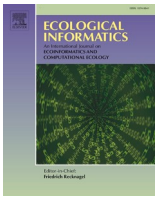
Ecological Informatics 95 (2026) 103747

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Machine learning-driven mapping of prokaryotic community diversity in the Mediterranean Sea using omics, earth observation, and model data

Christian Marchese [a,*], Maria Laura Zoffoli [b], Pierre Ramond [c,d], Ramiro Logares [c], François-Yves Bouget [e], Pierre E. Galand [d], Tinkara Tinta [f], Neža Orel [f], Gianluca Volpe [a], Angela Landolfi [a], Emanuele Organelli [a]

[a] National Research Council (CNR), Institute of Marine Sciences (ISMAR), Rome, Italy
[b] National Research Council (CNR), Institute of Marine Sciences (ISMAR), Trieste, Italy
[c] Institute of Marine Sciences (ICM), CSIC, Barcelona, Spain
[d] Sorbonne Université, CNRS, Laboratoire d'Ecogéochimie des Environnements Benthiques (LECOB), Observatoire Océanologique de Banyuls, Banyuls-sur-Mer, France
[e] Sorbonne Université, CNRS, Laboratoire d'Océanographie Microbienne (LOMIC), Observatoire Océanologique de Banyuls, Banyuls-sur-Mer, France
[f] Marine Biology Station Piran, National Institute of Biology (NIB), Piran, Slovenia

ARTICLE INFO

Keywords:
Prokaryotic community diversity
Machine learning
Mediterranean Sea
Omics
Shannon diversity index
Remote sensing

ABSTRACT

Marine prokaryotic communities are major contributors to oceanic food webs and global biogeochemical cycles. However, basin-scale diversity patterns and environmental drivers remain poorly understood. In this study, we applied a machine-learning framework to model the diversity of marine prokaryotic communities across the Mediterranean Sea. Diversity was quantified using the Shannon Diversity Index (SDI) derived from 16S rRNA gene sequencing. The *in situ* dataset included ~600 samples collected year-round from 2001 to 2023 at coastal and open-water sites, providing broad temporal coverage and multisite spatial sampling. We trained an XGBoost model using satellite-derived and modeled oceanographic variables matched to the SDI observations. The model achieved robust predictive performance (R^2 = 0.78 for training and 0.70 for testing, with RMSE = 0.31 and MAPE = 0.05 across both) and captured broad basin spatial and seasonal patterns in prokaryotic community diversity, with greater uncertainty in less-represented regions. Diversity was highest in nutrient-rich coastal areas and during winter mixing, and lowest in summer-stratified or oligotrophic waters. SHAP analysis identified photoperiod as the most significant predictor, underscoring the central role of seasonal light cycles in shaping prokaryotic community diversity. Other predictors exhibited significant season- and region-dependent effects, each contributing positively within specific environmental thresholds. Climatological diversity maps revealed consistent spatiotemporal patterns, highlighting a notable west-to-east decrease in diversity and coastal hotspots. These results demonstrate that machine learning can identify major environmental drivers of prokaryotic diversity and upscale discrete observations to basin-wide predictions. This approach is transferable to other planktonic groups and supports scalable ecosystem monitoring across environmental gradients.

1. Introduction

Approximately 71% of the Earth's surface is covered by the ocean, which strongly influences the global climate system (Bigg et al., 2003; Hays et al., 2005) and sustains vital ecosystem services and human well-being through its rich biodiversity (Marchese, 2015). Marine ecosystems are underpinned by a vast diversity of microscopic plankton, which, despite their tiny size, play a fundamental role in ocean functioning and act as sentinels of climate change through shifts in their composition and phenology (Hays et al., 2005; Ibarbalz et al., 2019). Plankton include photosynthetic microorganisms, both prokaryotic (e.g., cyanobacteria) and eukaryotic (e.g., diatoms and dinoflagellates), along with heterotrophic prokaryotes (bacteria and archaea) and zooplankton, all of which support major biogeochemical cycles (Falkowski et al., 2008). The photosynthetic fraction of plankton accounts for approximately half of global net primary production (Behrenfeld et al., 2006) and drives the biological carbon pump by transferring carbon from the surface to the deep ocean (Boyd et al., 2019). A substantial fraction of

* Corresponding author.
E-mail address: christianmarchese@cnr.it (C. Marchese).

https://doi.org/10.1016/j.ecoinf.2026.103747
Received 14 October 2025; Received in revised form 27 March 2026; Accepted 28 March 2026
Available online 2 April 2026

photosynthetically fixed carbon is released into the water column as dissolved organic matter (DOM), predominantly in the form of dissolved organic carbon (DOC), through biological processes such as excretion, exudation, and cell lysis. This DOM fuels heterotrophic prokaryotes, which represent approximately 30% of the water-column biomass and, on average, respire more than half of the surface net primary production (Heneghan et al., 2024). Heterotrophic prokaryotes subsequently remineralize DOM into inorganic nutrients, supporting new primary production, while simultaneously transforming labile organic carbon into more refractory forms that contribute to long-term carbon storage (Jiao et al., 2024). In addition, the assimilation of DOM into prokaryotic biomass enables the transfer of carbon and energy back into the food web when consumed by heterotrophic protists (Pomeroy et al., 2007). Through these tightly coupled processes, phytoplankton-derived DOM fuels the microbial loop, whereby heterotrophic prokaryotes and heterotrophic protists redirect DOM into the food web, supporting nutrient recycling and carbon transfer within marine ecosystems (Bunse and Pinhassi, 2017; Carlson et al., 2007; Giering et al., 2014; Heneghan et al., 2024; Pomeroy et al., 2007).

Variations in primary production alter the quantity and composition of organic matter released into seawater, thereby shaping prokaryotic community structure and function. Hence, understanding the diversity of eukaryotic phytoplankton and prokaryotic communities is essential for studying marine biogeochemical processes (Thompson et al., 2011). Prokaryotic community diversity, succession, and distribution are also strongly influenced by environmental factors, such as temperature, nutrient availability, light, and water-column stratification and mixing (Behrenfeld et al., 2008; Bunse and Pinhassi, 2017; Falkowski and Oliver, 2007), which exhibit pronounced seasonal variability in temperate marine systems (Junger et al., 2026; Lambert et al., 2019). Under these conditions, vertical mixing is a key physical driver of these dynamics. During winter, deep mixing brings nutrient-rich water to the surface, potentially enhancing diversity by disrupting summer oligotrophic conditions. However, winter mixing can also promote phytoplankton blooms (often dominated by eukaryotic taxa), leading to the temporary dominance of a few taxa and lower evenness. In spring, bloom-driven conditions often favor specialized microbial communities, whereas strong summer stratification promotes oligotrophy and is typically associated with reduced diversity. Autumn re-mixing can partially restore nutrient levels and community complexity. Together, these seasonal processes drive recurrent restructuring of prokaryotic communities in temperate regions (Bunse and Pinhassi, 2017; Teeling et al., 2012), while functional redundancy can buffer biogeochemical processes against this taxonomic turnover (Beauvais et al., 2023).

Advances in omics are increasingly enabling a quantitative understanding of biogeochemical processes driven by marine microbial communities (Levine et al., 2025). Marine metabarcoding studies have largely relied on ribosomal RNA gene markers (e.g., 16S/18S; Levine et al., 2025) to resolve plankton community composition and diversity, primarily in relative terms, and to track temporal changes in assemblage structure (Auladell et al., 2022; Celussi et al., 2024; Lambert et al., 2021; Yeh and Fuhrman, 2022). Such studies have identified recurring archaeal, bacterial, and eukaryotic taxa across years (Lambert et al., 2019), revealed potential biotic interactions (Deutschmann et al., 2023), and documented the long-term stability of microbial ecosystems under variable environmental conditions (Lambert et al., 2021). Beyond taxonomic patterns, microbial diversity is increasingly interpreted through the lens of functional diversity and redundancy, which modulate ecosystem efficiency, resistance, and resilience to environmental change (Ramond et al., 2025). Recent mesopelagic observations further show that diversity-function relationships are context-dependent, varying with microbial lifestyle and particle association, and reveal distinct roles for particle-associated versus free-living communities in regulating carbon cycling (Baumas et al., 2021). Accordingly, microbial diversity provides a useful proxy for evaluating ecosystem resilience and detecting climate-driven shifts in pelagic biodiversity and food web structure (Batten et al., 2019; Holland et al., 2025). Microbial diversity is often measured using ecological indices that integrate taxonomic richness and evenness, such as the Shannon Diversity Index (Feranchuk et al., 2018; Haegeman et al., 2013). For instance, in the northwestern Mediterranean Sea, Shannon diversity shows marked seasonal variability, with winter maxima (Lambert et al., 2019), and pronounced spatial gradients across basins and depth layers (Junger et al., 2023). However, these findings are largely based on point-based observations, which limit basin-scale inferences. To address this limitation, remote sensing products provide continuous observations of physical and biogeochemical proxies, enabling the analysis of microbial diversity over broader spatial and temporal scales.

Satellite observations provide comprehensive and frequent coverage of essential ocean variables (Miloslavich et al., 2018) at high temporal resolution over large scales. Examples include sea surface temperature, chlorophyll-a (Chl-a) concentration, and various physical and ocean color parameters (Polovina and Howell, 2005; Racault et al., 2014). In addition, model and reanalysis products, such as those from coupled physical-biogeochemical models, provide spatially and temporally continuous estimates of ocean state variables (Storto et al., 2019) that can be combined with satellite data. However, extrapolating valuable information from extensive datasets requires advanced algorithms to model the complex relationships between environmental factors and plankton dynamics. Machine learning (ML) models can combine satellite data, model outputs, and *in situ* observations to estimate key climate and biodiversity variables across large spatial and temporal scales (Hollmann et al., 2013; Kissling et al., 2018). This approach facilitates the creation of continuous fields that are particularly valuable for monitoring temporal changes, phenological patterns, and oceanic biodiversity (Rubbens et al., 2023). ML algorithms have been successfully used in diverse marine applications, such as retrieving the Chl-a concentration (Zoffoli et al., 2025), estimating the relative abundance of phytoplankton groups using remote sensing data combined with omics-based information (El Hourany et al., 2024), measuring dissolved organic carbon (Panaïotis et al., 2025), and in other areas such as atmospheric correction, carbon cycle analysis, and data reconstruction (Zhang et al., 2025). Therefore, using ML to combine omics datasets with synoptic environmental fields enables the generation of gap-filled maps and scenario projections of microbial diversity patterns, thereby supporting basin-scale monitoring.

Building on this potential, we developed an ML framework to estimate prokaryotic community diversity in the Mediterranean Sea using omics, remote sensing, and model-derived ocean parameters. Although the Mediterranean Sea is recognized for its high biodiversity (Coll et al., 2010), it is increasingly exposed to multiple stressors, including invasive species, habitat loss, and overfishing, with cascading effects on ecosystem services (Fasola et al., 2025). In addition, increasingly intense and frequent marine heatwave events are strengthening surface layer stratification (Marullo et al., 2023). These alterations may lead to a decline in plankton biomass, shifts in community composition and phenology, and ultimately affect higher trophic levels (El Hourany et al., 2021; Li et al., 2024), making the Mediterranean Sea both a particularly vulnerable ecosystem and a natural laboratory for studying plankton diversity and dynamics.

Specifically, this study aimed to (1) implement an ML pipeline that integrates satellite and model-derived oceanographic environmental variables with *in situ* prokaryotic community diversity indexed by 16S rRNA gene sequencing and quantified by the Shannon Diversity Index (SDI); (2) assess the influence of environmental predictors using SHapley Additive exPlanations (SHAP) analysis; and (3) generate annual and seasonal average maps to reveal regional patterns across the Mediterranean Sea and demonstrate the applicability of the algorithm. Using this approach, we mapped prokaryotic community diversity from daily to annual timescales. The resulting diversity maps, along with an evaluation of the importance of environmental predictors, revealed how basin-scale physical and biogeochemical gradients shape the

spatiotemporal patterns of prokaryotic community diversity, providing valuable support for biodiversity monitoring and management in a rapidly changing marine environment.

2. Materials and methods

2.1. Study area and sampling efforts

The Mediterranean Sea is a semi-enclosed basin characterized by a subtropical-like regime, marked by low variability in phytoplankton biomass throughout the year. It exhibits a strong west-to-east decline in nutrient concentrations and productivity (Barale et al., 2008). A secondary gradient in primary production is also evident from coastal to open waters. Primary production is generally higher in neritic areas due to terrestrial inputs and nutrient enrichment, and decreases toward offshore regions (Marchese et al., 2015). Temperate patterns can be observed in specific regions, with pronounced spring productivity peaks followed by very low values during the summer months (D'Ortenzio and Ribera d'Alcalà, 2009; Lavigne et al., 2013). Other areas experience intermittent or minor bloom events depending on local conditions. Most of the eastern and central Mediterranean basins exhibit non-blooming dynamics with slight seasonal variability, reflecting their ultra-oligotrophic regime (D'Ortenzio and Ribera d'Alcalà, 2009; Kotta and Kitsiou, 2019). Beyond its well-documented phytoplankton seasonality and trophic gradients, the Mediterranean Sea also exhibits distinct prokaryotic community dynamics across the basin, with pronounced seasonal changes in community composition driven by nutrient availability and physical mixing across coastal and offshore areas (Celussi et al., 2024; Ferrera et al., 2024; Pinhassi et al., 2006).

We computed the SDI based on 16S rRNA gene sequencing of *in situ* marine prokaryotic communities sampled at five fixed time-series stations and two oceanographic cruises across the Mediterranean Sea (Fig. 1). Throughout the manuscript, the term *prokaryotic community* refers to bacteria and archaea, including cyanobacteria, as inferred from 16S rRNA gene sequencing. Water samples were collected from the surface (0–3 m depth). The *in situ* database comprised approximately 600 samples spanning the period 2001–2023. Sampling was conducted year-round (Fig. 2), yielding a well-distributed temporal dataset; the main sites and cruises are summarized in Table 1. Fixed coastal stations were sampled weekly or monthly, depending on the site. In the northwestern Mediterranean Sea, the sampling sites included the Blanes Bay Microbial Observatory (BBMO; Gasol et al., 2016), Microbial Observatory Laboratory Arago (MOLA), and Service d'Observation du Laboratoire Arago (SOLA). The BBMO station (41°40′ N; 2°48′ E) is in the Bay of Blanes (Spain) and has been sampled monthly since 2001. SOLA (42°29′ N; 3°08′ E), located in the Bay of Banyuls-sur-Mer, has been sampled weekly since 1997, whereas MOLA (42°27′ N; 3°32′ E), situated ~37 km offshore on the continental slope, has been sampled monthly since 2003 (Lami et al., 2009). In the northern Adriatic Sea, samples were collected at the Marine Biology Station Piran oceanographic buoy VIDA (45°32′ N; 13°33′ E), located ~2.3 km offshore in the southern part of the Gulf of Trieste, and sampled monthly since 2018 (Celussi et al., 2024; Tinta et al., 2015). In the Tyrrhenian Sea, the NEREA observatory has operated three stations across the Gulf of Naples since April 2019 (Campese et al., 2024; Trano et al., 2024): the LTER-MC site (40°48′ N; 14°15′ E), located ~3.7 km offshore and sampled monthly; the NRS site (40°43′ N; 14°27′ E), near the mouth of the Sarno River and sampled seasonally; and the offshore NRC site (40°36′ N; 14°08′ E), located in the Dohrn Canyon near Capri and sampled biannually. Additional samples were obtained from the HOTMIX cruise, which collected surface seawater along an east–west transect of the Mediterranean Sea from late April to late May 2014 (Sebastián et al., 2021), and from the Mediterranean expedition of the TARA Oceans program, conducted between June and October 2014 (Dussud et al., 2018). The TARA cruise provided samples from the Ligurian-Provençal, Tyrrhenian, Balearic, Algerian, and Sardinian seas.

2.2. Data inputs and processing

2.2.1. Prokaryotic community diversity estimation via in situ omics data

Seawater was sequentially filtered through 3 μm (prefilter) and 0.2 μm filters to remove larger cells and aggregates and capture the free-living microbial portion. Unless otherwise specified, DNA was extracted solely from the 0.2 μm filters following the protocols described in the studies listed below. The 16S rRNA gene fragment was amplified using the universal 515F-926R primers (Parada et al., 2016) and sequenced on Illumina. Sequence reads were processed using DADA2 to infer amplicon sequence variants (ASVs) and generate a unified ASV table (Callahan et al., 2016). Taxonomy was assigned using the SILVA v138.2 database (Quast et al., 2013), and only ASVs classified as Bacteria (including cyanobacteria) and Archaea were retained for downstream analyses. Sequencing depth was standardized by rarefying the dataset to 10,000

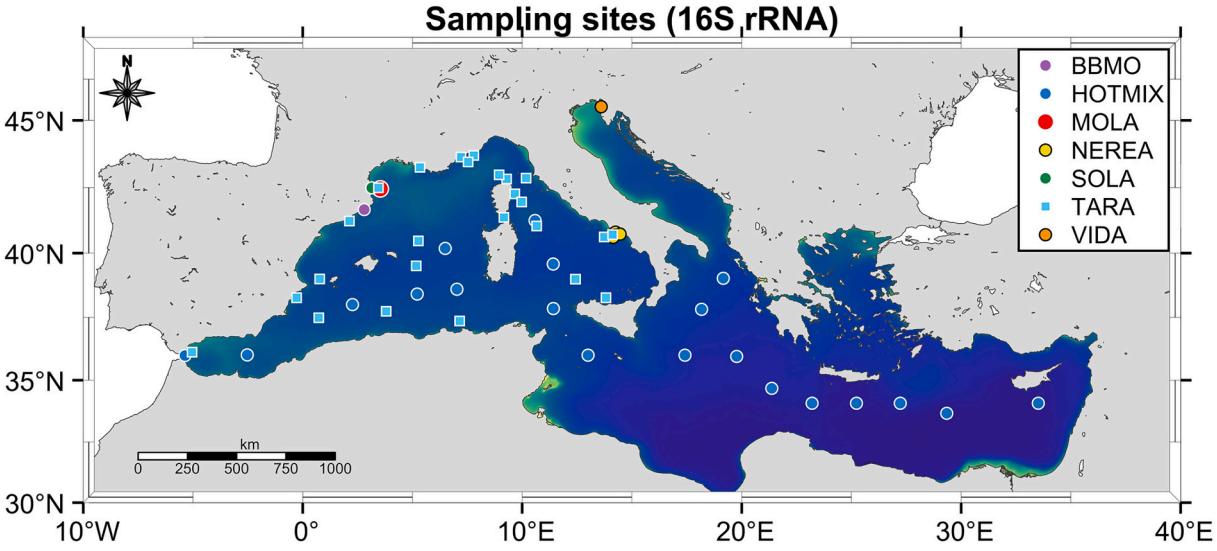


Fig. 1. Geographic distribution of sampling sites included in the prokaryotic community diversity dataset across the Mediterranean Sea. Colored points indicate the sampling stations BBMO, HOTMIX, MOLA, NEREA, SOLA, TARA, and VIDA. HOTMIX and TARA represent stations sampled during scientific cruises, whereas NEREA includes three sites within the Gulf of Naples (Italy). Data from these locations (see also Table 1) were used for training and testing the XGBoost model.

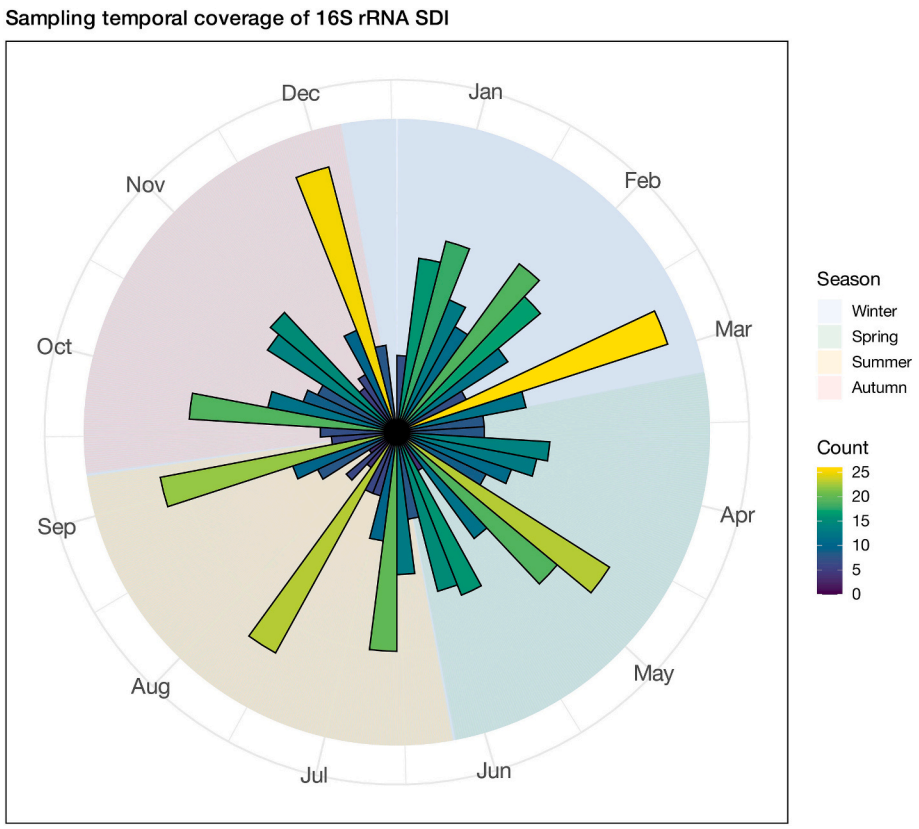


Fig. 2. Temporal sampling coverage of prokaryotic community diversity (16S rRNA gene SDI) across seasons. Radial bars represent samples collected on grouped days throughout the calendar year. Seasonal divisions are indicated by background shading, illustrating the distribution of sampling efforts during winter (JFM), spring (AMJ), summer (JAS), and autumn (OND).

Table 1
Summary of Sampling Stations and Campaigns (see also Fig. 1).

Station / Campaign	Location	Coordinates	Sampling frequency	Start year / Period
BBMO	Bay of Blanes, Spain	41°40'N-2°48'E	Monthly	2001-2023
MOLA	North side of Lacaze-Duthiers canyon, France	42°27'N-3°32'E	Monthly	2020-2023
SOLA	Banyuls-sur-Mer, France	42°29'N-3°08'E	Weekly	2008-2017
VIDA	Gulf of Trieste, Slovenia	45°32'N-13°33'E	Monthly	2018-2021
NEREA (LTER-MC)	Naples, Italy	40°48.5'N-14°15'E	Monthly	2019-2020
NEREA (NRS)	Near River Sarno mouth, Italy	40°43'N-14°27'E	Seasonal	2019-2020
NEREA (NRC)	Dohrn Canyon near Capri, Italy	40°36'N-14°08'E	Biannual	2019-2020
HOTMIX	Heraklion (Greece) to Las Palmas (Spain), across Eastern & Western Mediterranean	–	One-time transect	29 Apr-22 May 2014
TARA Mediterranean	Ligurian-Provençal, Tyrrhenian, Balearic, Algerian, Sardinian Seas	–	One-time transect	06 Jun-31 Oct 2014

reads per sample. Specific methodological details for the fixed-time series stations and sampling campaigns are provided in Lambert et al. (2021), Orel et al. (2022), Campese et al. (2024), Celussi et al. (2024), and Dussud et al. (2018).

Prokaryotic community diversity inferred from 16S rRNA gene sequencing was quantified using the SDI (Eq. 1), calculated from ASV abundance with the *vegan* package (Oksanen et al., 2025) in R. The SDI is one of the most widely applied metrics for estimating community diversity (de Vargas et al., 2015; Ibarbalz et al., 2019; Irigoien et al., 2004). It integrates both species richness and evenness (Henson et al., 2021) and, compared with richness alone, is generally less sensitive to variation in the detection of rare taxa arising from differences in sequencing depth and sampling effort (Chiarucci et al., 2011; Haegeman et al., 2013). Accordingly, we selected the SDI as the response metric because it captures complementary aspects of community diversity (Feranchuk et al., 2018) while facilitating comparisons with previous studies. It is defined as:

$$SDI = -\sum_{i=1}^{S} p_i * \ln(p_i) \quad (1)$$

where S corresponds to the total number of 16S rRNA gene ASVs observed in a given sample, and p_i is the proportion of sequences assigned to the i-th ASV. Higher SDI values indicate greater diversity, reflecting both the number of taxa and evenness of their distribution.

2.2.2. *Environmental predictors*

To investigate the environmental drivers of prokaryotic community diversity, we compiled a range of ocean color, physical, and biogeochemical predictors from publicly available reanalysis and satellite datasets provided by the Copernicus Marine Service (CMEMS). We retrieved daily data from 2001 to 2023 from three multiyear products specifically developed and validated for the Mediterranean Sea. Among the possible variables within the CMEMS products, we selected the following for modeling prokaryotic community diversity:

1) Satellite-derived Remote Sensing Reflectances (Rrs; in sr^{-1}) were obtained from the multi-year ocean-colour product *OCEANCOLOUR_MED_BGC_L3_MY_009_143*. This regional product corresponds to a merged dataset from multiple sensors (SeaWiFS, MODIS-Aqua, MERIS, VIIRS, and Sentinel-3 OLCI). Rrs are provided at 412, 443, 490, 510, 555, and 670 nm, with a 1-day and 1-km resolution, covering the entire Mediterranean Sea. To generate continuous gap-free fields, we applied the DINEOF interpolation method described by Marchese et al. (2024) to the multi-sensor Rrs product. Chlorophyll-a concentration (Chl-a; mg m^{-3}) was estimated from Rrs using a regional algorithm specifically developed and validated for the Mediterranean Sea (Volpe et al., 2019). This Chl-a algorithm considers single-pixel optical water-type membership by utilizing the water-type optical signatures found in the MedBiOp *in situ* dataset (Volpe et al., 2019). This product performed significantly better than any other satellite-derived product over the Mediterranean Sea, particularly for low Chl-a concentrations, which comprise the most open waters in the Mediterranean Sea (Volpe et al., 2019). Both Chl-a and Rrs were resampled to a 4 km spatial resolution to match the spatial grid of the physical and biogeochemical reanalysis products.
2) Nutrient (nitrate [NO_3^-], phosphate [PO_4^{3-}], and ammonium [NH_4^+]; in mmol m^{-3}) were derived from the biogeochemical reanalysis product *MEDSEA_MULTIYEAR_BGC_006_008* (Teruzzi et al., 2021a, 2021b). This gap-free product provides daily data with a spatial resolution of ~4 km and has been validated using *in situ* observations and BGC-Argo data (Cossarini et al., 2021). Here, we considered only the surface-layer nutrients. For simplicity, we refer to NO_3, PO_4, and NH_4 throughout the manuscript, although these notations do not strictly follow the standard nomenclature.
3) The water temperature (T; in °C) was obtained from the physical reanalysis product *MEDSEA_MULTIYEAR_PHY_006_004* (Escudier et al., 2020). This gap-free dataset provides daily fields with a spatial resolution of approximately 4 km (Escudier et al., 2021). It is based on hydrodynamic simulations of the Mediterranean Sea and assimilates satellite and *in situ* observations. This product was specifically developed and validated for the Mediterranean Sea to accurately represent ocean thermal conditions. In this study, we considered only temperature at the surface layer.

Each environmental variable was extracted based on the date and sampling station using a 3 × 3 pixel spatial window centered on the station, from which the mean value was calculated. In addition, we derived three additional variables from the matched environmental dataset using simple feature engineering. First, the photoperiod (in hours of daylight) was determined for each sampling date and latitude using the daylength function from the *geosphere* R package, based on the model of Forsythe et al. (1995). Second, we calculated the nitrate-to-phosphate ratio (NO_3: PO_4) using modeled nutrient data. Third, we estimated the Rrs ratio between the bands at 412 and 443 nm (Rrs_{412}:Rrs_{443}) using satellite-derived Rrs fields.

2.3. Machine learning model and workflow

2.3.1. The XGBoost model

Extreme Gradient Boosting (XGBoost; Chen and Guestrin, 2016) is an ensemble machine learning algorithm based on gradient-boosted decision trees. It builds several small decision trees and combines them to make accurate predictions. XGBoost uses a boosting method in which trees are added one after another. Each new tree focuses on the errors of the previous trees, thus steadily improving the model. The model improves by following the gradient of a loss function (e.g., Root Mean Squared Error - RMSE), which measures prediction errors (i.e., how far the predictions are from the actual values). Each new tree fits the negative gradient of the loss function (the residuals), thereby reducing the prediction error. By fitting new trees to these errors, XGBoost reduces bias and improves overall performance. XGBoost incorporates regularization to prevent overfitting and can handle missing data by learning optimal split directions (Chen and Guestrin, 2016). These features have made it a widely adopted algorithm in ecological studies (e.g., Kim et al., 2024; Mitra et al., 2024; Song et al., 2024). Furthermore, XGBoost is resilient to outliers because of its tree-based partitioning method and does not require feature scaling because splits are based on value rankings rather than absolute magnitudes. Specifically, the model was implemented in R using the *xgboost* package (Chen et al., 2025) based on the Extreme Gradient Boosting algorithm (Chen and Guestrin, 2016).

2.3.2. Modeling workflow and performance evaluation

Six environmental predictors: (1) photoperiod, (2) Chl-a, (3) NO_3:PO_4, (4) water temperature (T), (5) NH_4, and (6) Rrs_{412}:Rrs_{443}, served as inputs to our ML framework. Predictors were selected based on ecological relevance and retained only if they improved model performance during preliminary testing. Pairwise Spearman correlations among the six selected predictors did not exceed $|\rho| = 0.68$ (Fig. S1), indicating low redundancy within the predictor set. Specifically, photoperiod and temperature are well-established drivers of prokaryotic community structure (Fuhrman et al., 2006; Gilbert et al., 2012; Ibarbalz et al., 2019). Chl-a has been used as a proxy for phytoplankton biomass (Huot et al., 2007); NH_4 as an indicator of regenerated nutrient availability; NO_3:PO_4 as a measure of nutrient stoichiometry; and the band-ratio Rrs_{412}:Rrs_{443} as an optical proxy for the relative contribution of Colored Dissolved Organic Matter (CDOM; Morel and Gentili, 2009a). Geographic descriptors, such as distance from the coast and bathymetry, were not included because the selected predictors were considered to more directly represent the processes driving prokaryotic diversity and capture key coastal-offshore and east-west gradients.

We used a three-stage workflow to predict the SDI of the prokaryotic community inferred from 16S rRNA gene sequencing (Fig. 3). We split the data into 83% for training and 17% for testing (unseen data). We chose these percentages to maintain a large training set while ensuring a reliable independent test. We used a stratified split implemented with the *createDataPartition* function in the *caret* R package (Kuhn, 2008), ensuring similar SDI distributions in both datasets. After the train-test split, we addressed the imbalance in the SDI distribution, specifically the scarcity of low-diversity samples, which can negatively affect regression-based machine learning models. To mitigate this issue, we applied the K-Nearest Neighbors Over-sampling for Regression (KNNOR-Reg) method. This method creates synthetic samples through interpolation to improve model performance (Belhaouari et al., 2024; He and Garcia, 2009). Samples with an SDI of 3.5 or less in our dataset were classified as rare. For each rare point, synthetic data were created by interpolating between that point and one of its k-nearest neighbors in the predictor space. Random interpolation weights were used to create variations in the synthetic samples while keeping them within the low-diversity range. This approach increases the representation of low-SDI conditions while preserving the local structure of the data and avoiding unrealistic outliers. Importantly, stratified splitting was performed before oversampling to ensure that the SDI distribution remained comparable between the training and test sets (Fig. S2). KNNOR-Reg was applied exclusively to the training set, thereby preventing any information leakage from the test data into model development.

We tuned XGBoost using a random search on the training set (Fig. 3; *Stage 1*). Unlike grid search, which is computationally intensive, random search explores many settings efficiently, particularly when many hyperparameters are involved (Bergstra and Bengio, 2012). We defined a search space, performed random sampling, and evaluated each with 10-fold cross-validation repeated five times. We maintained the top 5% (lowest RMSE) and used the median of their values to set the final hyperparameters. Using these hyperparameters, we ran another 10-fold cross-validation (without further tuning) on the training set to assess RMSE stability and select the optimal number of boosting rounds (Fig. 3; *Stage 2*). Finally, we retrained the model from scratch on the full training

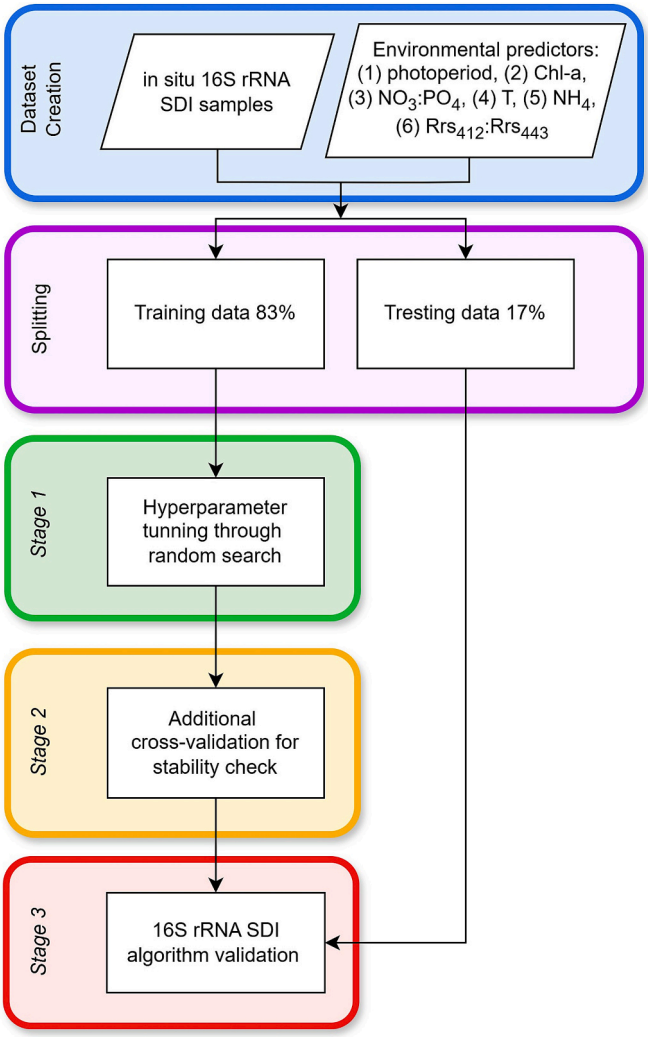


Fig. 3. Overview of the three-stage XGBoost workflow for predicting prokaryotic community diversity (16S rRNA gene SDI). The dataset was first divided into training (83%) and testing (17%) subsets. *Stage 1*, hyperparameter tuning was performed through random search within the training set. The top 5% of the configurations ranked by the lowest validation RMSE were selected, and their hyperparameters were combined using the median. In *Stage 2*, these selected hyperparameters were used for repeated 10-fold cross-validation on the training set to evaluate the stability of the RMSE and determine the optimal number of boosting rounds. *Stage 3*, the model was retrained from scratch with the entire training dataset and the chosen hyperparameters, followed by a final assessment of the untouched testing set to evaluate the predictive performance.

set using the chosen settings and rounds. We then evaluated it on the independent test set to obtain an unbiased performance estimate (Fig. 3; *Stage 3*).

Model accuracy was quantified using four metrics: Root Mean Squared Error (RMSE; Eq. 2), Mean Absolute Error (MAE, Eq. 3), Mean Absolute Percentage Error (MAPE, Eq. 4), and the coefficient of determination (R^2, Eq. 5). All metrics were computed between the predicted ($\hat{y}_i$) and actual (y_i) values, where n is the number of observations:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| (\text{excluding } y_i = 0) \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5}$$

Additionally, we used a nonparametric bootstrap approach to quantify the uncertainty in the estimates of the testing set based on paired actual and predicted values. We generated 1000 bootstrap samples (with replacement) from the testing set and calculated the RMSE and R^2 for each sample. The percentile method was applied to determine the 95% confidence intervals for both metrics. The stability of the final test RMSE and R^2 under resampling provided a reliable estimate of the prediction uncertainty. All analyses were conducted using R ver. 4.4.2 (R Core Team, 2024).

2.3.3. *Interpreting and mapping Mediterranean prokaryotic community diversity*

To interpret the model output, we conducted a SHAP analysis (Song et al., 2024; Zhu et al., 2024). The latter is a post-hoc technique that quantifies the predictor's influence, indicating its direction (positive or negative) and the strength of its effect, and clarifies whether a predictor contributes positively or negatively to the target variable. We used these contributions to rank predictor importance and interpret the mechanisms that may drive prokaryotic community SDI variability. The SHAP summary plot provides an overview of how all predictors contribute to the dataset. A SHAP dependence plot analyzes one predictor at a time, plotting the observed predictor values against the SHAP values for each sample to reveal nonlinear responses, interactions, and potential thresholds.

Finally, to create SDI spatial maps, we applied the final XGBoost model to daily georeferenced environmental data from 2021 for the six selected predictors. Model predictions were aggregated for the entire year (i.e., annual climatology) and into seasonal means: winter (January–March), spring (April–June), summer (July–September), and fall (October–December). We then used the K-means++ algorithm (Arthur and Vassilvitskii, 2007) on seasonal mean diversity data to identify five clusters that balanced spatial detail with the main gradients across the basin, labeled from *Very High Diversity* to *Very Low Diversity*.

To evaluate the spatial representativeness of the training data, we implemented an environmental distance metric adapted from the applicability domain framework of Meyer and Pebesma (2021). The six predictors were first standardized using the mean and standard deviation of the training set and then weighted by their relative importance in the XGBoost model. For each prediction grid cell, we calculated the Euclidean distance to the nearest training sample in this weighted, standardized predictor space, and scaled it to obtain a normalized environmental distance metric. Larger values indicate greater environmental dissimilarity from the training conditions and, therefore, greater extrapolation. Daily environmental distance values were averaged to produce annual and seasonal climatologies aligned with the SDI maps.

3. Results and discussion

This section is organized as follows: Section 3.1 describes the dataset; Section 3.2 evaluates model performance; Section 3.3 examines the contribution of environmental predictors; and Section 3.4 illustrates model application through spatial and seasonal biogeographic patterns in 2021.

3.1. *Prokaryotic community diversity: A dataset overview*

Prokaryotic community diversity across all samples (Fig. 4a) ranged from approximately 2.1 to 5.7, with most values between 4.4 and 4.9. Samples with low diversity ($\leq$ 3.5) were less common in the dataset. Furthermore, site-specific SDI distributions (Fig. 4b) showed significant spatial variation in diversity. The highest diversity was observed at sites near the coast, including NEREA, MOLA, BBMO, and SOLA, with VIDA

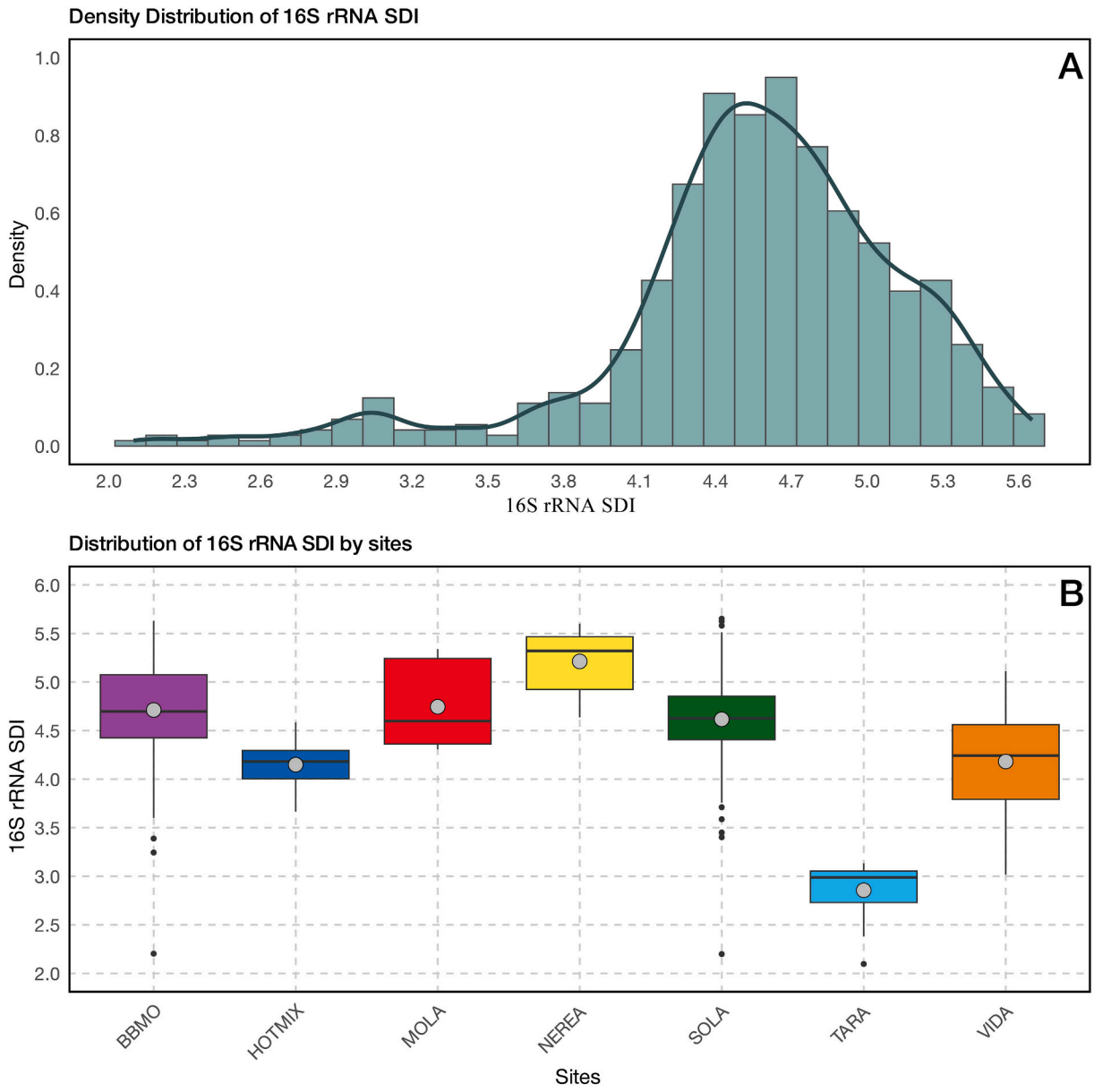


Fig. 4. (A) Frequency distribution of prokaryotic community diversity (16S rRNA gene SDI) for the entire dataset with an overlaid density curve. (B) Boxplots of individual datasets. The lines inside the boxes indicate the median, whereas the top and bottom edges mark the 75th and 25th percentiles, respectively, representing the interquartile range (IQR). Whiskers extend to the most extreme values within 1.5 times the IQR from the quartiles; individual black points beyond this range indicate outliers. The gray circles inside the boxplots show the mean values for each site. Note that the boxplot colors match the site colors in Fig. 1.

exhibiting slightly lower diversity. The lowest diversity levels were observed during the TARA cruise, whereas VIDA and HOTMIX exhibited comparable SDI values (Fig. 4b). These differences in diversity can be attributed to seasonal effects. Cruises occurred between spring and autumn, when prokaryotic community diversity is often low. Conversely, the time-series stations included samples collected in winter, when prokaryotic community diversity is known to be higher (Lambert et al., 2019). This seasonal contrast is characteristic of temperate latitudes (Ladau et al., 2013) and is consistent with globally recurrent seasonal patterns in prokaryotic community diversity documented across multiannual marine time series (Raes et al., 2024). Although derived from the SDI, these spatial and seasonal diversity patterns are consistent with previously documented large-scale environmental gradients in the marine environment, where variations in productivity, temperature, and trophic conditions are associated with systematic differences between coastal and open-ocean systems (Lefort and Gasol, 2013). Together, these observations provide a baseline for examining the environmental drivers explored in the following sections.

3.2. Model performance evaluation

The XGBoost model accurately predicted prokaryotic community diversity in both the training and testing datasets (Fig. 5). The training performance yielded an R^2 value of 0.78, accompanied by low errors (RMSE = 0.31, MAE = 0.22, and MAPE = 0.05), indicating that a significant portion (78%) of the variance was explained (Fig. 5a). Using RMSE and MAE together is generally advised because they offer complementary insights (i.e., with MAE reflecting the average error magnitude and RMSE giving greater weight to large deviations) into the model performance and error distribution (Chai and Draxler, 2014). The

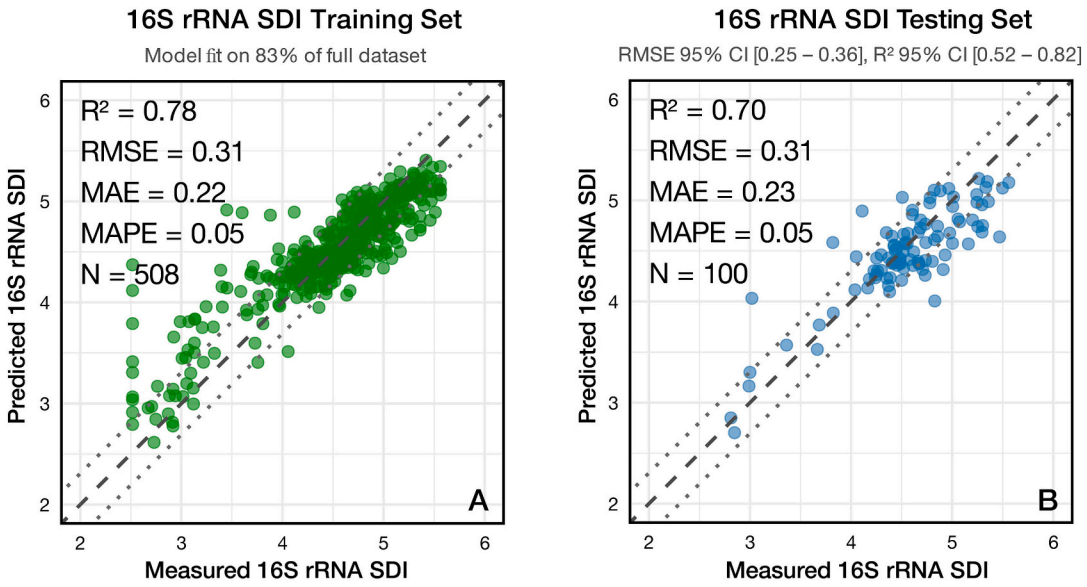


Fig. 5. Performance of the XGBoost model in predicting prokaryotic community diversity indexed by the 16S rRNA gene SDI. The plots show the predicted versus measured values for (A) the training set (83% of the data) and (B) the testing set (17% of the unseen data). The predictive performance, including 95% confidence intervals (CIs) for RMSE and R^2, is provided at the top, along with the number of samples (N) used in each subset. The dashed black lines represent the 1:1 line, whereas the dotted gray lines indicate ±1 standard deviation from the 1:1 line.

performance on the independent testing set (Fig. 5b) remained robust, with an R^2 value of 0.70 and nearly identical error metrics (RMSE = 0.31; MAE = 0.23; and MAPE = 0.05), indicating no overfitting. To assess model uncertainty in the test set, we calculated 95% confidence intervals using a bootstrap approach with 1000 repetitions (Tsamardinos et al., 2018). The resulting R^2 ranged from 0.52 to 0.82, while RMSE ranged from 0.25 to 0.36 (Fig. 5b), indicating that the model had moderate to high explanatory power on unseen data. The close alignment between the predicted and measured values of the SDI across both datasets suggests that the model generalizes well and can be confidently applied to estimate prokaryotic community diversity in unseen data. Additional evidence of the robustness of the model was indicated by the distribution of the RMSE values obtained across five repeated 10-fold cross-validation runs (Fig. S3). Using repeated cross-validation may help minimize variability and provide more dependable results than running the model only once, particularly when working with moderate-sized datasets (Jung, 2018; Krstajic et al., 2014). The model performed consistently across all repeats, as shown by the stable median RMSE and narrow interquartile range (Fig. S3). Additionally, the residuals from both the training and testing data were symmetrically centered around zero with similar ranges, demonstrating reliable predictive accuracy in both subsets (Fig. S4). Without heavy tails, the errors were evenly spread, indicating that the model did not systematically over- or under-predict any specific group of samples.

3.3. *Interpreting environmental drivers of prokaryotic community diversity*

At the Mediterranean basin scale, our findings differ from those of some global studies that identified temperature as the strongest predictor of prokaryotic community diversity (Fuhrman et al., 2008; Ibarbalz et al., 2019; Logares et al., 2020; Sunagawa et al., 2015). In our model, photoperiod emerged as the single most important predictor, although it accounted for only part of the explained variance, with the remaining 57.7% distributed across other environmental variables. A similar multivariate pattern was reported by Fuhrman et al. (2006), who showed that while photoperiod was the dominant predictor, combinations including temperature, nutrients, and Chl-a explained additional variability. Consistent with this, our results indicate comparable contributions from Chl-a, NO_3:PO_4, temperature, and NH_4 (11.3–14.6%; Fig. 6a), whereas Rrs_{412}:Rrs_{443} was the least influential predictor, although it accounted for 6.4% of the explained variance (Fig. 6a).

Among the predictors shown in Fig. 6 (see also Fig. S5 for their distribution), photoperiod alone accounted for 42.3% of the predictive power of the model (Fig. 6a). Our results indicated that day length was inversely related to diversity, as indicated by the high photoperiod values on the left side of the SHAP axis (Fig. 6b). SHAP values turned negative once day length exceeded ~11–13 h, indicating a summer decline in diversity when daylight was longest at these latitudes (Fig. 7a – note that in SHAP dependence plots, positive values indicate conditions that enhance diversity, whereas negative values indicate conditions that reduce it). This result is consistent with those of previous studies highlighting the primary role of seasonal light regimes in prokaryotic community diversity (Celussi et al., 2024; Fuhrman et al., 2015; Gilbert et al., 2010; Ladau et al., 2013; Lambert et al., 2019; Raes et al., 2024; Teeling et al., 2012). The diurnal cycle of solar radiation strongly shapes microbial physiology in the euphotic zone (Boyd and Van Mooy, 2025). For example, experiments have shown that extended light exposure can dampen prokaryotic activity and reshape communities, with some taxa being more sensitive to light than others (Alonso-Sáez et al., 2006; Merbt et al., 2012). Taken together, these findings from prior studies support our result that prokaryotic diversity decreases as the photoperiod lengthens, likely reflecting a combination of direct photic stress on light-sensitive taxa and indirect seasonal effects mediated by phytoplankton dynamics, stratification, nutrient availability, and grazing. Although endogenous circadian clocks have only been confirmed in cyanobacteria and remain unproven in heterotrophic bacteria and archaea (Lambert et al., 2019), photoperiod may act as a dominant external driver structuring microbial communities through tightly coupled physical and biological processes. However, disentangling the direct photic effects of photoperiod from its role as a proxy for broader seasonal environmental change is inherently challenging within a correlative ML framework, and the relative contribution of these processes remains uncertain.

The importance of Chl-a and NO_3:PO_4 suggests that phytoplankton biomass and nutrient stoichiometry are key drivers of prokaryotic community dynamics within specific ranges. The SHAP summary plot shows both the magnitude and direction of the influence of each

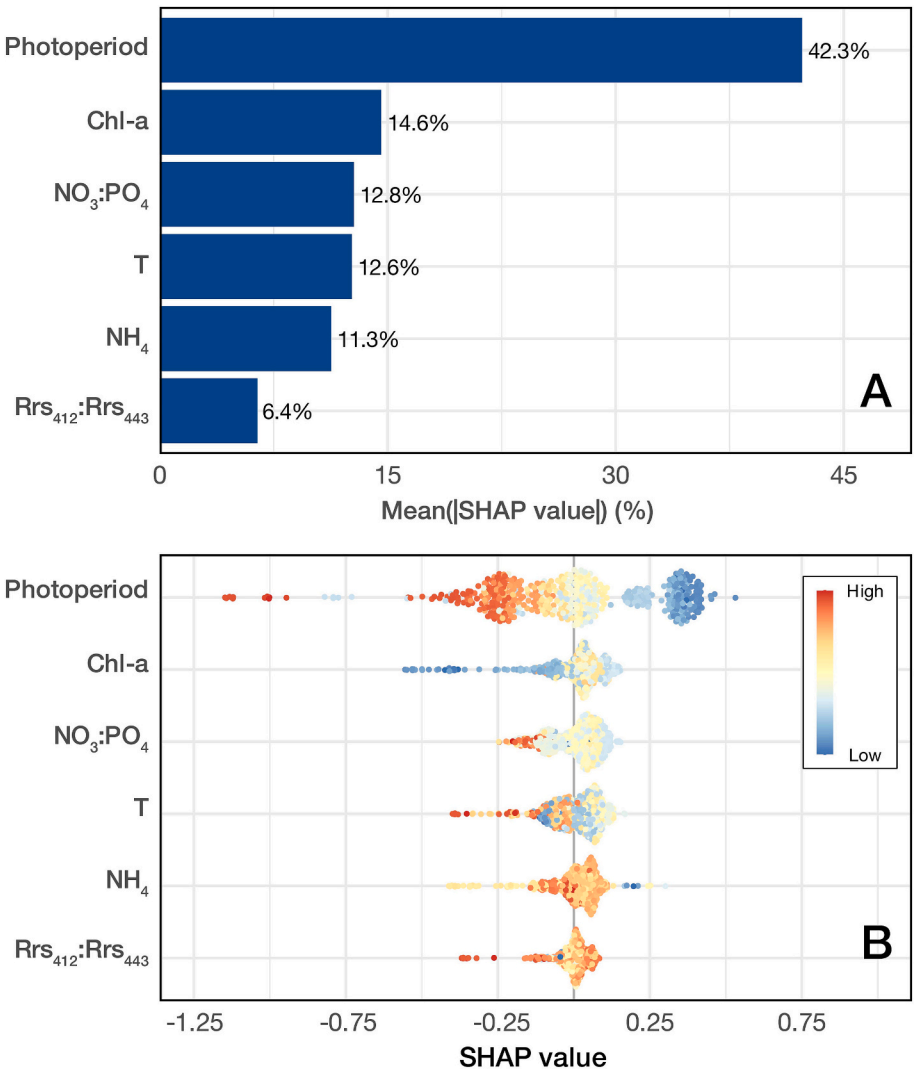


Fig. 6. SHAP analysis of predictor importance and effects on prokaryotic community diversity. (A) Summary plot of feature importance based on mean absolute SHAP values, showing each predictor's average contribution to the model output, expressed as a percentage; (B) SHAP summary plot presenting both the magnitude and direction of each feature's impact across all predictions. Each dot represents a single sample, and the color indicates the observed feature value (warmer colors = higher values). The horizontal spread shows the distribution of SHAP values for each feature, revealing whether high or low values tend to increase (positive SHAP values) or decrease (negative SHAP values) the predicted diversity.

predictor (Fig. 6b). The dependence plots went further, showing how the changes in Chl-a and NO_3:PO_4 either increased or decreased the predicted diversity (Fig. 7b and c).

Chl-a positively contributed to diversity, with values ranging from approximately 0.1 to 2 mg m^{-3} (Fig. 7b). Previous studies have shown strong connections between phytoplankton and heterotrophic bacteria through various types of interactions (Gomes et al., 2015; Seymour et al., 2017). Consistent with this, Krabberød et al. (2022) identified a core microbial community (defined as taxa present in more than 30% of monthly samples over 10 years) composed of bacteria and protists (pico- and nanoplankton) characterized by dense and mostly positive interactions in the northwestern Mediterranean. The network pattern aligns with other time-series data from the northwestern Mediterranean, which showed recurring co-occurrences between pico-eukaryotic phytoplankton and the abundant, free-living alphaproteobacterial clade SAR11 (Lambert et al., 2019). Such co-occurrence likely reflects metabolic dependencies and common nutrient needs. Similar phytoplankton–prokaryotic community associations have also been reported in other marine systems, where nutrient regimes structure community composition (Cordone et al., 2022). High-frequency observations further indicate that these relationships can be rapidly reorganized following disturbances and nutrient inputs, with microbial community composition shifting on timescales ranging from days to weeks after phytoplankton blooms (Lambert et al., 2021; Needham and Fuhrman, 2016).

The response of the model to the NO_3:PO_4 ratio was non-monotonic (Fig. 7c). The predictions peak near the Redfield values (~12–20), show fluctuations between ~30 and 100, and drop sharply at very high ratios (>100). Small-scale variability in the SHAP response could partly reflect spatial heterogeneity and interactions with co-varying environmental drivers. Overall, balanced or phosphorus (P)-replete conditions promoted higher predicted diversity, whereas P-limited (high NO_3:PO_4) conditions suppressed it. Across the Mediterranean, phosphate limitation intensified from west to east, whereas nitrate generally declined from north to south (Fig. S6, panels a and b). Consequently, surface NO_3:PO_4 ratios increased eastward, meaning that they were lower and closer to the Redfield ratio in the northwestern basin (indicating more balanced conditions) and were the highest in the ultra-oligotrophic Eastern Mediterranean (Fig. S6 panel c). In the latter, persistent phosphorus scarcity is reinforced by nitrogen-rich external inputs, especially atmospheric deposition and river runoff, as well as the near absence of denitrification (Lazzari et al., 2016; Ribera d'Alcalà et al., 2003; Krom

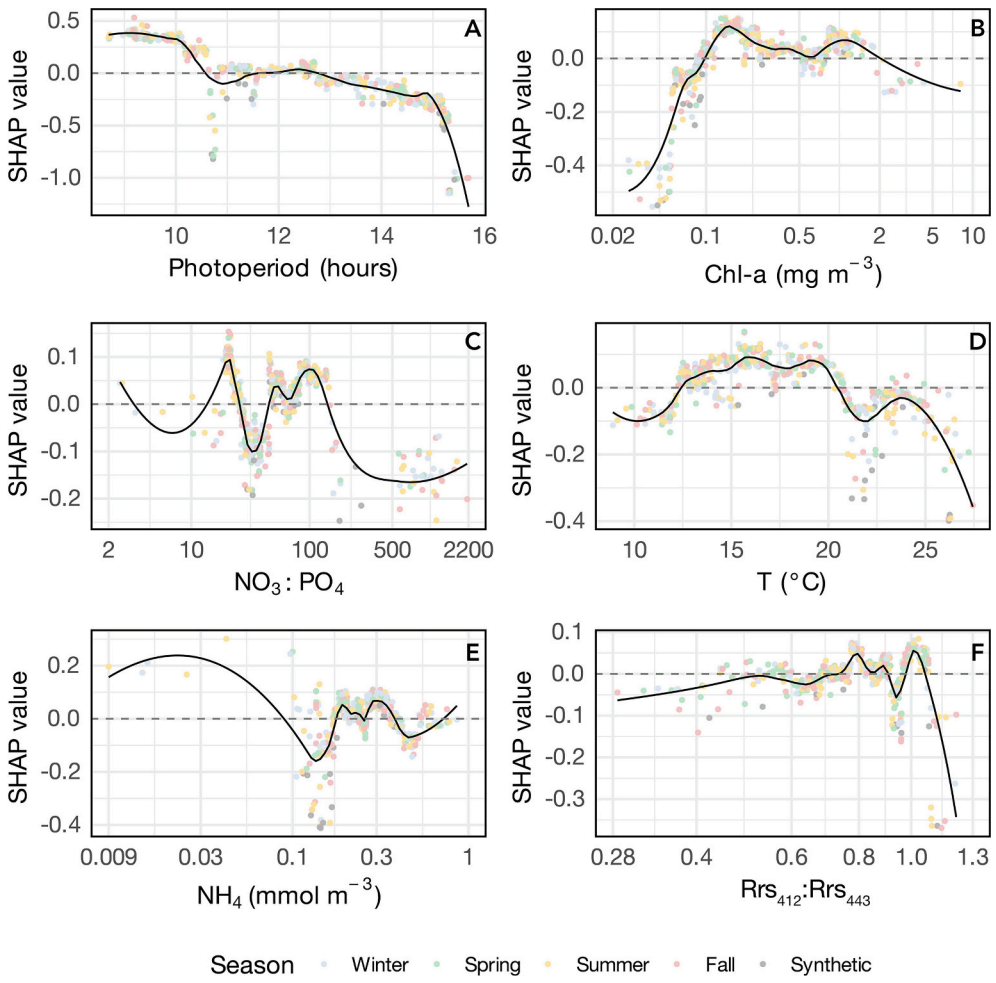


Fig. 7. SHAP dependence plots illustrating the influence of key predictors on prokaryotic community diversity. Each panel features a predictor with its SHAP value, which reveals the predictor's contribution to the predicted diversity. The predictors are as follows: (A) photoperiod, (B) Chl-a, (C) NO_3:PO_4, (D) Temperature, (E) NH_4, and (F) Rrs_{412}:Rrs_{443}. Points are colored by season to illustrate the seasonal distribution of observations along each environmental gradient. Positive SHAP values indicate an increase in predicted diversity, whereas negative values suggest a decrease. The dashed horizontal line marks SHAP = 0, and the black line shows a loess smoother summarizing the overall response pattern.

et al., 2010). In this sub-basin, prokaryotes display uneven nutrient limitation. Nutrient addition experiments have demonstrated that different bacterioplankton groups respond distinctly to phosphorus and nitrogen availability, highlighting strong bottom-up control of community structure in the Mediterranean Sea (Sebastián and Gasol, 2013). Overall, the Mediterranean Sea is a phosphorus-limited marine system, a condition that generally favors prokaryotes capable of competing for scarce phosphorus resources (Feingersch et al., 2010). Phosphorus limitation is a key structuring force shaping the microbial community composition in the northern Adriatic Sea. Long-term and high-resolution studies in the Gulf of Trieste have shown that microbial productivity, diversity, and succession are closely linked to P availability, with pulses of riverine phosphate driving shifts in both phytoplankton and prokaryotic community composition (Malfatti et al., 2014; Tinta et al., 2015). When phosphorus is scarce, small taxa adapted to low-nutrient conditions dominate. In contrast, increased phosphorus input can trigger diatom blooms and favor fast-growing prokaryotes (Tinta et al., 2015). Bayesian network analysis from the northern Adriatic region supports this finding, showing that phosphorus-driven microbial interactions are key regulators of particulate organic carbon (McDonald et al., 2017). The Western Mediterranean region shows a similar pattern. Nutrient addition experiments revealed year-round bacterial phosphorus limitation. The latter peaks in spring and summer, when phosphorus and Chl-a concentrations are very low and nutrient ratios are elevated (Pinhassi et al., 2006). Overall, seasonal shifts in nutrient availability can influence bacterioplankton in two ways. First, they directly regulate bacterial growth and activity. Second, nutrients indirectly affect prokaryotes by changing phytoplankton abundance and composition (Pinhassi et al., 2006).

Temperature (T) also plays a significant role in shaping seasonal prokaryotic community composition (Fuhrman et al., 2006; Ibarbalz et al., 2019; Lambert et al., 2019). On a global scale, analyses of the Tara Oceans dataset identified temperature as the strongest factor influencing epipelagic prokaryotic community composition (Sunagawa et al., 2015). Although this global pattern is not specifically driven by Mediterranean samples, it offers a broader context for understanding regional relationships. Consistent with this idea, our results show that temperature also affects prokaryotic diversity in the Mediterranean (Fig. 6b), with diversity increasing from approximately 12 °C to 20 °C (Fig. 7d). These relationships are often nonlinear and system-specific. For example, warming may promote slow-growing bacteria to adapt to low-nutrient conditions (Abreu et al., 2023), which can reshape the community structure and potentially decrease richness beyond a certain thermal threshold. Notably, this temperature range spans both late winter and autumn conditions in the Mediterranean, during which enhanced prokaryotic diversity may arise from different but complementary processes, including late-winter phytoplankton blooms and the breakdown of summer stratification during autumn mixing (see Section 3.4).

Our results also showed that NH_4 is a key factor influencing prokaryotic diversity (Fig. 6b). Its impact on diversity was non-monotonic (Fig. 7e), with the strongest negative effects observed at intermediate concentrations (~0.1–0.2 mmol m^{-3}) and near-neutral to positive effects at higher concentrations (~0.2–0.4 mmol m^{-3}). The direction and strength of this effect likely vary depending on the season, co-varying nutrients, and regional setting (e.g., coastal vs. open sea). For instance, in the North Pacific Ocean, NH_4 addition can alter heterotrophic prokaryotic community composition, revealing differential uptake across taxa (Shilova et al., 2017). Among photoautotrophs, small cyanobacteria, such as *Prochlorococcus,* also show taxon-specific preferences, preferentially assimilating NH_4 over nitrate (Olofsson et al., 2019; Zubkov et al., 2003). Overall, increased NH_4 availability can stimulate phytoplankton and net DOM production, yielding more labile DOM that enhances substrate availability for heterotrophic prokaryotes and promotes microbial growth (Goldberg et al., 2017). However, responses to elevated NH_4 are often taxon-specific, shifting the composition of eukaryotic phytoplankton and the prokaryotic community and intensifying competition for ammonium among key groups. In some settings, higher NH_4 concentrations coincide with lower alpha diversity, suggesting a shift toward fewer, better-adapted taxa (Doane et al., 2023; Glibert et al., 2016; Klawonn et al., 2019). In the northwestern Mediterranean, surface nutrient concentrations, including NH_4, are generally higher during winter mixing and lowest during the strongly stratified summer period; open-sea ammonium decreases from ~0.6 mmol m^{-3} in winter to ~0.1 mmol m^{-3} in summer before partially recovering (~0.4 mmol m^{-3}) in autumn (Segura-Noguera et al., 2016). Across the broader Mediterranean, NH_4 patterns are strongly shaped by spatial heterogeneity, particularly elevated coastal concentrations relative to the open sea, consistent with modulation by coastal inputs and regeneration processes and the persistence of coastal-offshore gradients across seasons (Fig. S7). These contrasts are expected to be amplified in river-influenced coastal waters, where freshwater–marine mixing creates sharp biogeochemical transitions. In river-plume mixing zones (e.g., the Rhône); bacterial production is highest where NH_4 uptake is maximal along the salinity gradient (Pujo-Pay et al., 2006). Such gradients could therefore modulate community composition and potentially influence diversity across coastal-offshore transitions.

Finally, although Rrs_{412}:Rrs_{443} was the least influential predictor, it still contributed 6.4% of the model output (Fig. 6a), likely reflecting the relative influence of CDOM in Mediterranean waters (Morel and Gentili, 2009a; Organelli et al., 2014; Organelli et al., 2016). The Mediterranean Sea harbors higher CDOM concentrations than the adjacent Atlantic waters for a given chlorophyll concentration, with particularly elevated values in areas such as the Gulf of Lion and the northern Aegean Sea (Morel and Gentili, 2009b). As shown in the SHAP plots (Figs. 6b and 7f), intermediate reflectance ratios (~0.7–0.9) positively influenced the predicted prokaryotic community diversity response. In our dataset, the Rrs_{412}:Rrs_{443} optical proxy showed a nonlinear association with diversity: very low ratios, potentially reflecting elevated CDOM absorption, coincided with reduced diversity, whereas very high ratios (clearer, strongly oligotrophic surface conditions) were also associated with reduced diversity. Consistent with previous Mediterranean analyses, surface CDOM typically peaks in winter-spring and is lowest in summer due to photobleaching and stratification (Lazzari et al., 2021; Massi et al., 2020; Xing et al., 2014). However, the optimal Rrs_{412}:Rrs_{443} range should be interpreted cautiously because of uncertainties in wavelength-dependent atmospheric correction, especially at 412 nm. Furthermore, the non-monotonic SHAP dependence pattern of Rrs_{412}:Rrs_{443} suggests that, at the basin scale, variability in this ratio could be driven by spatial heterogeneity, particularly coastal-offshore gradients, rather than seasonal cycling. In Mediterranean coastal and estuarine waters, terrestrially influenced CDOM often decreases with increasing salinity, consistent with conservative freshwater–marine mixing (Para et al., 2010; Vignudelli et al., 2004). Offshore, however, CDOM is increasingly shaped by autochthonous production and internal cycling processes rather than conservative freshwater–marine mixing (Galletti et al., 2019). Accordingly, spatial variability in Rrs_{412}:Rrs_{443} may partly reflect the same freshwater–marine gradients captured by salinity, particularly in river-influenced areas. Prokaryotes consume, produce, and transform CDOM, thereby shaping their composition while being structured by CDOM chemistry, making them central to CDOM dynamics (Kinsey et al., 2018; Lazzari et al., 2021; Organelli et al., 2014; Romera-Castillo et al., 2011). Together, these patterns suggest that CDOM-related optical properties integrate multiple biogeochemical processes that can indirectly shape prokaryotic community diversity, particularly in regions with strong coastal and riverine influences.

The dependence curves shown in Fig. 7 represent marginal effects and may include variability arising from interactions among the predictors. Seasonal coloring shows that predictors with a strong seasonal cycle, such as photoperiod and temperature, occupy feature-value ranges that closely align with the season. In contrast, for predictors with greater spatial and ecological heterogeneity, the seasonal overlap is much greater, likely because similar values can arise in different environmental contexts, including coastal and open-sea settings. Samples from the different datasets overlapped across the central portion of most predictor distributions (Fig. S8), suggesting that the observed non-monotonic SHAP patterns were not driven primarily by any single dataset. Although BBMO is generally the most represented, the other datasets contribute to most of the main feature space. Observations become sparse toward the extremes of some predictors; therefore, the SHAP values in these regions should be interpreted with caution. Numerous other environmental variables not considered in this study may also influence the structure of microbial communities (and hence, diversity), such as specific compounds (which we may have overlooked owing to the limited resolution of our analyses) and inter- and intra-species interactions.

3.4. *Biogeographic patterns and seasonal dynamics of prokaryotic community diversity*

The 2021 annual mean prokaryotic diversity revealed a distinct spatial pattern across the Mediterranean Sea, with higher diversity in coastal regions and the western basin and lower diversity in the eastern basin (Fig. 8a). Compared to open waters, coastal areas in the Mediterranean and other temperate coastal systems are often characterized by higher diversity values (Pommier et al., 2010), likely because riverine inflow and wastewater inputs enhance nutrient availability and phytoplankton biomass, thereby favoring fast-growing taxa and reshaping prokaryotic diversity, community structure, and ecosystem functioning (Orel et al., 2022; Orel et al., 2026; Thompson et al., 2011). Across the basin, the observed longitudinal gradient in diversity mirrors the well-documented west-to-east decrease in nutrient availability associated with increasing oligotrophy and phosphate limitation toward the eastern basin (Sebastián et al., 2021). Although nutrient availability is generally low throughout the Mediterranean Sea, particularly for PO_4, this limitation may be locally buffered or alleviated by terrestrial and atmospheric inputs, potentially supporting higher prokaryotic biomass and production (Siokou-Frangou et al., 2010). For instance, atmospheric dust inputs from the Sahara have been shown to influence heterotrophic prokaryotic dynamics in oligotrophic Mediterranean waters by supplying bioavailable phosphorus and nitrogen, stimulating bacterial activity, and reshaping community composition (Pulido-Villena et al., 2014). Furthermore, in the very oligotrophic Levantine Sea, PO_4 enrichment was followed by a rapid transfer of added P to particulate matter and a slight decrease in Chl-a, consistent with scenarios in which heterotrophic prokaryotes preferentially exploit the P pulse and increase competition with N-limited autotrophs (Siokou-Frangou et al., 2010). More generally, under nutrient limitation, phytoplankton carbon fixation can outpace heterotrophic prokaryotic processing of organic matter, promoting DOM accumulation and altering microbial carbon cycling, as shown in mesocosm experiments conducted in the P-limited

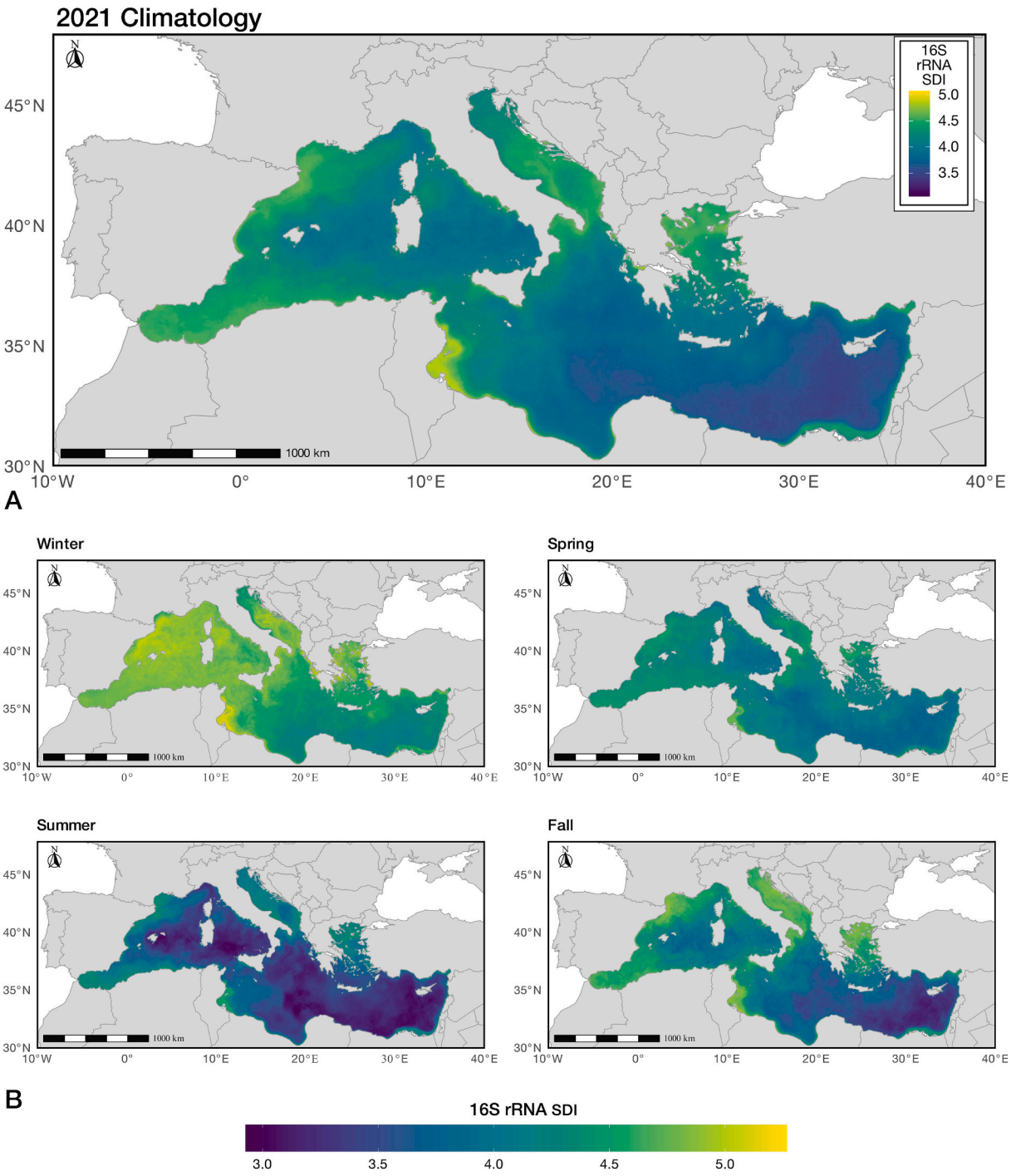


Fig. 8. Predicted Shannon diversity maps for 2021. (A) Annual mean map of prokaryotic community diversity (16S rRNA gene SDI) in the Mediterranean Sea. The map was generated by applying a trained XGBoost model to daily environmental predictors and averaging the predictions over the year. (B) Seasonal mean maps of prokaryotic community diversity computed using the same approach as in panel A, but averaged for winter (JFM; top left), spring (AMJ; top right), summer (JAS; bottom left), and fall (OND; bottom right). Warmer colors (yellow) indicate higher predicted diversity, whereas cooler colors (blue to purple) indicate lower predicted diversity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

northern Adriatic Sea (Malfatti et al., 2014). At the same time, the west-to-east decrease in microbial production broadly tracks the basin-scale decline in primary production. Because primary production is a major DOC source for heterotrophic prokaryotes, lower production toward the eastern basin likely constrains heterotrophic prokaryotic activity (Siokou-Frangou et al., 2010). DOC quantity and lability, along with the metabolic capacity of heterotrophic prokaryotic communities to utilize these carbon pools, can in turn shape community structure and diversity patterns (Nelson and Wear, 2014). Notably, surface DOC concentrations tend to increase eastward in the Mediterranean, consistent with stronger stratification and reduced vertical mixing that favor DOC accumulation above the pycnocline (Santinelli, 2015). This pattern indicates that DOC accumulation in the eastern basin reflects physical retention rather than enhanced production of labile organic carbon. Accordingly, the

eastward increase in bulk DOC does not correspond to higher prokaryotic diversity in our data, suggesting that total DOC concentration alone is a poor proxy for the carbon fraction effectively supporting heterotrophic prokaryotic diversity in the eastern basin. Additionally, DOC quality may also depend on phytoplankton community composition: when a few phytoplankton taxa dominate, they can produce DOC pools with altered composition and bioavailability, thereby selecting for specific heterotrophic prokaryotic groups through substrate preferences and resource partitioning (Landa et al., 2016; Sarmento and Gasol, 2012).

Beyond the annual patterns, the seasonal maps (Fig. 8b) showed consistent variations throughout the year. Diversity was highest in winter (January–March) and lowest in summer (July–September), broadly consistent with seasonal basin-scale changes in phytoplankton biomass, production, and water-column structure. This seasonal pattern appears to be primarily driven by deep vertical mixing in winter, which replenishes nutrients in the euphotic zone and fuels the late winter-to-early spring bloom as solar irradiance increases. As the bloom progresses, nutrients are gradually consumed, and intense summer stratification further limits their resupply, suppressing phytoplankton growth (D'Ortenzio and Ribera d'Alcalà, 2009; Lavigne et al., 2013; Lazzari et al., 2012). Together, mixing and stratification modulate phytoplankton primary production and the supply of phytoplankton-derived DOC, which, in turn, can influence the growth and composition of prokaryotic community assemblages (Lazzari et al., 2021; Organelli et al., 2014). However, heterotrophic prokaryotic production does not necessarily peak synchronously with phytoplankton dynamics. Instead, it often increases during the post-bloom period, as a delayed response to primary production as phytoplankton-derived organic matter accumulates and becomes available for heterotrophic processing (Gomes et al., 2015). Thus, the seasonal cycle in prokaryotic diversity likely reflects not only contemporaneous phytoplankton production but also delayed microbial responses to the production, accumulation, and subsequent transformation of organic matter. Experimental evidence from the northwestern Mediterranean has shown that prokaryotic production can be carbon-limited in autumn-winter and shifts toward phosphorus limitation in spring-summer, consistent with seasonal changes in resource availability and prokaryotic community dynamics (Pinhassi et al., 2006). Such seasonal shifts are well documented in temperate seas, where microbial succession closely tracks physical forcing and bloom progression (Bunse and Pinhassi, 2017; Celussi et al., 2024; Teeling et al., 2012; Tinta et al., 2015). Accordingly, winter mixing may favor taxa adapted to enhanced nutrient supply, whereas summer stratification supports oligotrophic taxa adapted to low nutrient availability (Bunse and Pinhassi, 2017).

To explore the spatial expression of these seasonal dynamics across the Mediterranean Sea, we performed a clustering analysis using the seasonal climatology of prokaryotic community diversity. The analysis identified five distinct subregions, classified from Very High to Very Low Diversity, summarizing the dominant biogeographic patterns across the Mediterranean (Fig. 9). Clusters provide a basin-wide view of prokaryotic community diversity regimes in the Mediterranean, highlighting both an evident west-east decline in diversity and localized coastal diversity hotspots, particularly in the northwestern Mediterranean, the Adriatic Sea, the Gulf of Lion, and the Alboran Sea. Such spatial regionalization aligns with previous research, which demonstrated that prokaryotic diversity largely reflects the oligotrophic gradient of the basin (Sebastián et al., 2021) and its well-known biogeochemical regimes (D'Ortenzio and Ribera d'Alcalà, 2009; Siokou-Frangou et al., 2010). Gradients in nutrient availability, primary production, and physical dynamics shape these regimes and structure prokaryotic community diversity across basins (Sebastián et al., 2021).

Finally, the spatial representativeness of these predictions was evaluated using environmental distance maps (Fig. 10). Lower environmental-distance values denote areas where predictions are based on environmental conditions better represented in the training data, whereas higher values indicate greater extrapolation and lower confidence. Across the basin, the western Mediterranean and several coastal areas showed relatively low environmental distances, suggesting that predictions in these regions are made under better-represented conditions. The environmental distance increased toward the southern-central Mediterranean and the eastern basin, where the coverage of training data was more limited (Fig. 10a). These patterns were generally consistent across seasons, although slightly higher distances appeared during summer and fall, including in parts of the western basin (Fig. 10b). Overall, the predicted SDI patterns in the western basin and northern coastal regions (Fig. 8) are more strongly supported by the training data, whereas predictions in the southern central

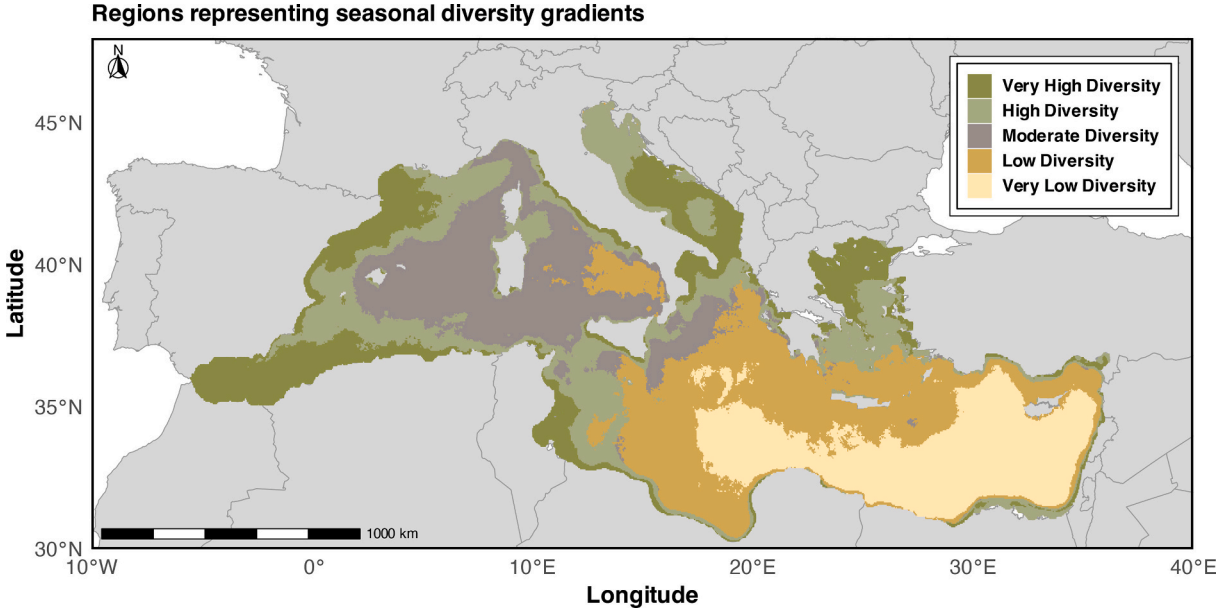


Fig. 9. Spatial distribution of the five clusters representing seasonal gradients in prokaryotic community diversity (16S rRNA gene SDI) across the Mediterranean Sea. The clusters were derived from the 2021 seasonal averages. Regions are classified from Very High Diversity to Very Low Diversity based on their relative magnitude and seasonal patterns.

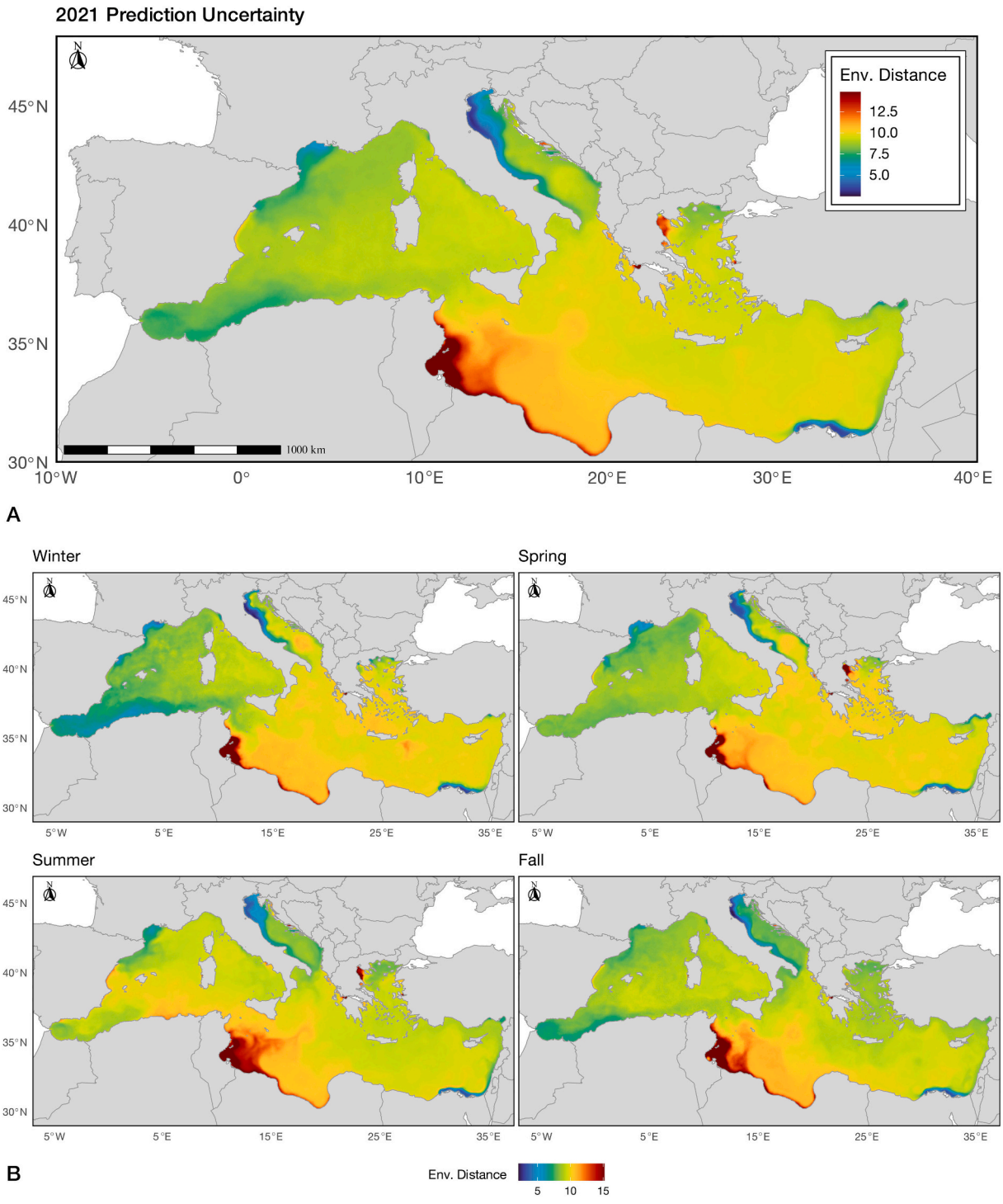


Fig. 10. Environmental distance maps for 2021. (A) Annual mean environmental distance across the Mediterranean Sea. Environmental distance was defined as the Euclidean distance in standardized predictor space between each prediction pixel and its nearest neighbor in the training dataset. (B) Seasonal mean environmental distance computed as in panel A, but averaged for winter (JFM; top left), spring (AMJ; top right), summer (JAS; bottom left), and fall (OND; bottom right). Lower values (blue to green) indicate environmental conditions well represented in the training data, whereas higher values (orange to red) indicate increasing extrapolation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Mediterranean and parts of the eastern Mediterranean should be interpreted with greater caution.

4. Study limitations and outlook

Our results demonstrate that prokaryotic community diversity, as measured by the 16S rRNA gene SDI, can be effectively predicted using a combination of biogeochemical and environmental drivers merged through an ML model. However, it is essential to acknowledge certain limitations and opportunities for future improvement.

A significant issue was uneven sampling and the resulting spatial imbalance, particularly in oligotrophic areas of the Mediterranean Sea. Nevertheless, the model captured broad seasonal patterns of prokaryotic community diversity, likely supported by the broad temporal coverage

of the dataset and the use of synthetic samples that partially mitigated this imbalance. Therefore, future efforts should focus on broader sampling to improve environmental coverage and enhance dataset quality, thereby improving the generalizability of ML models. The environmental distance analysis confirmed that the model extrapolates with greater uncertainty in areas with sparse training data. Expanding the sampling effort in the eastern Mediterranean and open-sea environments would be essential to reduce this uncertainty and improve the reliability of basin-wide predictions. In addition, variation in dataset representation across predictors should be considered when interpreting SHAP-based predictor–diversity relationships, particularly at predictor extremes, where observations are sparse. Capturing all environmental variability of the ocean for use in training datasets is inherently complex, if not impossible. Extensive and consistent datasets remain the best practice for improving the robustness of ML models because they expose models to a wide range of oceanographic conditions (Gray et al., 2024). Although prokaryotic biomass appears to be resilient to climate change (Heneghan et al., 2024), a shift toward more extreme conditions may compromise ML predictions (Meunier et al., 2025). In this context, ongoing monitoring programs and research cruises are essential, as they provide critical data inputs for sustained model improvement. Continuous sampling and further retraining of algorithms are therefore necessary to keep ML models up to date and ensure the accurate detection of spatiotemporal dynamics.

A potential limitation relates to differences in the volume of water filtered across cruises and stations, which could influence the detection of rare taxa and thus affect richness estimates. However, results from a large intercalibration exercise indicate that variations in filtration volume do not significantly influence composite diversity indices for prokaryotic or protist communities across different sequencing approaches (Pascoal et al., 2023). Moreover, because Shannon diversity integrates both taxon richness and evenness and is primarily driven by dominant taxa, it is inherently less sensitive to incomplete sampling than simple richness metrics. Therefore, although absolute richness values should be interpreted with caution when comparing samples from different cruises and stations, the spatial and seasonal patterns of SDI observed here are robust for interpreting large-scale trends.

An additional limitation is the potential mismatch between the timing or depth of environmental predictors and the timing of biological sampling. We addressed this issue by using only 16S rRNA gene samples gathered from the top three meters of the water column and performing daily matchups to ensure consistency with satellite and modeled surface data. Nevertheless, slight mismatches can occur, particularly in highly dynamic environments. However, prokaryotic community composition is generally structured by changes occurring over timescales of days to weeks, whereas diel-scale variation appears to be comparatively limited (Needham et al., 2013). The daily resolution of our environmental predictors is therefore appropriate for capturing the dominant temporal dynamics relevant to this study. Furthermore, the surface-focused approach used in this study did not account for diversity in the deeper layers. Sebastián et al. (2021) reported the same west-east decline in prokaryotic diversity; however, higher diversity was found in the meso- and bathypelagic layers than in surface waters.

Finally, although ML models, such as XGBoost, are strong tools for detecting complex patterns and making accurate predictions from specific datasets, they do not explain why these patterns occur. In contrast, mechanistic models explain patterns based on known ecological processes; however, they may oversimplify the complexity of natural systems. Biotic interactions, such as grazing and viral lysis, if included in the ML model, could help capture key ecological processes in prokaryotic community structure (Bunse and Pinhassi, 2017; Doane et al., 2023; Landolfi et al., 2021). Such an approach could combine the predictive power of ML models with improved ecological interpretability, although this remains challenging because of limited data and methodological complexity. As hybrid modeling approaches develop and increasingly heterogeneous datasets become available, systematic variable selection methods, such as hierarchical clustering of predictors or forward/backward selection, could help identify alternative yet informative combinations of environmental drivers and further improve model interpretability.

5. Concluding remarks

This study demonstrated that ML is effective for modeling prokaryotic community diversity in the Mediterranean Sea. By integrating satellite-derived and modeled biogeochemical variables with *in situ* 16S rRNA gene data, we predicted SDI patterns with promising predictive accuracy, which could become more robust with the availability of larger and more balanced training datasets. Applying this model to georeferenced data provides an opportunity to extend the monitoring of prokaryotic community diversity to the basin scale, revealing distinct spatiotemporal patterns that have not been previously explored. In agreement with previous studies, photoperiod emerged as the strongest predictor, indicating that seasonal light cycles can strongly influence prokaryotic community diversity (Gilbert et al., 2012). Nevertheless, other predictors also play important roles, with each exhibiting positive effects on diversity under specific environmental conditions. Because environmental controls on a global scale may differ across regions and latitudes, these results emphasize the importance of including multiple area-specific environmental variables in ML models to gain a more comprehensive understanding of plankton dynamics (Bunse and Pinhassi, 2017). The workflow developed in this study can be extended to other microbial groups, including eukaryotic plankton, thereby broadening its applicability. Although this model cannot account for specific biological interactions, it remains a valuable tool for monitoring microbial diversity. Given the rapid pace of environmental change, scientific projects such as Biodiversa+ PETRI-MED, which combine complementary observation methods and modeling techniques, are crucial for equipping ecosystem managers and policymakers with the knowledge required to protect vulnerable marine ecosystems such as the Mediterranean Sea.

CRediT authorship contribution statement

Christian Marchese: Conceptualization, Methodology, Data curation, Software, Formal analysis, Visualization, Investigation, Writing – review & editing, Writing – original draft. **Maria Laura Zoffoli**: Methodology, Visualization, Investigation, Writing – review & editing, Writing – original draft. **Pierre Ramond**: Data curation, Validation, Writing – review & editing. **Ramiro Logares**: Data curation, Writing – review & editing. **François-Yves Bouget**: Data curation, Writing – review & editing. **Pierre E. Galand**: Data curation, Writing – review & editing. **Tinkara Tinta**: Data curation, Writing – review & editing. **Neža Orel**: Data curation, Writing – review & editing. **Gianluca Volpe**: Data curation, Writing – review & editing. **Angela Landolfi**: Data curation, Writing – review & editing. **Emanuele Organelli**: Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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, co-funded by the European Commission (GA

N°101052342) and with the funding organizations: the Italian Ministry of University and Research (MUR), 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 data and support provided by the VIDA program, funded through the EU Research Infrastructure projects eLTER ERIC and LifeWatch ERIC, and by the Slovenian Research Agency (Research Core Funding No. P1-0237). We thank the four reviewers for their careful evaluation of the manuscript and for their constructive and insightful comments, which helped improve the clarity, rigor, and overall quality of this work. We thank the scientists and crew aboard the R/V Sarmiento de Gamboa and the schooner Tara for collecting and sequencing DNA samples during the HOTMIX and Tara Mediterranean expeditions. We also thank the teams and platforms that collected and curated 16S rRNA gene sequencing data across the Mediterranean Sea. Finally, we thank Josep M. Gasol, Vanessa Balagué, Clara Cardelus, and the whole BBMO sampling team for their contributions to generating the LTER BBMO 16S rRNA gene dataset.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.103747.

Data availability

All environmental data used in this study are publicly available from the Copernicus Marine Service (CMEMS) at https://marine.copernicus.eu. The complete 16S rRNA gene dataset used in this study to compute the Shannon Diversity Index is not publicly available but can be made available by the authors upon reasonable request. The 16S rRNA gene amplicon sequences generated at the VIDA station between 2018 and 2021 are available in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJEB60871. The 16S rRNA gene amplicon sequences generated during the Tara Mediterranean expedition in 2014 are available in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA380761. All 16S rRNA amplicon sequence variant (ASV) data from NEREA are publicly available in the Zenodo Sample Registry "NEREA – Naples Ecological REsearch for Augmented observatories" (https://zenodo.org/communities/nerea/records). The processing scripts for 16S ASV generation and taxonomic assignment are archived at Doi: https://doi.org/10.5281/zenodo.12801913.

References

Abreu, C.I., Dal Bello, M., Bunse, C., Pinhassi, J., Gore, J., 2023. Warmer temperatures favor slower-growing bacteria in natural marine communities. Sci. Adv. 9, eade8352. https://doi.org/10.1126/sciadv.ade8352.

Alonso-Sáez, L., Gasol, J.M., Lefort, T., Hofer, J., Sommaruga, R., 2006. Effect of natural sunlight on bacterial activity and differential sensitivity of natural bacterioplankton groups in northwestern mediterranean coastal waters. Appl. Environ. Microbiol. 72, 5806–5813. https://doi.org/10.1128/AEM.00597-06.

Arthur, D., Vassilvitskii, S., 2007. K-means++: the advantages of careful seeding. In: proceedings of the eighteenth annual ACM–SIAM symposium on discrete algorithms (SODA '07). Society for industrial and applied mathematics, Philadelphia, pp. 1027–1035.

Auladell, A., Barberán, A., Logares, R., Garcés, E., Gasol, J.M., Ferrera, I., 2022. Seasonal niche differentiation among closely related marine bacteria. ISME J. 16, 178–189. https://doi.org/10.1038/s41396-021-01053-2.

Barale, V., Jaquet, J.-M., Ndiaye, M., 2008. Algal blooming patterns and anomalies in the mediterranean sea as derived from the SeaWiFS data set (1998-2003). Remote Sens. Environ. 112, 3300–3313. https://doi.org/10.1016/j.rse.2007.10.014.

Batten, S.D., Abu-Alhaija, R., Chiba, S., Edwards, M., Graham, G., Jyothibabu, R., Kamykowski, D., Kang, J.-J., Kang, S., Karnataka, S., Lasley-Rasher, R., Lenz, P.H., Mazzocchi, M.G., Rakhesh, M., Shinada, A., Sun, S., 2019. A global plankton diversity monitoring program. Front. Mar. Sci. 6, 321. https://doi.org/10.3389/fmars.2019.00321.

Baumas, C.M.J., Le Moigne, F.A.C., Garel, M., Bhairy, N., Guasco, S., Riou, V., Armougom, F., Grossi, V., Tamburini, C., 2021. Mesopelagic microbial carbon production correlates with diversity across different marine particle fractions. ISME J. 15, 1695–1708. https://doi.org/10.1038/s41396-020-00880-z.

Beauvais, M., Schatt, P., Montiel, L., Logares, R., Galand, P.E., Bouget, F.-Y., 2023. Functional redundancy of seasonal vitamin b12 biosynthesis pathways in coastal marine microbial communities. Environ. Microbiol. 25, 3753–3770. https://doi.org/10.1111/1462-2920.16545.

Behrenfeld, M.J., O'Malley, R.T., Siegel, D.A., McClain, C.R., Sarmiento, J.L., Feldman, G. C., Milligan, A.J., Falkowski, P.G., Letelier, R.M., Boss, E.S., 2006. Climate-driven trends in contemporary ocean productivity. Nature 444, 752–755. https://doi.org/10.1038/nature05317.

Behrenfeld, M.J., Halsey, K.H., Milligan, A.J., 2008. Evolved physiological responses of phytoplankton to their integrated growth environment. Philos. Trans. R. Soc. B 363, 2687–2703. https://doi.org/10.1098/rstb.2008.0019.

Belhaouari, S.B., Islam, A., Kassoul, K., Al-Fuqaha, A., Bouzerdoum, A., 2024. Oversampling techniques for imbalanced data in regression. Expert Syst. Appl. 252, 124118. https://doi.org/10.1016/j.eswa.2024.124118.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Bigg, G.R., Jickells, T.D., Liss, P.S., Osborn, T.J., 2003. The role of the oceans in climate. Int. J. Climatol. 23, 1127–1159. https://doi.org/10.1002/joc.926.

Boyd, P.W., Van Mooy, B.A.S., 2025. Using the diel cycle of ocean microbes to better understand their biogeochemical functions. Limnol. Oceanogr. Lett. 10, 434–447. https://doi.org/10.1002/lol2.70027.

Boyd, P.W., Claustre, H., Levy, M., Siegel, D.A., Weber, T., 2019. Multi-faceted particle pumps drive carbon sequestration in the ocean. Nature 568, 327–335. https://doi.org/10.1038/s41586-019-1098-2.

Bunse, C., Pinhassi, J., 2017. Marine bacterioplankton seasonal succession dynamics. Trends Microbiol. 25, 494–505. https://doi.org/10.1016/j.tim.2016.12.013.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high-resolution sample inference from illumina amplicon data. Nat. Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869.

Campese, L., Russo, L., Abagnale, M., Alberti, A., Bachi, G., Balestra, C., Bellardini, D., Buondonno, A., Cardini, U., Carotenuto, Y., Checcucci, G., Chiusano, M.L., D'Ambra, I., d'Ippolito, G., Di Capua, I., Donnarumma, V., Fontana, A., Furia, M., Galarza-Verkovitch, D., Gallia, R., Labadie, K., Leone, S., Licandro, P., Longo, A., Maselli, M., Merquiol, L., Murano, C., Oliveira, P.H., Passarelli, A., Percopo, I., Perdereau, A., Piredda, R., Raffini, F., Roncalli, V., Ruscheweyh, H.-J., Russo, E., Saggiomo, M., Santinelli, C., Sarno, D., Sunagawa, S., Tramontano, F., Trano, A.C., Uttieri, M., Wincker, P., Zampicinini, G., Casotti, R., Conversano, F., D'Alelio, D., Iudicone, D., Margiotta, F., Montresor, M., 2024. The NEREA augmented observatory: an integrative approach to marine coastal ecology. Sci. Data 11, 989. https://doi.org/10.1038/s41597-024-03787-y.

Carlson, C.A., Del Giorgio, P.A., Herndl, G.J., 2007. Microbes and the dissipation of energy and respiration: from cells to ecosystems. Oceanography 20, 89–100. https://doi.org/10.5670/oceanog.2007.52.

Celussi, M., Manna, V., Banchi, E., Fonti, V., Bazzaro, M., Flander-Putrle, V., Francé, J., Janeković, I., Kružić, P., Neri, F., Tinta, T., Del Negro, P., 2024. Annual recurrence of prokaryotic climax communities in shallow waters of the north mediterranean. Environ. Microbiol. 26, e16595. https://doi.org/10.1111/1462-2920.16595.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2025. Xgboost: extreme gradient boosting. R package version 1.7.11.1. https://CRAN.R-project.org/package=xgboost.

Chiarucci, A., Bacaro, G., Scheiner, S.M., 2011. Old and new challenges in using species diversity for assessing biodiversity. Philos. Trans. R. Soc. B 366, 2426–2437. https://doi.org/10.1098/rstb.2011.0065.

Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Ben Rais Lasram, F., Aguzzi, J., Ballesteros, E., Bianchi, C.N., Corbera, J., Dailianis, T., Danovaro, R., Estrada, M., Froglia, C., Galil, B.S., Gasol, J.M., Gertwagen, R., Gil, J., Guilhaumon, F., Kesner-Reyes, K., Kitsos, M.S., Koukouras, A., Lampadariou, N., Laxamana, E., López-Fé de la Cuadra, C.M., Lotze, H.K., Martin, D., Mouillot, D., Oro, D., Raicevich, S., Rius-Barile, J., Saiz-Salinas, J.I., San Vicente, C., Somot, S., Templado, J., Turon, X., Vafidis, D., Villanueva, R., Voultsiadou, E., 2010. The biodiversity of the Mediterranean Sea: estimates, patterns, and threats. PLoS One 5, e11842. https://doi.org/10.1371/journal.pone.0011842.

Cordone, A., D'Errico, G., Magliulo, M., Bolinesi, F., Selci, M., Basili, M., Caruso, G., Mangoni, O., Saggiomo, M., Rivaro, P., Giordano, M., Saggiomo, V., 2022. Bacterioplankton diversity and distribution in relation to phytoplankton community structure in the ross sea surface waters. Front. Microbiol. 13, 722900. https://doi.org/10.3389/fmicb.2022.722900.

Cossarini, G., Feudale, L., Teruzzi, A., Bolzon, G., Coidessa, G., Solidoro, C., Di Biagio, V., Amadio, C., Lazzari, P., Brosich, A., Salon, S., 2021. High-resolution reanalysis of the

mediterranean sea biogeochemistry (1999-2019). Front. Mar. Sci. 8, 741486. https://doi.org/10.3389/fmars.2021.741486.
de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J. M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605. https://doi.org/10.1126/science.1261605.
Deutschmann, I.M., Krabberød, A.K., Latorre, F., Hyöky, V., Marotz, C., Tanber, C., Sundberg, P., Münch, S., Rahkonen, M., Özsezen, S., Kauska, R., Mendes, L.W., Lara, E., Krabberød, K.A., Logares, R., 2023. Disentangling temporal associations in marine microbial networks. Microbiome 11, 83. https://doi.org/10.1186/s40168-023-01523-z.
Doane, M.P., Ostrowski, M., Brown, M., Bramucci, A., Bodrossy, L., Van De Kamp, J., Kahlke, T., Seymour, J.R., 2023. Defining marine bacterioplankton community assembly rules by contrasting the importance of environmental determinants and biotic interactions. Environ. Microbiol. 25, 1084–1098. https://doi.org/10.1111/1462-2920.16341.
D'Ortenzio, F., Ribera d'Alcalà, M., 2009. On the trophic regimes of the mediterranean sea: a satellite analysis. Biogeosciences 6, 139–148. https://doi.org/10.5194/bg-6-139-2009.
Dussud, C., Meistertzheim, A.L., Conan, P., Pujo-Pay, M., George, M., Fabre, P., Couchet, N., Puissant, A., Stahl, A., Ghiglione, J.F., 2018. Evidence of niche partitioning among bacteria living on plastics, organic particles and surrounding seawaters. Environ. Pollut. 236, 807–816. https://doi.org/10.1016/j.envpol.2017.12.027.
El Hourany, R., Mejia, C., Faour, G., Crépon, M., Thiria, S., 2021. Evidencing the impact of climate change on the phytoplankton community of the mediterranean sea through a bioregionalization approach. J. Geophys. Res. Oceans 126, e2020JC016808. https://doi.org/10.1029/2020JC016808.
El Hourany, R., Pierella Karlusich, J., Zinger, L., Loisel, H., Levy, M., Bowler, C., 2024. Linking satellites to genes with machine learning to estimate phytoplankton community structure from space. Ocean Sci. 20, 217–239. https://doi.org/10.5194/os-20-217-2024.
Escudier, R., Clementi, E., Omar, M., Cipollone, A., Pistoia, J., Aydogdu, A., Drudi, M., Grandi, A., Lyubartsev, V., Lecci, R., Cretí, S., Masina, S., Coppini, G., Pinardi, N., 2020. Mediterranean Sea physical reanalysis (CMEMS MED-currents), version 1 [data set]. Copernicus Monitoring Environment Marine Service (CMEMS). https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_PHY_006_004_E3R1.
Escudier, R., Clementi, E., Cipollone, A., Pistoia, J., Drudi, M., Grandi, A., Lyubartsev, V., Lecci, R., Aydogdu, A., Delrosso, D., Omar, M., Masina, S., Coppini, G., Pinardi, N., 2021. A high-resolution reanalysis for the mediterranean sea. Front. Earth Sci. 9, 702285. https://doi.org/10.3389/feart.2021.702285.
Falkowski, P.G., Oliver, M.J., 2007. Mix and match: how climate selects phytoplankton. Nat. Rev. Microbiol. 5, 813–819. https://doi.org/10.1038/nrmicro1751.
Falkowski, P.G., Fenchel, T., Delong, E.F., 2008. The microbial engines that drive earth's biogeochemical cycles. Science 320, 1034–1039. https://doi.org/10.1126/science.1153213.
Fasola, E., Santolini, C., Villa, B., Zanoletti, A., Magni, G., Pachner, J., Stefani, F., Boldrocchi, G., Bettinetti, R., 2025. Integrating traditional and innovative monitoring approaches to monitor the marine biodiversity in the tyrrhenian sea (mediterranean sea). Mar. Environ. Res. 208, 107160. https://doi.org/10.1016/j.marenvres.2025.107160.
Feingersch, R., Suzuki, M.T., Shmoish, M., Sharon, I., Sabehi, G., Partensky, F., Béjà, O., 2010. Microbial community genomics in eastern mediterranean sea surface waters. ISME J. 4, 78–87. https://doi.org/10.1038/ismej.2009.92.
Feranchuk, S., Belkova, N., Potapova, U., Kuzmin, D., Belikov, S., 2018. Evaluating the use of diversity indices to distinguish between microbial communities with different traits. Res. Microbiol. 169, 254–261. https://doi.org/10.1016/j.resmic.2018.03.004.
Ferrera, I., Auladell, A., Balagué, V., Reñé, A., Garcés, E., Massana, R., Gasol, J.M., 2024. Seasonal and interannual variability of the free-living and particle-associated bacteria of a coastal microbiome. Environ. Microbiol. Rep. 16, e13299. https://doi.org/10.1111/1758-2229.13299.
Forsythe, W.C., Rykiel, E.J., Stahl, R.S., Wu, H., Schoolfield, R.M., 1995. A model comparison for daylength as a function of latitude and day of year. Ecol. Model. 80, 87–95. https://doi.org/10.1016/0304-3800(94)00034-F.
Fuhrman, J.A., Hewson, I., Schwalbach, M.S., Steele, J.A., Brown, M.V., Naeem, S., 2006. Annually reoccurring bacterial communities are predictable from ocean conditions. Proc. Natl. Acad. Sci. U. S. A. 103, 13104–13109. https://doi.org/10.1073/pnas.0602399103.
Fuhrman, J.A., Steele, J.A., Hewson, I., Schwalbach, M.S., Brown, M.V., Green, J.L., Brown, J.H., 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc. Natl. Acad. Sci. U. S. A. 105, 7774–7778. https://doi.org/10.1073/pnas.0803070105.
Fuhrman, J.A., Cram, J.A., Needham, D.M., 2015. Marine microbial community dynamics and their ecological interpretation. Nat. Rev. Microbiol. 13, 133–146. https://doi.org/10.1038/nrmicro3417.
Galletti, Y., Gonnelli, M., Retelletti Brogi, S., Vestri, S., Santinelli, C., 2019. DOM dynamics in open waters of the mediterranean sea: new insights from optical properties. Deep-Sea Res. I 144, 95–114. https://doi.org/10.1016/j.dsr.2019.01.007.
Gasol, J.M., Cardelús, C., Morán, X.A.G., Balagué, V., Forn, I., Marrasé, C., Massana, R., Pedrós-Alió, C., Sala, M.M., Simó, R., Vaqué, D., Estrada, M., 2016. Seasonal patterns in phytoplankton photosynthetic parameters and primary production at coastal NW mediterranean site. Sci. Mar. 80 (S1), 63–77. https://doi.org/10.3989/scimar.04480.06E.
Giering, S.L.C., Sanders, R., Lampitt, R.S., Anderson, T.R., Tamburini, C., Boutrif, M., Zubkov, M.V., Marsay, C.M., Henson, S.A., Saw, K., Cook, K., Mayor, D.J., 2014. Reconciliation of the carbon budget in the ocean's twilight zone. Nature 507, 480–483. https://doi.org/10.1038/nature13123.
Gilbert, J.A., Somerfield, P.J., Temperton, B., Huse, S., Joint, I., Field, D., 2010. Day-length is central to maintaining consistent seasonal diversity in marine bacterioplankton. Nat. Preced. https://doi.org/10.1038/npre.2010.4406.1.
Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbrück, L., Reeder, J., Temperton, B., Huse, S., McHardy, A.C., Knight, R., Joint, I., Somerfield, P., Fuhrman, J.A., Field, D., 2012. Defining seasonal marine microbial community dynamics. ISME J. 6, 298–308. https://doi.org/10.1038/ismej.2011.107.
Glibert, P.M., Wilkerson, F.P., Dugdale, R.C., Raven, J.A., Dupont, C.L., Leavitt, P.R., Parker, A.E., Burkholder, J.M., Kana, T.M., 2016. Pluses and minuses of ammonium and nitrate uptake and assimilation by phytoplankton and implications for productivity and community composition, with emphasis on nitrogen-enriched conditions. Limnol. Oceanogr. 61, 165–197. https://doi.org/10.1002/lno.10203.
Goldberg, S.J., Nelson, C.E., Viviani, D.A., Shulse, C.N., Church, M.J., 2017. Cascading influence of inorganic nitrogen sources on DOM production, composition, lability and microbial community structure in the open ocean. Environ. Microbiol. 19, 3450–3464. https://doi.org/10.1111/1462-2920.13825.
Gomes, A., Gasol, J.M., Estrada, M., Franco-Vidal, L., Díaz-Pérez, L., Ferrera, I., Morán, X. A.G., 2015. Heterotrophic bacterial responses to the winter-spring phytoplankton bloom in open waters of the NW mediterranean. Deep-Sea Res. Part I 96, 59–68. https://doi.org/10.1016/j.dsr.2014.11.007.
Gray, P.C., Boss, E., Prochaska, J.X., Kerner, H., Demeaux, C.B., Lehahn, Y., 2024. The promise and pitfalls of machine learning in ocean remote sensing. Oceanography 37. https://doi.org/10.5670/oceanog.2024.511.
Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., Weitz, J.S., 2013. Robust estimation of microbial diversity in theory and in practice. ISME J. 7, 1092–1101. https://doi.org/10.1038/ismej.2013.10.
Hays, G.C., Richardson, A.J., Robinson, C., 2005. Climate change and marine plankton. Trends Ecol. Evol. 20 (6), 337–344. https://doi.org/10.1016/j.tree.2005.03.004.
He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239.
Heneghan, R.F., Holloway-Brown, J., Gasol, J.M., Herndl, G.J., Morán, X.A.G., Galbraith, E.D., 2024. The global distribution and climate resilience of marine heterotrophic prokaryotes. Nat. Commun. 15 (1), 6943. https://doi.org/10.1038/s41467-024-50635-z.
Henson, S.A., Cael, B.B., Allen, S.R., Dutkiewicz, S., 2021. Future phytoplankton diversity in a changing climate. Nat. Commun. 12, 5372. https://doi.org/10.1038/s41467-021-25699-w.
Holland, M.M., Artigas, L.F., Atkinson, A., Best, M., Bresnan, E., Devlin, M., Eerkes-Medrano, D., Johansen, M., Johns, D.G., Machairopoulou, M., Pitois, S., Scott, J., Schilder, J., Stern, R., Tait, K., Whyte, C., Widdicombe, C., McQuatters-Gollop, A., 2025. Mind the gap - the need to integrate novel plankton methods alongside ongoing long-term monitoring. Ocean Coast. Manag. 262, 107542. https://doi.org/10.1016/j.ocecoaman.2025.107542.
Hollmann, R., Merchant, C.J., Saunders, R., Downy, C., Buchwitz, M., Cazenave, A., Chuvieco, E., Defourny, P., de Leeuw, G., Forsberg, R., Holzer-Popp, T., Paul, F., Sandven, S., Sathyendranath, S., van Roozendael, M., Wagner, W., 2013. The ESA climate change initiative: satellite data records for essential climate variables. Bull. Am. Meteorol. Soc. 94, 1541–1552. https://doi.org/10.1175/BAMS-D-11-00254.1.
Huot, Y., Babin, M., Bruyant, F., Grob, C., Twardowski, M.S., Claustre, H., 2007. Relationship between photosynthetic parameters and proxies of phytoplankton biomass in the subtropical ocean. Biogeosciences 4, 853–868. https://doi.org/10.5194/bg-4-853-2007.
Ibarbalz, F.M., Henry, N., Brandão, M.C., Martini, S., Busseni, G., Byrne, H., Coelho, L.P., Endo, H., Gasol, J.M., Gregory, A.C., Mahé, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Sáez, I., Scalco, E., Soviadan, D., Zayed, A.A., Zingone, A., Labadie, K., Ferland, J., Marec, C., Kandels, S., Picheral, M., Dimier, C., Poulain, J., Pisarev, S., Carmichael, M., Pesant, S., Babin, M., Boss, E., Iudicone, D., Jaillon, O., Acinas, S.G., Ogata, H., Pelletier, E., Stemmann, L., Sullivan, M.B., Sunagawa, S., Bopp, L., de Vargas, C., Karp-Boss, L., Wincker, P., Lombard, F., Bowler, C., Zinger, L., 2019. Global trends in marine plankton diversity across kingdoms of life. Cell 179, 1084–1097. https://doi.org/10.1016/j.cell.2019.10.008.
Irigoien, X., Huisman, J., Harris, R.P., 2004. Global biodiversity patterns of marine phytoplankton and zooplankton. Nature 429, 863–867. https://doi.org/10.1038/nature02593.
Jiao, N., Luo, T., Chen, Q., Zhao, Z., Xiao, X., Liu, J., Jian, Z., Xie, S., Thomas, H., Herndl, G.J., Benner, R., Gonsior, M., Chen, F., Cai, W.-J., Robinson, C., 2024. The microbial carbon pump and climate change. Nat. Rev. Microbiol. 22, 408–419. https://doi.org/10.1038/s41579-024-01018-0.
Jung, Y., 2018. Multiple predicting k-fold cross-validation for model selection. Journal of Nonparametric Statistics 30 (1), 197–215. https://doi.org/10.1080/10485252.2017.1404598.
Junger, P.C., Sarmento, H., Giner, C.R., Mestre, M., Sebastián, M., Morán, X., Arístegui, J., Agustí, S., Duarte, C.M., Acinas, S.G., Massana, R., Gasol, J.M., Logares, R., 2023. Global biogeography of the smallest plankton across ocean depths. Sci. Adv. 9, eadg9763. https://doi.org/10.1126/sciadv.adg9763.
Junger, P.C., Kavagutti, V.S., Deutschmann, I.M., Gazulla, C.R., Huber, P., Menezes, M., Paranhos, R., Amado, A.M., Ferrera, I., Rigonato, J., Chaffron, S., Gasol, J.M.,

Logares, R., Sarmento, H., 2026. Ecological processes shaping marine microbial assemblages diverge between equatorial and temperate time-series. Mol. Ecol. 35, e70241. https://doi.org/10.1111/mec.70241.
Kim, D., Lee, K., Jeong, S., Song, M., Kim, B., Park, J., Heo, T-Y., 2024. Real-time chlorophyll-a forecasting using machine learning framework with dimension reduction and hyperspectral data. Environ. Res. 262, 119823. Doi: https://doi.org/10.1016/j.envres.2024.119823.
Kinsey, J.D., Corradino, G., Ziervogel, K., Schnetzer, A., Osburn, C.L., 2018. Formation of CDOM by bacterial degradation of phytoplankton-derived aggregates. Front. Mar. Sci. 4, 430. https://doi.org/10.3389/fmars.2017.00430.
Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick, R.P., Isaac, N.J.B., Kelling, S., Los, W., McRae, L., Mihoub, J.-B., Obst, M., Santamaria, M., Skidmore, A.K., Williams, K.J., Agosti, D., Amariles, D., Arvanitidis, C., Bastin, L., De Leo, F., Egloff, W., Elith, J., Hobern, D., Martin, D., Pereira, H.M., Pesole, G., Peterseil, J., Saarenmaa, H., Schigel, D., Schmeller, D.S., Segata, N., Turak, E., Uhlir, P.F., Wee, B., Hardisty, A.R., 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biol. Rev. 93, 600–625. https://doi.org/10.1111/brv.12359.
Klawonn, I., Bonaglia, S., Whitehouse, M.J., Littmann, S., Tienken, D., Kuypers, M.M.M., Brüchert, V., Ploug, H., 2019. Untangling hidden nutrient dynamics: rapid ammonium cycling and single-cell ammonium assimilation in marine plankton communities. ISME J. 13, 1960–1974. https://doi.org/10.1038/s41396-019-0386-z.
Kotta, D., Kitsiou, D., 2019. Chlorophyll in the eastern mediterranean sea: correlations with environmental factors and trends. Environments 6 (8), 98. https://doi.org/10.3390/environments6080098.
Krabberød, A.K., Deutschmann, I.M., Bjorbækmo, M.F.M., Jakobsen, K.S., Jardillier, L., Kormas, K., Maloy, A.P., Metfies, K., Morán, X.A.G., Richards, T.A., Yau, S., Logares, R., 2022. Long-term patterns of an interconnected core marine microbiota. Environ. Microbiome 17, 22. https://doi.org/10.1186/s40793-022-00417-1.
Krom, M.D., Emeis, K.-C., Van Cappellen, P., 2010. Why is the eastern mediterranean phosphorus limited? Prog. Oceanogr. 85, 236–244. https://doi.org/10.1016/j.pocean.2010.03.003.
Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminform. 6, 10. https://doi.org/10.1186/1758-2946-6-10.
Kuhn, M., 2008. Building predictive models in r using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05.
Ladau, J., Sharpton, T.J., Finucane, M.M., Jospin, G., Kembel, S.W., O'Dwyer, J., Koeppel, A.F., Green, J.L., Pollard, K.S., 2013. Global marine bacterial diversity peaks at high latitudes in winter. ISME J. 7, 1669–1677. https://doi.org/10.1038/ismej.2013.37.
Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., Galand, P.E., 2019. Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. ISME J. 13, 388–401. https://doi.org/10.1038/s41396-018-0281-z.
Lambert, S., Lozano, J.-C., Bouget, F.-Y., Galand, P.E., 2021. Seasonal marine microorganisms change neighbours under contrasting environmental conditions. Environ. Microbiol. 23, 2592–2604. https://doi.org/10.1111/1462-2920.15482.
Lami, R., Cuperová, Z., Ras, J., Lebaron, P., Koblízek, M., 2009. Distribution of free-living and particle-attached aerobic anoxygenic phototrophic bacteria in marine environments. Aquat. Microb. Ecol. 55, 31–38. https://doi.org/10.3354/ame01282.
Landa, M., Blain, S., Christaki, U., Monchy, S., Obernosterer, I., 2016. Shifts in bacterial community composition associated with increased carbon cycling in a mosaic of phytoplankton blooms. ISME J. 10, 39–50. https://doi.org/10.1038/ismej.2015.105.
Landolfi, A., Prowe, A.E.F., Pahlow, M., Somes, C.J., Chien, C.-T., Schartau, M., Koeve, W., Oschlies, A., 2021. Can top-down controls expand the ecological niche of marine n_2 fixers? Front. Microbiol. 12, 690200. https://doi.org/10.3389/fmicb.2021.690200.
Lavigne, H., D'Ortenzio, F., Migon, C., Claustre, H., Testor, P., d'Alcalà, M.R., Lavezza, R., Houpert, L., Prieur, L., 2013. Enhancing the comprehension of mixed layer depth control on the mediterranean phytoplankton phenology. J. Geophys. Res. Oceans 118, 3416–3430. https://doi.org/10.1002/jgrc.20251.
Lazzari, P., Solidoro, C., Ibello, V., Salon, S., Teruzzi, A., Béranger, K., Colella, S., Crise, A., 2012. Seasonal and inter-annual variability of plankton chlorophyll and primary production in the mediterranean sea: a modelling approach. Biogeosciences 9, 217–233. https://doi.org/10.5194/bg-9-217-2012.
Lazzari, P., Solidoro, C., Salon, S., Bolzon, G., 2016. Spatial variability of phosphate and nitrate in the mediterranean sea: a modeling approach. Deep Sea Res. Part I Oceanogr. Res. Pap. 108, 39–52. Doi: https://doi.org/10.1016/j.dsr.2015.12.006.
Lazzari, P., Álvarez, E., Terzić, E., Cossarini, G., Chernov, I., D'Ortenzio, F., Organelli, E., 2021. CDOM spatiotemporal variability in the mediterranean sea: a modelling study. J. Mar. Sci. Eng. 9, 176. https://doi.org/10.3390/jmse9020176.
Lefort, T., Gasol, J.M., 2013. Global-scale distributions of marine surface bacterioplankton groups along gradients of salinity, temperature, and chlorophyll: a meta-analysis of fluorescence in situ hybridization studies. Aquat. Microb. Ecol. 70, 111–130. https://doi.org/10.3354/ame01643.
Levine, N.M., Alexander, H., Bertrand, E.M., Coles, V.J., Dutkiewicz, S., Leles, S.G., Zakem, E.J., 2025. Microbial ecology to ocean carbon cycling: from genomes to numerical models. Annu. Rev. Earth Planet. Sci. 53, 595–624. https://doi.org/10.1146/annurev-earth-040523-020630.
Li, M., Organelli, E., Serva, F., Bellacicco, M., Landolfi, A., Pisano, A., Marullo, S., Shen, F., Mignot, A., van Gennip, S., Santoleri, R., 2024. Phytoplankton spring bloom inhibited by marine heatwaves in the North-Western mediterranean sea. Geophys. Res. Lett. 51, e2024GL109141. https://doi.org/10.1029/2024GL109141.
Logares, R., Deutschmann, I.M., Junger, P.C., Giner, C.R., Krabberød, A.K., Schmidt, T.S. B., Rubinat-Ripoll, L., Mestre, M., Salazar, G., Ruiz-González, C., Sebastián, M., de Vargas, C., Acinas, S.G., Duarte, C.M., Gasol, J.M., Massana, R., 2020. Disentangling the mechanisms shaping the surface ocean microbiota. Microbiome 8, 55. https://doi.org/10.1186/s40168-020-00827-8.
Malfatti, F., Turk, V., Tinta, T., Mozetič, P., Manganelli, M., Samo, T.J., Ugalde, J.A., Kovač, N., Stefanelli, M., Antonioli, M., Fonda Umani, S., Del Negro, P., Cataletto, B., Hozić, A., Ivošević DeNardis, N., Žutić, V., Svetličić, V., Mišić Radić, Tea, Radić, T., Fuks, D., Azam, F., 2014. Microbial mechanisms coupling carbon and phosphorus cycles in phosphorus-limited northern adriatic sea. Sci. Total Environ. 470-471, 1173–1183. https://doi.org/10.1016/j.scitotenv.2013.10.040.
Marchese, C., 2015. Biodiversity hotspots: a shortcut for a more complicated concept. Glob. Ecol. Conserv. 3, 297–309. https://doi.org/10.1016/j.gecco.2014.12.008.
Marchese, C., Lazzara, L., Pieri, M., Massi, L., Nuccio, C., Santini, C., Maselli F., 2015. Analysis of chlorophyll-a and primary production dynamics in North Tyrrhenian and Ligurian coastal–neritic and oceanic waters. J. Coast. Res. 31 (3), 690–701. https://doi.org/10.2112/JCOASTRES-D-13-00210.1.
Marchese, C., Colella, S., Brando, V.E., Zoffoli, M.L., Volpe, G., 2024. Towards accurate l4 ocean colour products: interpolating remote sensing reflectance via DINEOF. Int. J. Appl. Earth Obs. Geoinf. 135, 104270. https://doi.org/10.1016/j.jag.2024.104270.
Marullo, S., Serva, F., Iacono, R., Napolitano, E., Di Sarra, A., Meloni, D., Monteleone, F., Sferlazzo, D., De Silvestri, L., De Toma, V., Pisano, A., Bellacicco, M., Landolfi, A., Organelli, E., Yang, C., Santoleri, R., 2023. Record-breaking persistence of the 2022/23 marine heatwave in the mediterranean sea. Environ. Res. Lett. 18, 114041. https://doi.org/10.1088/1748-9326/ad02ae.
Massi, L., Frittitta, L., Melillo, C., Polonelli, F., Bianchi, V., De Biasi, A.M., Nuccio, C., 2020. Seasonal dynamic of CDOM in a shelf site of the south-eastern ligurian sea (western mediterranean). J. Mar. Sci. Eng. 8, 703. https://doi.org/10.3390/jmse8090703.
McDonald, K.S., Turk, V., Mozetič, P., Tinta, T., Malfatti, F., Hannah, D.M., Krause, S., 2017. Integrated network models for predicting ecological thresholds: microbial-carbon interactions in coastal marine systems. Environ. Model. Software 91, 156–167. https://doi.org/10.1016/j.envsoft.2017.01.017.
Merbt, S.N., Stahl, D.A., Casamayor, E.O., Martí, E., Nicol, G.W., Prosser, J.I., 2012. Differential photoinhibition of bacterial and archaeal ammonia oxidation. FEMS Microbiol. Lett. 327, 41–46. https://doi.org/10.1111/j.1574-6968.2011.02457.x.
Meunier, C.L., Schmidt, J., Ahme, A., Balkoni, A., Berg, K., Blum, L., Boersma, M., Brüwer, J.D., Fuchs, B.M., Gimenez, L., Guignard, M., Schulte-Hillen, R., Krock, B., Rick, J., Stibor, H., Stockenreiter, M., Tulatz, S., Weber, F., Wichels, A., Wiltshire, K. H., Wohlrab, S., Kirstein, I.V., 2025. Plankton communities today and tomorrow-potential impacts of multiple global change drivers and marine heatwaves. Limnol. Oceanogr. 70, S225–S241. https://doi.org/10.1002/lno.70042.
Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650.
Miloslavich, P., Bax, N.J., Simmons, S.E., Klein, E., Appeltans, W., Aburto-Oropeza, O., Andersen Garcia, M., Batten, S.D., Benedetti-Cecchi, L., Checkley Jr., D.M., Chiba, S., Duffy, J.E., Dunn, D.C., Fischer, A., Gunn, J., Kudela, R., Marsac, F., Muller-Karger, F.E., Obura, D., Shin, Y.-J., 2018. Essential Ocean variables for global sustained observations of biodiversity and ecosystem changes. Glob. Chang. Biol. 24, 2416–2433. https://doi.org/10.1111/gcb.14108.
Mitra, B., Tiwari, S.P., Uddin, M.S., Mahmud, K., 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. https://doi.org/10.1016/j.marpolbul.2023.115945.
Morel, A., Gentili, B., 2009a. A simple band ratio technique to quantify the colored dissolved and detrital organic material from ocean color remotely sensed data. Remote Sens. Environ. 113, 998–1011. https://doi.org/10.1016/j.rse.2009.01.008.
Morel, A., Gentili, B., 2009b. The dissolved yellow substance and the shades of blue in the mediterranean sea. Biogeosciences 6, 2625–2636. https://doi.org/10.5194/bg-6-2625-2009.
Needham, D.M., Fuhrman, J.A., 2016. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. Nat. Microbiol. 1, 16005. https://doi.org/10.1038/nmicrobiol.2016.5.
Needham, D.M., Chow, C.-E.T., Cram, J.A., Sachdeva, R., Parada, A., Fuhrman, J.A., 2013. Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. ISME J. 7, 1274–1285. https://doi.org/10.1038/ismej.2013.19.
Nelson, C.E., Wear, E.K., 2014. Microbial diversity and the lability of dissolved organic carbon. Proc. Natl. Acad. Sci. U. S. A. 111 (20), 7166–7167. https://doi.org/10.1073/pnas.1405751111.
Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Borman, T., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H.B.A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., Martino, C., McGlinn, D., Ouellette, M.-H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C.J.F., Weedon, J., 2025.vegan: Community Ecology Package. R package version 2.7–1. https://CRAN.R-project.org/package=vegan.
Olofsson, M., Robertson, E.K., Edler, L., Arneborg, L., Whitehouse, M.J., Ploug, H., 2019. Nitrate and ammonium fluxes to diatoms and dinoflagellates at a single cell level in mixed field communities in the sea. Sci. Rep. 9, 1424. https://doi.org/10.1038/s41598-018-38059-4.
Orel, N., Fadeev, E., Klun, K., Ličer, M., Tinta, T., Turk, V., 2022. Bacterial indicators are ubiquitous members of pelagic microbiome in anthropogenically impacted coastal

ecosystem. Front. Microbiol. 12, 765091. https://doi.org/10.3389/fmicb.2021.765091.
Orel, N., Fadeev, E., Celussi, M., Turk, V., Klun, K., Afjehi-Sadat, L., Herndl, G.J., Tinta, T., 2026. Down the drain: exploring wastewaters role in coastal microbiome transformations. Microbiome 14, 46. https://doi.org/10.1186/s40168-025-02298-1.
Organelli, E., Bricaud, A., Antoine, D., Matsuoka, A., 2014. Seasonal dynamics of light absorption by chromophoric dissolved organic matter (CDOM) in the NW mediterranean sea (BOUSSOLE site). Deep Sea Res. Part I Oceanogr. Res. Pap. 91, 72–85. https://doi.org/10.1016/j.dsr.2014.05.003.
Organelli, E., Bricaud, A., Gentili, B., Antoine, D., Vellucci, V., 2016. Retrieval of colored detrital matter (CDM) light absorption coefficients in the mediterranean sea using field and satellite ocean color radiometry: evaluation of bio-optical inversion models. Remote Sens. Environ. 186, 297–310. https://doi.org/10.1016/j.rse.2016.08.028.
Panaïotis, T., Wilson, J., Cael, B., 2025. A machine learning-based dissolved organic carbon climatology. Geophys. Res. Lett. 52, e2024GL112792. https://doi.org/10.1029/2024GL112792.
Para, J., Coble, P.G., Charrière, B., Tedetti, M., Fontana, C., Sempéré, R., 2010. Fluorescence and absorption properties of chromophoric dissolved organic matter (CDOM) in coastal surface waters of the northwestern mediterranean sea, influence of the rhône river. Biogeosciences 7, 4083–4103. https://doi.org/10.5194/bg-7-4083-2010.
Parada, A.E., Needham, D.M., Fuhrman, J.A., 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples environ. Microbiol 18, 1403–1414. https://doi.org/10.1111/1462-2920.13023.
Pascoal, F., Tomasino, M.P., Piredda, R., Quero, G.M., Torgo, L., Poulain, J., Galand, P.E., Fuhrman, J.A., Mitchell, A., Tinta, T., Turk Dermastia, T., Fernandez-Guerra, A., Vezzi, A., Logares, R., Malfatti, F., Endo, H., Dąbrowska, A.M., De Pascale, F., Sánchez, P., Henry, N., Fosso, B., Wilson, B., Toshchakov, S., Ferrant, G.K., Grigorov, I., Vieira, F.R.J., Costa, R., Pesant, S., Magalhães, C., 2023. Inter-comparison of marine microbiome sampling protocols. ISME Commun. 3, 84. https://doi.org/10.1038/s43705-023-00278-w.
Pinhassi, J., Gómez-Consarnau, L., Alonso-Sáez, L., Sala, M.M., Vidal, M., Pedrós-Alió, C., Gasol, J.M., 2006. Seasonal changes in bacterioplankton nutrient limitation and effects on community composition in the NW mediterranean sea. Aquat. Microb. Ecol. 44, 241–252. https://doi.org/10.3354/ame044241.
Polovina, J.J., Howell, E.A., 2005. Ecosystem indicators derived from satellite oceanographic data for the north pacific. ICES J. Mar. Sci. 62, 319–327. https://doi.org/10.1016/j.icesjms.2004.07.031.
Pomeroy, L.R., Williams, P.J.leB., Azam, F., Hobbie, J.E., 2007. The microbial loop. Oceanography 20 (2), 28–33. https://doi.org/10.5670/oceanog.2007.45.
Pommier, T., Neal, P.R., Gasol, J.M., Coll, M., Acinas, S.G., Pedrós-Alió, C., 2010. Spatial patterns of bacterial richness and evenness in the NW mediterranean sea explored by pyrosequencing of the 16S rRNA. Aquat. Microb. Ecol. 61, 221–233. https://doi.org/10.3354/ame01484.
Pujo-Pay, M., Conan, P., Joux, F., Oriol, L., Naudin, J.J., Cauwet, G., 2006. Impact of phytoplankton and bacterial production on nutrient and DOM uptake in the rhône river plume (NW mediterranean). Mar. Ecol. Prog. Ser. 315, 43–54. https://doi.org/10.3354/meps315043.
Pulido-Villena, E., Baudoux, A.-C., Obernosterer, I., Landa, M., Caparros, J., Catala, P., Georges, C., Harmand, J., Guieu, C., 2014. Microbial food web dynamics in response to a saharan dust event: mesocosm study in the oligotrophic mediterranean sea. Biogeosciences 11, 5607–5619. https://doi.org/10.5194/bg-11-5607-2014.
Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596. https://doi.org/10.1093/nar/gks1219.
R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Racault, M.-F., Platt, T., Sathyendranath, S., Ağirbaş, E., Martinez Vicente, V., Brewin, R., 2014. Plankton indicators and ocean observing systems: support to marine ecosystem state assessment. J. Plankton Res. 36, 621–629. https://doi.org/10.1093/plankt/fbu016.
Raes, E.J., Myles, S., MacNeil, L., Wietz, M., Bienhold, C., Tait, K., Somerfield, P.J., Bissett, A., van de Kamp, J., Gasol, J.M., Massana, R., Yeh, Y.-C., Fuhrman, J.A., LaRoche, J., 2024. Seasonal patterns of microbial diversity across the world oceans. Limnol. Oceanogr. Lett. 9, 512–523. https://doi.org/10.1002/lol2.10422.
Ramond, P., Galand, P.E., Logares, R., 2025. Microbial functional diversity and redundancy: moving forward. FEMS Microbiol. Rev. 49, fuae031. https://doi.org/10.1093/femsre/fuae031.
Ribera d'Alcalà, M., Civitarese, G., Conversano, F., Lavezza, R., 2003. Nutrient ratios and fluxes hint at overlooked processes in the mediterranean sea. J. Geophys. Res. 108, 8106. https://doi.org/10.1029/2002JC001650.
Romera-Castillo, C., Sarmento, H., Álvarez-Salgado, X.A., Gasol, J.M., Marrasé, C., 2011. Net production and consumption of fluorescent CDOM by bacteria growing on phytoplankton exudates. Appl. Environ. Microbiol. 77, 7490–7498. https://doi.org/10.1128/AEM.00200-11.
Rubbens, P., Brodie, S., Cordier, T., Destro Barcellos, D., Devos, P., Fernandes-Salvador, J.A., Fincham, J.I., Gomes, A., Handegard, N.O., Howell, K., Jamet, C., Kartveit, K.H., Moustahfid, H., Parcerisas, C., Politikos, D., Sauzède, R., Sokolova, M., Uusitalo, L., Van den Bulcke, L., van Helmond, A.T.M., Watson, J.T., Welch, H., Beltran-Perez, O., Chaffron, S., Greenberg, D.S., Kühn, B., Kiko, R., Lo, M., Lopes, R.M., Möller, K.O., Michaels, W., Pala, A., Romagnan, J.-B., Schuchert, P., Seydi, V., Villasante, S., Malde, K., Irisson, J.-O., 2023. Machine learning in marine ecology: an overview of techniques and applications. ICES J. Mar. Sci. 80, 1829–1853. https://doi.org/10.1093/icesjms/fsad100.
Santinelli, C., 2015. DOC in the mediterranean sea. In: Hansell, D.A., Carlson, C.A. (Eds.), Biogeochemistry of Marine Dissolved Organic Matter, 2nd ed. Elsevier, Amsterdam, pp. 579–608.
Sarmento, H., Gasol, J.M., 2012. Use of phytoplankton-derived dissolved organic carbon by different types of bacterioplankton. Environ. Microbiol. 14 (9), 2348–2360. https://doi.org/10.1111/j.1462-2920.2012.02787.x.
Sebastián, M., Gasol, J.M., 2013. Heterogeneity in the nutrient limitation of different bacterioplankton groups in the eastern mediterranean sea. ISME J. 7, 1665–1668. https://doi.org/10.1038/ismej.2013.42.
Sebastián, M., Ortega-Retuerta, E., Gómez-Consarnau, L., Zamanillo, M., Álvarez, M., Arístegui, J., Gasol, J.M., 2021. Environmental gradients and physical barriers drive the basin-wide spatial structuring of mediterranean sea and adjacent eastern Atlantic Ocean prokaryotic communities. Limnol. Oceanogr. 66, 4077–4095. https://doi.org/10.1002/lno.11944.
Segura-Noguera, M., Cruzado, A., Blasco, D., 2016. The biogeochemistry of nutrients, dissolved oxygen and chlorophyll a in the catalan sea (NW mediterranean sea). Sci. Mar. 80 (S1), 39–56. https://doi.org/10.3989/scimar.04309.20A.
Seymour, J.R., Amin, S.A., Raina, J.B., Stocker, R., 2017. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. Nat. Microbiol. 2, 17065. https://doi.org/10.1038/nmicrobiol.2017.65.
Shilova, I.N., Mills, M.M., Robidart, J.C., Turk-Kubo, K.A., Björkman, K.M., Kolber, Z., Rapp, I., Van Mooy, B.A.S., Church, M.J., Zehr, J.P., 2017. Differential effects of nitrate, ammonium, and urea as n sources for microbial communities in the north pacific ocean. Limnol. Oceanogr. 62, 2550–2574. https://doi.org/10.1002/lno.10590.
Siokou-Frangou, I., Christaki, U., Mazzocchi, M.G., Montresor, M., Ribera d'Alcalà, M., Vaqué, D., Zingone, A., 2010. Plankton in the open mediterranean sea: a review. Biogeosciences 7, 1543–1586. https://doi.org/10.5194/bg-7-1543-2010.
Song, J., Jiang, W., Xin, L., 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. https://doi.org/10.1016/j.ecss.2024.108820.
Storto, A., Alvera-Azcárate, A., Balmaseda, M.A., Barth, A., Chevallier, M., Counillon, F., Domingues, C.M., Drevillon, M., Drillet, Y., Forget, G., Garric, G., Haines, K., Hernandez, F., Iovino, D., Jackson, L.C., Lellouche, J.M., Masina, S., Mayer, M., Oke, P.R., Penny, S.G., Peterson, K.A., Yang, C., Zuo, H., 2019. Ocean reanalyses: recent advances and unsolved challenges. Front. Mar. Sci. 6, 418. https://doi.org/10.3389/fmars.2019.00418.
Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P., 2015. Structure and function of the global ocean microbiome. Science 348, 1261359. https://doi.org/10.1126/science.1261359.
Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., Kassabgy, M., Huang, S., Mann, A.J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F. D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K.H., Glöckner, F.O., Schweder, T., Amann, R., 2012. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 336, 608–611. https://doi.org/10.1126/science.1218344.
Teruzzi, A., Di Biagio, V., Feudale, L., Bolzon, G., Lazzari, P., Salon, S., Coidessa, G., Cossarini, G., 2021a. Mediterranean Sea biogeochemical reanalysis (CMEMS MED-biogeochemistry, MedBFM3 system), version 1 [data set]. Copernicus Monitoring Environment Marine Service (CMEMS). https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_BGC_006_008_MEDBFM3.
Teruzzi, A., Feudale, L., Bolzon, G., Lazzari, P., Salon, S., Di Biagio, V., Coidessa, G., Cossarini, G., 2021b. Mediterranean Sea biogeochemical reanalysis INTERIM (CMEMS MED-biogeochemistry, MedBFM3i system), version 1 [data set]. Copernicus Monitoring Environment Marine Service (CMEMS). https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_BGC_006_008_MEDBFM3I.
Thompson, F.L., Bruce, T., Gonzalez, A., Cardoso, A., Clementino, M., Costagliola, M., Hozbor, C., Otero, E., Piccini, C., Peressutti, S., Schmieder, R., Edwards, R., Smith, M., Takiyama, L.R., Vieira, R., Paranhos, R., Artigas, L.F., 2011. Coastal bacterioplankton community diversity along a latitudinal gradient in latin america by means of v6 tag pyrosequencing. Arch. Microbiol. 193, 105–114. https://doi.org/10.1007/s00203-010-0644-y.
Tinta, T., Vojvoda, J., Mozetič, P., Talaber, I., Vodopivec, M., Malfatti, F., Turk, V., 2015. Bacterial community shift induced by dynamic environmental parameters in a changing coastal ecosystem (northern adriatic): a 2-year time series. Environ. Microbiol. 17, 3581–3596. https://doi.org/10.1111/1462-2920.12519.
Trano, A.C., Casotti, R., Raffini, F., Piredda, R., 2024. 16S amplicon sequence variants (ASVs) data of NEREA augmented observatory. Zenodo (Dataset). https://doi.org/10.5281/zenodo.12801913.
Tsamardinos, I., Greasidou, E., Borboudakis, G., 2018. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach. Learn. 107, 1895–1922. https://doi.org/10.1007/s10994-018-5714-4.
Vignudelli, S., Santinelli, C., Murru, E., Nannicini, L., Seritti, A., 2004. Distributions of dissolved organic carbon (DOC) and chromophoric dissolved organic matter (CDOM) in coastal waters of the northern tyrrhenian sea (Italy). Estuar. Coast. Shelf Sci. 60, 133–149. https://doi.org/10.1016/j.ecss.2003.11.023.

Volpe, G., Colella, S., Brando, V.E., Forneris, V., La Padula, F., Di Cicco, A., Sammartino, M., Bracaglia, M., Artuso, F., Santoleri, R., 2019. Mediterranean Ocean colour level 3 operational multi-sensor processing. Ocean Sci. 15, 127–146. https://doi.org/10.5194/os-15-127-2019.

Xing, X., Claustre, H., Wang, H., Poteau, A., D'Ortenzio, F., 2014. Seasonal dynamics in colored dissolved organic matter in the mediterranean sea: patterns and drivers. Deep Sea Res. Part 1 Oceanogr. Res. Pap. 83, 93–101. https://doi.org/10.1016/j.dsr.2013.09.008.

Yeh, Y.-C., Fuhrman, J.A., 2022. Effects of phytoplankton, viral communities, and warming on free-living and particle-associated marine prokaryotic community structure. Nat. Commun. 13, 7905. https://doi.org/10.1038/s41467-022-35551-4.

Zhang, Z., Chen, P., Zhang, S., Huang, H., Pan, Y., Pan, D., 2025. A review of machine learning applications in ocean color remote sensing. Remote Sens. 17, 1776. https://doi.org/10.3390/rs17101776.

Zhu, L., Cui, T., Runa, A., Pan, X., Zhao, W., Xiang, J., Cao, M., 2024. Robust remote sensing retrieval of key eutrophication indicators in coastal waters based on explainable machine learning. ISPRS J. Photogramm. Remote Sens. 211, 262–280. https://doi.org/10.1016/j.isprsjprs.2024.04.007.

Zoffoli, M.L., Brando, V.E., Volpe, G., Vilas, L.G., Davies, B.F.R., Frouin, R., Pitarch, J., Oiry, S., Tan, J., Colella, S., Marchese, C., 2025. CIAO: a machine-learning algorithm for mapping arctic ocean chlorophyll-a from space. Sci. Remote Sens. 11, 100212. https://doi.org/10.1016/j.srs.2025.100212.

Zubkov, M.V., Fuchs, B.M., Tarran, G.A., Burkill, P.H., Amann, R., 2003. High rate of uptake of organic nitrogen compounds by prochlorococcus cyanobacteria as a key to their dominance in oligotrophic oceanic waters. Appl. Environ. Microbiol. 69, 1299–1304. https://doi.org/10.1128/AEM.69.2.1299-1304.2003.