coasts

MDPI

Article

Disentangling Environmental Drivers of Marine Primary Production Using Explainable Machine Learning

Teodoro Semeraro [1,2,*], Jessica Titocci [2,*], Lorenzo Liberatore [2], Antonio Calisi [3], Monica De Caroli [1,4] and Alberto Basset [1,4]

[1] Department of Biological and Environmental Sciences and Technologies, University of Salento, via Monteroni 165, 73100 Lecce, Italy; monica.decaroli@unisalento.it (M.D.C.); alberto.basset@unisalento.it (A.B.)
[2] Research Institute on Terrestrial Ecosystems (IRET), National Research Council of Italy (CNR), Campus Ecotekne, 73100 Lecce, Italy; lorenzo.liberatore@cnr.it (L.L.)
[3] Department of Science and Technological Innovation, University of Eastern Piedmont, viale Michel 11, I-15121 Alessandria, Italy; antonio.calisi@uniupo.it (A.C.)
[4] NBCF National Biodiversity Future Center, 90133 Palermo, Italy
* Correspondence: teodoro.semeraro@unisalento.it (T.S.); jessica.titocci@cnr.it (J.T.)

Abstract

Marine primary production is regulated by interacting environmental processes, making it challenging to distinguish biomass-associated variability from changes in production independent of phytoplankton biomass. We combined Random Forest models with SHAP-based explainable artificial intelligence in comparison with linear regression model to investigate the influence of environmental drivers on chlorophyll-*a* (Chl-a, used as a proxy of phytoplankton biomass) and primary production (PP) in a coastal marine ecosystem. The analysis was based on 2201 observations collected during 314 sampling campaigns between 1992 and 2018. Predictor variables included physical and chemical parameters and event time and sampling depth. Random Forest consistently outperformed linear regression, with the biomass-informed model explaining 58.8% of the variance in primary production (R^2 = 0.588). SHAP analysis identified sampling depth, nitrate, and temperature as the strongest environmental correlates of chlorophyll-*a*, whereas Chl-a dominated PP predictions, followed by depth and photosynthetically active radiation (PAR). The predictive contribution of temperature declined markedly after chlorophyll-*a* was included (Temperature Mediation Index = 0.642), indicating that much of the temperature-associated predictive signal was shared with biomass variability. The mean absolute SHAP contribution of temperature decreased markedly after chlorophyll-*a* was included as a predictor of PP (Temperature Mediation Index, TMI = 0.642). This index quantifies the relative attenuation of temperature's predictive importance between the two model configurations and does not constitute evidence of a causal mediation pathway through phytoplankton biomass. To distinguish the environmental drivers of total primary production from those associated with the component of production not explained by phytoplankton biomass we performed a decoupling analysis. Residuals from a chlorophyll-*a*-only predictive model were analysed using Random Forest regression and SHAP and identified depth and PAR as the dominant predictors of PP variability unexplained by Chl-a, potentially reflecting multiple sources of variation, including photoacclimation, phytoplankton community composition, unmeasured processes and environmental drivers, and observational error. Therefore, they should not be interpreted as a direct measure of biomass-specific photosynthetic efficiency. Our results demonstrate how explainable machine learning can disentangle complementary pathways linking environmental

Academic Editors: Kam W. Tang

Received: 9 August 2026
Revised: 4 November 2026
Accepted: 9 November 2026
Published: 12 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

variability, phytoplankton biomass, and marine primary production while providing ecologically interpretable predictive models.

Keywords: phytoplankton; machine learning; coastal monitoring; explanatory variables; field data; explainable artificial intelligence and SHAP

1. Introduction

Marine primary production underpins pelagic food webs and plays a fundamental role in the global carbon cycle by regulating the transfer of energy and carbon through marine ecosystems. In coastal environments, phytoplankton primary production is regulated by the interplay of multiple environmental drivers, including light availability, nutrient supply, water-column structure, temperature, and seasonal variability [1,2].

Environmental variability regulates primary production through two complementary mechanisms. First, it controls the accumulation of phytoplankton biomass, thereby determining the amount of photosynthetically active biomass available within the water column. Second, it influences the physiological performance of phytoplankton by modifying the efficiency with which existing biomass converts absorbed light energy into organic carbon through photosynthesis, independently of biomass changes [3,4]. Consequently, increases in primary production may arise either from greater phytoplankton biomass or from enhanced biomass-specific photosynthetic activity under similar environmental conditions. Although these mechanisms are conceptually distinct, they are strongly interconnected because the same environmental drivers frequently affect both biomass accumulation and photosynthetic efficiency simultaneously. Distinguishing their relative contributions is therefore essential for understanding how environmental variability regulates marine primary production [5,6].

The environmental regulation of marine primary production has been investigated extensively using both mechanistic and empirical approaches. Light and nutrient availability are widely recognized as the principal bottom-up controls of phytoplankton growth, whereas water-column stratification and vertical mixing regulate nutrient delivery to the euphotic zone while simultaneously modifying the underwater light environment experienced by phytoplankton cells [7–9]. Temperature represents a more complex environmental driver because its influence extends beyond direct physiological effects on metabolic processes. In natural ecosystems, temperature is often closely associated with seasonal succession, bloom phenology, water-column stability, nutrient availability, and circulation patterns. Consequently, observed relationships between temperature and primary production frequently reflect a combination of direct physiological responses and indirect environmental associations, making ecological interpretation particularly challenging.

Chlorophyll-*a* is commonly used as a proxy for phytoplankton biomass and has become one of the most widely adopted indicators of phytoplankton dynamics in aquatic ecosystems. Nevertheless, chlorophyll concentration does not directly quantify carbon fixation or phytoplankton carbon biomass. The relationship between chlorophyll-*a* and primary production varies substantially as a consequence of photoacclimation, nutrient availability, community composition, pigment packaging, and the recent light history experienced by phytoplankton populations [10–14]. As a result, similar chlorophyll concentrations may correspond to markedly different rates of carbon fixation under contrasting environmental conditions. In field observations, separating biomass-driven changes in primary production from variations in phytoplankton physiological performance is particularly challenging, as chlorophyll content and photosynthetic activity can both vary

with environmental conditions [4,10]. Investigating deviations from the expected chlorophyll–primary production relationship therefore provides an opportunity to investigate the environmental factors that regulate biomass-specific productivity, revealing conditions under which phytoplankton biomass exhibits higher or lower photosynthetic performance than expected.

Recent advances in machine learning (ML), together with the growing availability of environmental datasets, have expanded opportunities to investigate complex ecological processes by capturing nonlinear relationships and interactions among environmental drivers [15,16]. Unlike conventional statistical approaches, ML algorithms can model these relationships without requiring predefined functional forms [17]. Regression-based ML methods have therefore gained increasing application in environmental sciences [18–22], particularly in studies of marine primary production, where multiple correlated drivers may exert nonlinear and interacting effects.

Although ML models often achieve high predictive accuracy, the ML internal decision-making process is difficult to interpret, giving rise to the well-known "black-box" problem that limits ecological inference from model predictions [23]. This limitation is particularly relevant in ecological studies, where understanding the contribution of individual environmental drivers is often as important as obtaining accurate predictions.

Recent advances in explainable artificial intelligence (XAI) have improved the interpretability of complex ML models. Among these approaches, SHapley Additive exPlanations (SHAP) enables global and local interpretation by quantifying individual predictor contributions to model predictions, facilitating the identification of dominant environmental drivers and nonlinear relationships captured by the model [24–27]. However, SHAP values reflect predictive contributions within a fitted model and should not be interpreted as direct evidence of causal ecological mechanisms.

Although the environmental regulation of marine primary production has been extensively investigated, most studies have focused either on phytoplankton biomass or on primary production itself. The environmental drivers controlling biomass accumulation and biomass-specific productivity are often analysed separately, despite increasing evidence that the chlorophyll–primary production relationship is highly variable across environmental conditions and phytoplankton communities [1,4,28,29]. However, the simultaneous influence of environmental drivers on phytoplankton biomass and physiological activity, together with the strong covariation among environmental variables, makes it difficult to identify their relative contributions from field observations. Consequently, there is a need for integrated approaches that examine not only the environmental drivers of chlorophyll-*a* and primary production, but also the variability in production that cannot be explained by chlorophyll-*a* alone. Relatively few studies have attempted to disentangle these complementary components within a unified predictive framework. Distinguishing these complementary components is essential for identifying whether environmental variability primarily influences the amount of phytoplankton biomass available for photosynthesis or the physiological efficiency with which that biomass fixes carbon.

In this study, we combined Random Forest regression with SHAP-based model interpretation to investigate the environmental drivers of chlorophyll-*a*, primary production and chlorophyll–production decoupling in a coastal marine ecosystem. Specifically, we addressed four questions: (i) which environmental variables are most strongly associated with chlorophyll-*a* variability; (ii) which variables best explain primary production when considering environmental predictors alone or together with phytoplankton biomass; (iii) how does the predictive contribution of temperature change when phytoplankton biomass is explicitly accounted for; and (iv) which environmental variables are associated with variation in primary production that is not explained by the average chlorophyll–primary production relationship.

2. Materials and Methods

2.1. Dataset and Study Variables

The analyses were based on a long-term coastal monitoring dataset collected at the E2CO station (43°25.320′ N, 8°26.220′ W; 80 m depth), located on the Galician shelf (NW Spain), within the RADIALES monitoring program and available in the PANGEA repository [30,31]. The original dataset comprises vertical-profile observations collected at a single sampling station between 1990 and 2018. As the dataset originates from a single continental-shelf station, the ecological relationships investigated in this study are specific to the environmental conditions represented at E2CO and should not be assumed to be representative of other coastal ecosystems.

For the present study, only the period 1992–2018 was considered to exclude the initial years, which contained a high proportion of missing observations.

The final dataset comprised 2201 observations from 314 unique sampling campaigns, identified by sampling date and time. Chlorophyll-*a* (Chl-a) measurements were available for 2183 observations, whereas primary production (PP) estimates were available for 1530 observations.

The predictor variables included Month, Water depth, temperature (Temp), salinity (Sal), photosynthetically active radiation (PAR), and dissolved inorganic nutrient concentrations, including nitrate (NO_3^-), nitrite (NO_2^-), ammonium (NH_4^+), phosphate (PO_4^{3-}), and silicate ($Si(OH)_4$). Chlorophyll-*a* and primary production were used as response variables in different stages of the analysis [32,33].

The predictor set was selected to represent the main physical and biogeochemical drivers expected to influence phytoplankton biomass and productivity in coastal waters. Specifically, the variables describe seasonal forcing (Month), vertical position within the water column (Water depth), hydrographic conditions (temperature and salinity), light availability (PAR), and nutrient availability (NO_3^-, NO_2^-, NH_4^+, PO_4^{3-}, and $Si(OH)_4$). Throughout this manuscript, the term environmental predictors refers collectively to this set of physical and chemical variables.

Chlorophyll-*a* concentrations and primary production were determined following the standard protocols of the RADIALES monitoring program, using fluorometric pigment analysis and the ^{14}C-bicarbonate uptake method, respectively. Environmental measurements were collected from vertical CTD casts and water samples following the standard RADIALES monitoring protocols. Temperature, salinity and PAR were measured in situ, whereas dissolved inorganic nutrients were determined by standard colorimetric methods. Detailed descriptions of the sampling and analytical procedures are provided by Bode et al. (2019) [32].

2.2. Hierarchical Modelling Framework

To isolate the environmental drivers of phytoplankton biomass from those associated with biomass-specific primary production, we adopted a hierarchical modelling framework comprising three complementary models (Figure 1):

- M1: chlorophyll-*a* predicted exclusively from environmental variables, Chl-a = f(Environment).
- M2: primary production predicted exclusively from environmental variables, PP = f(Environment).
- M3: primary production predicted from environmental variables and observed chlorophyll-*a*, PP = f(Environment, Chl-a).

This hierarchical framework enabled environmental drivers on primary production to be examined both before and after explicitly accounting for phytoplankton biomass. Model M1 identified the environmental drivers associated with variability in chlorophyll-

a, M2 quantified the environmental predictors of primary production without direct information on biomass, whereas M3 evaluated the extent to which environmental effects persisted after phytoplankton biomass had been explicitly incorporated into the model.

Each model was implemented using both Linear Regression (LR) and Random Forest (RF), allowing the predictive performance of a conventional linear approach to be directly compared with that of a nonlinear machine-learning algorithm.

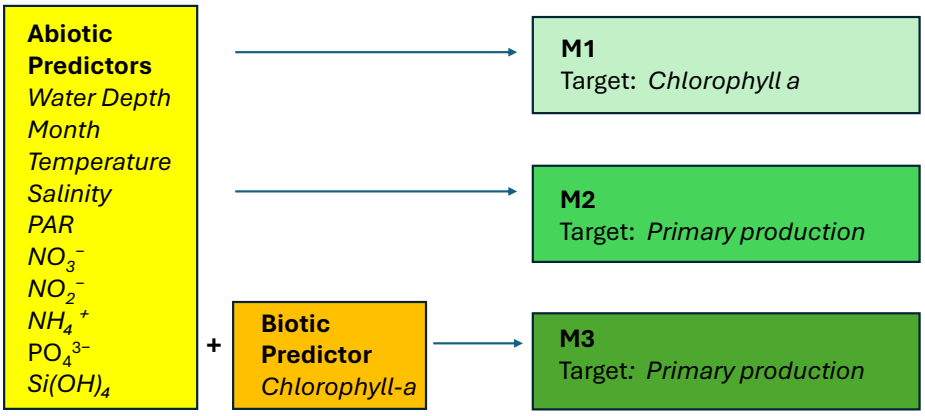


Figure 1. Schematic overview of the hierarchical modelling framework.

2.3. Data Pre-Processing

Strictly positive variables were $\log_{10}$-transformed prior to modelling to reduce skewness and improve model stability [34]. Specifically, PAR, chlorophyll-*a* and primary production were analysed on the $\log_{10}$ scale whenever they appeared as response variables or predictors in the modelling framework. For ease of interpretation, model effects are discussed in terms of the ecological meaning of the variables, while the statistical modelling was performed on the transformed scale where indicated.

Missing predictor values were reconstructed using multivariate iterative imputation implemented through the IterativeImputer algorithm in scikit-learn (max_iter = 20, sample_posterior = False, random_state = 42), which performs chained multivariate regression to estimate missing values from the observed relationships among predictors [35,36]. To avoid information leakage, imputation was performed within the modelling pipeline and therefore fitted separately within each training fold during cross-validation. The proportion of missing values for each predictor and comparisons of variable distributions before and after imputation are provided in the supplementary materials (Table S1, Figures S1 and S2).

Predictor variables were not standardised prior to model fitting because Random Forest is based on recursive binary partitioning and is therefore insensitive to differences in predictor scale [22,37]. For the linear models, used as baselines, predictors were entered on the transformed scale where appropriate, using the same response/predictor definitions as in the corresponding Random Forest formulation.

2.4. Cross-Validation and Model Evaluation

Model performance was evaluated using five-fold GroupKFold cross-validation, with sampling Date/Time used as the grouping factor to ensure that all observations collected during the same sampling campaign were assigned to the same fold, thereby preventing information leakage between training and validation sets and reducing overly optimistic estimates of predictive performance arising from within-cruise temporal autocorrelation [38]. However, this grouping strategy does not fully account for temporal

dependence across different sampling campaigns, including seasonal and interannual autocorrelation, which remains a limitation of the model evaluation.

Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) [14,39,40]. For each model, performance metrics were calculated from cross-validated out-of-fold predictions.

The predictive performance of Random Forest models was compared with that of corresponding linear regression models to determine whether nonlinear model structures provided additional predictive skill. Linear regression was used as the baseline approach for all hierarchical models.

2.5. Random Forest Specification and Hyperparameter Tuning

Linear Regression (LR) and Random Forest (RF) models were implemented in Python (version 3.11) using the scikit-learn library [36]. Random Forest was selected over other nonlinear regression models because the primary objective was robust ecological interpretation rather than maximising predictive accuracy. Random Forest is widely used for its predictive performance, flexibility, and robustness to outliers and multicollinearity [22]. In addition to its compatibility with the TreeSHAP framework for model interpretation [41], Random Forest has been shown to provide stable inference in ecological applications and to be relatively robust to correlated predictors [42,43]. Hyperparameters for the RF models were optimized separately for each modelling formulation using an exhaustive grid search implemented through GridSearchCV within the GroupKFold cross-validation framework. The hyperparameter search explored combinations of the number of trees, maximum tree depth, number of predictors considered at each split, minimum number of samples required for node splitting, minimum number of samples per terminal leaf, and bootstrap resampling (Table 1).

Candidate hyperparameter combinations were evaluated using the same grouped cross-validation strategy adopted for model assessment. Model performance was quantified using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), while the optimal hyperparameter set was selected according to the highest cross-validated R^2. Once the optimal configuration had been identified for each model, the corresponding RF was refitted using the complete dataset and subsequently used for SHAP interpretation, threshold analysis, and sensitivity analyses. The optimal hyperparameter configurations selected for each Random Forest model are reported in supplementary materials (Table S2).

Table 1. Random Forest hyperparameter grid used during GridSearchCV. Final selected parameters were stored for each base model and reused in the fixed-parameter sensitivity analysis.

Hyperparameter	Grid Values
n_estimators	300, 500, 800
max_depth	8, 10, 12, None
max_features	0.5, 0.7, 1
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
bootstrap	True
random_state	42

The same cross-validation strategy, missing-data imputation procedure, and performance metrics were applied consistently across all chlorophyll-*a*, primary production, and residual-decoupling models to ensure methodological comparability. The modelling workflow was intentionally standardised across all model formulations so that differences in predictive performance and variable importance could be attributed to the information

content of the predictor sets rather than to differences in model configuration or optimisation.

2.6. SHAP Analysis and Temperature Mediation Index (TMI)

Model interpretation was performed using SHapley Additive exPlanations (SHAP) [24], which quantify the contribution of each predictor to individual model predictions. SHAP values represent the additive contribution of each predictor to the model output relative to the expected model prediction (baseline). Positive and negative values indicate contributions that increase or decrease the predicted output, respectively. In this study, SHAP values are expressed on the $\log_{10}$-transformed scale of the response variable and do not represent percentages of contribution. The main SHAP analyses were performed on the final Random Forest (RF) models after hyperparameter optimisation. For each modelling formulation, the optimal RF was refitted using the complete dataset and subsequently interpreted using the TreeExplainer algorithm implemented in the shap Python library. Consequently, SHAP analyses complemented the cross-validated evaluation of predictive performance by providing an interpretation of the final fitted models rather than serving as an alternative measure of model accuracy.

Global predictor importance was quantified as the mean absolute SHAP value across all analysed observations, providing a measure of the overall contribution of each predictor to model predictions. SHAP summary and bar plots were generated for the main chlorophyll-*a* (M1) and primary production (M3) models, whereas temperature-specific SHAP values were additionally extracted from the primary production models to quantify the contribution of temperature before and after explicitly accounting for phytoplankton biomass. Throughout this study, SHAP values are interpreted as measures of predictor contribution within the fitted Random Forest models and should therefore be regarded as indicators of predictive importance rather than evidence of causal ecological relationships [24, 25].

A central objective of the study was to quantify how the predictive contribution of temperature to primary production changed after explicitly accounting for phytoplankton biomass. To summarise this attenuation, we defined a Temperature Mediation Index (TMI) as the relative reduction in the mean absolute SHAP contribution of temperature between the environmental-only primary-production model (M2) and the biomass-informed model (M3):

$$\mathrm{TMI} = 1 - |\mathrm{SHAP}_{Temp,M3}| / |\mathrm{SHAP}_{Temp,M2}|$$

where $|\mathrm{SHAP}_{Temp,M2}|$ and $|\mathrm{SHAP}_{Temp,M3}|$ denote the mean absolute SHAP importance of temperature in M2 and M3, respectively. Positive TMI values indicate reduced temperature importance after including chlorophyll-*a*, values near zero indicate little change, and negative values indicate increased importance. The term "mediation" is used solely in a predictive sense, as the TMI describes changes in temperature's SHAP importance rather than causal mediation. Such changes may reflect predictor correlations, shared information, interactions, or redistribution of feature importance. Therefore, the TMI should be interpreted as a descriptive measure of predictive attenuation, not as evidence of a biological mediation pathway.

2.7. Threshold Analysis from SHAP Dependence Curves

Threshold-like transitions in environmental responses were explored using SHAP dependence relationships for selected environmental predictors included in models M1 and M3, namely temperature, $\log_{10}$(PAR), NO_3^-, NO_2^-, PO_4^{3-}, NH_4^+ and $Si(OH)_4$. For each predictor, SHAP values were plotted against the corresponding predictor values, and

smoothed dependence curves were obtained using a centred rolling mean to reduce local variability while preserving the overall response pattern.

Two complementary metrics were extracted from each smoothed SHAP curve to characterise potential transition points. The first corresponded to the initial zero crossing of the SHAP response, indicating the point at which the predictor contribution changed sign. The second was defined as the point of maximum absolute slope, representing the location of the most rapid change in predictor influence along the environmental gradient.

The robustness of these threshold estimates was assessed using 500 bootstrap resamples. Threshold values were recalculated for each bootstrap replicate, and the median together with the 95% percentile interval was used to summarise the uncertainty associated with each transition estimate. Bootstrap medians and corresponding 95% percentile intervals for the SHAP-derived transition points are provided in Supplementary Materials (Table S3).

These analyses were intended as an exploratory approach to identify potential transitions in the relationships between environmental variables and model predictions. The identified transition points should be interpreted as model-derived patterns rather than experimentally validated ecological or physiological thresholds. Therefore, these results should be regarded as a basis for generating ecological hypotheses within the context of the specific modelling framework adopted in this study.

2.8. Residual Decoupling Between Chlorophyll-a and Primary Production

To characterize the baseline relationship between phytoplankton biomass and primary production, simplified empirical models were developed to quantify the direct relationship between Chl-a and PP.

First, an ordinary least-squares (OLS) regression was fitted by relating $\log_{10}$-transformed PP to $\log_{10}$-transformed Chl-*a*:

$$log_{10}(PP) = \alpha + \beta\ log_{10}(Chl\text{-}a) + \varepsilon.$$

A Random Forest, using chlorophyll-*a* as the sole predictor, was then fitted to assess potential nonlinear departures from the linear relationship.

To distinguish the environmental controls on total primary production from those affecting the component of production not explained by phytoplankton biomass, the chlorophyll-*a*-only OLS and Random Forest models were compared using out-of-fold predictive performance (R^2, RMSE and MAE). The OLS model showed lower prediction errors and was therefore used to calculate the residuals for the subsequent environmental decoupling analysis. This residual, hereafter referred to as the PP–Chl-*a* decoupling, was defined as the difference between the observed and model-predicted $\log_{10}$-transformed primary production:

$$Residual_{}_{PP|Chl} = \log_{10}(PPobs) - \log_{10}(PPpred)$$

where PP*pred* denotes the primary production predicted by the chlorophyll-*a*-only OLS regression model. Positive residuals indicate observations in which primary production was higher than expected for a given chlorophyll-*a* concentration, whereas negative residuals indicate lower-than-expected production. The residuals were derived from a chlorophyll-*a*-only OLS regression model rather than from the environmental Random Forest models. As temperature and the other environmental predictors were not included in this baseline regression, the subsequent Random Forest analysis was designed to identify environmental variables associated with deviations from the expected chlorophyll-*a*–primary production relationship, rather than to detect effects missed by an initial environmental Random Forest.

To investigate the environmental drivers of this decoupling, a Random Forest model was fitted using the environmental predictor set:

$$Residual_{PP/Chl} = f(Environment)$$

Model interpretation was performed using SHAP values to identify the environmental variables most strongly associated with departures from the baseline chlorophyll-*a*–primary production relationship. The PP–Chl-*a* decoupling should be interpreted as an empirical index of deviation from the expected biomass–production relationship. Although positive values are consistent with enhanced biomass-specific production and negative values with reduced biomass-specific production, this metric does not represent a direct physiological measure of photosynthetic efficiency.

2.9. Scope of the Random-Seed Sensitivity Analysis

To assess the robustness of the Random Forest results to stochastic model initialisation, the analyses for M1, M2 and M3 were repeated across five random seeds (1, 10, 42, 100 and 999). The sensitivity analysis was run only for the base configuration, with the hyperparameter configuration fixed to the best parameters selected from the base GridSearchCV runs (SENSITIVITY_USE_GRIDSEARCH = False; SENSITIVITY_BASE_ONLY = True). Thus, the analysis tested seed-dependent reproducibility of a fixed model structure rather than uncertainty arising from repeated hyperparameter optimisation.

For each seed and model, out-of-fold predictions were generated with the same grouped cross-validation workflow and evaluated using R^2, RMSE and MAE. Mean absolute SHAP importance of temperature was then recomputed for each model realisation. When both M2 and M3 were available, the TMI was recalculated for each seed as 1-SHAP_Temp_with_Chl_a/SHAP_Temp_without_Chl_a. Sensitivity summaries were reported as medians, percentile intervals, interquartile ranges and sample counts across seeds. Accordingly, the sensitivity analysis addresses stochastic reproducibility of the fitted Random Forest workflow rather than the full uncertainty associated with hyperparameter selection or alternative model specifications.

Importantly, the analysis did not quantify uncertainty associated with alternative cross-validation grouping schemes or data preprocessing choices, including missing-data imputation and predictor transformations. Therefore, stability across random seeds under the selected configuration should not be interpreted as evidence of global robustness of the entire analytical workflow. Rather, this analysis was intended to assess the sensitivity of model performance to variations in the random seed while keeping the overall analytical configuration unchanged.

2.10. Statistical Software

All statistical analyses were conducted in Python using the scikit-learn and SHAP libraries for model development and interpretation [24, 36]. Data pre-processing, model fitting, hyperparameter optimisation, cross-validation, SHAP-based model interpretation, threshold analysis, and sensitivity analyses were integrated into a fully reproducible analytical workflow to ensure methodological consistency across all modelling formulations.

3. Results

3.1. Model Performance and Predictive Skill

Random Forest models showed moderate to high predictive skill and generally outperformed the corresponding linear models, although the magnitude of the improvement varied among model formulations (Table 2). Among all models, M3 provided the best

performance for primary production, explaining 58.8% of the observed variance in the Random Forest implementation (R^2 = 0.588), compared with 55.1% explained by the corresponding linear model (R^2 = 0.551). This indicates that a substantial fraction of the relationship between primary production and its predictors can be captured by a linear formulation, but that the nonlinear model still accounts for an additional component of variability.

The inclusion of observed chlorophyll-*a* markedly improved the prediction of primary production relative to M2, which relied exclusively on environmental predictors. In the Random Forest framework, R^2 increased from 0.438 in M2 to 0.588 in M3, while in the linear framework it increased from 0.321 to 0.551. Thus, the gain associated with including chlorophyll-*a* was larger than the gain associated with switching from a linear to a nonlinear algorithm, indicating that phytoplankton biomass captures a major fraction of the information required to explain variability in primary production.

The advantage of Random Forest over linear regression was strongest for M2, i.e., the model predicting primary production from environmental variables alone. In this case, R^2 increased by 0.117 (from 0.321 to 0.438), with concurrent reductions in both RMSE and MAE. This pattern suggests that, in the absence of chlorophyll-*a*, the relationship between environmental forcing and primary production includes nonlinear responses and interactions that are not adequately represented by a linear model. For M1, which predicts phytoplankton biomass, the improvement associated with Random Forest was smaller but still evident, with R^2 increasing from 0.324 to 0.390.

Table 2. Cross-validated performance of base linear and Random Forest models. Performance metrics are based on grouped out-of-fold predictions.

Model	Algorithm	Response	Predictor Set	R^2	RMSE	MAE
M1	Linear regression	$\log_{10}$(Chl-a)	Environment	0.324	0.378	0.299
M1	Random Forest	$\log_{10}$(Chl-a)	Environment	0.390	0.359	0.278
M2	Linear regression	$\log_{10}$(PP)	Environment	0.321	0.610	0.477
M2	Random Forest	$\log_{10}$(PP)	Environment	0.438	0.555	0.423
M3	Linear regression	$\log_{10}$(PP)	Environment + Chl-a	0.551	0.496	0.371
M3	Random Forest	$\log_{10}$(PP)	Environment + Chl-a	0.588	0.475	0.358

Feature-removal configurations further supported the role of sampling depth and seasonality in the environmental models (Table 3). Removing Month or Water depth reduced performance in M1, M2 and M3, with the largest decrease generally associated with removing Water depth. This pattern indicates that the vertical position of samples within the water column contains substantial predictive information for both chlorophyll-*a* and primary production.

Table 3. Random Forest performance under feature-removal configurations. These runs used the base-model hyperparameters fixed from the corresponding GridSearchCV result.

Model	Configuration	R^2	RMSE	MAE
M1	Base	0.390	0.359	0.278
M1	NoMonth	0.342	0.373	0.290
M1	NoDepth	0.269	0.394	0.304
M1	NoMonth_NoDepth	0.230	0.404	0.313
M2	Base	0.438	0.555	0.423
M2	NoMonth	0.377	0.584	0.451
M2	NoDepth	0.292	0.623	0.473
M2	NoMonth_NoDepth	0.253	0.639	0.486

M3	Base	0.588	0.475	0.358
M3	NoMonth	0.573	0.484	0.368
M3	NoDepth	0.543	0.500	0.380
M3	NoMonth_NoDepth	0.536	0.504	0.384

3.2. Environmental Correlates of Phytoplankton Biomass (M1)

SHAP analysis of M1 identified Water depth as the main correlate of chlorophyll-*a* variability, followed by nitrate, temperature, month and nitrite (Figure 2). Increasing sampling depth was associated with progressively more negative SHAP contributions to chlorophyll-*a*, whereas shallower conditions yielded positive contributions. This pattern is consistent with a decline in phytoplankton biomass with increasing depth, as expected from reduced light availability and from the tendency for productive biomass to accumulate in shallower, more illuminated waters.

Nitrate showed a clear inverse association: high concentrations were generally associated with negative or near-zero SHAP values, whereas lower concentrations tended to contribute positively to predicted chlorophyll-*a*. This inverse association should not be interpreted as a direct inhibitory effect of nitrate on phytoplankton biomass. Within the observed dataset, it likely reflects the co-occurrence of elevated nitrate with deeper, less illuminated or pre-bloom conditions, whereas low nitrate may mark periods of active nutrient drawdown by phytoplankton biomass. Temperature likewise exhibited an inverse association, with warmer conditions contributing negatively and lower temperatures positively. The importance of Month among the top predictors indicates that chlorophyll-*a* variability includes a marked seasonal component not fully captured by the other environmental variables.

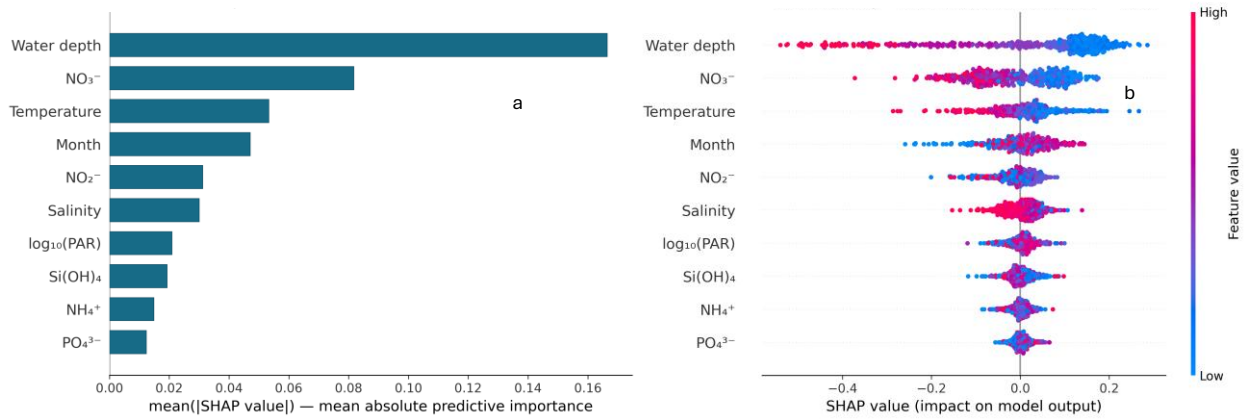


Figure 2. SHAP summary plot for M1, the Random Forest model predicting $\log_{10}$ chlorophyll-*a* from environmental predictors. (**a**) Mean absolute SHAP values showing the relative predictive importance of each variable. (**b**) SHAP summary plot, where each point represents an observation and its colour indicates the corresponding predictor value, ranging from low (blue) to high (red). The horizontal position represents the SHAP contribution to the model prediction relative to the expected model output (baseline). Positive and negative SHAP values indicate contributions that increase or decrease the predicted $\log_{10}$ chlorophyll-*a*.

3.3. Environmental Correlates of Primary Production Without and with Chlorophyll-a

In the environmental-only primary-production model (M2), the dominant predictors were PAR, Water depth and temperature, indicating that when chlorophyll-*a* is not supplied to the model, primary production is associated primarily with the physical and radiative environment. By contrast, in M3, which combines environmental predictors with observed chlorophyll-*a*, chlorophyll-*a* emerged as the dominant predictor of primary

production by a wide margin (Figure 3). The SHAP summary plot showed a nearly monotonic relationship between chlorophyll-*a* and primary production, with high chlorophyll-*a* values contributing strongly and positively to predicted production and low values contributing negatively. This indicates that phytoplankton biomass captures a major fraction of the variability in ecosystem productivity.

In addition to chlorophyll-*a*, Water depth and PAR retained substantial importance in M3, showing that primary production depends not only on the amount of biomass present but also on the physical and radiative context in which that biomass operates. High Water depth values were associated with negative contributions to primary production, whereas shallower conditions were associated with positive contributions. PAR showed predominantly positive contributions, consistent with its role in sustaining photosynthetic carbon fixation. By contrast, the importance of temperature declined markedly once chlorophyll-*a* was included in the model.

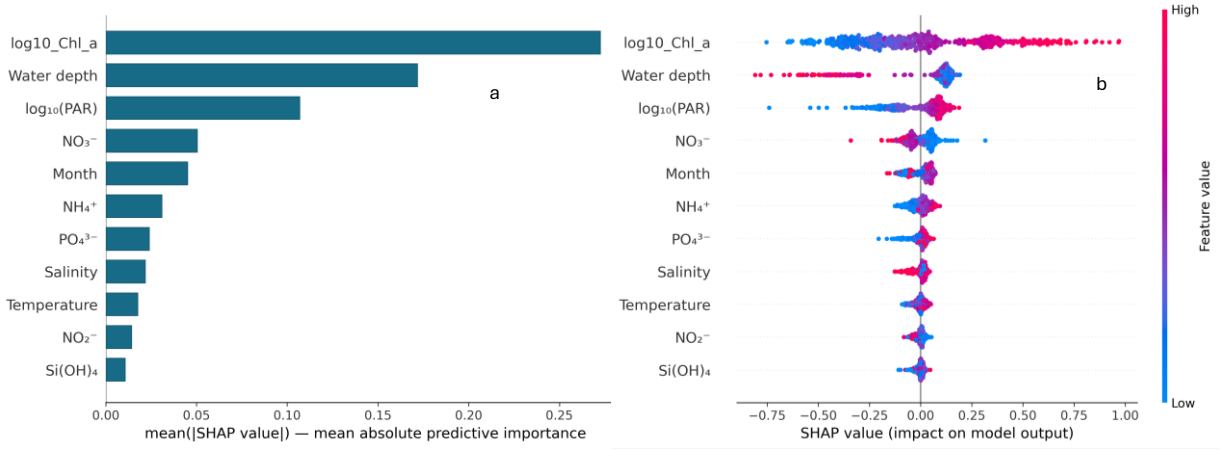


Figure 3. SHAP analysis for M3, the Random Forest model predicting $\log_{10}$-transformed primary production from environmental predictors and observed $\log_{10}$-transformed chlorophyll-*a*. (**a**) Mean absolute SHAP values showing the relative predictive importance of each variable. (**b**) SHAP summary plot, where each point represents an observation and its colour indicates the corresponding predictor value, ranging from low (blue) to high (red). The horizontal position represents the SHAP contribution to the model prediction relative to the expected model output (baseline). Positive and negative SHAP values indicate contributions that increase or decrease the predicted $\log_{10}$ primary production, respectively.

3.4. Temperature Signal After Accounting for Phytoplankton Biomass

The mean absolute SHAP contribution of temperature declined markedly from the environmental-only production model (M2) to the biomass-informed model (M3), from 0.049 to 0.018. The TMI was 0.642, indicating that the predictive contribution of temperature was substantially attenuated after chlorophyll-*a* was included in the model. This pattern is consistent with a large fraction of the temperature-associated predictive signal being shared with, or captured by, phytoplankton biomass.

This result does not imply formal causal mediation, but it does indicate that the predictive association between temperature and primary production is strongly reduced after chlorophyll-*a* is supplied to the model. The most conservative interpretation is therefore that a substantial fraction of the temperature-related predictive information in the environmental-only model overlaps with information already captured by chlorophyll-*a* in the biomass-informed model.

3.5. Model-Derived Transition Points-like Responses Along Environmental Gradients

A comparison of the SHAP dependence plots revealed markedly different temperature responses between the chlorophyll-*a* model (M1) and the primary production model including observed chlorophyll-*a* (M3) (Figure 4). In M1, temperature exerted a strong, nearly monotonic negative influence on chlorophyll-*a*, with SHAP values decreasing from positive contributions at lower temperatures to increasingly negative values above the model-derived transition point of approximately 14.2 °C. This pattern indicates that higher temperatures were associated with progressively lower predicted phytoplankton biomass. In contrast, the residual effect of temperature on primary production after accounting for chlorophyll-*a* was substantially weaker and distinctly non-linear. In M3, temperature shifted from slightly negative to positive contributions around 14.3 °C, reached a modest optimum between approximately 15 and 16 °C, and then gradually declined, becoming slightly negative again above 17 °C. The much smaller SHAP amplitude observed in M3 compared with M1 indicates that most of the predictive information carried by temperature is already captured through chlorophyll-*a*, leaving only a limited residual contribution of temperature to primary production. This interpretation is consistent with the high Temperature Mediation Index (TMI = 0.642), suggesting that a substantial proportion of the temperature-associated predictive signal is shared with phytoplankton biomass rather than representing an independent effect on primary production.

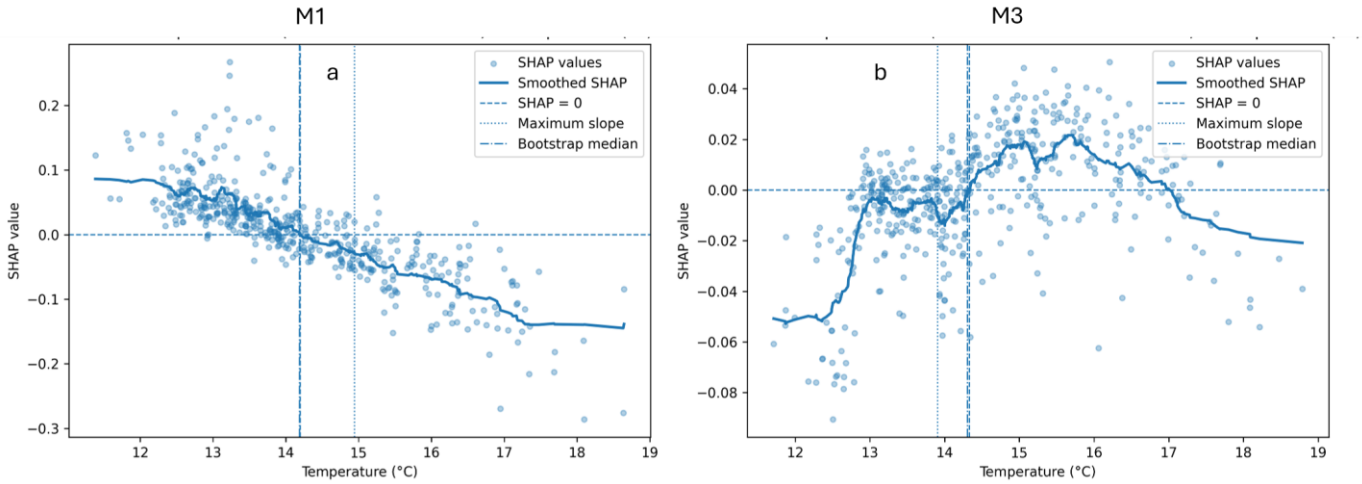


Figure 4. SHAP dependence plots illustrate model-derived temperature responses for predicted chlorophyll-*a* in model M1 (**a**) and predicted primary production in model M3 (**b**). The y-axis ranges of SHAP values differ between the M1 and M3 panels. Because the models predict different response variables, absolute SHAP magnitudes are not directly comparable. Therefore, comparisons should focus on the shapes and directions of the SHAP dependence relationships rather than on the absolute sizes of SHAP contributions.

In M1, nitrate exhibited a second transition at approximately 3.6 µM, beyond which its SHAP contribution also became predominantly negative. Together, these patterns suggest that, within the observed environmental range, higher chlorophyll-*a* concentrations were associated with relatively cooler temperatures and lower nitrate concentrations (Figure 5a).

In M3, the most pronounced model-derived transition point was observed for PAR, with SHAP values shifting from negative to positive at approximately $\log_{10}(\text{PAR}) = 1.30$, indicating a transition from less favourable to more favourable light conditions for predicted primary production (Figure 5b).

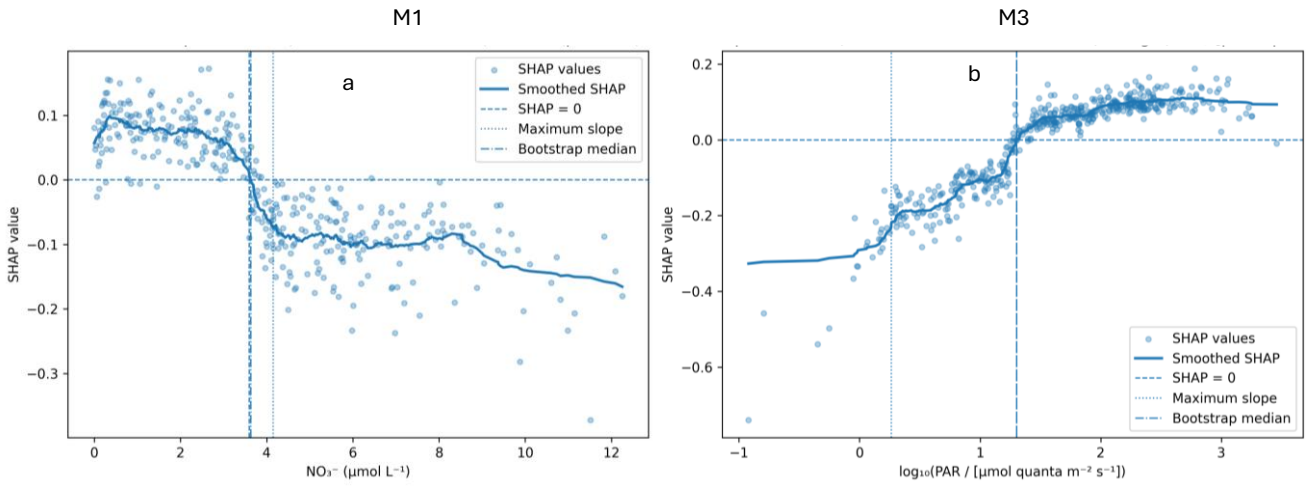


Figure 5. SHAP dependence plots illustrating model-derived relationships between nitrate ($NO3^-$) and log_{10}(PAR) on predicted chlorophyll-*a* in model M1 (**a**) and on predicted primary production in model M3 (**b**). The y-axis ranges differ between panels, and absolute SHAP magnitudes are not directly comparable because the models predict different response variables.

These transition points represent model-derived features of the SHAP dependence relationships rather than experimentally validated ecological or physiological thresholds. Their estimated locations may depend on model specification, predictor correlations, and the observed environmental range. Accordingly, they should be regarded as exploratory indicators of potentially relevant environmental ranges that may inform future ecological hypotheses and require independent validation to establish their biological significance.

3.6. Environmental Correlates of Chlorophyll-Production Decoupling

The chlorophyll-*a*-only OLS model showed better out-of-fold predictive performance (R^2 = 0.393, RMSE = 0.577, MAE = 0.437) than the corresponding Random Forest model (R^2 = 0.341, RMSE = 0.601, MAE = 0.457). Therefore, residuals from the OLS model were used for the subsequent environmental decoupling analysis.

To investigate the environmental factors associated with deviations from the average chlorophyll-*a*–primary production relationship, we analysed the residual variability remaining after accounting for phytoplankton biomass. SHAP analysis of the Random Forest model fitted to this residual identified Water depth and PAR as the dominant predictors by a wide margin (Figure 6). Secondary contributions were observed for temperature, nitrate, ammonium and salinity, followed by phosphate, silicate, month and nitrite.

Unlike the primary production model, which predicts total primary production from environmental variables, the residual model focuses on variability that remains unexplained by chlorophyll-*a* alone. The prominence of water depth and light availability therefore indicates that these variables are associated with deviations from the expected chlorophyll-*a*–primary production relationship. However, these residuals should be interpreted as an empirical indicator of chlorophyll–production decoupling rather than as a direct measure of biomass-specific photosynthetic efficiency.

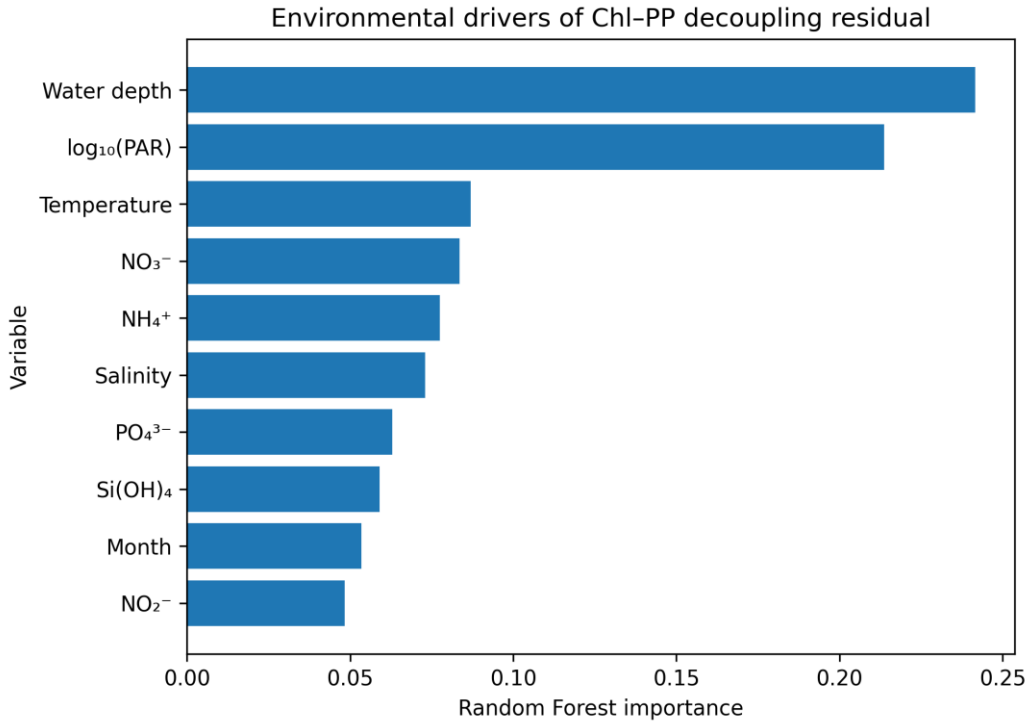


Figure 6. Random Forest variable importance for the model predicting residual chlorophyll-production decoupling from environmental predictors. Higher values indicate greater contribution to explaining departures from the mean chlorophyll-*a*-primary-production relationship.

3.7. Sensitivity Analysis

Sensitivity analyses showed that the Random Forest results were fully robust to stochastic model initialisation under the fixed base-model hyperparameter configuration. Repeating the analyses for M1, M2 and M3 across five random seeds (1, 10, 42, 100 and 999) yielded identical values of R^2, RMSE and MAE for each model, indicating no detectable seed-dependent variability in predictive performance under this setup. The same stability was observed for model interpretation: the mean absolute SHAP contribution of temperature remained unchanged across repetitions in all three models (M1 = 0.053, M2 = 0.049, M3 = 0.018).

Consequently, the TMI was also invariant across runs, with a median value of 0.642 and no variation across seeds. These results confirm that both predictive skill and SHAP-based ecological interpretation were highly reproducible under the selected Random Forest configuration. Because the sensitivity analysis was conducted under fixed hyperparameters and a fixed cross-validation design, it specifically demonstrates the stability of the fitted Random Forest solution to stochastic initialization rather than the full uncertainty associated with alternative model specifications.

4. Discussion

Although Random Forest consistently outperformed the corresponding linear regression models, the ecological significance of these results extends beyond the modest improvement in predictive performance. The largest advantage of nonlinear modelling was observed when primary production was predicted exclusively from environmental variables, indicating that environmental forcing operates through nonlinear responses and interactions that are only partially captured by linear models [22,44]. However, once phytoplankton biomass was explicitly incorporated into the model, the difference between Random Forest and linear regression became comparatively small (R^2 = 0.588

versus 0.551), whereas the inclusion of chlorophyll-*a* itself produced a much larger increase in predictive skill than the choice of modelling algorithm. This finding indicates that the ecological information carried by phytoplankton biomass is substantially more important than algorithmic complexity for explaining primary production in this coastal ecosystem. Similarly, feature-removal analyses showed that excluding sampling depth caused a greater reduction in predictive performance than excluding month, suggesting that the vertical position of the sampled water parcel integrates multiple ecological processes, including light availability, water-column structure and nutrient gradients, which are more directly linked to phytoplankton functioning than seasonal timing alone [6,45]. Consequently, the principal value of the Random Forest framework lies not simply in achieving higher predictive accuracy, but in providing sufficient flexibility to capture complex environmental relationships while remaining fully interpretable through SHAP [24]. This combination makes it possible to distinguish environmental effects mediated through phytoplankton biomass from those acting independently of biomass, thereby offering an interpretable hierarchical framework for disentangling complementary pathways linking environmental variability, phytoplankton biomass and marine primary production [24].

4.1. Phytoplankton Biomass as the Main Link Between Environment and Primary Production

The strong improvement from M2 to M3, together with the dominant SHAP contribution of chlorophyll-*a* in the biomass-informed production model, indicates that chlorophyll-*a* provides substantial predictive information about primary production that is not captured by the environmental predictors alone [10]. This pattern is consistent with the broader view that primary production is jointly structured by biomass and physiological activity, but that biomass often explains a large share of production variability at ecosystem scales [5, 6, 10].

The M1–M2–M3 hierarchy further suggests that environmental conditions contain predictive information about primary production both in association with chlorophyll-*a* variability and through environmental predictors that retain predictive contributions after chlorophyll-*a* is included. The modest predictive advantage of Random Forest over linear regression in M3, compared with its larger advantage in M2, suggests that nonlinear environmental relationships become particularly relevant when chlorophyll-*a* is unavailable as a predictor. This distinction highlights the value of hierarchical model comparisons for identifying where nonlinear modelling provides additional predictive information [6,45].

4.2. Water-Column Structure and Light as Cross-Scale Controls

Among the environmental predictors, Water depth and PAR emerged as the most consistent controls across the modelling framework. Because Water depth represents the vertical position of the sampled water parcel, its dominant role likely reflects a combination of light attenuation, vertical structuring of phytoplankton biomass, and differences in the recent light history experienced by the cells [6, 45]. Its importance in M1, M3 and the residual-decoupling model therefore suggests that vertical position is not only associated with biomass accumulation, but also with how effectively biomass is translated into production under different light regimes.

The importance of PAR in M3 and in the residual-decoupling model suggests that the component of production not explained by biomass alone is strongly associated with the physical and radiative context in which phytoplankton biomass is embedded [9]. In practical terms, chlorophyll-*a* describes how much pigment biomass is present, whereas residual chlorophyll-production decoupling reflects whether that biomass is associated with higher or lower production than expected from the mean biomass-production relationship. The strong role of PAR in the residual model is particularly consistent with the

idea that samples with similar chlorophyll-*a* may differ in realized carbon fixation because of differences in instantaneous light supply, photoacclimation state or recent vertical mixing history [5,46].

4.3. Interpreting the Temperature Influence

Temperature emerged as an important correlate of chlorophyll-*a* and of primary production in the environmental-only model, but its contribution declined sharply once chlorophyll-*a* was included. The TMI supports the interpretation that a large share of the temperature-related predictive information in the environmental-only model overlaps with information already captured by chlorophyll-*a* in the biomass-informed model. This pattern is consistent with biomass-associated structuring of the temperature signal, rather than a strong independent temperature contribution in the biomass-informed production model [5,6]. The nonlinear temperature response observed in Figure 4b may reflect the combined influence of phytoplankton physiological processes and environmental variability. The alternating flat and sloped segments could indicate changes in the relative importance of temperature under different ecological conditions, potentially associated with variations in nutrient availability, water-column stratification, and phytoplankton community composition. The decline observed at higher temperatures may reflect increasingly limiting environmental conditions for primary production. Overall, this pattern suggests that the temperature-associated contribution to primary production is context-dependent and may result from interactions among multiple ecological processes rather than a single physiological mechanism.

However, this interpretation should remain cautious. Because the analysis is based on observational data and predictive machine-learning models, the TMI does not demonstrate causal mediation in the formal statistical sense [25,47]. Temperature likely covaries with several unmeasured or partially measured processes, including stratification, mixing, bloom timing and nutrient supply. The present results therefore support a predictive interpretation-namely, that the temperature signal is strongly attenuated after biomass is accounted for-rather than a mechanistic claim that temperature acts exclusively or directly through phytoplankton biomass. Moreover, the reduction in temperature's SHAP importance may partly reflect the redistribution of predictive information among correlated predictors after chlorophyll-*a* is introduced. Potential nonlinear interactions between temperature and other environmental variables may also contribute to this pattern. Therefore, the observed attenuation should not be interpreted as demonstrating that the direct physiological influence of temperature on primary production is negligible.

Therefore, the comparison between environmental-only and biomass-informed models provides useful ecological information by revealing how the predictive relevance of temperature changes when chlorophyll-*a* variability is explicitly considered. The TMI should therefore be regarded as a descriptive indicator of predictive attenuation rather than as an estimator of causal mediation.

4.4. Residual Chlorophyll-Production Decoupling as an Ecological Diagnostic

The residual-decoupling analysis provides an additional ecological perspective by examining deviations from the expected chlorophyll-*a*–primary production relationship. The importance of Water depth and PAR suggests that these deviations are associated with the underwater light environment and water-column position. This interpretation is ecologically plausible because samples with similar chlorophyll-*a* concentrations may exhibit different production rates under contrasting light conditions, mixing regimes or nutrient histories [9].

However, the residual metric may reflect multiple sources of variability, including differences in biomass-specific production, photoacclimation, species composition,

unmeasured environmental forcing, and observational error. Accordingly, it should not be interpreted as a direct proxy for photosynthetic efficiency *sensu stricto*, but rather as an integrated diagnostic of departures from the average biomass–production relationship [9,46]. By identifying environmental predictors associated with higher or lower production than expected from chlorophyll-*a* alone, the residual-decoupling approach provides a basis for generating ecological hypotheses about the conditions associated with these departures, without establishing specific physiological mechanisms.

The methodological contribution of this study lies in the integration of complementary predictive analyses rather than in the introduction of a new machine-learning algorithm. Conventional multiple regression provides a useful benchmark for assessing environmental associations, but its standard linear formulation may not capture complex nonlinear relationships without explicitly specified terms. Similarly, simple residual analysis quantifies departures from a fitted relationship but does not, by itself, characterize their multivariate environmental associations. By combining hierarchical model comparisons, Random Forest predictions, SHAP-based attribution, the predictive Temperature Mediation Index and environmental modelling of biomass–production residuals, the proposed framework provides a more integrated assessment of the environmental information associated with phytoplankton biomass, primary production and deviations from their expected relationship. Although these methods are individually established, their coordinated application to the long-term dataset constitutes the principal incremental contribution of the study. The resulting interpretations remain predictive and hypothesis-generating rather than causal.

4.5. Nonlinearity and Threshold-like Transitions

The threshold analysis suggests that the environmental controls on chlorophyll-*a* and primary production are not purely monotonic [48,49]. Temperature and nitrate in M1, and PAR in M3, displayed transition points beyond which the direction or magnitude of their SHAP contributions changed appreciably. Such transition points are useful because they identify regions of environmental space where the fitted model becomes particularly sensitive to small changes in predictor values.

These threshold estimates should, however, be interpreted conservatively. They are derived from smoothed SHAP dependence curves and therefore represent model-based transitions rather than experimentally validated physiological thresholds [25]. Their main value lies in highlighting candidate ranges for future hypothesis testing and in illustrating the nonlinear structure captured by the Random Forest models [24]. Accordingly, the threshold analysis should be viewed primarily as a descriptive complement to the main hierarchical modelling results rather than as a standalone source of mechanistic inference [44].

4.6. Strengths and Limitations of the Approach

A major strength of the study is the hierarchical modelling design, which separates the environmental correlates of phytoplankton biomass from those of primary production conditional on biomass, and then isolates the residual component of production not explained by the biomass-production relationship. This framework provides a more process-oriented interpretation than a single black-box production model. The use of GroupKFold cross-validation also reduces inflation of predictive skill caused by within-campaign autocorrelation, although it does not eliminate potential seasonal or interannual temporal dependence. The sensitivity analysis indicates that the main Random Forest and SHAP results are stable across the tested random seeds under the selected Base configuration.

The study also has important limitations. First, the analysis is observational and therefore cannot establish causal pathways. SHAP values quantify predictive contributions within a fitted model, not mechanistic effects [24,39]. Second, chlorophyll-*a* is an imperfect proxy for phytoplankton biomass and may vary independently of carbon biomass because of photoacclimation and taxonomic shifts in pigment content. Third, several environmental predictors covary seasonally and physically; therefore, SHAP importance values should be interpreted as model-specific attributions under correlated predictors rather than as fully separable ecological effects [25]. Although Random Forest can capture nonlinear relationships and interactions among predictors, the present SHAP analysis focused on individual feature attributions and did not explicitly quantify pairwise SHAP interaction effects. Consequently, the relative contributions of interactions among environmental variables, such as temperature, light availability, and water depth, remain unresolved and represent an avenue for further investigation. Fourth, some important processes relevant to primary production, such as species composition, grazing pressure, mixed-layer depth, short-term weather forcing or unmeasured nutrient forms, were not explicitly represented [50].

An additional limitation concerns the spatial representativeness of the dataset, which is restricted to the single E2CO coastal monitoring station. Consequently, the ecological relationships identified in this study should be interpreted within this specific environmental context. Although the hierarchical modelling framework can be applied to datasets from other coastal ecosystems, the numerical TMI value, SHAP-based predictor importance rankings, and model-derived transition points are site- and dataset-specific and cannot be directly extrapolated to other regions without independent evaluation. Furthermore, the Temperature Mediation Index (TMI) quantifies the attenuation of temperature's predictive contribution after including chlorophyll-*a* in the model and should not be interpreted as a formal causal-mediation estimator or as evidence of a biological mediation pathway.

An additional methodological limitation concerns the scope of the sensitivity analysis. Although predictive performance and SHAP-based interpretation were stable across the tested random seeds, this assessment was conducted under fixed hyperparameters and a fixed cross-validation design. Consequently, it primarily evaluates sensitivity to stochastic model fitting and does not quantify uncertainty associated with hyperparameter optimization, alternative cross-validation grouping schemes, or data preprocessing choices, including missing-value imputation and predictor transformations. Stability under this fixed configuration should therefore not be interpreted as evidence of global robustness of the entire analytical workflow.

Finally, transition points derived from SHAP dependence curves represent exploratory, model-derived features rather than experimentally validated physiological thresholds or universal ecological breakpoints, and the residual-decoupling metric should be interpreted as an empirical diagnostic rather than as a direct measure of physiological efficiency [44].

4.7. Potential Integration with Satellite Remote Sensing and Spatial Applications

The proposed analytical workflow could be extended by incorporating satellite remote-sensing products, potentially enabling the investigation of environmental controls on coastal primary production over broader spatial scales. Satellite-derived chlorophyll-*a* concentrations, sea surface temperature, and light-related variables could complement in situ observations and serve as predictors within the Random Forest–SHAP framework, building on established satellite-based approaches to marine primary production [1,5]. Previous work has also demonstrated the feasibility of integrating field observations and satellite-derived sea surface temperature within a machine-learning workflow to

investigate phytoplankton biomass and productivity patterns [13]. In particular, spatially averaged satellite-derived variables could characterize environmental conditions over a defined coastal area, allowing the proposed workflow to investigate temporal variability at an aggregated spatial scale without substantial changes to its computational structure. Under this configuration, the date-grouped cross-validation adopted in the present study could be retained to evaluate predictive performance on held-out temporal groups, provided that the response variable is representative of the same spatial scale. However, because the present analysis is based on a single monitoring station, applications across multiple coastal areas or points would require independent validation across locations and spatially structured cross-validation to assess geographic generalization [38]. Particular attention should also be given to spatial consistency between predictors and response variables, satellite retrieval uncertainties in optically complex coastal waters [51], and environmental variables that cannot be directly observed from space. Potential collinearity among satellite-derived predictors, particularly those based on shared spectral information or retrieval algorithms, should also be assessed, as dependence among predictors may complicate the attribution and interpretation of individual SHAP contributions [52].

5. Conclusions

Our results indicate that environmental forcing influences primary production predominantly through its effects on phytoplankton biomass, while light availability and water-column structure account for much of the remaining variability after biomass has been considered. The hierarchical modelling framework revealed that chlorophyll-*a* is the dominant predictor of primary production, whereas sampling depth and PAR remain important secondary predictors. The residual-decoupling analysis further identified sampling depth and PAR as the main environmental variables associated with deviations from the expected chlorophyll-*a*–primary production relationship. However, these residuals should be regarded as empirical diagnostics integrating multiple sources of variability, including photoacclimation, phytoplankton community composition, unmeasured environmental drivers, and observational error, rather than as a direct proxy for biomass-specific photosynthetic efficiency.

The strong attenuation of the temperature signal after including chlorophyll-*a* suggests that a substantial fraction of the predictive information associated with temperature is mediated through phytoplankton biomass rather than acting independently. More broadly, these findings highlight that future changes in coastal primary production are likely to depend not only on environmental controls over phytoplankton biomass, but also on how light climate and water-column structure regulate biomass-specific productivity. The integration of hierarchical modelling with interpretable machine-learning techniques provides a transparent approach for disentangling the complementary pathways linking environmental variability, phytoplankton biomass and marine primary production. From an operational perspective, this approach can help translate routinely collected monitoring data into ecologically meaningful information on changes in ecosystem functioning by coupling predictive capability with an interpretable assessment of environmental drivers. Its application to long-term monitoring programmes may support the identification of environmental conditions associated with shifts in phytoplankton biomass and primary production, providing a transferable tool for data-driven coastal assessment and adaptive management. The approach also offers potential for integration with operational remote-sensing products, particularly for variables such as chlorophyll-*a*, sea-surface temperature and light-related conditions, extending its applicability towards spatially explicit and near-real-time assessments of coastal ecosystem productivity.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/coasts6040052/s1. Figure S1: Missing-data percentages; Figure S2: Predictor distributions before and after imputation; Table S1: Missing-data summary and imputation diagnostics; Table S2: Optimal_RF_hyperparameters; Table S3: SHAP_transition_bootstrap.

Author Contributions: Conceptualization, T.S.; methodology, T.S. and J.T.; software, T.S.; validation, T.S., J.T., L.L., A.C. and M.D.C.; formal analysis, T.S.; investigation, T.S., J.T. and L.L.; resources, T.S., J.T. and L.L.; data curation, T.S., J.T. and L.L.; writing—original draft preparation, T.S.; writing—review and editing, T.S., J.T., L.L., A.C., M.D.C. and A.B.; visualization, T.S., J.T., L.L., A.C., M.D.C. and A.B.; supervision, A.C., M.D.C. and A.B.; project administration, A.B.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: The publication has been funded by EU—Next Generation EU Mission 4 "Education and Research"—Component 2: "From research to business"—Investment 3.1: "Fund for the realization of an integrated system of research and innovation infrastructures"—Project IR0000032—ITINERIS—Italian Integrated Environmental Research Infrastructures System—CUP B53C22002150006 and by the DTO-BioFlow–HORIZON-MISS-2022-OCEAN-01–101112823-FSTP Grant Agreement Projects "KAIROS-ZooplanKton data from Arctic marIne time-seRies to understand biOdiversity dynamicS".

Data Availability Statement: The data used in this study are available in PANGEA repository at the following link: https://doi.org/10.1594/PANGAEA.885413 and https://doi.org/10.1594/PANGAEA.919082.

Acknowledgments: The authors acknowledge the Research Infrastructures participating in the ITINERIS project with their Italian nodes: ACTRIS, ANAEE, ATLaS, CeTRA, DANUBIUS, DISSCO, e-LTER, ECORD, EMPHASIS, EMSO, EUFAR, Euro-Argo, EuroFleets, Geoscience, IBISBA, ICOS, JERICO, LIFEWATCH, LNS, N/R Laura Bassi, SIOS and SMINO.The authors acknowledge the use of AI tool (ChatGPT) for English language editing and grammatical improvements. The authors reviewed and approved all suggested changes and take full responsibility for the final content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor European Commission can be held responsible for them.

References

1. Westberry, T.K.; Behrenfeld, M.J.; Siegel, D.A.; Boss, E. Gross and net primary production in the global ocean: An ocean color remote sensing perspective. *Earth-Sci. Rev.* **2023**, *237*, 104322. https://doi.org/10.1016/j.earscirev.2023.104322.
2. Izett, R.W.; Fennel, K.; Stoer, A.C.; Nicholson, D.P. Reviews and syntheses: Expanding the global coverage of gross primary production and net community production measurements using Biogeochemical-Argo floats. *Biogeosciences* **2024**, *21*, 13–47. https://doi.org/10.5194/bg-21-13-2024.
3. Dokulil, M.T.; Qian, K. Photosynthesis, carbon acquisition and primary productivity of phytoplankton: A review dedicated to Colin Reynolds. *Hydrobiologia* **2021**, *848*, 77–94. https://doi.org/10.1007/s10750-020-04321-y.
4. Halsey, K.H.; Jones, B.M. Phytoplankton Strategies for Photosynthetic Energy Allocation. *Annu. Rev. Mar. Sci.* **2015**, *7*, 265–297. https://doi.org/10.1146/annurev-marine-010814-015813.
5. Behrenfeld Michael, J.; Falkowski Paul, G. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.* **1997**, *42*. https://doi.org/10.4319/lo.1997.42.1.0001.
6. Falkowski, P.G.; Raven, J.A. *Aquatic Photosynthesis*, 2nd ed.; Princeton University Press: Princeton, NJ, USA, 2007.

7. 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. Climate-driven trends in contemporary ocean productivity. *Nature* **2006**, *444*, 752–755. https://doi.org/10.1038/nature05317.
8. Chavez, F.P.; Messié, M.; Pennington, J.T. Marine primary production in relation to climate variability and change. *Annu. Rev. Mar. Sci.* **2011**, *3*, 227–260. https://doi.org/10.1146/annurev.marine.010908.163917.
9. Cloern, J.E.; Foster, S.Q.; Kleckner, A.E. Phytoplankton primary production in the world's estuarine–coastal ecosystems. *Biogeosciences* **2014**, *11*, 2477–2501. https://doi.org/10.5194/bg-11-2477-2014.
10. Behrenfeld, M.J.; Boss, E.; Siegel, D.A.; Shea, D.M. Carbon-based ocean productivity and phytoplankton physiology from space, Global Biogeochem. *Cycles* **2005**, *19*, GB1006. https://doi.org/10.1029/2004GB002299.
11. Sathyendranath, S.; Platt, T.; Kovac, Z.; Dingle, J.; Jackson, T.; Brewin, R.J.W.; Franks, P.; Maranon, E.; Kulk, G.; Bouman, H.A. Reconciling models of primary production and photoacclimation. *Appl. Opt.* **2020**, *59*, C100–C114. https://doi.org/10.1364/AO.386252.
12. Marañón, E.; Van Wambeke, F.; Uitz, J.; Boss, E.S.; Dimier, C.; Dinasquet, J.; Engel, A.; Haëntjens, N.; Pérez-Lorenzo, M.; Taillandier, V.; et al. Deep maxima of phytoplankton biomass, primary production and photoacclimation in the ocean. *Biogeosciences* **2021**, *18*, 1749–1767. https://doi.org/10.5194/bg-18-1749-2021.
13. Semeraro, T.; Titocci, J.; Liberatore, L.; Monti, F.; De Leo, F.; Ingrosso, G.; Shokri, M.; Basset, A. Analytical Workflow for Tracking Aquatic Biomass Responses to Sea Surface Temperature Changes. *Environments* **2025**, *12*, 210. https://doi.org/10.3390/environments12070210.
14. Semeraro, T.; Titocci, J.; Liberatore, L.; Monti, F.; Cazzetta, A.; Pinna, M.; Shokri, M.; Basset, A. A Machine Learning Framework for Assessing the Sensitivity of Regional Ocean Productivity to Climate Change. *Environments* **2026**, *13*, 137. https://doi.org/10.3390/environments13030137.
15. Pichler, M.; Hartig, F. Machine learning and deep learning—A review for ecologists. *Methods Ecol. Evol.* **2023**, *14*, 994–1016. https://doi.org/10.1111/2041-210X.14061.
16. Cazzetta, A.; Zangaro, F.; Marcucci, F.; Julius, O.T.; Rainò, M.; Shauer, M.; Massaro, R.; Semeraro, T.; Basset, A.; Pinna, M. Applying Artificial Intelligence (AI) Innovative Tools for Ecological Research and Monitoring of Transitional Water Ecosystems: A Systematic Review. *Environments* **2026**, *13*, 193. https://doi.org/10.3390/environments13040193.
17. Jemeļjanova, M.; Kmoch, A.; Uuemaa, E. Adapting machine learning for environmental spatial data—A review. *Ecol. Inform.* **2024**, *81*, 102634. https://doi.org/10.1016/j.ecoinf.2024.102634. ISSN 1574-9541.
18. Huang, J.C.; Ko, K.M.; Shu, M.H.; Hsu, B.M. Application and comparison of several machine learning algorithms and their integration algorithms in regression problems. *Neural Comput. Appl.* **2020**, *32*, 5461–5469. https://doi.org/10.1007/s00521-019-04644-5.
19. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. https://doi.org/10.1007/s42979-021-00592-x.
20. Cao, Y.; Nakhjiri, A.T.; Ghadiri, M. Different applications of machine learning approaches in materials science and engineering: Comprehensive review. *Eng. Appl. Artif. Intell.* **2024**, *135*, 108783. https://doi.org/10.1016/j.engappai.2024.108783.
21. Sebt, M.V.; Sadati-Keneti, Y.; Rahbari, M.; Gholipour, Z.; Mehri, H. Regression Method in Data Mining: A Systematic Literature Review. *Arch. Comput. Methods Eng.* **2024**, *31*, 3515–3534. https://doi.org/10.1007/s11831-024-10088-5.
22. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
23. Welchowski, T.; Maloney, K.O.; Mitchell, R.M.; Schmid, M. Techniques to improve ecological interpretability of black box machine learning models. *J. Agric. Biol. Environ. Stat.* **2021**, *27*, 175–197. https://doi.org/10.1007/s13253-021-00479-7.
24. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS), Long Beach, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4766–4777.
25. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 3rd ed.; Lulu Press, Inc.: Durham, NC, USA 2025. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 6 June 2026).
26. Retzlaff, C.O.; Angerschmid, A.; Saranti, A.; Schneeberger, D.; Röttger, R.; Müller, H.; Holzinger, A. Post-Hoc vs. Ante-Hoc Explanations: XAI Design Guidelines for Data Scientists. *Cogn. Syst. Res.* **2024**, *86*, 101243. https://doi.org/10.1016/j.cogsys.2024.101243.
27. Bennetot, A.; Donadello, I.; El Qadi El Haouari, A.; Dragoni, M.; Frossard, T.; Wagner, B.; Sarranti, A.; Tulli, S.; Trocan, M.; Chatila, R.; et al. A practical tutorial on explainable AI techniques. *ACM Comput. Surv.* **2025**, *57*, 50. https://doi.org/10.1145/3670685.

28. Behrenfeld, M.J.; O'Malley, R.T.; Boss, E.S.; Westberry, T.K.; Graff, J.R.; Halsey, K.H.; Milligan, A.J.; Siegel, D.A.; Brown, M.B. Revaluating ocean warming impacts on global phytoplankton. *Nat. Clim. Change* **2016**, *6*, 323–330. https://doi.org/10.1038/nclimate2838.
29. Fontaine, D.N.; Marrec, P.; Menden-Deuer, S.; Sosik, H.M.; Rynearson, T.A. Time series of phytoplankton net primary production reveals intense interannual variability and size-dependent chlorophyll-specific productivity on a continental shelf. *Limnol. Oceanogr.* **2025**, *70*, 203–216. https://doi.org/10.1002/lno.12749.
30. Bode, A.; Alvarez, M.; Ruiz-Villareal, M.; Varela, M. Time series of hydrographic, chemical and primary production variables for a shelf station off A Coruña (NW Spain): 1989–2016 [dataset]. PANGAEA. *Ocean Dyn.* **2018**, *69*, 861–873. https://doi.org/10.1594/PANGAEA.885413.
31. Bode, A.; Alvarez, M.; Castaño, M.; García, L.; Louro, M.A.; Nieto-Cid, M.; Ruiz-Villarreal, M.; Varela, M.M. Time series of hydrographic, chemical and primary production variables for a shelf station off A Coruña (NW Spain): 2017–2018 [dataset]. PANGAEA. 2020.. https://doi.pangaea.de/10.1594/PANGAEA.919087
32. Bode, A.; Álvarez, M.; Ruíz-Villarreal, M.; Varela, M.M. Changes in phytoplankton production and upwelling intensity off A Coruña (NW Spain) for the last 28 years. *Ocean Dyn.* **2019**, *69*, 861–873. https://doi.org/10.1007/s10236-019-01278-y.
33. Bode, A.; Álvarez, M.; García, L.; Louro, M.Á.; Nieto-Cid, M.; Ruiz-Villarreal, M.; Varela, M.M. Climate and Local Hydrography Underlie Recent Regime Shifts in Plankton Communities off Galicia (NW Spain). *Oceans* **2020**, *1*, 181–197. https://doi.org/10.3390/oceans1040014.
34. Zuur, A.F.; Ieno, E.N.; Elphick, C.S. A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **2010**, *1*, 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x.
35. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03.
36. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
37. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 3rd ed.; O'Reilly Media: Sebastopol, CA, USA, 2022.
38. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. https://doi.org/10.1111/ecog.02881.
39. Were, K.; Bui, D.T.; Dick, O.B.; Singh, B.R. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol. Indic.* **2015**, *52*, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028.
40. Balasubramaniam, T.; Mohotti, W.A.; Sabir, K.; Nayak, R. Feature engineering on climate data with machine learning to understand time-lagging effects in pasture yield prediction. *Ecol. Inform.* **2025**, *86*, 103011. https://doi.org/10.1016/j.ecoinf.2025.103011.
41. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
42. Fox, E.W.; Hill, R.A.; Leibowitz, S.G.; Olsen, A.R.; Thornbrugh, D.J.; Weber, M.H. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environ. Monit. Assess.* **2017**, *189*, 316. https://doi.org/10.1007/s10661-017-6025-0.
43. Hanberry, B.B. Practical guide for retaining correlated climate variables and unthinned samples in species distribution modeling, using random forests. *Ecol. Inform.* **2024**, *79*, 102406. https://doi.org/10.1016/j.ecoinf.2023.102406.
44. Olden, J.D.; Lawler, J.J.; Poff, N.L. Machine learning methods without tears: A primer for ecologists. *Q. Rev. Biol.* **2008**, *83*, 171–193. https://doi.org/10.1086/587826.
45. Kirk, J.T.O. *Light and Photosynthesis in Aquatic Ecosystems*, 3rd ed.; Cambridge University Press: Cambridge, MA, USA, 2011.
46. Westberry, T.; Behrenfeld, M.J.; Siegel, D.A.; Boss, E. Carbon-based primary productivity modeling with vertically resolved photoacclimation. *Glob. Biogeochem. Cycles* **2008**, *22*, GB2024. https://doi.org/10.1029/2007GB003078.
47. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, MA, USA, 2009.
48. Reynolds, C.S. *The Ecology of Phytoplankton*; Cambridge University Press: Cambridge, MA, USA, 2006.
49. Litchman, E.; Klausmeier, C.A. Trait-based community ecology of phytoplankton. *Annu. Rev. Ecol. Evol. Syst.* **2008**, *39*, 615–639. https://doi.org/10.1146/annurev.ecolsys.39.110 707.173549.

50. Calbet, A.; Landry, M.R. Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. *Limnol. Oceanogr.* **2004**, *49*, 51–57. https://doi.org/10.4319/lo.2004.49.1.0051.
51. Mouw, C.B.; Greb, S.; Aurin, D.; DiGiacomo, P.M.; Lee, Z.; Twardowski, M.; Binding, C.; Hu, C.; Ma, R.; Moore, T.; et al. Aquatic color radiometry remote sensing of coastal and inland waters: Challenges and recommendations for future satellite missions. *Remote Sens. Environ.* **2015**, *160*, 15–30. https://doi.org/10.1016/j.rse.2015.02.001.
52. Aas, K.; Jullum, M.; Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.* **2021**, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.