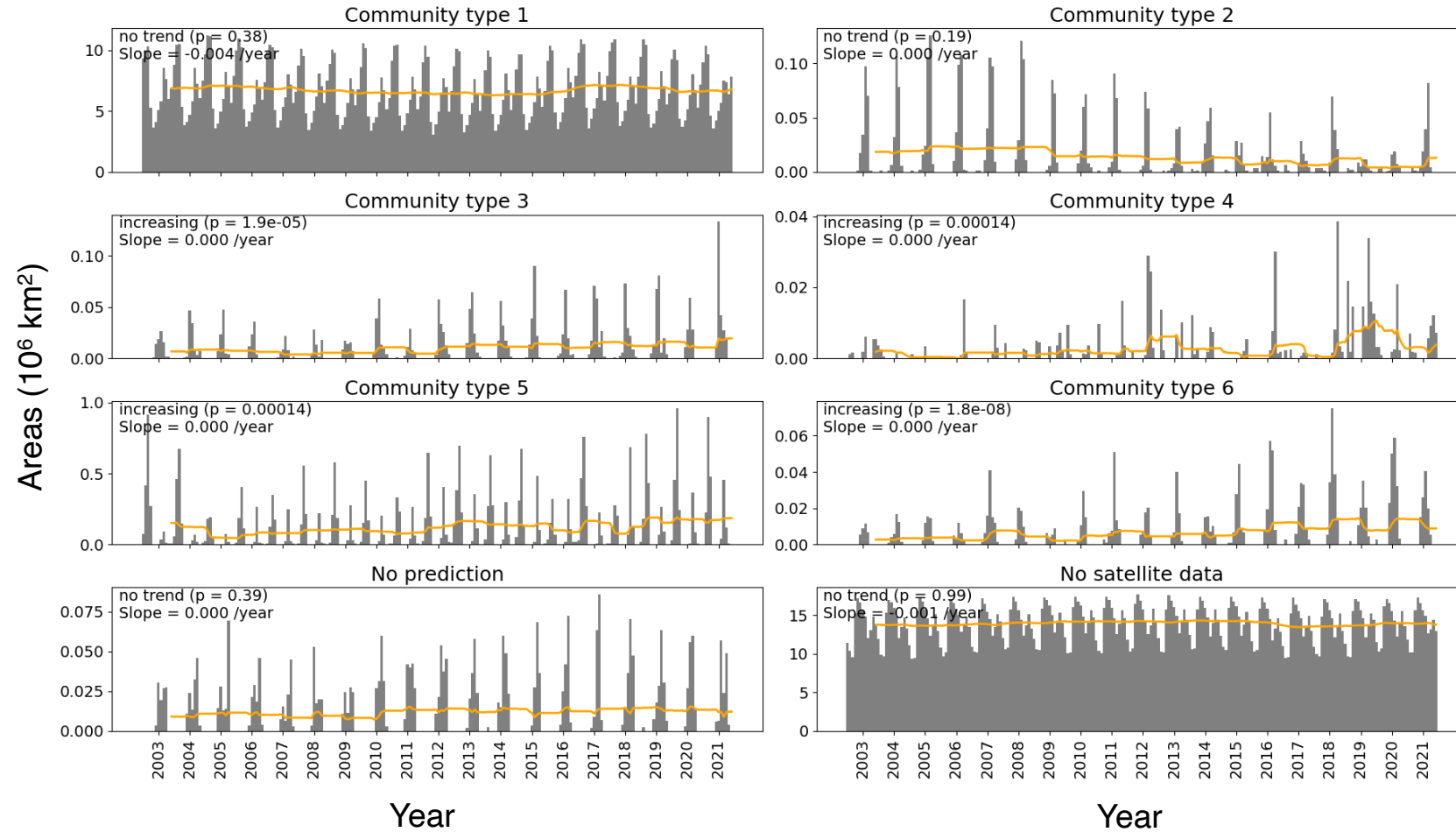


Figure S14. Areas of community types in Polar biome predicted based on satellite data from 2003 to 2021.

Gray bars show the areas in each month and orange lines show the seasonal rolling mean curve of them. Trends were tested by the seasonal Mann-Kendall test and slopes were estimated by the seasonal Theil-Sen's slope estimator.

Trades

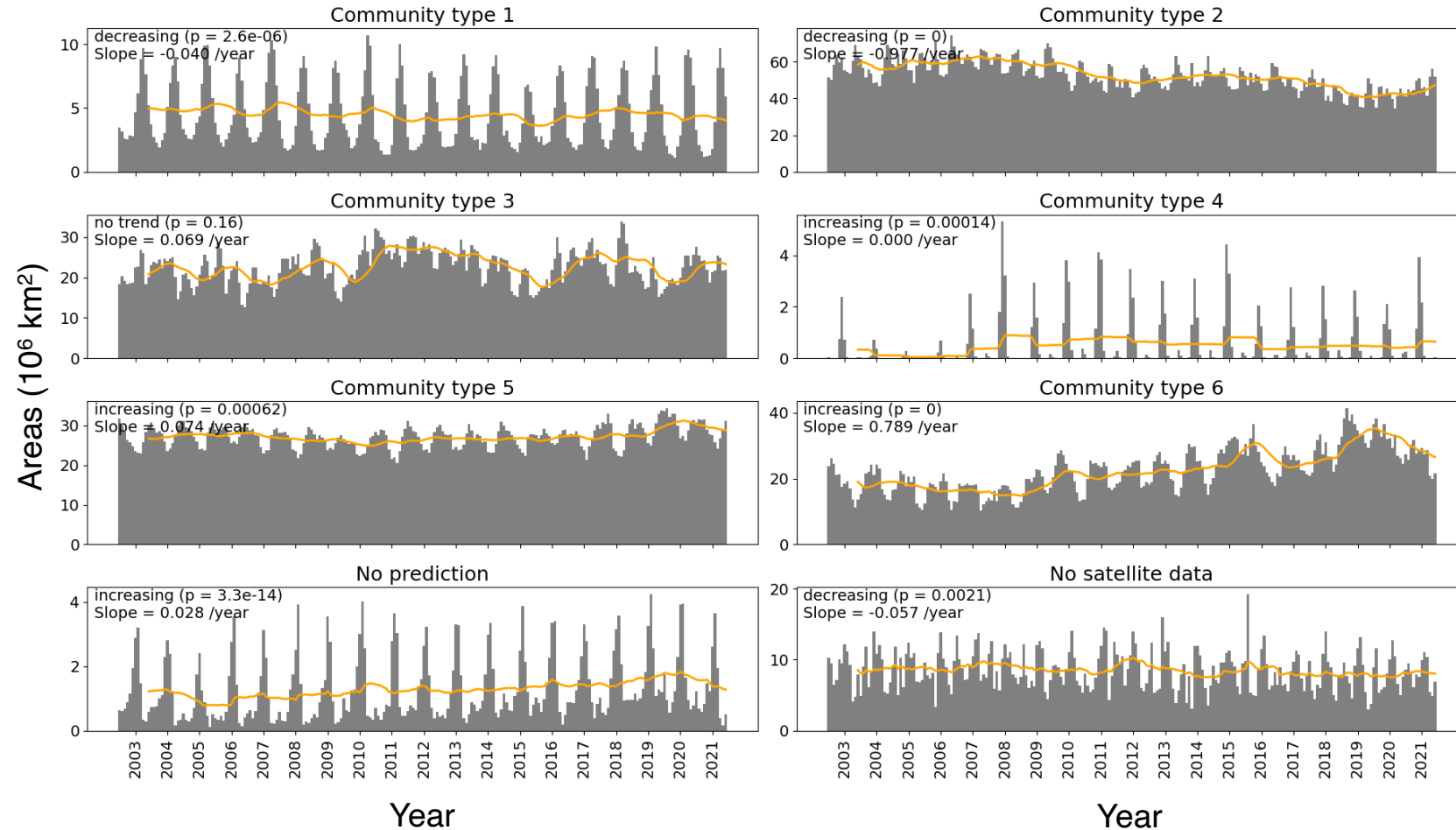


Figure S15. Areas of community types in Trades biome predicted based on satellite data from 2003 to 2021.

Gray bars show the areas in each month and orange lines show the seasonal rolling mean curve of them. Trends were tested by the seasonal Mann-Kendall test and slopes were estimated by the seasonal Theil-Sen's slope estimator.

Westerlies

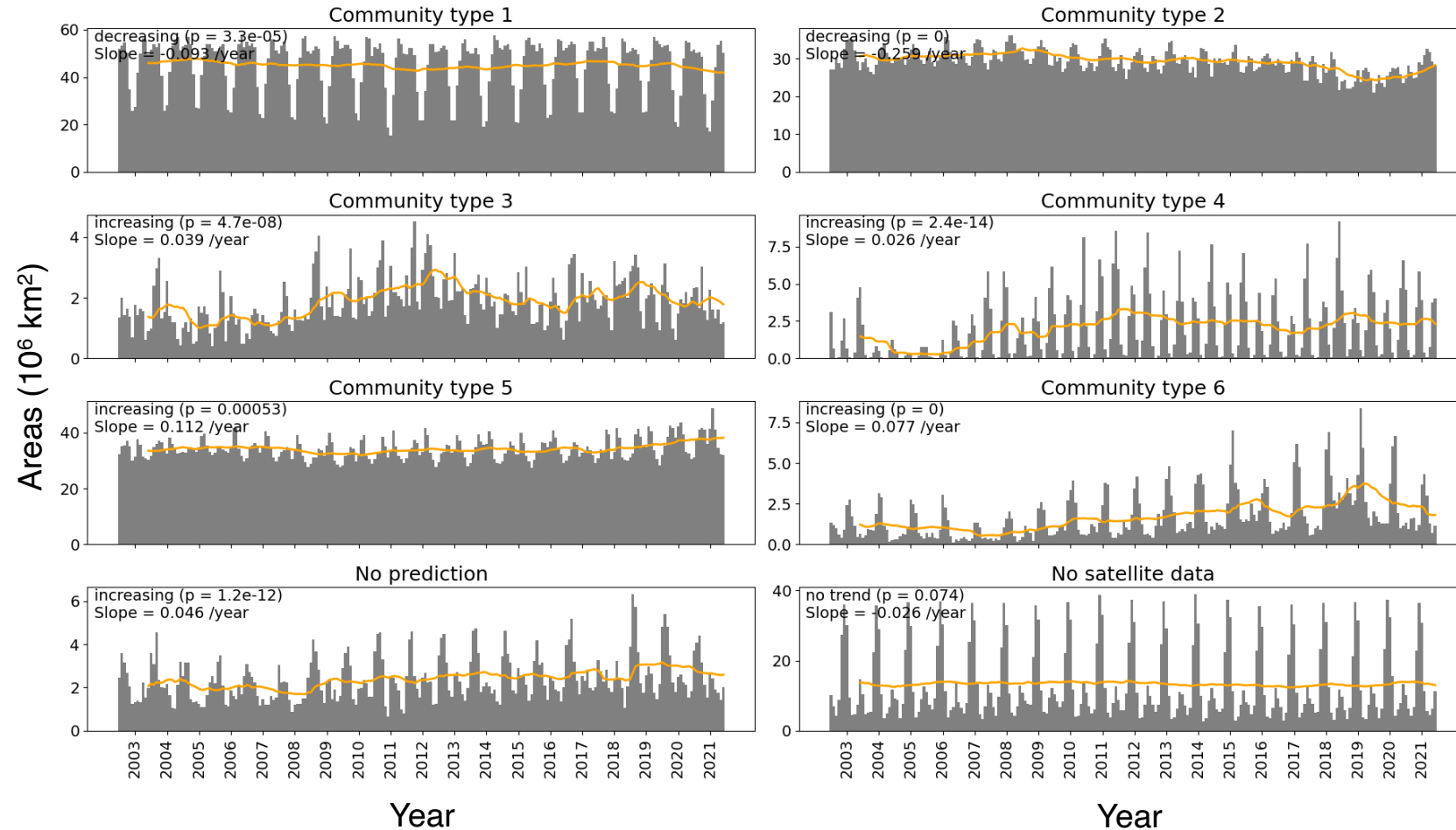


Figure S16. Areas of community types in Westerlies biome predicted based on satellite data from 2003 to 2021.

Gray bars show the areas in each month and orange lines show the seasonal rolling mean curve of them. Trends were tested by the seasonal Mann-Kendall test and slopes were estimated by the seasonal Theil-Sen's slope estimator.

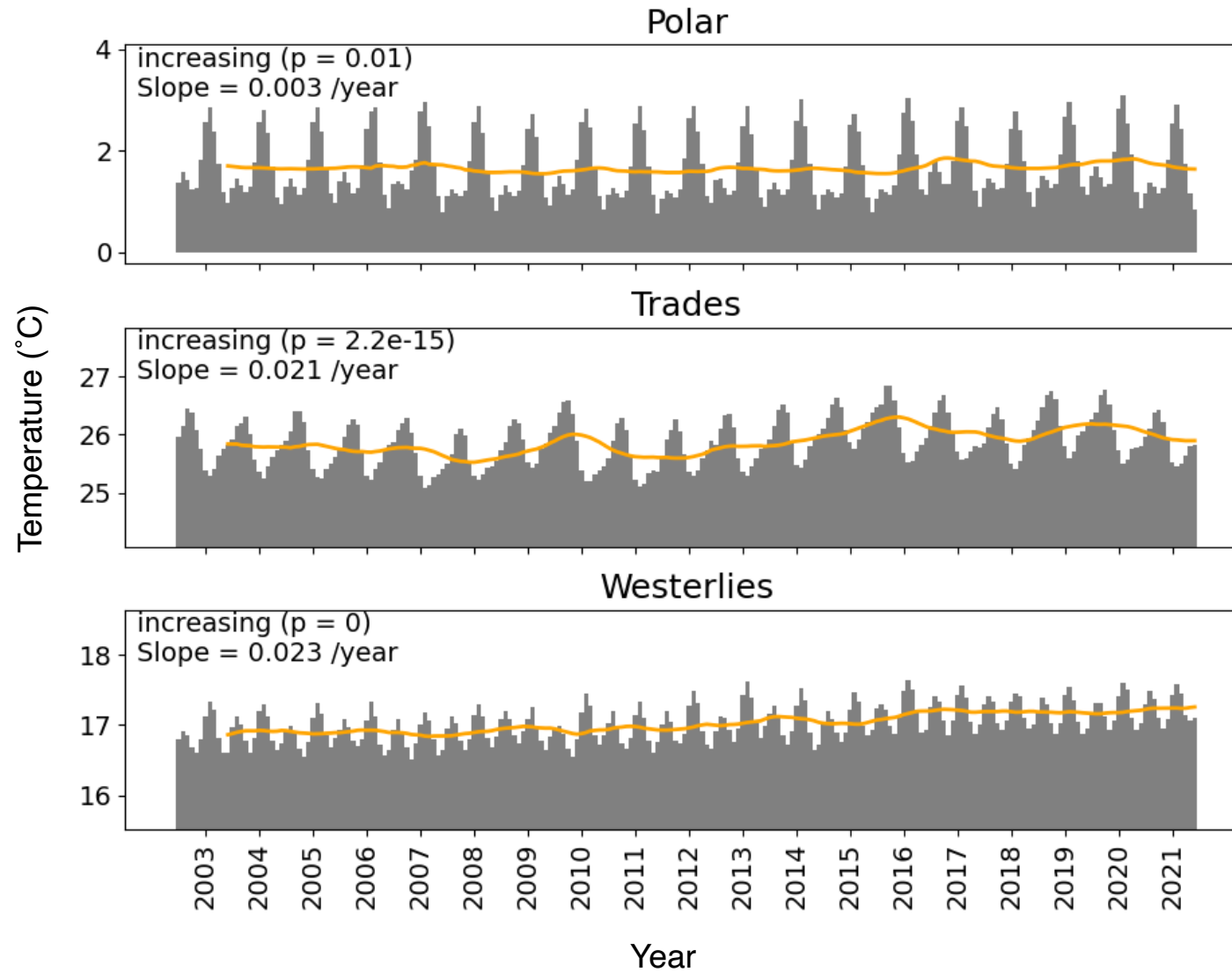


Figure S17. Average sea surface temperature (SST) in each biome observed by satellite from 2003 to 2021.

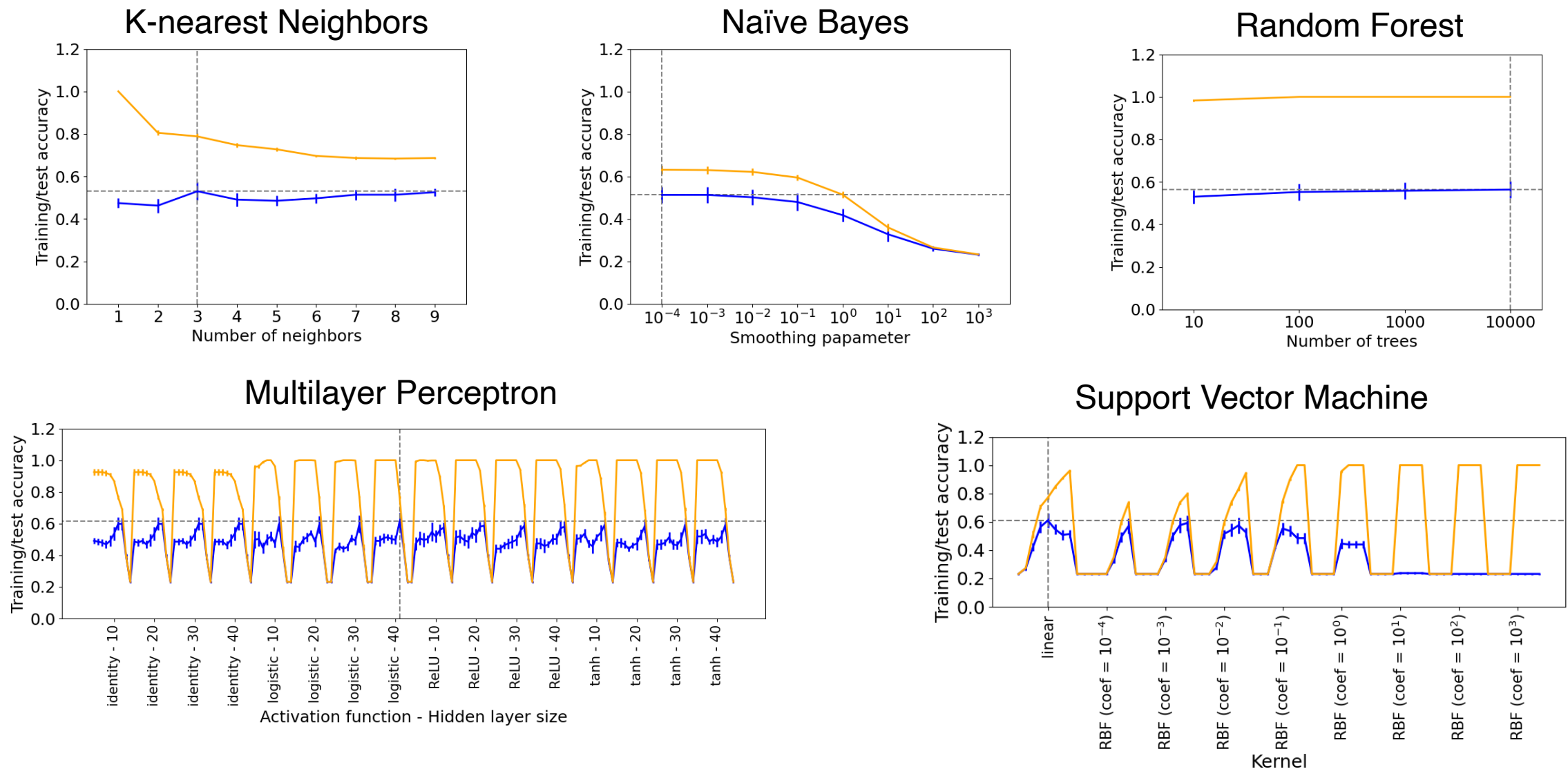


Figure S18. Grid search results in the training of predictive models with all samples. Orange and blue lines show training and test accuracy, respectively. Gray dashed lines show the parameter with the best test accuracy. Ten L2 penalty parameters (10^{-6} , 10^{-5} , ..., 10^3 ; from left to right) were tested for each setting of Multilayer Perceptron and eight (10^{-4} , 10^{-3} , ..., 10^3) were tested for Support Vector Machine.

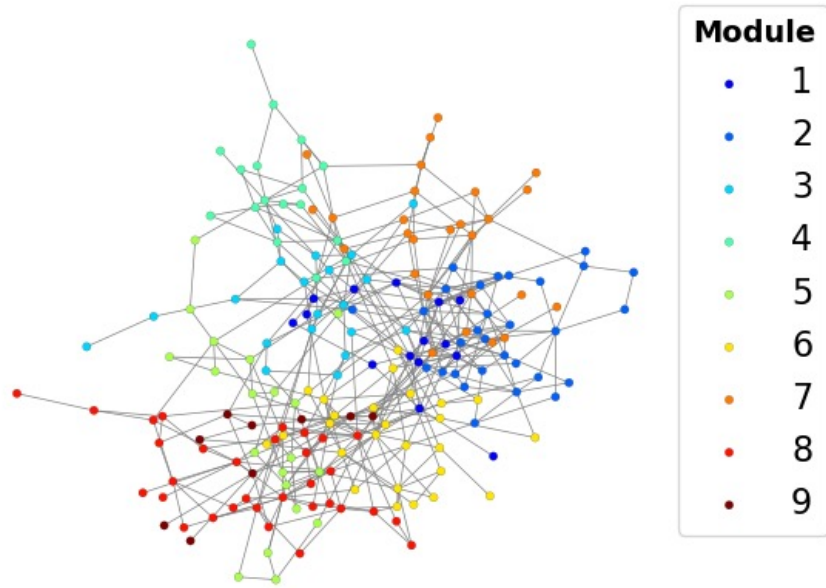
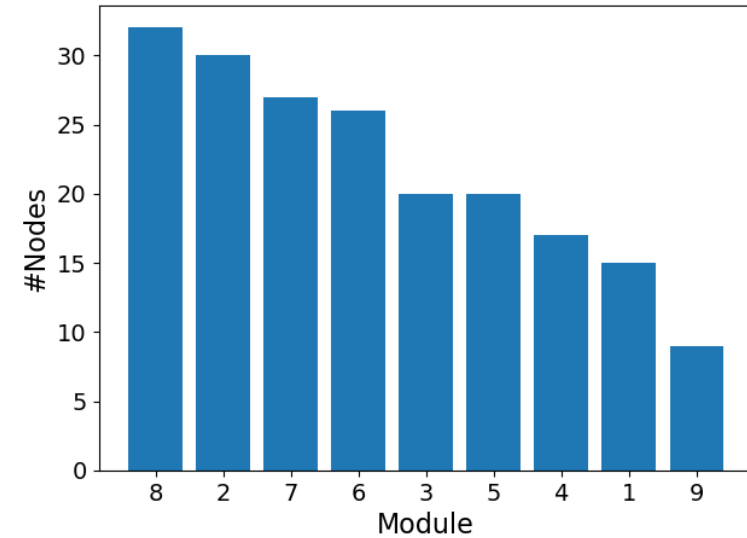
A**B**

Figure S19. Plankton network inferred with a “heterogeneous=True” option in FlashWeave.

A A force-directed representation of the network. Nodes (plankton OTUs) are colored by belonging module. 415 positive edges (correlation coefficients > 0) between 196 OTUs were detected. Twelve nodes were not included in the network because they had no edges. **B** Size of detected modules by Leiden algorithm. Modularity index was 0.57.