frontiers in Marine Science

TYPE Brief Research Report
PUBLISHED 23 July 2026
DOI 10.3389/fmars.2026.1875929

OPEN ACCESS

EDITED BY
Lin Liu,
Ministry of Natural Resources, China

REVIEWED BY
Vinícius W. Salazar,
The University of Melbourne, Australia
Luciana Frazão,
University of São Paulo, Brazil

*CORRESPONDENCE
Christian Marchese
christianmarchese@cnr.it

RECEIVED 08 May 2026
REVISED 19 June 2026
ACCEPTED 07 July 2026
PUBLISHED 23 July 2026

CITATION
Marchese C, Zoffoli ML, Ramond P, Turk Dermastia T, Tinta T, Logares R, Galand PE and Organelli E (2026) Machine learning predictions for microbial eukaryotic plankton: implications from unevenly structured data.
Front. Mar. Sci. 13:1875929.
doi: 10.3389/fmars.2026.1875929

Machine learning predictions for microbial eukaryotic plankton: implications from unevenly structured data

Christian Marchese[1]*, Maria Laura Zoffoli[2], Pierre Ramond[3,4], Timotej Turk Dermastia[5], Tinkara Tinta[5], Ramiro Logares[3], Pierre E. Galand[4] and Emanuele Organelli[1]

[1]Institute of Marine Sciences (ISMAR), National Research Council (CNR), Rome, Italy, [2]Institute of Marine Sciences (ISMAR), National Research Council (CNR), Trieste, Italy, [3]Institute of Marine Sciences (ICM), Spanish National Research Council (CSIC), Barcelona, Spain, [4]Sorbonne Université, CNRS, Laboratoire d'Ecogéochimie des Environnements Benthiques (LECOB), Observatoire Océanologique de Banyuls, Banyuls-sur-Mer, France, [5]National Institute of Biology (NIB), Marine Biology Station Piran, Piran, Slovenia

Machine learning models provide a scalable approach for predicting the diversity of eukaryotic microbial plankton from environmental predictors. However, the extent to which these models generalize to data outside the training set remains poorly quantified. In this study, XGBoost was used to predict the 18S rRNA gene Shannon Diversity Index (SDI) from seven environmental predictors derived from satellite and model data. Surface samples were collected between 2001 and 2025 at two fixed stations in the northwestern Mediterranean (BBMO and SOLA), one fixed station in the northern Adriatic Sea (VIDA), and during the HOTMIX expedition, which sampled an east-west open-sea transect across the Mediterranean Sea. Model performance was assessed using standard repeated K-fold cross-validation (CV), Leave-One-Dataset-Out CV (LODO-CV), and a blocked spatio-temporal CV that combined LODO with temporal forward chaining. Under standard K-fold CV, the model showed moderate performance (R^2 = 0.44, RMSE = 0.59). In contrast, performance declined substantially under LODO-CV (R^2 = 0.09, RMSE = 0.73), with uniformly low per-dataset generalization, a pattern also observed with blocked spatiotemporal CV. VIDA and HOTMIX sampled environmental regimes distinct from those at BBMO and SOLA, which may partly explain their poor transferability. Additionally, BBMO and SOLA, despite similar environmental conditions, exhibited poor transferability, indicating that technical differences among independently collected 18S rRNA datasets likely constrain transferability, although their effects cannot be disentangled from environmental variation. Overall, these results highlight the limitations of imbalanced training data and underscore the importance of spatially explicit evaluation, protocol standardization, and environmentally representative coverage.

KEYWORDS

cross-validation, data leakage, eukaryotic plankton, machine learning, marine biodiversity, Mediterranean Sea, omics, shannon diversity index

1 Introduction

Marine biodiversity underpins ecosystem function and biogeochemical cycling (Worm et al., 2006; Cardinale et al., 2012), so limitations in its assessment have consequences that extend well beyond the mere description of community structure. Biodiversity loss can alter ecosystem functioning and biogeochemical processes through changes within and across trophic levels (Duffy et al., 2007). Nevertheless, marine biodiversity remains largely uncharacterized, with current estimates indicating that only 10-30% of marine species diversity has been described (Danovaro, 2024). Microbial communities exemplify this disparity; although they drive marine biogeochemical cycles, regulate greenhouse gas fluxes, and mediate ecosystem responses to environmental change, they remain disproportionately understudied relative to their ecological importance (Cavicchioli et al., 2019). Resolving the spatial and temporal patterns of marine biodiversity under ongoing climate change therefore requires sustained observation across the full extent of the ocean.

Capturing biodiversity at this scale is difficult because marine environments are a mosaic of ecosystems shaped by distinct physical, chemical, and geological factors that support diverse biological communities (Longhurst, 2007). The ocean's scale and variability require observations from multiple complementary platforms, since no single method can capture the full range of spatial and temporal variability. Time-series stations, moored observatories, and ship-based surveys provide high-frequency observations at fixed locations and high-accuracy measurements along repeat transects. Complementing these approaches, autonomous platforms such as gliders and Argo (including Biogeochemical-Argo) floats continuously sample the ocean interior, and satellite remote sensing offers synoptic coverage of the surface ocean (Miloslavich et al., 2018; Testor et al., 2019; Claustre et al., 2020). However, this expansion of monitoring is uneven. While autonomous platforms and satellite remote sensing enable near-continuous, large-scale observations of physical and biogeochemical variables, *in situ* biodiversity sampling remains spatially sparse and temporally discontinuous, mainly tied to cruises and fixed monitoring stations (Miloslavich et al., 2018; Bax et al., 2019). *In situ* biodiversity observations remain indispensable as ground truth, but their limited coverage can constrain the generalizability of data-driven models trained on them.

These large, high-dimensional, and still-growing datasets, together with the heterogeneity and autocorrelation typical of ecological observations (Crisci et al., 2012), exceed the capacity of traditional statistical techniques and motivate the use of machine learning (ML) (Malde et al., 2020; Goodwin et al., 2022; Rubbens et al., 2023). ML algorithms can learn complex relationships without predefined assumptions about data distributions or underlying processes (Lou et al., 2023; Rubbens et al., 2023). Critically, by linking *in situ* biodiversity measurements with satellite- and model-derived environmental predictors, ML can extend discrete observations into spatially continuous estimates across ocean basins. Marine protists, including diatoms, dinoflagellates, ciliates, and small flagellates, represent a particularly informative group for this type of analysis. They encompass diverse trophic modes and life-history strategies, respond rapidly to environmental change, and play central roles in marine food webs (Arrigo, 2005; Hays et al., 2005). Molecular surveys, particularly 18S rRNA gene metabarcoding, offer a scalable approach for characterizing these communities across broad taxonomic and size ranges, including nano- and picoplanktonic fractions that are challenging to distinguish morphologically (de Vargas et al., 2015). Recent studies have combined omics and environmental data to upscale community patterns. El Hourany et al. (2024) integrated remotely sensed predictors with omics-derived data to estimate the global phytoplankton community structure, and Marchese et al. (2026) applied a similar framework to map prokaryotic communities across the Mediterranean Sea. More broadly, integrative approaches are increasingly utilized in marine ecology for upscaling, classification, and pattern detection (Rubbens et al., 2023).

Despite these advances, methodological limitations persist. ML models are susceptible to overfitting, learning patterns specific to the training data that fail to generalize to new observations (Gray et al., 2024). When spatial and temporal autocorrelation is ignored during model evaluation, dependence between training and test data can inflate performance estimates through data leakage (Roberts et al., 2017; Gray et al., 2024). This issue is particularly relevant in marine ecology, where uneven spatiotemporal sampling often underrepresents environmental gradients and temporal variability. Here, we quantified the effects of data leakage using an unevenly distributed omics dataset from the Mediterranean Sea. An XGBoost model was trained to predict microbial eukaryotic plankton diversity, quantified by the Shannon Diversity Index (SDI) from 18S rRNA gene sequencing data, using seven satellite- and model-derived environmental variables. We compared model performance under standard repeated K-fold cross-validation (CV), Leave-One-Dataset-Out (LODO) CV, and a blocked spatiotemporal CV that combines LODO with temporal forward chaining. Our findings highlight the necessity of making leakage-aware evaluations standard practice before scaling from discrete *in situ* observations to basin-wide predictions.

2 Methods

2.1 Shannon diversity index and environmental data

Microbial eukaryotic community diversity in surface seawater was quantified using the Shannon Diversity Index (SDI; Shannon, 1948), computed from Amplicon Sequence Variant (ASV) abundances for 474 samples. Details on seawater sampling, laboratory processing, and SDI calculation are provided in the Supplementary Text S1. SDI combines richness and evenness and places greater weight on dominant than on rare taxa, making it less sensitive to incomplete sampling than richness-based metrics (Haegeman et al., 2013; Feranchuk et al., 2018). However, SDI does not capture compositional differences, as two samples with identical SDI values may contain entirely different taxa. Despite this limitation, the SDI remains a robust summary statistic for large-scale

biodiversity assessments and is well-suited for modeling broad spatiotemporal patterns (Marchese et al., 2026).

Environmental predictors were obtained from the Copernicus Marine Service (CMEMS) and selected based on evidence that environmental gradients and oceanographic features shape eukaryotic plankton diversity, community composition, and biogeography (Ibarbalz et al., 2019; Sommeria-Klein et al., 2021; Lévy et al., 2025). Net primary production (NPP; g C m^{-2} d^{-1}) was retrieved from the ocean-color product *OCEANCOLOUR_MED_BGC_L4_MY_009_144* (Volpe et al., 2019). Concentrations of nitrate (NO_3^-; mmol m^{-3}), phosphate (PO_4^{3-}; mmol m^{-3}), and dissolved inorganic carbon (DIC; mol m^{-3}) were derived from the biogeochemical reanalysis *MEDSEA_MULTIYEAR_BGC_006_008* (Cossarini et al., 2021). Temperature (°C) and salinity (PSU) were obtained from the physical reanalysis *MEDSEA_MULTIYEAR_PHY_006_004* (Escudier et al., 2021). The euphotic depth (i.e., the sunlit layer supporting photosynthesis; in meters) was retrieved from the global biogeochemical product *GLOBAL_MULTIYEAR_BGC_001_033* (Lehodey et al., 2010, 2015).

Two additional predictors were derived: photoperiod (daylength, in hours), calculated from date and latitude following Forsythe et al. (1995) to capture seasonality, and the nitrate-to-phosphate ratio (N:P), computed from modeled nutrient fields to represent nutrient balance, a key driver of plankton community structure. Daily environmental data (2001-2025) were extracted for each sampling location (Figure 1A) using a 3 × 3-pixel window and averaged. Matching these data to SDI observations by location and sampling date, and retaining samples with complete environmental predictors, yielded a modeling dataset of 469 of the 474 samples.

2.2 Machine learning model and cross-validation frameworks

XGBoost (Chen and Guestrin, 2016) is a gradient-boosted ensemble ML algorithm that sequentially builds decision trees while incorporating regularization to reduce overfitting. It has become a widely adopted algorithm in ecological research (e.g., Kim et al., 2024; Mitra et al., 2024; Song et al., 2024). For this study, hyperparameters were tuned using a space-filling grid search with the root-mean-square error (RMSE) as the selection criterion, and early stopping was used to prevent overfitting. To reduce the influence of unequal dataset sizes during training, inverse-frequency case weights were applied at the dataset level. Stratified partitioning preserved each dataset's proportional representation across folds, whereas inverse-frequency weighting prevented larger datasets from dominating model fitting. The predictor set, hyperparameter search space, and tuning process were kept identical across all CV frameworks, ensuring that observed performance differences were attributable solely to the data-partitioning strategy.

Model performance was evaluated using three CV methods that imposed progressively stricter constraints on spatial and temporal independence to assess generalizability (Roberts et al., 2017). The first method, standard repeated k-fold CV, partitioned the full dataset into a training set (84%) and a held-out test set (16%), maintaining proportional representation of each source dataset. The training set was then divided into four dataset-stratified folds and repeated three times (12 resamples), with hyperparameters selected by minimizing mean validation RMSE. Final model performance was evaluated on the held-out test set, with 95% confidence intervals for R^2 and RMSE estimated from 1,000 bootstrap resamples. The second method, Leave-One-Dataset-Out CV (LODO-CV), was used to evaluate spatial generalizability. Each of the four datasets was excluded in turn, while the model was tuned and trained on the remaining three datasets, generating predictions for the withheld dataset. Because no samples from the target dataset were used during training, LODO-CV provides a conservative estimate of spatial predictive performance (Pohjankukka et al., 2017; Roberts et al., 2017). The third method, blocked spatiotemporal CV, combined the LODO outer loop with forward-chaining temporal splits for inner-loop hyperparameter tuning (using 2-year steps and a 3-year window), ensuring that future observations were not used during model selection. The final model was then trained on all samples from the remaining datasets and evaluated on the withheld dataset, so temporal blocking affected tuning but not the final training set. This design eliminated spatial leakage and reduced temporal leakage during tuning (Roberts et al., 2017). However, because temporal coverage is highly uneven across datasets, it should be read as a stricter spatial test rather than a clean separation of spatial and temporal effects. The three CV strategies and their data-partitioning schemes for evaluating model generalizability are illustrated in Supplementary Figure 1. All models and associated workflows were implemented in R (R Core Team, 2024).

3 Results

The four Mediterranean datasets differ in spatial and temporal sampling structure: BBMO, SOLA, and VIDA are coastal time-series stations, whereas HOTMIX sampled an east-west open-sea transect (Figure 1A). The SDI frequency distribution showed that most values clustered at the upper end of the range (median = 4.46, mean = 4.35), indicating that most samples supported highly diverse eukaryotic plankton communities (Figure 1B). BBMO displayed the widest range of SDI values at the dataset level. VIDA and HOTMIX were heavily underrepresented, accounting for only 14% of the samples, compared with 86% from BBMO and SOLA (Figures 1C, D). Seasonal coverage was broadly comparable across datasets, except for HOTMIX, which was sampled exclusively in spring (Figure 1E). Across the 2001–2025 record, the SDI exhibited a strong, persistent seasonal cycle (Figure 1F), although temporal coverage was highly uneven across datasets, further constraining the environmental coverage of the combined dataset.

SDI distributions differed significantly across the datasets (Kruskal-Wallis $p = 1.38 \times 10^{-7}$; Figure 2A). Pairwise comparisons (Figure 2B) showed small to moderate effect sizes, with the largest differences between SOLA and BBMO (Cohen's d = 0.57, $p < 0.001$), followed by SOLA and HOTMIX (d = 0.51, $p < 0.05$) and VIDA and BBMO (d = 0.41, $p < 0.05$). The remaining comparisons (VIDA-HOTMIX; SOLA-VIDA; HOTMIX-BBMO) were not significant. These differences in SDI distributions may reflect ecological variability, methodological inconsistencies, or both.

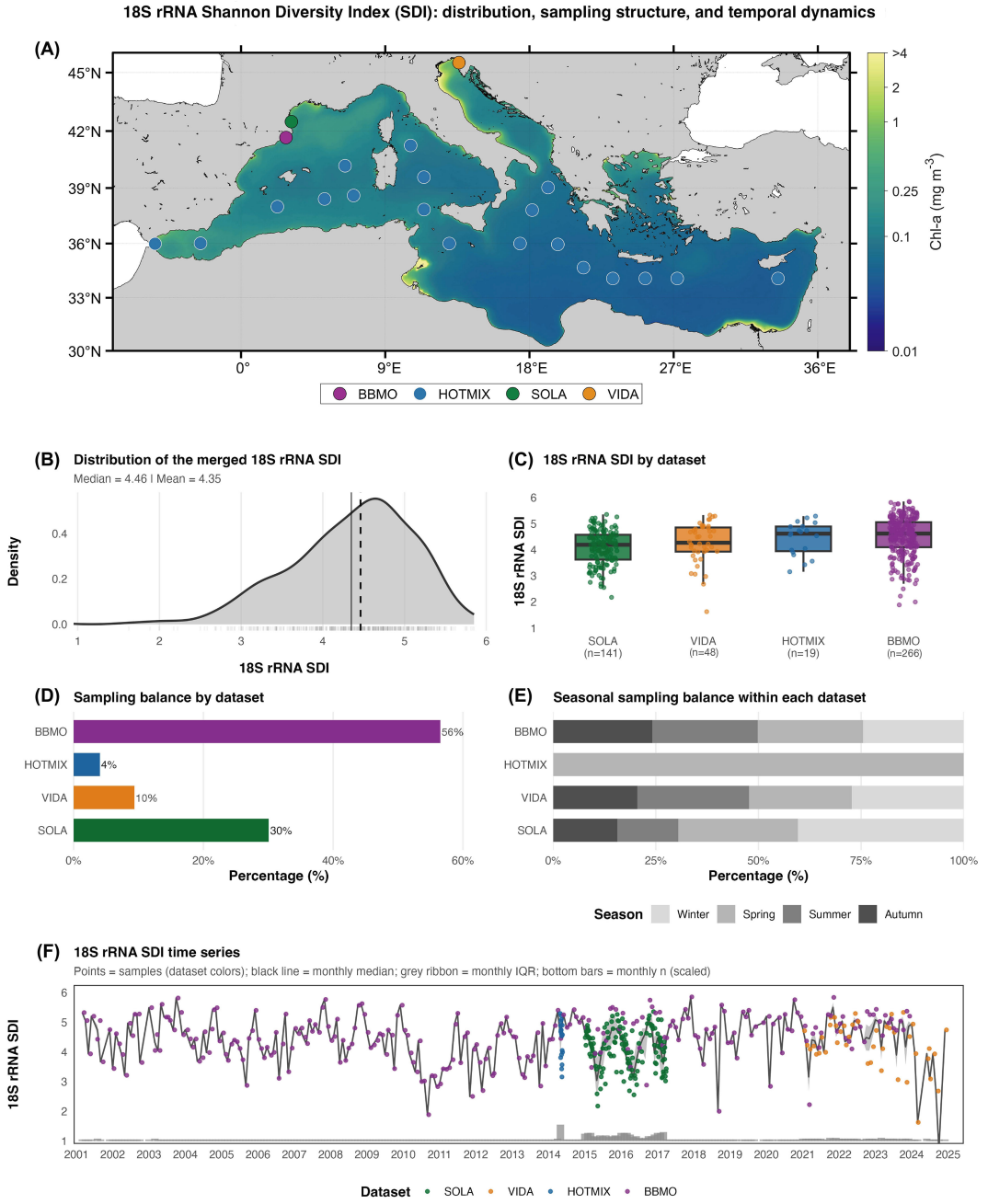

FIGURE 1
Microbial eukaryotic diversity (18S rRNA Shannon Diversity Index, SDI) across Mediterranean monitoring sites (n = 474 SDI observations). **(A)** Sampling locations: BBMO, SOLA, HOTMIX, and VIDA, overlaid on climatological surface Chlorophyll-a (Chl-a; mg m^{-3}). **(B)** Density distribution of the merged SDI values. **(C)** SDI by dataset with sample size. **(D)** Proportional contribution of each dataset to total observations. **(E)** Seasonal sampling balance within each dataset. **(F)** SDI time series (2001-2025) showing individual samples (colored by dataset), monthly median (black line), interquartile range (grey ribbon), and scaled monthly sample counts (bottom bars).

The PCA of the seven predictors showed substantial overlap across datasets, with VIDA showing the clearest separation along PC1 (Figure 2C). The correlation circle indicated that temperature, photoperiod, and euphotic depth contributed strongly to PC1, whereas dissolved inorganic carbon (DIC) contributed most strongly to PC2, with salinity loading negatively on both axes (Figure 2D). Overall, the first two components explained 60.7% of the total variance. Univariate comparisons confirmed that all seven predictors differed significantly across datasets (Kruskal-Wallis $p < 0.05$; Supplementary Figure 2). DIC and euphotic depth showed the most widespread pairwise differences, with significant contrasts in nearly all dataset pairs, whereas N:P ratio differences were concentrated between VIDA and the remaining three datasets (all $p < 0.001$; Supplementary Table 1).

Under repeated K-fold CV, the XGBoost model yielded R^2 = 0.50, RMSE = 0.58, MAE = 0.43, and MAPE = 12.3% on the training set (N = 393; Figure 3A). On the held-out test set (N = 76; Figure 3B), the model achieved R^2 = 0.44, RMSE = 0.59, MAE = 0.48, and MAPE = 11.1%. Bootstrap resampling of the test predictions produced 95% confidence intervals of 0.28-0.61 for

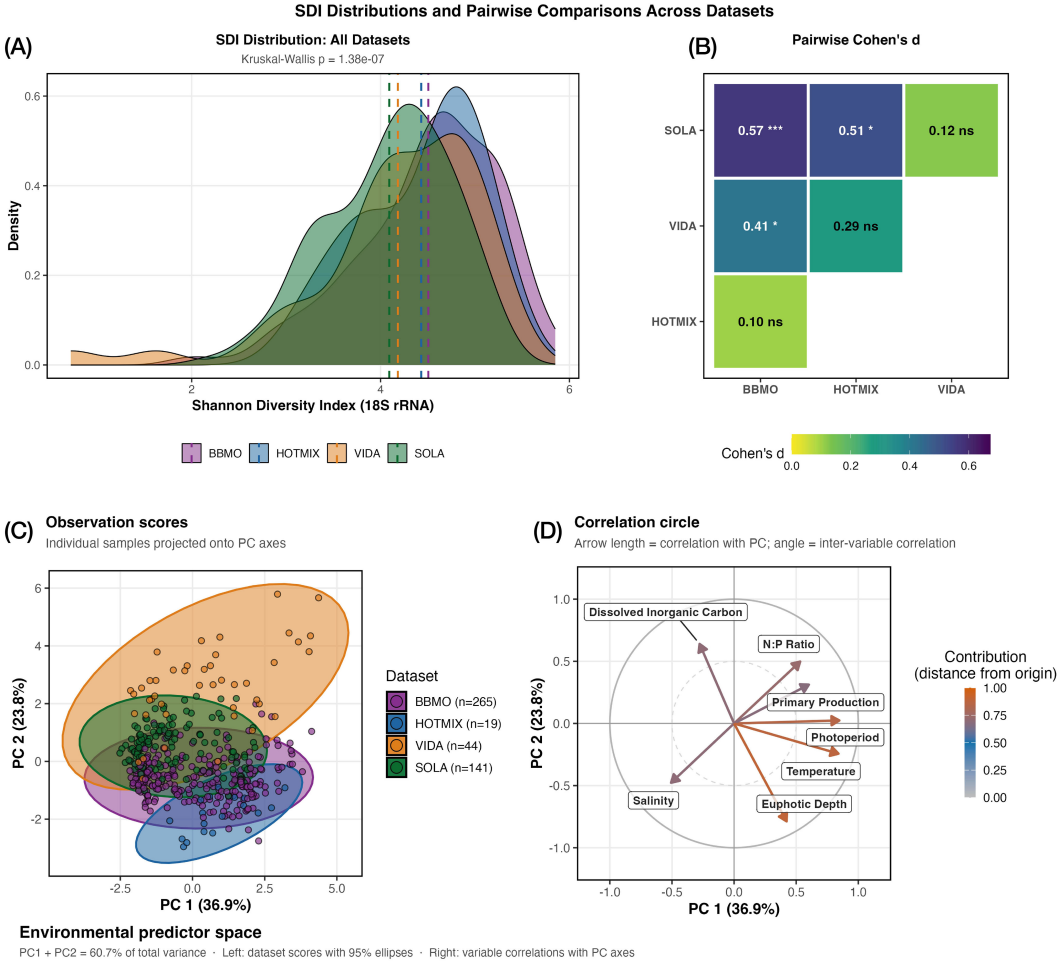

FIGURE 2
SDI distributions and environmental predictor space across the datasets. **(A)** Overlapping density distributions of microbial eukaryotic diversity, expressed as the 18S rRNA Shannon Diversity Index (SDI), with dashed lines indicating dataset medians (Kruskal–Wallis p = 1.38 × 10^{-7}). **(B)** Pairwise Cohen's d effect sizes with significance levels (***p < 0.001, *p < 0.05, ns = not significant). **(C)** PCA observation scores based on seven environmental predictors, with 95% confidence ellipses for each dataset. **(D)** PCA correlation circle showing variable loadings on PC1 and PC2; arrow length indicates correlation strength, direction indicates loading sign, and color indicates the variable contribution. The first two principal components explained ~61% of the total variance.

R^2 and 0.51-0.66 for RMSE, indicating substantial uncertainty in performance estimates and only moderate stability under random partitioning (Figure 3B). The stratified split preserved each dataset's relative proportions across the 84% training and 16% test sets (Supplementary Figure 3A). Fold composition was stable for BBMO and SOLA but varied for the smaller VIDA and HOTMIX (Supplementary Figure 3B). Because stratification preserved the underlying imbalance, model tuning was effectively driven by the two dominant datasets, with VIDA and HOTMIX contributing minimally to each fold. Salinity was the most important predictor, followed by DIC, photoperiod, N:P ratio, euphotic depth, temperature, and NPP (Supplementary Figure 3C). These results suggest an environmental influence on eukaryotic microbial diversity. However, whether K-fold performance reflects true generalization to independent datasets remains unclear.

When the evaluation shifted to LODO-CV, in which each dataset was predicted solely by a model trained on the remaining three datasets, overall predictive performance decreased sharply (R^2 = 0.09, 95% CI [0.05-0.15]; RMSE = 0.73, 95% CI [0.68-0.79]; MAE = 0.56, MAPE = 15.1%; Figure 3C; Supplementary Table 2). Per-dataset performance was uniformly low, though HOTMIX and SOLA retained marginally higher predictive performance (R^2 = 0.23 and 0.20, respectively) than BBMO and VIDA (R^2 = 0.07 and 0.06, respectively; Figure 3D).

Under the blocked spatiotemporal CV, which combined spatial hold-out with temporal forward chaining, predictive accuracy remained comparable to LODO-CV (R^2 = 0.07, 95% CI [0.04-0.12]; RMSE = 0.74, 95% CI [0.68-0.79]; MAE = 0.58, MAPE = 15.4%; Figure 3E; Supplementary Table 2). Per-dataset skill was similarly low, with SOLA performing best (R^2 = 0.22), whereas VIDA and HOTMIX showed lower point estimates (R^2 = 0.03 and 0.05, respectively; Figure 3F). These per-dataset estimates have wide bootstrap intervals for the smaller datasets (Supplementary Table 3). In particular, the apparent difference between LODO-CV (R^2 = 0.23) and spatiotemporal CV (R^2 = 0.05) for HOTMIX is not interpreted further, as the sample size (n = 19) is too small to support a reliable comparison. The K-fold and spatially blocked R^2 intervals did not overlap (K-fold [0.28-0.61] versus LODO [0.05-0.15] and spatiotemporal [0.04-0.12]),

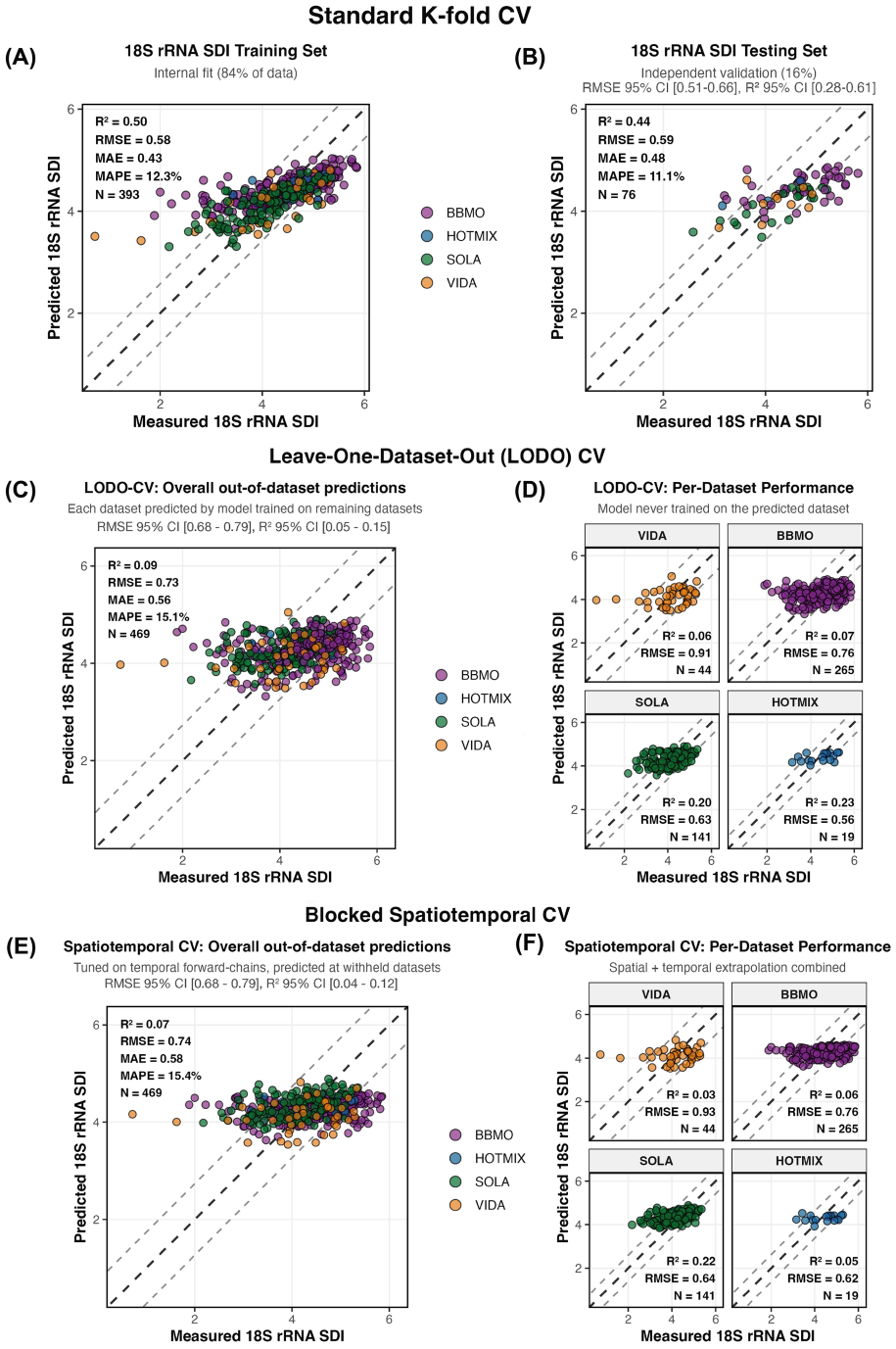

FIGURE 3
XGBoost model performance for predicting microbial eukaryotic diversity (18S rRNA SDI) under three cross-validation (CV) strategies. **(A, B)** Standard K-fold CV: training set fit (R^2 = 0.50) and independent test set validation (R^2 = 0.44). **(C, D)** Leave-One-Dataset-Out (LODO) CV: overall out-of-dataset predictions (R^2 = 0.09) and per-dataset performance for testing spatial generalizability. **(E, F)** Blocked spatiotemporal CV: overall predictions (R^2 = 0.07) and per-dataset performance combining spatial and temporal extrapolation. Dashed lines indicate the 1:1 line and ±1 SD of the prediction residuals, computed separately within each panel.

indicating that the performance gap exceeds sampling uncertainty (Figures 3B, C, E).

Across all three CV strategies, prediction errors increased progressively, while R^2 decreased from standard K-fold to LODO to spatiotemporal CV (Figure 4A), confirming that each additional blocking step yielded a more stringent estimate of generalization performance. A direct comparison between standard K-fold and LODO-CV revealed that, depending on the dataset, random partitioning underestimated prediction error, with per-dataset RMSE increasing by 0% to +82% (Figure 4B). The RMSE increase attributable to temporal blocking relative to LODO-CV was small across all datasets (-1% to +11%; Figure 4C).

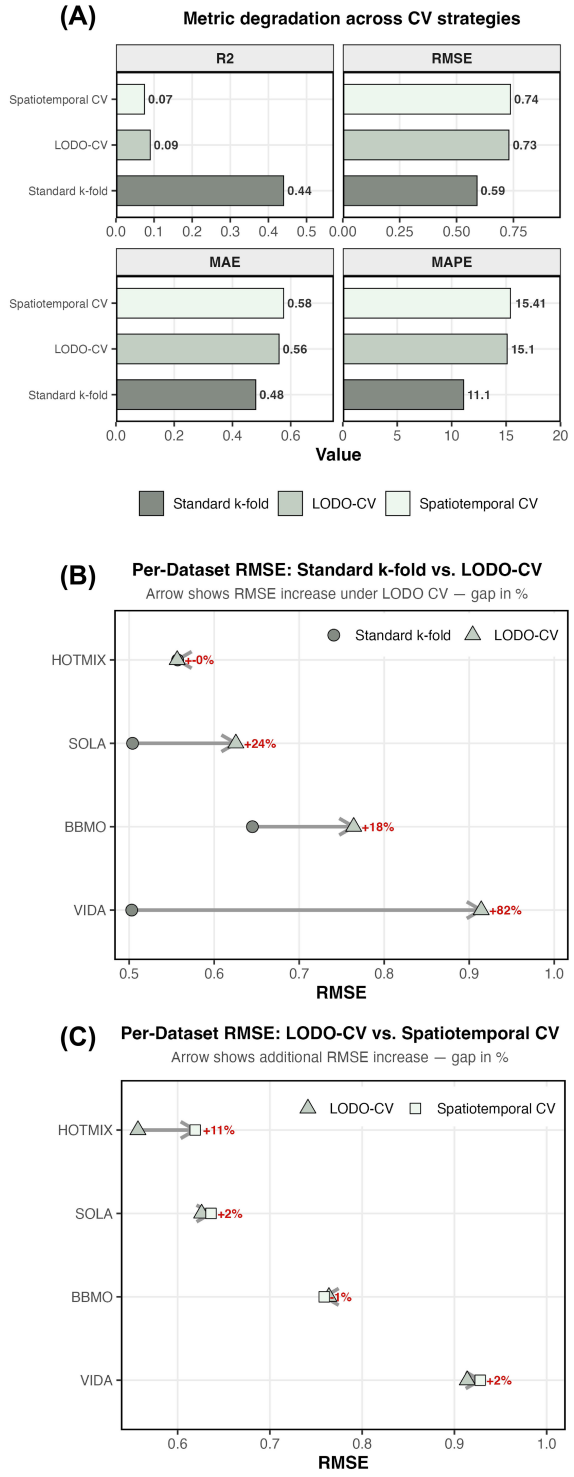

FIGURE 4
Performance degradation of the XGBoost model across the cross-validation (CV) strategies. **(A)** Overall metrics (R^2, RMSE, MAE, and MAPE) for standard K-fold, LODO-CV, and spatiotemporal CV, showing marked declines in R^2 and increases in error from standard to more stringent CV schemes. **(B)** Per-dataset RMSE increases from standard K-fold to LODO-CV. **(C)** Additional RMSE increase from LODO-CV to spatiotemporal CV, showing modest further degradation across datasets.

4 Discussion and conclusion

XGBoost has shown considerable potential for predictive modeling in marine ecological applications (Marchese et al., 2026). However, its performance may be overstated when evaluated solely using standard CVs in the presence of spatially uneven observations (Stock, 2022; Misiuk and Brown, 2023). This overestimation of apparent skill in spatially structured data is well documented (Roberts et al., 2017; Meyer et al., 2019; Ploton et al., 2020). We demonstrate its consequences for marine eukaryotic

omics, where leakage-aware evaluation is not yet routine, and show that transfer can fail even across environmentally similar datasets. In our analysis, standard K-fold CV based on stratified random partitioning substantially overestimated the ML model's accuracy in predicting eukaryotic plankton diversity when applied to the unbalanced merged dataset, yielding an R^2 of 0.44. However, the R^2 value decreased to 0.09 under LODO-CV, while the RMSE increased from 0.59 to 0.73, a 24% increase. The apparent model performance therefore reflected a failure of cross-dataset transfer rather than genuine predictive generalization. These findings are consistent with growing evidence that random partitioning does not produce truly independent test sets when spatial imbalance causes similar samples to occur in both the training and test sets, leading to overly optimistic estimates of model performance (Pohjankukka et al., 2017; Roberts et al., 2017; Gray et al., 2024). In our dataset, both training and test sets were dominated by BBMO and SOLA, resulting in substantial spatial overlap between partitions and limiting what held-out evaluations could reveal about model performance on the minority datasets. CV based on random splitting evaluates a model's ability to reproduce the training data rather than its spatial prediction skill (Meyer et al., 2019).

Per-dataset LODO-CV performance was consistently low, with R^2 values ranging from 0.06 (VIDA) to 0.23 (HOTMIX), indicating two distinct failure modes. VIDA occupies a distinct region at the edge of the predictor space and was processed using a different size-fraction protocol (see Supplementary Text S1), so it partly samples a different organismal size range rather than only a different environment. HOTMIX, with a small sample size, provides limited information during training and leaves open-sea conditions unrepresented in the held-out set. Both datasets likely fall outside the environmental range well represented by the training data (Meyer and Pebesma, 2021). Per-dataset R^2 cannot be reliably estimated for the small datasets (i.e., VIDA and HOTMIX; Supplementary Table 3), so per-dataset RMSE values and the overall pattern of poor transferability are more informative than individual R^2 point estimates. In contrast, although BBMO and SOLA overlapped in predictor space, they exhibited the largest pairwise difference in SDI distributions (Cohen's $d = 0.57$, $p < 0.001$), and cross-dataset predictions between them remained poor. These results indicate that overlap in predictor space alone is insufficient to ensure transferability. Dataset-specific factors, including variation in sampling protocols, community composition, or local processes not captured by the environmental predictors, may also limit transferability. Such methodological heterogeneity (Pascoal et al., 2023) is a plausible but unproven contributor. This interpretation is consistent with evidence that plankton community structure is shaped by both abiotic gradients and biotic interactions, suggesting that environmental predictors alone may not fully explain observed community patterns (Lima-Mendez et al., 2015). As Gray et al. (2024) observed, ML models are fundamentally interpolators; overlap in the predictor space does not ensure transferability beyond the training distribution. The BBMO-SOLA comparison is consistent with this interpretation because strong environmental overlap coincides with poor mutual transferability, suggesting that dataset-specific technical effects may be more important than environmental differences, although unmeasured local processes could also contribute. The modest dataset sample size and the strong imbalance toward BBMO and SOLA may further limit transferability by biasing model tuning toward the dominant datasets and increasing uncertainty for small held-out datasets such as HOTMIX. Importantly, dataset identity is inseparable from technical factors such as size fraction, extraction method, primers, sequencing platform, and filtered volume. Consequently, our results demonstrate that cross-dataset transferability is poor, but we were unable to resolve the relative contributions of sample size, environmental mismatch, and technical differences.

Blocked spatiotemporal CV introduced only minor additional error relative to LODO-CV (~1% in overall RMSE), indicating that performance degradation was driven primarily by cross-dataset transfer failure rather than temporal leakage, although the two effects are difficult to disentangle because temporal overlap among datasets is limited. We interpret the observed change in HOTMIX between the LODO and spatiotemporal CVs ($R^2 = 0.23$ versus 0.05) with caution, given the extremely wide bootstrap intervals (Supplementary Table 3). Since the final model in each fold was trained on all remaining datasets regardless of temporal chaining, this difference likely reflects variation in tuned hyperparameters rather than a temporal effect on HOTMIX itself. Similar patterns have been documented across aquatic, marine, and terrestrial ecosystems. Bauknecht et al. (2025) reported reduced performance when fish biodiversity Random Forest models were transferred to previously unseen rivers. Similarly, Stock (2022) showed that random hold-out and 10-fold CV underestimated prediction error in marine remote-sensing applications when observations were spatially and temporally structured, and Ploton et al. (2020) reported comparable overestimation in terrestrial biomass mapping. Our findings thus suggest that cross-dataset transfer reflects a combination of incomplete environmental coverage and technical differences across independently generated datasets, which this design cannot disentangle. This limitation persists regardless of model type or complexity, reflecting the training distribution rather than the learning algorithm. Standard K-fold CV is insufficient for assessing spatial generalizability when *in situ* sampling is unevenly distributed. In such cases, future studies on marine biodiversity prediction should report both standard and spatially blocked performance metrics to explicitly assess the extent of spatial leakage and facilitate direct comparisons across validation strategies (Pohjankukka et al., 2017; Roberts et al., 2017; Meyer et al., 2019; Valavi et al., 2019). The differences among the metrics presented here provide a concrete example of the overestimation that can result from ignoring spatial structure during evaluation.

Overall, under pronounced sampling imbalance, standard K-fold CV substantially overestimates cross-dataset generalizability. Achieving more uniform spatial coverage and employing predictors that better capture underlying biological variability would likely lead to convergence in performance estimates across CV strategies, thereby reflecting genuine predictive generalization rather than validation bias. More broadly, basin-scale diversity maps based on random-CV skill estimates should be interpreted as optimistic when training data are spatially uneven and are most robust when paired with an area-of-applicability assessment (Meyer and Pebesma, 2021), which identifies where predictions are reliable. For instance,

the prokaryotic mapping of Marchese et al. (2026) used random K-fold CV but included such an assessment, partly offsetting the limitation identified here. While extending point-specific observations to basin-scale predictions of eukaryotic microbial diversity in the Mediterranean Sea is both ecologically relevant and technically feasible, our findings highlight concerns regarding the current spatial distribution of sampling. Robust generalization requires two complementary measures: standardizing data collection and processing protocols (Pascoal et al., 2023) to reduce technical batch effects, and expanding the monitoring network to encompass a wider range of trophic regimes, including both productive coastal and shelf systems and oligotrophic offshore waters, to enhance representativeness. Until comprehensive coverage is achieved, spatial evaluation strategies remain crucial for preventing overestimation of model transferability and should be standard practice.

While basin-wide predictions remain limited by current sampling coverage, site-specific applications are achievable, particularly at BBMO. This time series is long enough to resolve seasonal, interannual, and long-term variability. At this site, models trained with temporally blocked CV can estimate SDI dynamics using satellite- and model-derived predictors, enabling gap filling during unsampled periods and establishing a baseline for detecting anomalous diversity events. Model performance may improve through alternative predictors or time-lagged environmental variables that account for delayed community responses to physical forcing (Edwards and Richardson, 2004). However, the goal was not to maximize predictive accuracy but to evaluate how different CV strategies reveal spatial leakage and constrain model transferability.

Data availability statement

All environmental data used in this study are publicly available from the Copernicus Marine Service (CMEMS; https://marine.copernicus.eu). The analysis code and the modeling dataset (Shannon Diversity Index values and their associated environmental matchups) are available via Zenodo (https://doi.org/10.5281/zenodo.20718853). Sample sequencing runs from the VIDA dataset are publicly available through the European Nucleotide Archive (ENA) under accession number PRJEB111742. The harmonized 18S rRNA gene dataset used to compute the diversity index is not yet publicly archived; it may be made available from the authors upon reasonable request.

Author contributions

CM: Methodology, Writing – review & editing, Software, Data curation, Conceptualization, Investigation, Writing – original draft, Formal analysis, Visualization. MZ: Methodology, Investigation, Writing – review & editing, Visualization. PR: Data curation, Validation, Writing – review & editing. TD: Data curation, Writing – review & editing, Validation. TT: Data curation, Writing – review & editing. RL: Writing – review & editing, Data curation. PG: Data curation, Writing – review & editing. EO: Project administration, Funding acquisition, Writing – review & editing, Supervision, Investigation, Resources, Conceptualization.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the PETRI-MED project (Plankton Biodiversity Through Remote Sensing and Omics in the Mediterranean Sea; CUP B53C23003650001), funded by Biodiversa+, the European Biodiversity Partnership under the 2021-2022 BiodivProtect joint call for research proposals. It was co-funded by the European Commission (GA N°101052342) and funding organizations: the Italian Ministry of University and Research (MUR), the Spanish Ministry for the Ecological Transition and the Demographic Challenge (MITECO), the Spanish Fundación Biodiversidad, the Slovenian Ministry of Education, Science, and Sport, and the French National Research Agency (ANR). This research also benefited from the data and support provided by the VIDA program, which was funded through the EU Research Infrastructure projects eLTER ERIC and LifeWatch ERIC and the Slovenian Research Agency (Research Core Funding No. P1-0237). Sequencing of the LTER-BBMO datasets was supported by project MAORI (PID2022-136281NB-I00, MICINN, Spain) to R.L. The SOLA work was supported by funding from the French Agence Nationale de la Recherche (ANR) Seasoning project (ANR-24-CE02-7681) to P.E.G.

Acknowledgments

We thank Ramon Massana, Vanessa Balagué, and Clara Cardelús for their contributions to the generation of the LTER BBMO 18S rRNA gene dataset. We also thank the scientists and crew aboard the R/V Sarmiento de Gamboa for collecting and sequencing DNA samples during the HOTMIX expeditions, and the teams from other platforms for collecting and curating the 18S rRNA gene sequencing data used in this study. We would like to thank the MARBITS platform of the Institut de Ciencies del Mar (ICM; http://marbits.icm.csic.es) and the DRAGO computing cluster of the CSIC for supporting bioinformatics analyses. We thank the two reviewers for their thoughtful and constructive comments, which have substantially improved the quality of this manuscript.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author EO declared that they were aneditorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2026.1875929/full#supplementary-material

References

Arrigo, K. R. (2005). Marine microorganisms and global nutrient cycles. *Nature* 437, 349–355. doi: 10.1038/nature04159

Bauknecht, R., Pellissier, L., Brosse, S., Prié, V., Lopes-Lima, M., Beja, P., et al. (2025). Combining environmental DNA and remote sensing variables to model fish biodiversity in tropical river ecosystems. *Ecol. Inf.* 90, 103251. doi: 10.1016/j.ecoinf.2025.103251

Bax, N. J., Miloslavich, P., Muller-Karger, F. E., Allain, V., Appeltans, W., Batten, S. D., et al. (2019). A response to scientific and societal needs for marine biological observations. *Front. Mar. Sci.* 6, 395. doi: 10.3389/fmars.2019.00395

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., et al. (2012). Biodiversity loss and its impact on humanity. *Nature* 486, 59–67. doi: 10.1038/nature11148

Cavicchioli, R., Ripple, W. J., Timmis, K. N., Azam, F., Bakken, L. R., Baylis, M., et al. (2019). Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* 17, 569–586. doi: 10.1038/s41579-019-0222-5

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system", in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco California USA: ACM), 785–794. doi: 10.1145/2939672.2939785

Claustre, H., Johnson, K. S., and Takeshita, Y. (2020). Observing the global ocean with Biogeochemical-Argo. *Annu. Rev. Mar. Sci.* 12, annurev-marine-010419-010956. doi: 10.1146/annurev-marine-010419-010956

Cossarini, G., Feudale, L., Teruzzi, A., Bolzon, G., Coidessa, G., Solidoro, C., et al. (2021). High-resolution reanalysis of the Mediterranean Sea biogeochemistry, (1999–2019). *Front. Mar. Sci.* 8, 741486. doi: 10.3389/fmars.2021.741486

Crisci, C., Ghattas, B., and Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Modell.* 240, 113–122. doi: 10.1016/j.ecolmodel.2012.03.001

Danovaro, R. (2024). Understanding marine biodiversity patterns and drivers: the fall of Icarus. *Mar. Ecol.* 00, e12814. doi: 10.1111/maec.12814

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* 348, 1261605–1261605. doi: 10.1126/science.1261605

Duffy, J. E., Cardinale, B. J., France, K. E., McIntyre, P. B., Thébault, E., and Loreau, M. (2007). The functional role of biodiversity in ecosystems: incorporating trophic complexity. *Ecol. Lett.* 10, 522–538. doi: 10.1111/j.1461-0248.2007.01037.x

Edwards, M., and Richardson, A. J. (2004). Impact of climate change on marine pelagic phenology and trophic mismatch. *Nature* 430, 881–884. doi: 10.1038/nature02808

El Hourany, R., Pierella Karlusich, J., Zinger, L., Loisel, H., Levy, M., and Bowler, C. (2024). Linking satellites to genes with machine learning to estimate phytoplankton community structure from space. *Ocean Sci.* 20, 217–239. doi: 10.5194/os-20-217-2024

Escudier, R., Clementi, E., Cipollone, A., Pistoia, J., Drudi, M., Grandi, A., et al. (2021). A high resolution reanalysis for the Mediterranean Sea. *Front. Earth Sci.* 9, 702285. doi: 10.3389/feart.2021.702285

Feranchuk, S., Belkova, N., Potapova, U., Kuzmin, D., and Belikov, S. (2018). Evaluating the use of diversity indices to distinguish between microbial communities with different traits. *Res. Microbiol.* 169, 254–261. doi: 10.1016/j.resmic.2018.03.004

Forsythe, W. C., Rykiel, E. J., Stahl, R. S., Wu, H., and Schoolfield, R. M. (1995). A model comparison for daylength as a function of latitude and day of year. *Ecol. Modell.* 80, 87–95. doi: 10.1016/0304-3800(94)00034-F

Goodwin, M., Halvorsen, K. T., Jiao, L., Knausgård, K. M., Martin, A. H., Moyano, M., et al. (2022). Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. *ICES J. Mar. Sci.* 79, 319–336. doi: 10.1093/icesjms/fsab255

Gray, P. C., Boss, E., Prochaska, J. X., Kerner, H., Demeaux, C. B., and Lehahn, Y. (2024). The promise and pitfalls of machine learning in ocean remote sensing. *Oceanography* 37, 52–63. doi: 10.5670/oceanog.2024.511

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., and Weitz, J. S. (2013). Robust estimation of microbial diversity in theory and in practice. *ISME J.* 7, 1092–1101. doi: 10.1038/ismej.2013.10

Hays, G., Richardson, A., and Robinson, C. (2005). Climate change and marine plankton. *Trends Ecol. Evol.* 20, 337–344. doi: 10.1016/j.tree.2005.03.004

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., et al. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell* 179, 1084–1097.e21. doi: 10.1016/j.cell.2019.10.008

Kim, D., Lee, K., Jeong, S., Song, M., Kim, B., Park, J., et al. (2024). Real-time chlorophyll-a forecasting using machine learning framework with dimension reduction and hyperspectral data. *Environ. Res.* 262, 119823. doi: 10.1016/j.envres.2024.119823

Lehodey, P., Conchon, A., Senina, I., Domokos, R., Calmettes, B., Jouanno, J., et al. (2015). Optimization of a micronekton model with acoustic data. *ICES J. Mar. Sci.* 72, 1399–1412. doi: 10.1093/icesjms/fsu233

Lehodey, P., Murtugudde, R., and Senina, I. (2010). Bridging the gap from ocean models to population dynamics of large marine predators: a model of mid-trophic functional groups. *Prog. Oceanogr.* 84, 69–84. doi: 10.1016/j.pocean.2009.09.008

Lévy, M., Haëck, C., Mangolte, I., Cassianides, A., and El Hourany, R. (2025). Shift in phytoplankton community composition over fronts. *Commun. Earth Environ.* 6, 591. doi: 10.1038/s43247-025-02553-1

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., et al. (2015). Determinants of community structure in the global plankton interactome. *Science* 348, 1262073. doi: 10.1126/science.1262073

Longhurst, A. (2007). *Ecological Geography of the Sea. 2nd edn* (Burlington, MA: Academic Press).

Lou, R., Lv, Z., Dang, S., Su, T., and Li, X. (2023). Application of machine learning in ocean data. *Multimedia Syst.* 29, 1815–1824. doi: 10.1007/s00530-020-00733-x

Malde, K., Handegard, N. O., Eikvil, L., and Salberg, A.-B. (2020). Machine intelligence and the data-driven future of marine science. *ICES J. Mar. Sci.* 77, 1274–1285. doi: 10.1093/icesjms/fsz057

Marchese, C., Zoffoli, M. L., Ramond, P., Logares, R., Bouget, F.-Y., Galand, P. E., et al. (2026). Machine learning-driven mapping of prokaryotic community diversity in the Mediterranean Sea using omics, earth observation, and model data. *Ecol. Inf.* 95, 103747. doi: 10.1016/j.ecoinf.2026.103747

Meyer, H., and Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* 12, 1620–1633. doi: 10.1111/2041-210X.13650

Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. *Ecol. Modell.* 411, 108815. doi: 10.1016/j.ecolmodel.2019.108815

Miloslavich, P., Bax, N. J., Simmons, S. E., Klein, E., Appeltans, W., Aburto-Oropeza, O., et al. (2018). Essential ocean variables for global sustained observations of

biodiversity and ecosystem changes. *Global Change Biol.* 24, 2416–2433. doi: 10.1111/gcb.14108

Misiuk, B., and Brown, C. J. (2023). Improved environmental mapping and validation using bagging models with spatially clustered data. *Ecol. Inf.* 77, 102181. doi: 10.1016/j.ecoinf.2023.102181

Mitra, B., Tiwari, S. P., Uddin, M. S., Mahmud, K., and Rahman, S. M. (2024). Decision tree ensemble with Bayesian optimization to predict the spatial dynamics of chlorophyll-a concentration: a case study in Bay of Bengal. *Mar. pollut. Bull.* 199, 115945. doi: 10.1016/j.marpolbul.2023.115945

Pascoal, F., Tomasino, M. P., Piredda, R., Quero, G. M., Torgo, L., Poulain, J., et al. (2023). Inter-comparison of marine microbiome sampling protocols. *ISME Commun.* 3, 84. doi: 10.1038/s43705-023-00278-w

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* 11, 4540. doi: 10.1038/s41467-020-18321-y

Pohjankukka, J., Pahikkala, T., Nevalainen, P., and Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *Int. J. Geogr. Inf. Sci.* 31, 2001–2019. doi: 10.1080/13658816.2017.1346255

R Core Team (2024). *R: A Language and Environment for Statistical Computing_* (Vienna, Austria: R Foundation for Statistical Computing). Available online at: https://www.R-project.org/ (Accessed October 31, 2024).

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 913–929. doi: 10.1111/ecog.02881

Rubbens, P., Brodie, S., Cordier, T., Destro Barcellos, D., Devos, P., Fernandes-Salvador, J. A., et al. (2023). Machine learning in marine ecology: an overview of techniques and applications. *ICES J. Mar. Sci.* 80, 1829–1853. doi: 10.1093/icesjms/fsad100

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Sommeria-Klein, G., Watteaux, R., Ibarbalz, F. M., Pierella Karlusich, J. J., Iudicone, D., Bowler, C., et al. (2021). Global drivers of eukaryotic plankton biogeography in the sunlit ocean. *Science* 374, 594–599. doi: 10.1126/science.abb3717

Song, J., Jiang, W., Xin, L., and Zhang, X. (2024). Predicting the temporal-spatial distribution of chlorophyll-a in the Yellow River estuary using explainable machine learning. *Estuar. Coast. Shelf Sci.* 304, 108820. doi: 10.1016/j.ecss.2024.108820

Stock, A. (2022). Spatiotemporal distribution of labeled data can bias the validation and selection of supervised learning algorithms: a marine remote sensing example. *ISPRS J. Photogramm. Remote Sens.* 187, 46–60. doi: 10.1016/j.isprsjprs.2022.02.023

Testor, P., De Young, B., Rudnick, D. L., Glenn, S., Hayes, D., Lee, C. M., et al. (2019). OceanGliders: a component of the integrated GOOS. *Front. Mar. Sci.* 6, 422. doi: 10.3389/fmars.2019.00422

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2019). block CV: an r package for generating spatially or environmentally separated folds for k -fold cross-validation of species distribution models. *Methods Ecol. Evol.* 10, 225–232. doi: 10.1111/2041-210X.13107

Volpe, G., Colella, S., Brando, V. E., Forneris, V., La Padula, F., Di Cicco, A., et al. (2019). Mediterranean ocean colour Level 3 operational multi-sensor processing. *Ocean Sci.* 15, 127–146. doi: 10.5194/os-15-127-2019

Worm, B., Barbier, E. B., Beaumont, N., Duffy, J. E., Folke, C., Halpern, B. S., et al. (2006). Impacts of biodiversity loss on ocean ecosystem services. *Science* 314, 787–790. doi: 10.1126/science.1132294