MISSIONE 4
ISTRUZIONE
RICERCA

Deliverable D8.17 – Activity 8.2

CNR IRET LE

BIOMASS VRE

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Finanziato dall'Unione europea
NextGenerationEU

ITINERIS

Deliverable number:	D8.17
Work package:	WP8
Intermediate Objective:	IO8.10
Deliverable type:	X Document, report
	☐ Websites, patent filings, videos, etc.
	☐ Other: please specify .....
Dissemination level:	X Public
	☐ Restricted
Estimated delivery (bimester):	B19
Actual delivery date:	31/12/2025
Author(s) (Partner-OU):	Teodoro Semeraro, Jessica Titocci, Lorenzo Liberatore, Flavio Monti (CNR IRETLE), Alberto Basset (UNISAL),
Reviewed by:	ITINERIS Executive Board
Note:	

Sommario

1. OVERVIEW OF THE WORK PACKAGE 8.2

The Work Package (WP) 8.2 - Virtual Research Environment for aquatic Biomass services (BIOMASS VRE) is part of WP 8 (Virtual Research Environments and Cross-disciplinary Activities) of the Italian Integrated Environmental Research Infrastructures System (ITINERIS). ITINERIS is a project funded by EU - Next Generation EU PNRR- 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".

The main goal of WP 8.2 is to strengthen the collaboration between the e-Science European infrastructure for biodiversity and ecosystem research of LifeWatch ERIC, through its national node of LifeWatch Italy, and the CNR-IRET with the creation of a specific Virtual Research Environment: Biomass VRE. The planned activity is to design, implement and validate a VRE on the responses of aquatic and terrestrial ecosystems to climate change, based on earlier developments by LifeWatch ERIC.

This document represents the final report and deliverable for the VRE BIOMASS within the ITINERIS project, related to the last development and implementation of innovative technological solutions adopted in the VRE for research on aquatic and terrestrial biomass.

2. STATE OF THE ART

EquiUp S.R.L., the company that won the contract to develop the virtual research environment (VRE), has completed the development of the BIOMASS VRE. The BIOMASS VRE has been released and is now installed on the Data Centre of the LifeWatch Italy Research Infrastructure and is accessible through the dedicated webpage developed for navigation of all its e-Services and workflow.

The BIOMASS VRE website includes the following sections:

- **About Us** (A web page which includes an overview of the BIOMASS VRE project, objectives and main features);
- **e-Services** (a web page which describes all workflows and additional e-Services developed components);
- **e-Training** (a web page linked to the LifeWatch Italy e-learning platform containing all training resources);
- **Use cases** (a web page which describes all use cases developed for each workflow),
- **Contact Us** (a web page with contact details for project information and support).

The website has been developed using WordPress and is in compliance with accessibility and user experience guidelines defined in the project call.

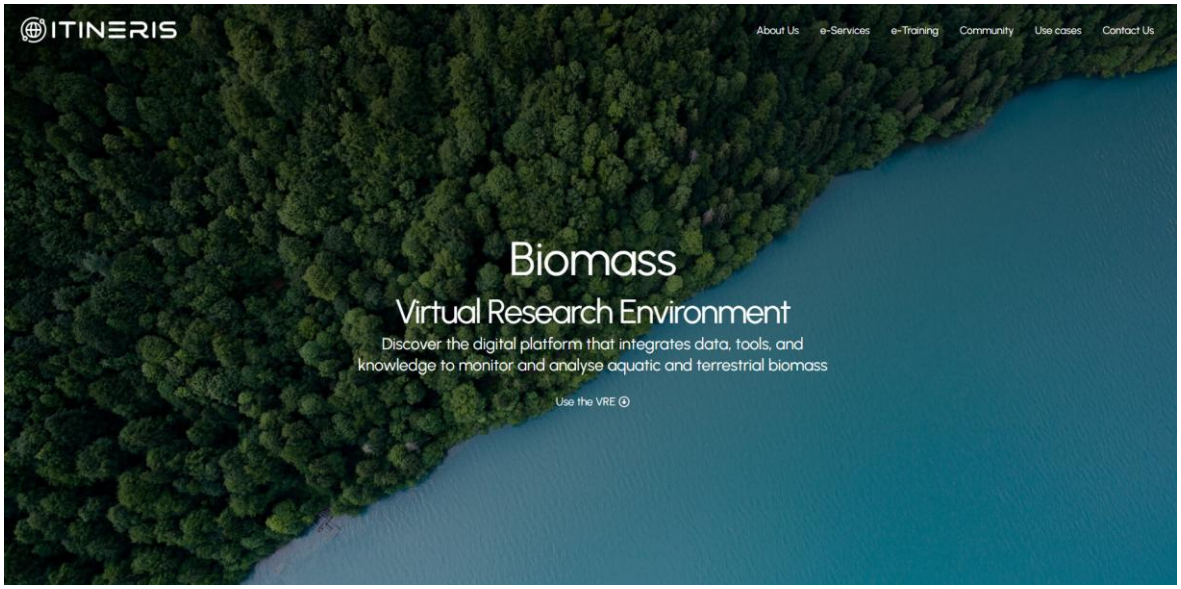


Figure 1. The BIOMASS VRE website.

VRE Architecture

The implementation of the Virtual Research Environment (VRE) is based on an integration of the LifeWatch ERIC VREs (i.e., NaaVRE and Tesseract), as an infrastructure designed to manage, access, and process heterogeneous datasets in a scalable and interoperable manner. This architecture allows**:**

1. Integration of Scientific Workflows – The implemented workflows enable the analysis of aquatic and terrestrial biomass and its relationship with global warming
2. User Authentication and Access Management – Configuration of an authentication system to ensure security and controlled access to data.
3. Automated Data Analysis – Setting up pipelines for preprocessing, analysis, and visualization of environmental and climate data.
4. Interoperability with Other Research Systems – Connection with platforms such as LifeWatch ERIC, ECOPORTAL.

3. E-SERVICES

The available services of the Biomass VRE are:

- The *Biomass-VRE Thesaurus*:
- Workflow 1: *Variation of net primary productivity in aquatic environments in relation to global warming*
- Workflow 2: *Analysis of changes in net primary production in terrestrial ecosystems*.
- Workflow 3: *Trait response of consumers to intrinsic and extrinsic factors*
- Workflow 4: *Application of machine learning for ecological analysis*
- Workflow 5: *Effects of net primary production on biodiversity in aquatic ecosystem*

- Workflow 6: *Analysis of the relationship between biotic and abiotic factors*
- Additional e-Services: These are indipendent e-services supporting workflow analysis and can be integrated into workflows.

Biomass-VRE Thesaurus

The Biomass-VRE Thesaurus provides a standardised vocabulary to support navigation, data harmonisation, interoperability, and data comparability within the Biomass Virtual Research Environment (VRE) and across research platforms (Figure 2).

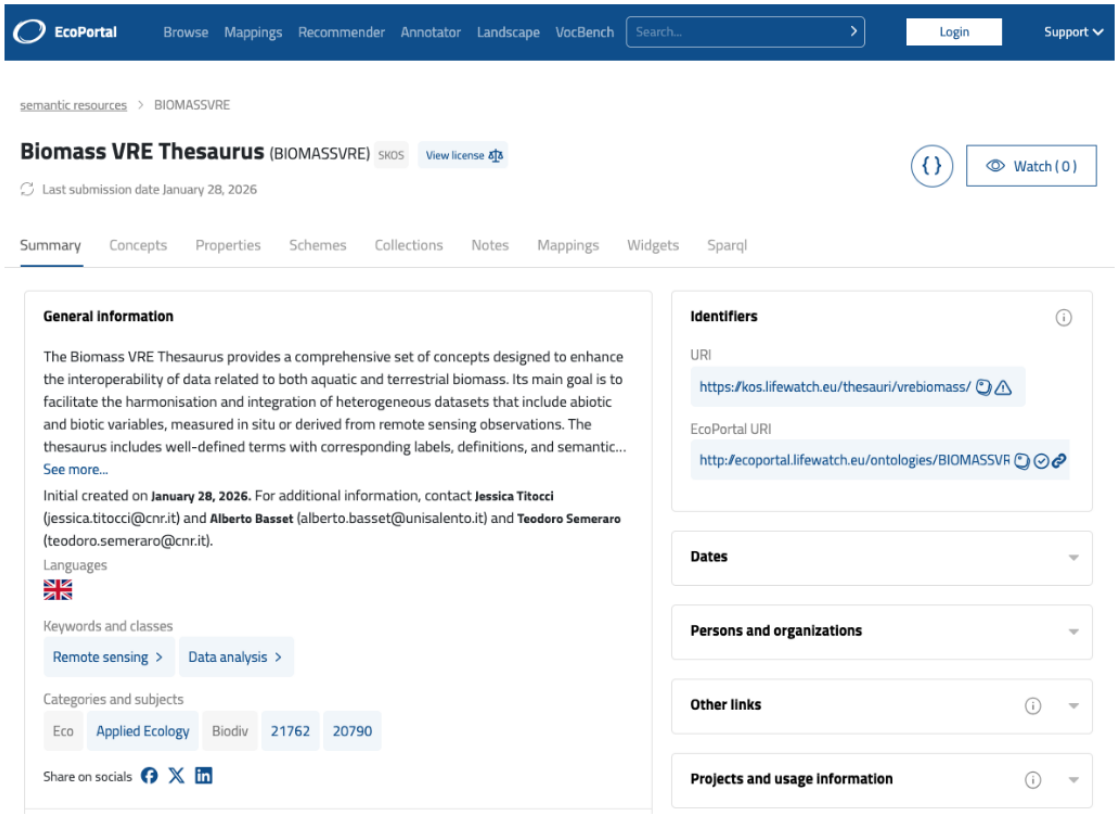
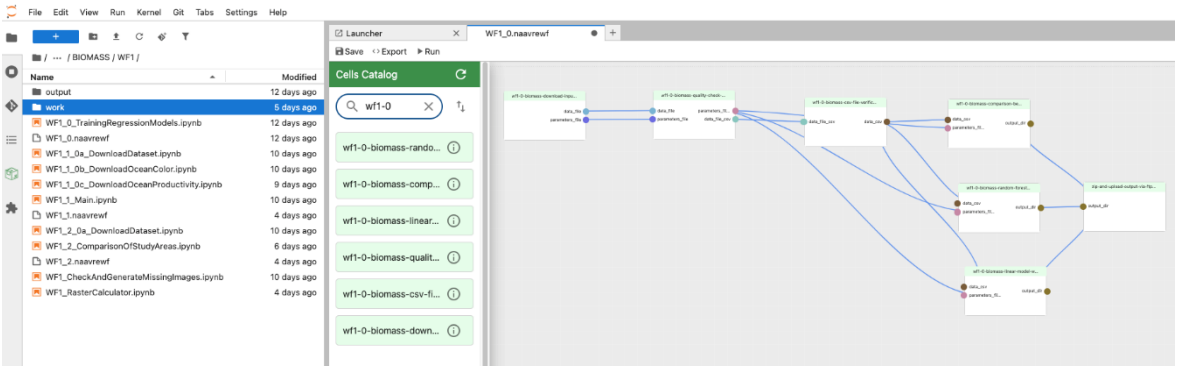


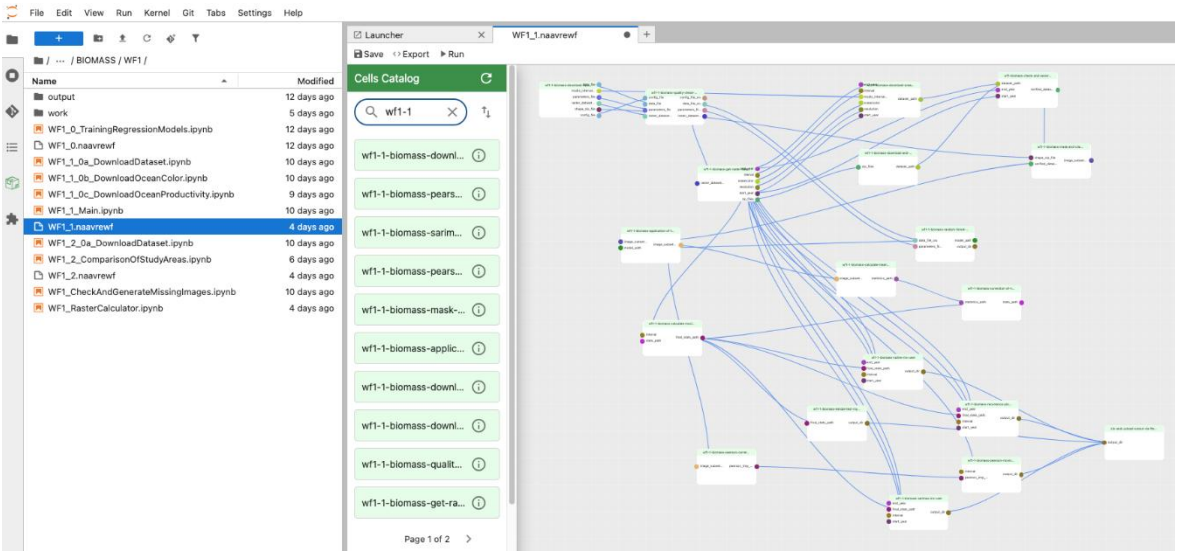
Figure 2. Biomass-VRE Thesaurus.

Workflow 1: Variation of net primary productivity in aquatic environments in relation to global warming

This analytical workflow (Figure 3) investigates the relationships between global warming, primary production, and other key environmental variables in aquatic ecosystems. It enables the integration of in situ aquatic monitoring data with remote sensing observations through the application of regression-based approaches, such as random forest and multiple linear regression models. Multiple linear regression and random forest algorithms can be trained using in situ measurements and

subsequently applied to remote sensing imagery, which are used as predictor variables consistent with the trained models, to generate new datasets. The workflow supports the construction of time series for both the target variable and the associated remote sensing datasets used, enabling the analysis of temporal trends and recurrence patterns. In addition, it incorporates predictive modeling techniques to assess ecosystem responses and to forecast future dynamics under changing environmental conditions.





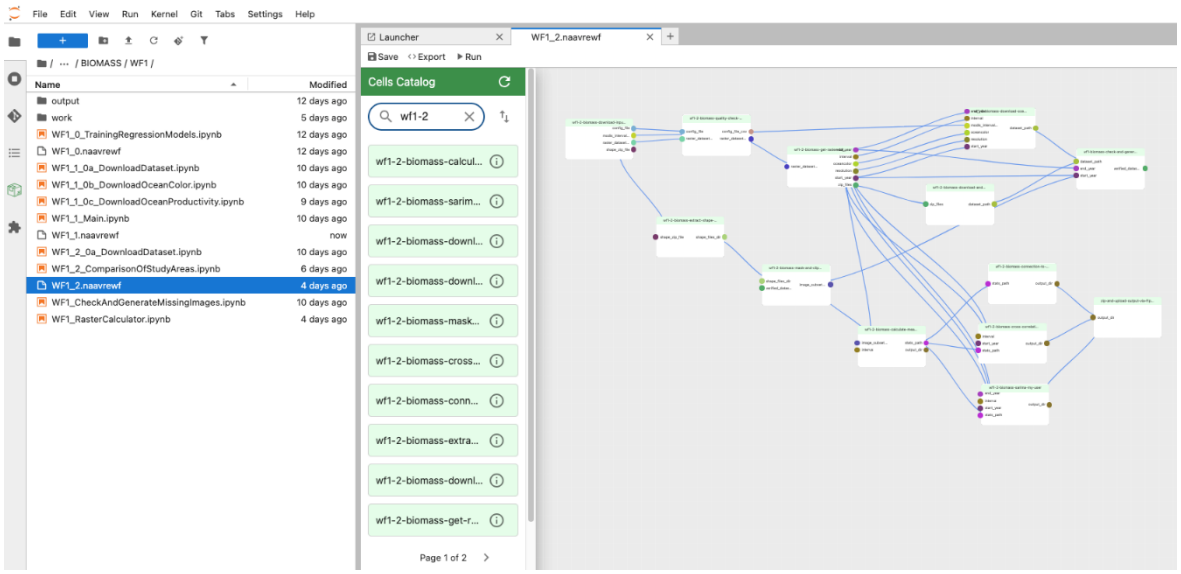


Figure 3: Schematic illustration of workflow 1 for Biomass VRE

Workflow 2: Analysis of changes in net primary production in terrestrial ecosystems

This analytical workflow (Figure 4) investigates the relationship between global warming and primary production (or its proxies) and other variables in terrestrial environments using MODIS products.

The workflow supports the construction of time series, the analysis of temporal trends and recurrence plots, and the application of predictive modeling approaches to vegetation indices and other MODIS-derived products in terrestrial ecosystems.

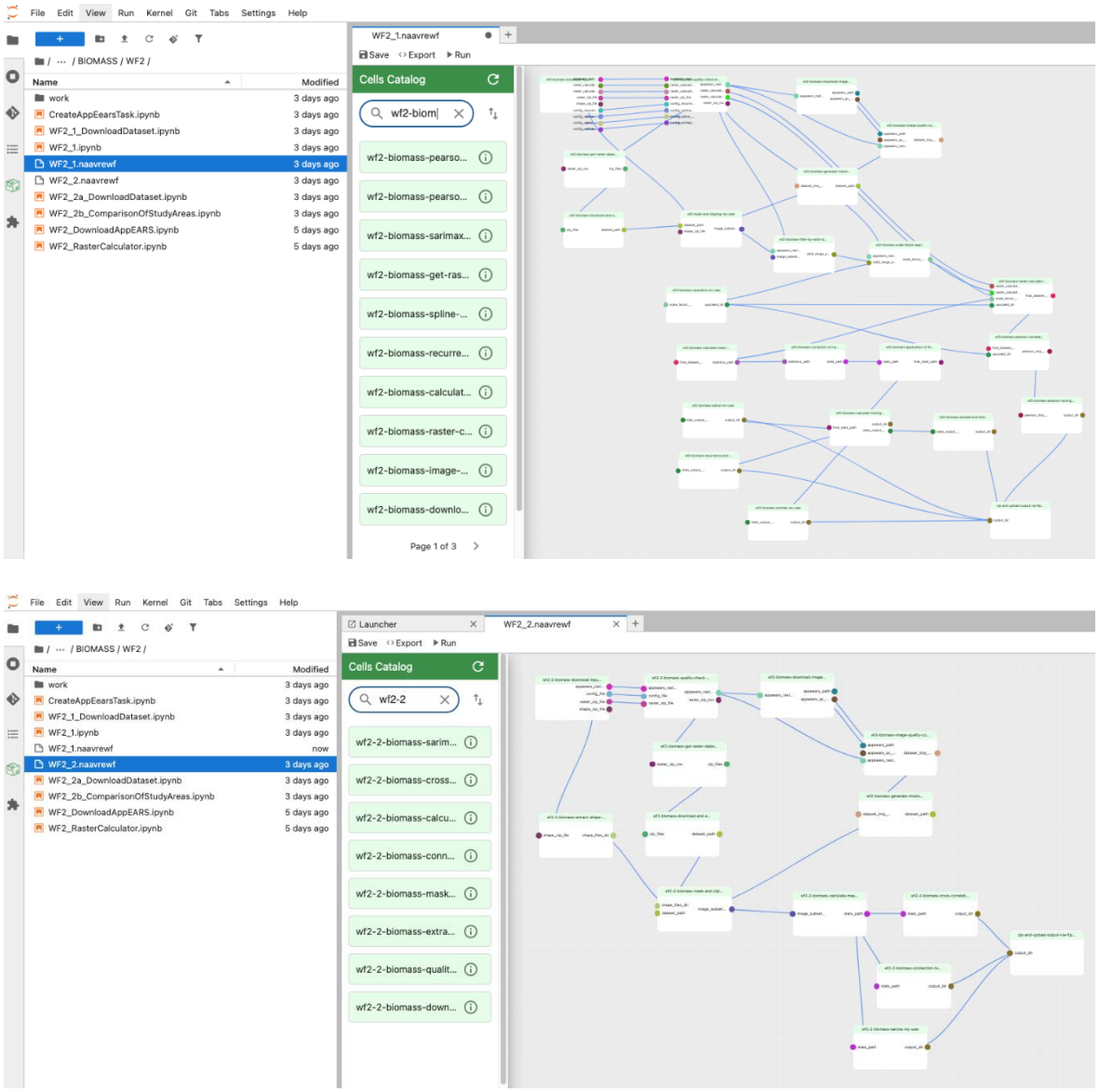


Figure 4: Schematic illustration of Workflow 2 for Biomass VRE.

Workflow 3: Trait response of consumers to intrinsic and extrinsic factors: implications on standing biomass in coastal ecosystems

This analytical workflow (Figure 5) examines the response of stable biomass consumers to intrinsic and extrinsic factors in coastal ecosystems; it is based on the assessment of individual and population quantitative responses and their scaling up to the community and ecosystem levels. Workflow 3 (WF3) applies a range of analytical methods to a curated dataset of behavioral and functional traits of vertebrate species, with the aim of evaluating how these traits affect the relationship between home range and body mass and the implication to the higher levels of the ecological hierarchy. This

workflow can be used to analyse how temperature increases affect the metabolic costs and spatial use of trophic resources by heterotrophs.

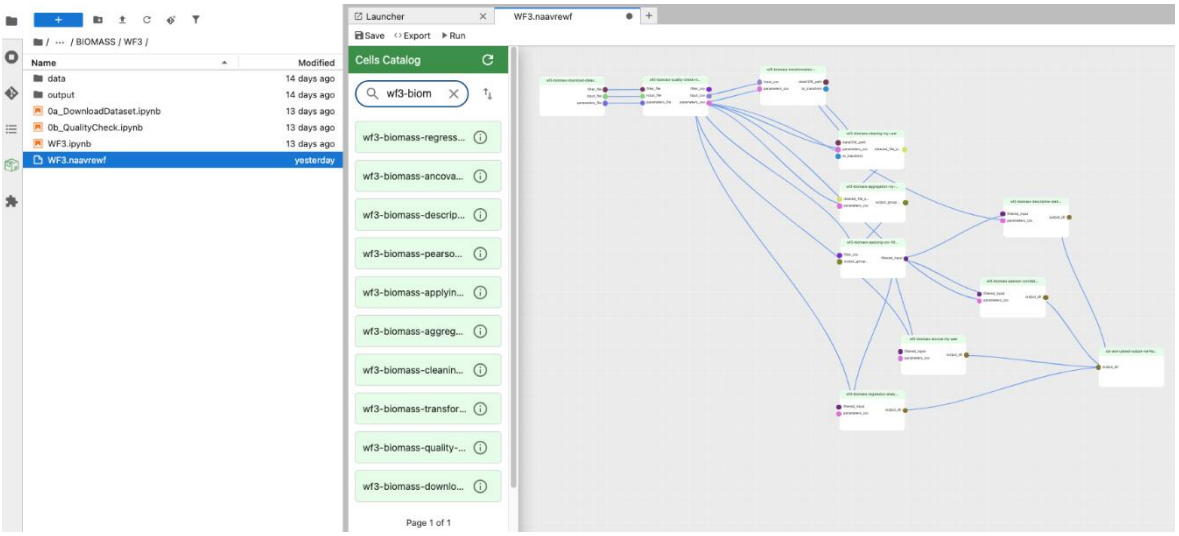


Figure 5: Schematic illustration of Workflow 3 for Biomass VRE

Workflow 4: Application of machine learning for ecological analysis

This analytical workflow (Figure 6) studies the application of machine learning models for ecological analysis. This analytical workflow explores a range of machine learning algorithms with a regression approach, such as eXtreme Gradient Boosting (XGBoost), multiple linear regression, neural networks, random forests and support vector machines, that can be flexibly applied to diverse ecological analyses. The workflow incorporates SHAP (SHapley Additive exPlanations) analysis to interpret individual predictions and determine the significance of global features by attributing a model's output to its input features.

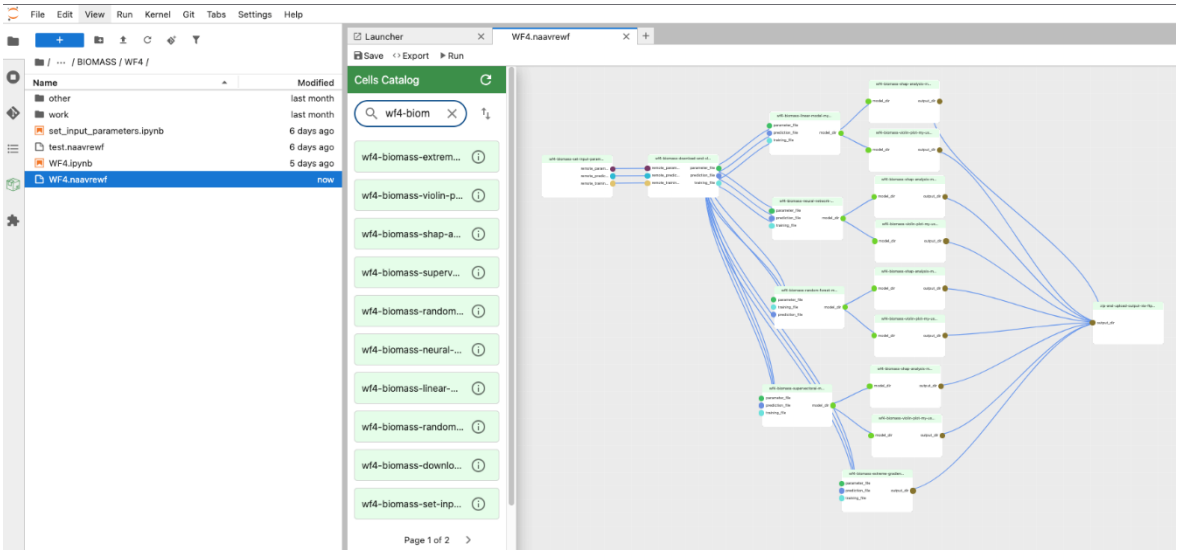


Figure 6: Schematic illustration of Workflow 4 for Biomass VRE

Workflow 5: Effects of net primary production on biodiversity in aquatic ecosystem

WF5 is a workflow designed to support the management and analysis of geospatial monitoring data. It does this by integrating in situ observations with information derived from satellite sensors (Ocean Productivity and Ocean Colour) (Figure 7). This allows users to explore the relationships between aquatic biodiversity indicators or diversity indices, derived from in situ measurements, with net primary production, chlorophyll concentration, as well as other environmental data obtained from satellite products. The system also incorporates data on water depth and distance from the coast. WF5 provides an analytical framework combining multivariate analyses to extract key environmental gradients (PCA), assess the influence of environmental variables on biodiversity patterns (multiple linear regression and random forest), and group sites with similar biodiversity–productivity characteristics (clustering).

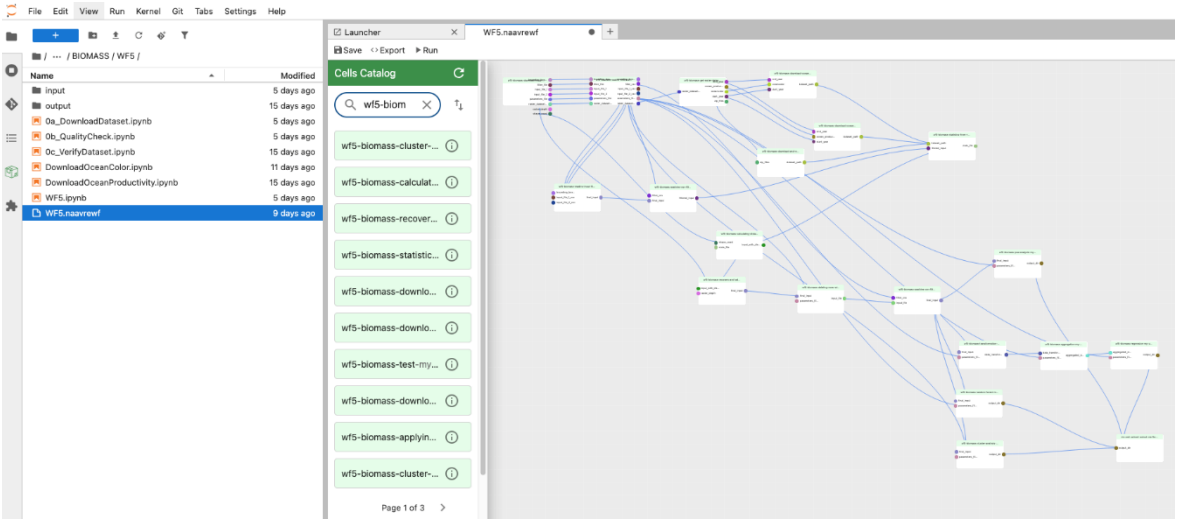


Figure 7: Schematic illustration of Workflow 5 for Biomass VRE

Workflow 6: Analysis of the relationship between biotic and abiotic factors

This analytical workflow (Figure 8) uses multivariate analysis to explore the interaction between biotic and abiotic factors. It allows multiple response variables to be examined simultaneously in relation to several environmental drivers using statistical tools such as redundancy analysis (RDA) and analysis of covariance (ANCOVA). When applied to the study of aquatic biomass, for example, this approach can clarify how biological components such as net primary production and chlorophyll concentration respond to environmental conditions including water temperature and nutrient availability.

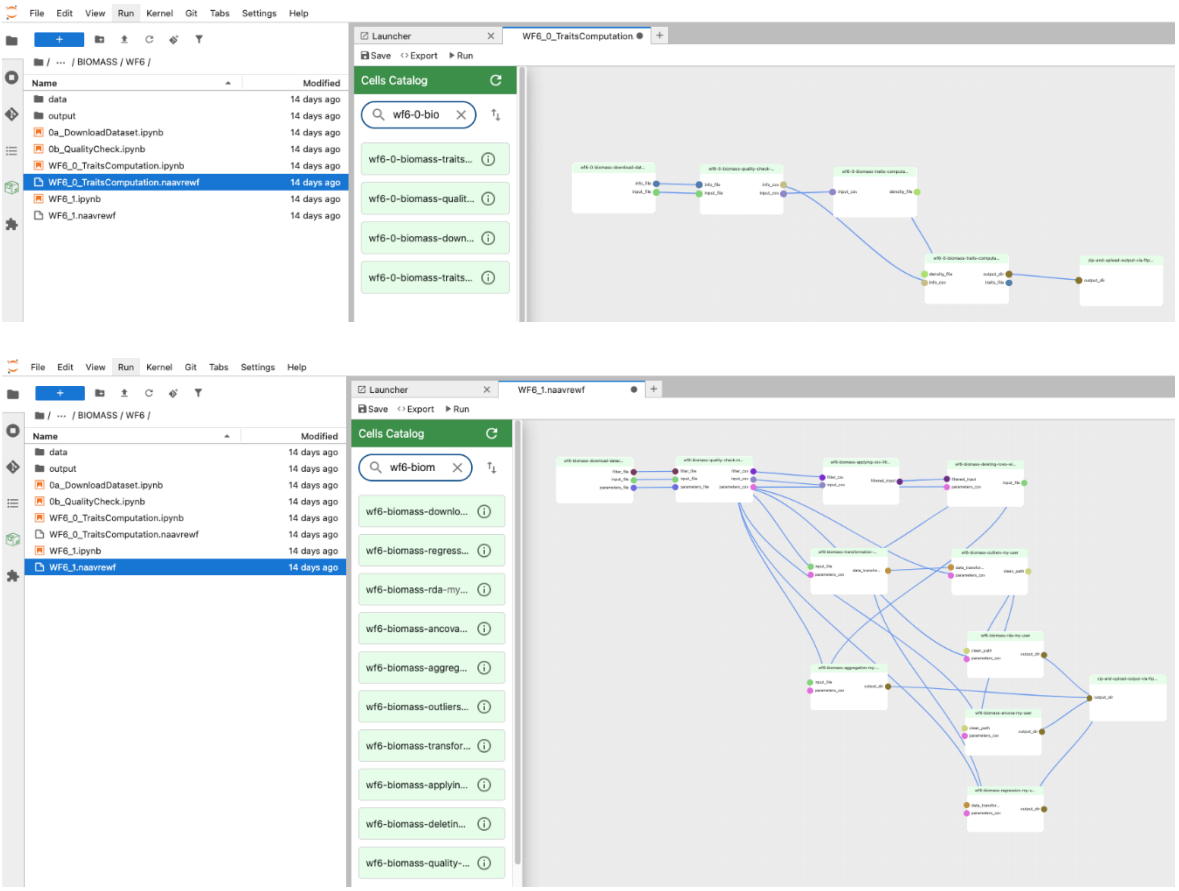


Figure 8: Schematic illustration of Workflow 6 for Biomass VRE

4. EXAMPLE OF USE CASES.

In this section three examples of use cases are reported.

WF 1 - Relationship between primary productivity and temperature in the marine environment

1. Relationship between primary productivity and temperature in the marine environment

The case study investigates how Primary Production (PP) is influenced by variability in Sea Surface Temperature (SST) and chlorophyll-a (Chl-a). It is based on an integrated framework that combines *in situ* ocean observations, satellite remote sensing data, and regression machine learning techniques to explore the links between these key variables.

This approach makes it possible to analyse how SST-driven changes propagate through phytoplankton dynamics to influence PP over long periods. Ultimately, the study aims to improve our understanding of the response of marine ecosystems to climate-driven changes in ocean

temperature and to provide more robust information on the evolution of oceanic primary production in a warming climate.

2. Dataset

The dataset used for this case study was derived from field measurements available on the *Ocean Productivity* website. The oceanographic dataset includes water temperature, chlorophyll-a concentration, and primary productivity estimates obtained using the ^{14}C method. For this analysis, we selected measurements taken at the sea surface within the equatorial zone. Further details on the original data types and processing methods are available at:https://orca.science.oregonstate.edu/field.data.c14.online.php

The dataset also includes sea surface temperature and chlorophyll-a concentration, derived from Ocean Color's remote sensing imagery, at an 8-day temporal resolution (https://oceandata.sci.gsfc.nasa.gov/l3/). In particular, the WF offers the ability to download various ocean colour products directly from the virtual research environment.

3. Method

The aim of this study is to develop an advanced analytical framework with which to understand the dynamics of Primary Production (PP) in relation to variability in Sea Surface Temperature (SST) and chlorophyll-a (Chl-a). The framework integrates *in situ* ocean measurements, remote sensing imagery and machine learning techniques to investigate the relationships between these key variables.

The methodology consists of the following main steps:

- Regression model development: Linear regression and Random Forest models were implemented to estimate PP using Chl-a and SST variables derived from field-based oceanographic data.
- Application of the regression model to satellite data: The model with the best performance was applied to a selected portion of the study area using 8-day Chl-a concentration and SST imagery derived from MODIS. This produced a new PP image product covering the period from 2004 to the end of 2023.
- Statistical analysis of time series: Statistical analyses were carried out on the SST, Chl-a and PP datasets to compute mean time series for each variable, and to assess temporal trends using the Kendall test, which identifies the direction and statistical significance of trends.
- Recurrence plot analysis: Recurrence plot analysis was applied to the SST, Chl-a and PP time series to investigate short-term variability and detect potential perturbations or regime shifts over time.
- Forecasting future scenarios: A SARIMAX model was implemented to forecast future PP dynamics in relation to Chl-a and SST. This provided possible future scenarios of primary production under varying oceanographic conditions.

4. Results

4.1. Regression Model Estimation

The models were trained and evaluated using cross-validation to limit overfitting and obtain a robust estimate of their predictive ability. This procedure allowed for a reliable assessment of model performance and their ability to generalize to independent data. The models were developed using a logarithmic transformation (ln) applied to the entire dataset. The results show that, in our case study, the Random Forest algorithm outperformed the Linear Model in predicting PP from Chl-a and SST (Table 1). Specifically, the Random Forest model achieved the highest R^2 and the lowest MAE and RMSE values, confirming its superior ability to capture variability in the dataset and to provide more accurate PP estimates with fewer prediction errors.

Table 1. Evaluation of the regression and random forest model accuracy and performance.

Results	R^2	RMSE	MAE
Linear Regression	0.8088	0.7256	0.5795
Random Forest	0.8264	0.6724	0.5116

4.2. Time series analysis of SST, Chl-a

The portion of the study area was extracted from each 8-day SST and Chl-a remote sensing image. For each 8-day image, the mean value of each variable was calculated to construct a time series of SST and Chl-a of the study area. A spline was applied to the time series derived for SST, Chl-a, to obtain a trend that visually describes the temporal evolution of each series (Figure 1).

Figure 1a illustrates the time series of mean Chl-a, which is characterised by high short-term variability and a weak long-term trend. The smoothing spline curve highlights a slight decrease until the mid-2010s, followed by an increase in more recent years.

Figure 1b displays the time series of mean SST, which exhibits a pronounced seasonal cycle superimposed on a long-term trend that is more regular than that observed for Chl-a. The smoothing spline indicates a gradual temperature increase from the mid-2000s to 2018–2020, followed by a slight decline in subsequent years.

For the missing scenes Chl_2022_0407_2022_0414 and SST_2018_0101_2018_0108, the gap created in the time series was filled by replacing the missing values with the average of the observations immediately before and after each gap.

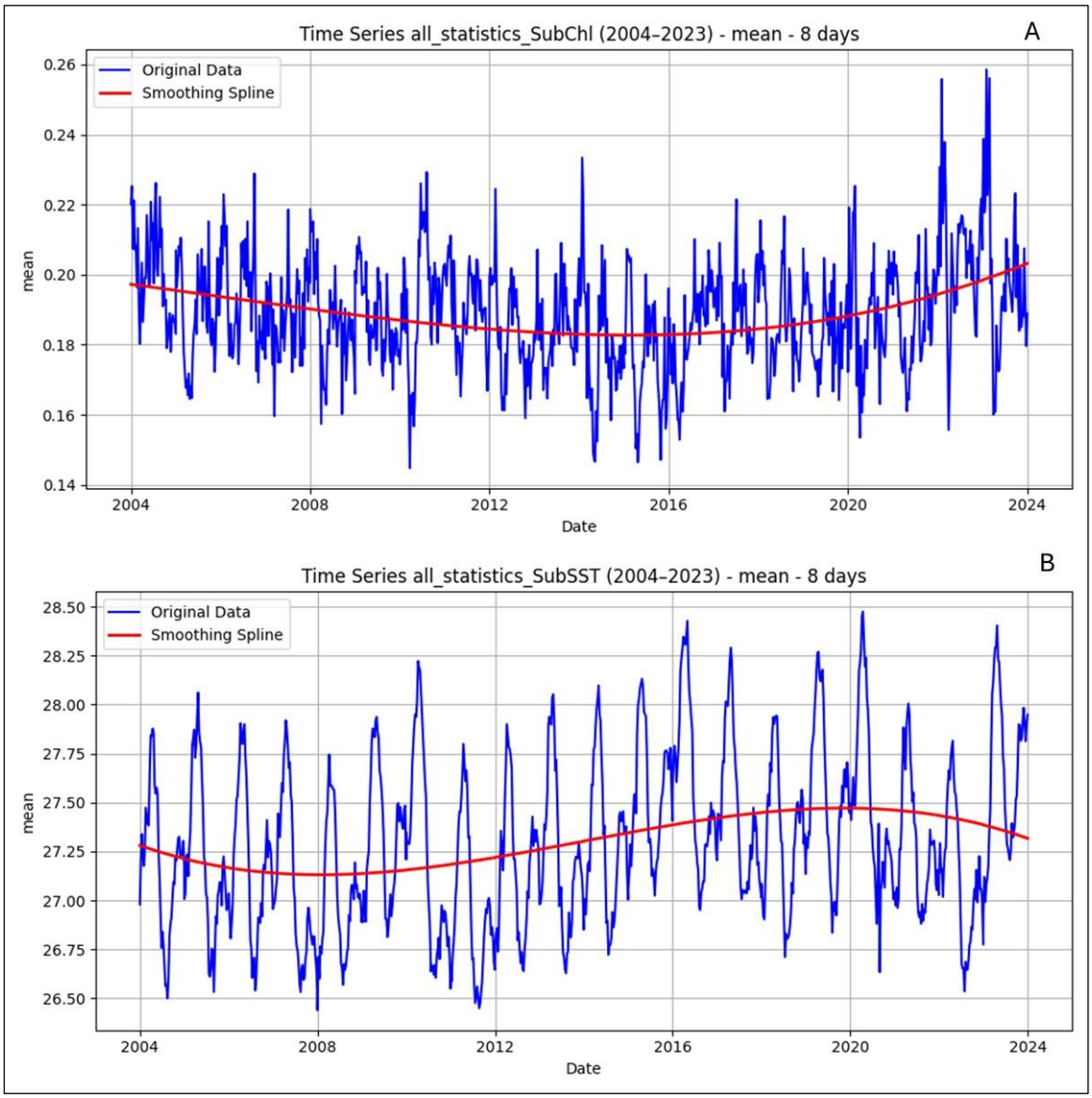


Figure 1. Temporal evolution of Chl-a (a) and SST (b) over the period 2004–2023, using 8-day composites.

The long-term trend analysis based on the Kendall test and the Theil–Sen estimator revealed markedly different behaviours between SST and Chl-a. For SST, a statistically significant increasing trend was detected ($\tau = 0.1594$, $p = 4.47\times10^{-13}$), with a positive slope estimated by the Theil–Sen method (0.000598). This result confirms a monotonic increase in temperature over the study period, consistent with warming signals reported for many marine regions globally. In contrast, the Chl-a series does not show any significant trend ($\tau = -0.0149$, $p = 0.4997$), and the estimated slope is near zero (-8.5×10^{-7}). This indicates that, despite substantial intra- and interannual variability, no persistent direction of change emerges.

The relationship between Chl-a and SST was analysed by computing the Pearson correlation coefficient for each pair of temporal images covering the study area for all time series. The results

indicate a scenario characterised by generally weak and predominantly negative correlations between the two variables. The distribution of correlation coefficients shows moderate variability, with values ranging from a minimum of –0.175 to a maximum of 0.087. However, the inverse relationship between Chl-a and SST, although weak, is a persistent feature of the system rather than the result of individual events or seasonal anomalies.

4.3. Recurrence Plot Analysis

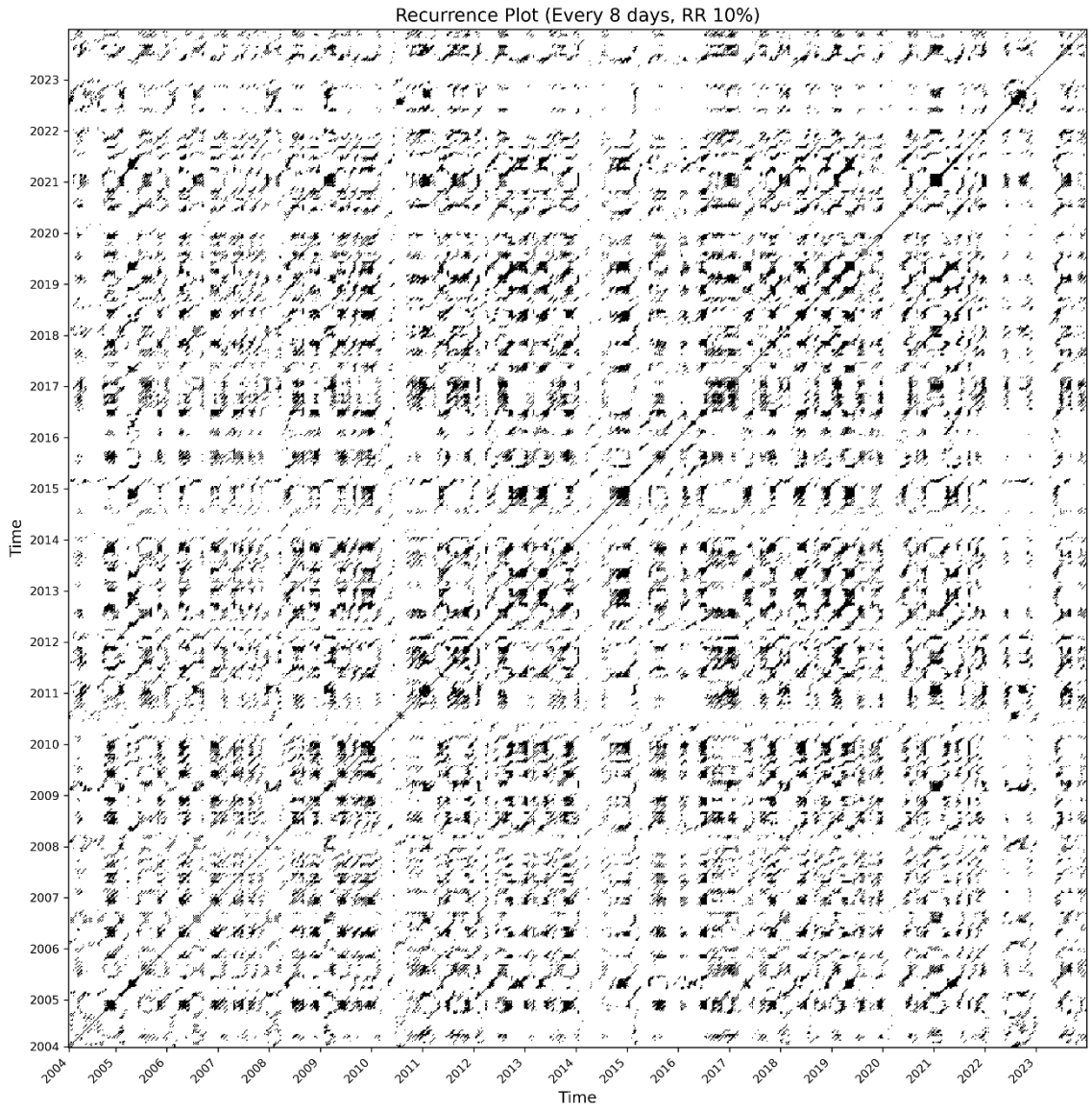
Recurrence plots (RPs) computed for SST and Chl-a reveal markedly different recurring structures between the two series, highlighting the dynamic and complex nature of the oceanographic processes involved. The SST series RP displays a highly organised structure characterised by regular diagonal patterns reflecting a strong cyclical component and an almost periodic dynamic (Figure 2a). In contrast, the Chl-a series exhibits a much more disordered RP with fragmented patterns that lack a dominant recurring structure (Figure 2b).

In addition to providing information on the system's cyclicity, RPs also allow potential disturbance events to be identified. The Chl-a RP shows interruptions to the diagonal structures, areas of low recurrence and irregular patterns arranged in 'patches' or short vertical or horizontal bands. These features are indicative of phases in which the system's trajectory in phase space temporarily deviates from its 'average' seasonal state, potentially representing signatures of episodic disturbance events (e.g. extreme meteorological conditions, intense mixing episodes, thermal anomalies, or abrupt changes in nutrient availability).

While the RP cannot identify the specific nature of individual events, analysis indicates that the seasonal regime is frequently modulated by disturbance episodes that interrupt the system's quasi-periodic dynamics.


Recurrence Plot (Every 8 days, RR 10%)
Time
Time
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

A



B

Figure 2. Comparison between two recurrence plots constructed using data sampled every 8 days. A) Recurrence plot for the SST time series. B) Recurrence plot for the Chl time series.

4.4. Estimation of PP time series

The Random Forest model was then used to predict PP using the 8-day Chl-a and SST images for the study area from 2004 to 2023. Since the Random Forest had been trained on log-transformed data, the same logarithmic transformations were applied to the imagery to ensure consistency with the model configuration. A spline was applied to the time series derived for PP to obtain a trend that visually describes the temporal evolution of PP (Figure 3).

Figure 3 shows the time series of mean primary productivity (PP) for the period 2004–2023, computed using 8-day composites. The series is characterized by marked short-term variability, with

regular oscillations that reflect the seasonal dynamics of the system. The smoothing spline curve highlights a weak long-term trend: after relatively higher values in the early part of the series, a slight decrease is observed up to around the mid-2010s, followed by a partial increase in more recent years.

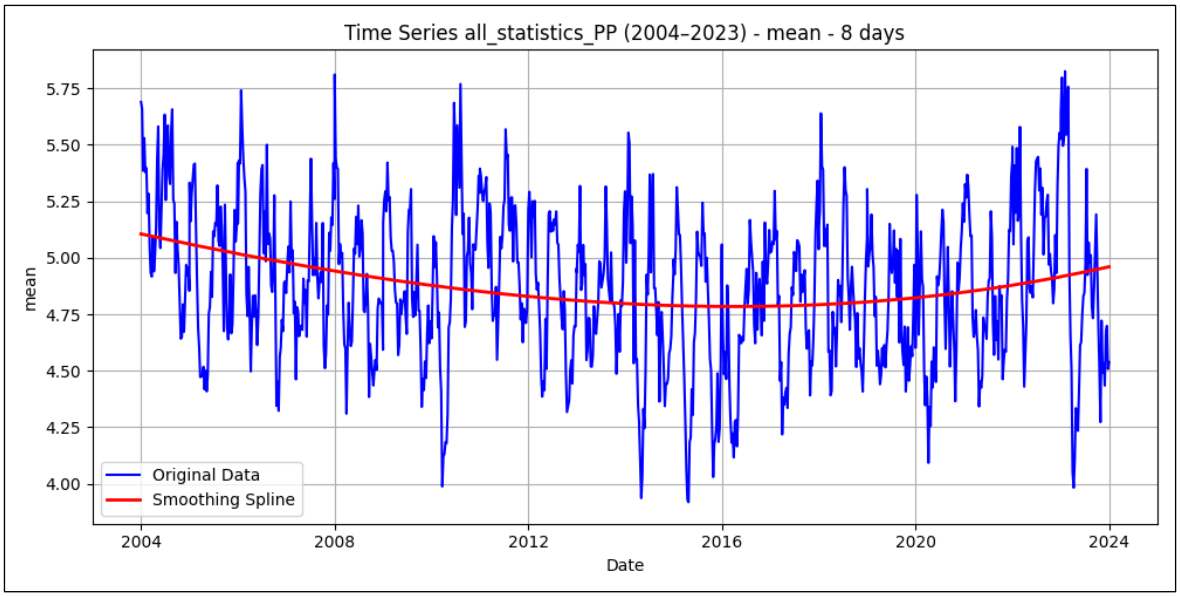


Figure 3. Temporal evolution of PP estimated using a random forest model applied to Chl-a and SST imagery over the period 2004–2023, based on 8-day composites.

Analysis of the long-term trend in PP, conducted using the Kendall test and the Theil–Sen estimator, reveals statistically significant behaviour over the examined period. The Kendall coefficient ($\tau = -0.1029$, $p = 2.98\times10^{-6}$) indicates a monotonic decrease, which is further confirmed by the negative slope estimated using the Theil–Sen method (–0.0002245).

4.5 future development scenario of PP

Most climate-change projections focus on long time horizons (50–100 years), whereas in this work we provide a short-term forecast (5 years). The SARIMAX model for PP was developed using SST and Chl-a as exogenous predictors, projected forward in time to support near-term estimates of PP. A 16-day temporal resolution was adopted, since the 8-day frequency did not yield stable or statistically reliable forecasts (Figure 4a).

Diagnostic checks indicate a good model fit: residuals are centred around zero, show no evident temporal structure, and display an approximately symmetric, only slightly platykurtic distribution (Figure 4a and Figure 4c). The Ljung–Box test confirms the absence of significant residual autocorrelation ($p > 0.05$), suggesting that temporal dependence is adequately captured.

The forecasts indicate that, in the coming years, primary productivity will maintain a seasonal pattern similar to that historically observed. No marked trends or signals of structural changes emerge within the forecast period, suggesting a relative stability of the system in the medium term. Seasonal oscillations remain well defined, and the amplitude of variability falls within the limits of past fluctuations; the 95% confidence intervals remain relatively narrow, supporting the reliability of the forecasts. In the absence of anomalous external forcing, the projections suggest that PP will continue

to oscillate within historical ranges, with no evidence of substantial increases or decreases over the period considered.

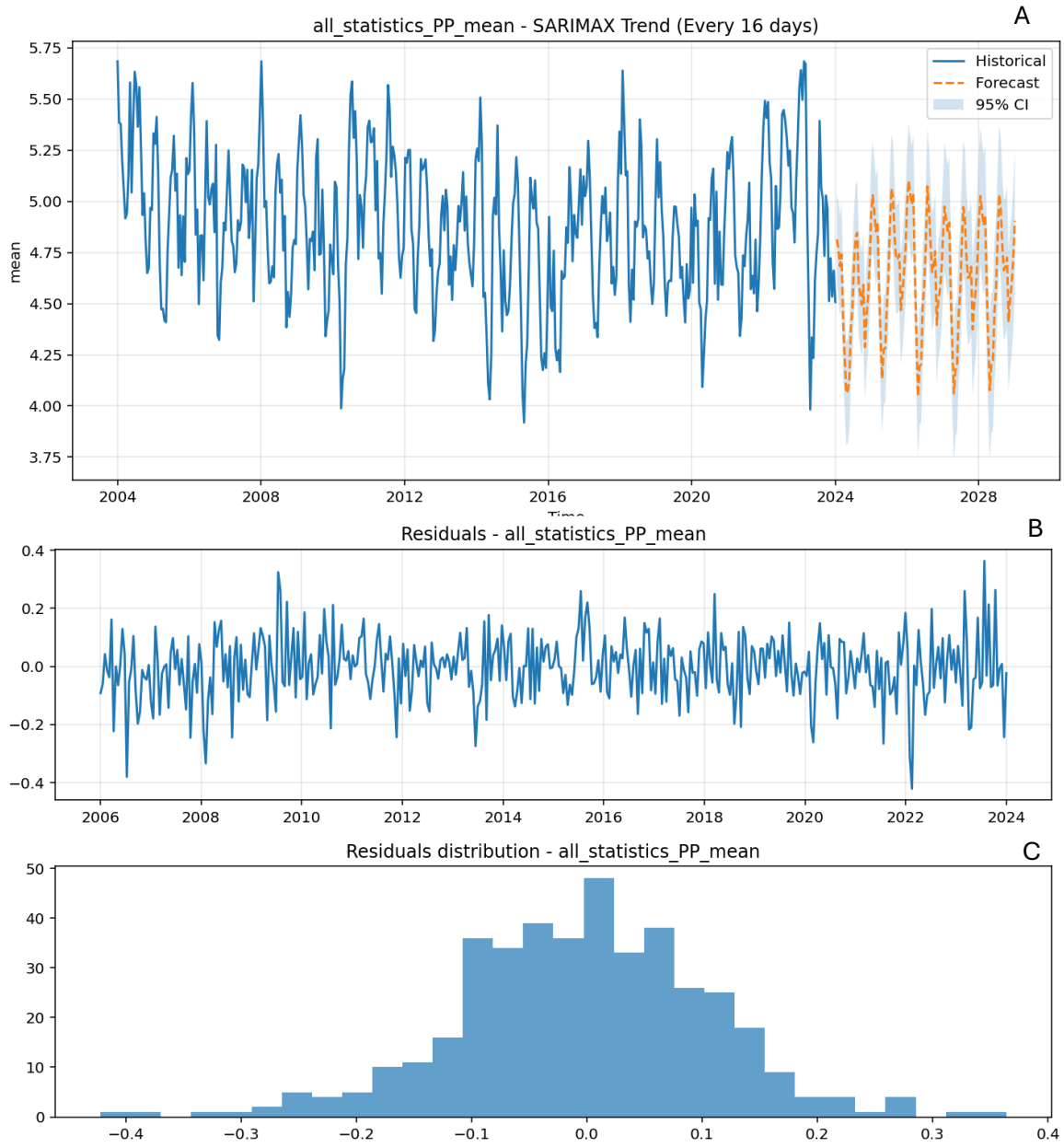


Figure 4. **A)** Time series of mean primary productivity (PP) and 5-year forecast obtained using the SARIMAX model (16-day resolution), with 95% confidence interval. **B)** Model residuals, showing no evident temporal patterns. **C)** Distribution of the residuals, approximately symmetric and centered around zero, confirming the good adequacy of the model.

WF 3 - Influence of behavioral and functional traits on the home range-body mass relationships in consumer species.

1. Workflow 3 Case study: Influence of behavioral and functional traits on the home range-body mass relationships in consumer species.

Body size and home range are fundamental attributes that reflect the energetic demands, locomotor capacity and spatial strategies of consumer species. Allometric and movement ecology theories predict a positive allometric scaling of home range with body mass. Nevertheless, the strength and functional form of this scaling can vary widely among taxa as a result of differences in behavioural, ecological, and physiological traits. For instance, aquatic and terrestrial species encounter distinct spatial and environmental constraints that can shape their movement patterns and space-use strategies. Primary diet may further influence home-range size through contrasts in foraging ecology, dietary energy content, and resource distribution. Locomotion mode also affects movement efficiency and the energetic costs associated with space use. Similarly, the thermoregulatory strategy of organisms (endothermy vs ectothermy) influences metabolic rates and the allocation of energy to movement and resource acquisition. All these variables, may hence potentially influence the home range-body mass relationship.

Using a trait-annotated species database that includes information on habitat, primary diet, thermoregulatory strategy, and locomotion mode, the analytical workflow tests the home range-body mass scaling relationships within each trait class and evaluates between-groups differences in scaling parameters. This approach allows the user to visualize and assess how functional and behavioural traits mediate space-use patterns across vertebrate taxa and to test whether these traits systematically modify fundamental allometric expectations.

This case study, as a part of WF3, focuses on testing various analytical operations on a designed dataset describing the behavioural and functional traits of vertebrate species, with the specific aim of assessing how these traits influence the relationship between home-range and body mass. It should be noted that, within the same workflow, users may perform different or additional analyses involving other variables, depending on the structure and content of their own input files. In addition, users may perform similar operations by importing an individual-level dataset of the same species, including other categorical variables for specific tests. The dataset reported below is therefore provided as an example.

2. Dataset

The provisional dataset includes behavioural and life-history traits for a representative selection of aquatic and terrestrial vertebrate species compiled from the published literature. Estimates of home-range size and body mass were gathered from three primary sources: Tamburello et al. (2015), McCauley et al. (2015), and Udyawer et al. (2023). These datasets were merged into a single harmonised dataset. For species represented in more than one source, home-range values (expressed in m^2) and body-mass values (expressed in g) were averaged to obtain a single estimate per species.

This procedure was implemented to identify and correct misspelled names, outdated nomenclature, and unaccepted synonyms. The resulting dataset comprises 1,164 species-level records, spanning fishes (n = 191, including elasmobranchs), reptiles (n = 136), mammals (n = 642), and birds (n = 195) from across the globe.

For each species, the following key parameters were compiled: body mass (bodyMass) and home-range area (meanHomeRange), both treated as numerical variables. In addition, several categorical traits were included: habitat (terrestrial or aquatic), thermoregulatory strategy (endothermic or ectothermic), primary locomotion mode (walking, swimming, flying, crawling), and primary food type (carnivore or herbivore). These trait data were obtained from published sources and associated datasets.

3. Methods

The analysis was conducted on the entire dataset, after applying a logarithmic transformation (log10(x)) to all numeric variables. The analysis was performed by setting *meanHomeRange* as the dependent variable (Y) and *bodyMass* as the independent variable (X), to assess the presence and strength of linear relationship between the two. Model specifications were provided in a configuration file (Parameter.csv),, which enabled the automatic distinction between regression and analysis of covariance (ANCOVA) analyses. The ANCOVA models incorporated additional categorical predictors (trait variables), enabling the evaluation of how behavioural and functional traits influence the home range-body mass relationship.

A schematic summary of the analytical steps is provided below:

- Regression model:
  - ✓ Dependent variable (Y): *meanHomeRange.*
  - ✓ Covariate (X): *bodyMass (this variable was mean-centered).*
- ANCOVA model:
  - ✓ Dependent variable (Y): *meanHomeRange.*
  - ✓ Covariate (X): *bodyMass (this variable was mean-centered).*
  - ✓ Factors (f): *habitat*, *thermoregulation*, *locomotion*, *foodType.*

This framework enables the assessment of variation in *meanHomeRange* as a function of *bodyMass* , while accounting for differences associated with habitat and other ecological or beahvioural traits.

The data regression and ANCOVA analysis were performed using linear models implemented in Python (the statsmodels package). Prior to model fitting, continuous covariates were mean-centred to facilitate interpretation of the main effects in the presence of interactions. A full ordinary least squares (OLS) model was then estimated, including all main effects of covariates and factors, as well as all covariate × factor interactions in ANCOVA analysis. Sum-to-zero (Sum) contrasts were applied to categorical factors and estimates were computed with robust standard errors (HC3). Factors with only one observed level were automatically excluded. The final model was selected using a backward selection procedure within an information-theoretic framework. Terms were iteratively dropped, while preserving the hierarchy between main effects and interactions, until the AIC improved. During this process, multicollinearity constraints were enforced using the variance

inflation factor (VIF< 3), with terms having VIF values above a predefined threshold being removed. Residual diagnostics were conducted for both the full and final models, together with Shapiro–Wilk tests for normality and Breusch–Pagan tests for heteroskedasticity. Inference on effects was summarised using Type III ANCOVA tables, with effect sizes expressed as partial η^2. For each retained factor, estimated marginal means (EMMs/LS-means) were computed with Holm-adjusted pairwise comparisons. Finally, regression lines for key covariates and ANCOVA lines, stratified by factor levels, were plotted, including 95% confidence intervals.

3.1. Dataset transformation

At this stage, a check is performed for non-values, and if any are found, they are removed. In addition, species for which the original value equal 1 were removed, as this would result in zero after the log transformation. After these steps, the final dataset used for the analyses consisted of 1,162 species (configurable action). These filters can be adjusted by the user according to specific analytical needs.

4. Results

4.1. Linear regression: the relationship between *home range* and *body mass*.

Regression analysis confirmed a positive and highly significant relationship between mean home range size and body mass. The model explains a substantial proportion of the observed variation in mean home range ($R^2 = 0.557$) and indicates that a one-unit increase in centred body mass is associated, on average, with a 1.24-unit increase in home-range size ($\beta = 1.2419$, $p < 0.001$).

The nearly perfect linear relationship between the observed and fitted values further illustrates model fit, suggesting high consistency between the model and the data (Figure 1). However, inspection of the residuals reveals some deviations from the assumptions of a classical linear model. The diagnostic deviations may reflect extreme values that are biologically meaningful (e.g. ecologically atypical species) or an inherently skewed distribution of home-range sizes across taxa within the dataset. This finding is consistent with the Shapiro-Wilk test, which strongly rejects the null hypothesis of normality ($p = 5.3 \times 10^{-16}$). The Breusch-Pagan test also revealed significant heteroskedasticity ($p = 0.036$), suggesting that the variance of the residuals increases with the fitted values. To address this, HC3 heteroskedasticity-robust standard errors were employed, confirming the significance and robustness of the estimated coefficients.

Overall, the model demonstrates a strong fit and a robust effect of body mass on home range size. This indicates that body mass is a strong predictor of home-range size. Nonetheless, additional ecological and behavioural variables may contribute to variation in home-range size across species, accounting for part of the unexplained variation beyond body mass alone.

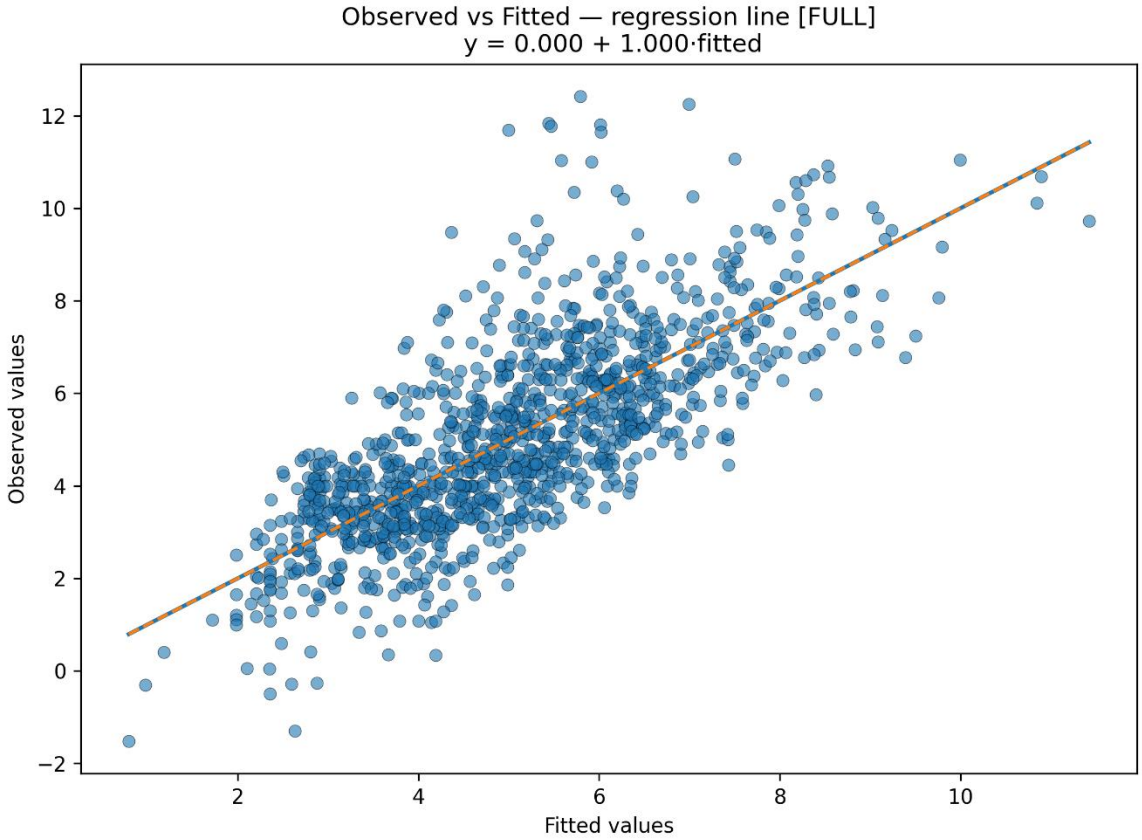


Figure 1. Comparison plot between observed values and values predicted by the full linear regression model. The strong concentration of points around the regression line (overlapping the bisector) indicates good predictive performance of the model.

4.2. ANCOVA – pairwise comparison

Subsequent ANCOVA analysis incorporating additional categorical factors as fixed effects, were conducted to evaluate the extent to which behavioural and ecological traits influence both the intercepts (mean body mass at a given home range) and the slopes (body mass scaling). This approach allows for the separation of general allometric patterns from functional differences among ecological groups. It is important to note that, at this stage, the user can choose to include or exclude one or more fixed factors when running the model, depending on the research questions and the relationships among the variables.

In detail: To examine the relationship between home range and body mass more thoroughly, we considered four categorical factors that could potentially modulate this relationship. A factorial analysis was conducted, incorporating either all factors or a selected subset, depending on the degree of multicollinearity observed among the variables and their interaction terms.

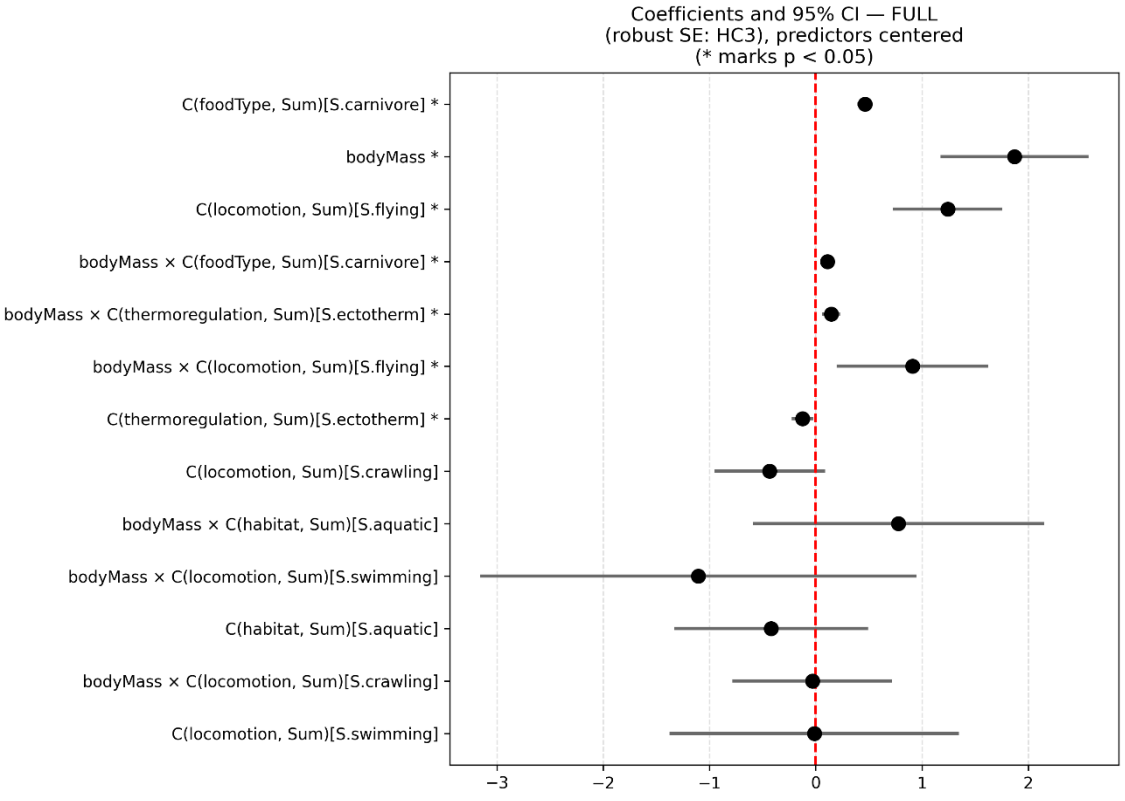
The global model, which included all main effects and their interactions, demonstrated strong explanatory power ($R^2 = 0.754$; adj. $R^2 = 0.751$). Body mass, locomotion type, thermoregulation strategy, and primary diet significantly influence mean home range across species. Flying animals and carnivores also exhibit larger mean home ranges. Some predictors, such as aquatic habitat and

swimming locomotion, were not statistically significant. Overall, the results highlight the combined effect of body mass, ecological traits, and interactions on home range size. However, residual diagnostics indicate moderate skewness and kurtosis. The residuals significantly deviated from normality (Shapiro-Wilk $p = 3.6 \times 10^{-14}$) and strong heteroskedasticity was evident (Breusch-Pagan $p = 3.7 \times 10^{-15}$), and standard errors were corrected for heteroscedasticity (HC3). The very small eigenvalue suggests potential multicollinearity in the model, which should be considered when interpreting coefficients.

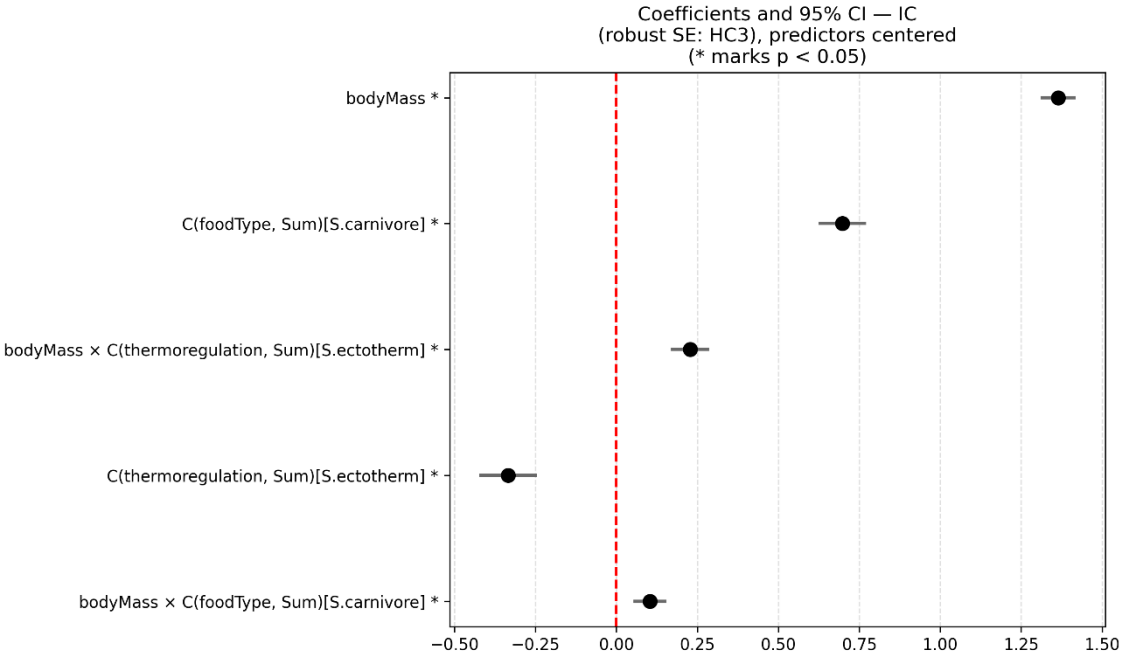
Consequently, the analysis proceeded with a model simplification strategy. Non-significant predictors and interactions were systematically removed, guided by information criteria (AIC) and multicollinearity checks (VIF < 3), to obtain a more parsimonious model that retains explanatory power while satisfying classical linear model assumptions.

The *selected model*, derived through AIC-based model selection and VIF filtering, indicates that body mass, thermoregulation strategy, and primary diet were significant predictors of mean home range. As expected, its explanatory power is slightly lower, approximately 68% ($R^2 = 0.685$; adjusted $R^2 = 0.683$), but the model is more parsimonious and its coefficients are more stable. Ectothermic species tend to have smaller home ranges, while carnivores have larger ones, with body mass amplifying these effects. Diagnostic tests still indicate deviations from residual normality (Shapiro-Wilk $p = 7.7 \times 10^{-18}$) and persistent heteroskedasticity (Breusch–Pagan $p = 1.9 \times 10^{-6}$), although these issues are less pronounced compared to the global model.

Panels in Figure 2, compare effects from the global and selected model (Figure 2). The selected model represents a more balanced compromise between complexity, coefficient significance, and statistical parsimony. Although it does not fully resolve the issues related to the error structure, it reduces the over-specification of the full model and is preferable in terms of interpretability and inferential stability. Overall, this finding supports the notion that body mass is the primary determinant of home-range size, while its effect on space use may be modulated by other key biological traits.



A



B

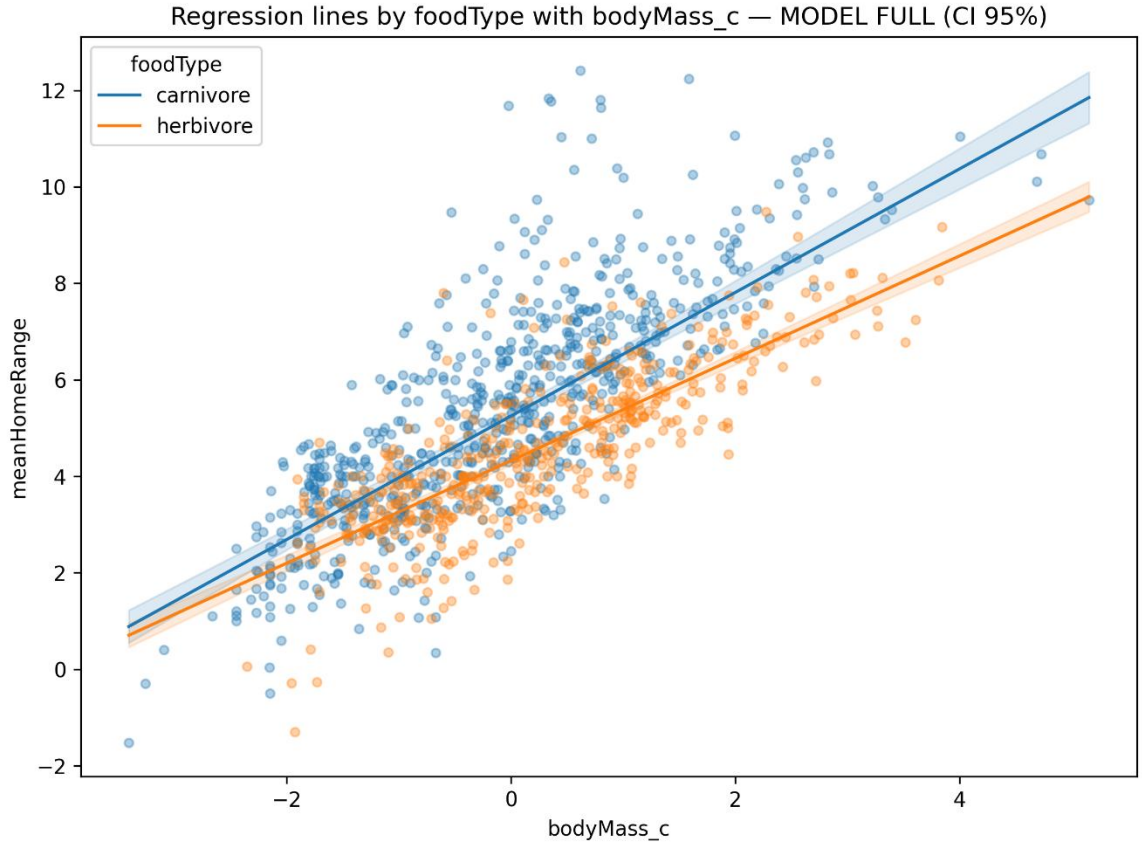
Figure 2: The two panels show the estimated coefficients (with 95% confidence intervals and robust standard errors, HC3) for the global model (Panel A) and the model selected using information criteria with multicollinearity control (Panel B).

In summary, the ANCOVA model for the selected specification (after applying the VIF criterion) identified body mass as the strongest predictor of home-range size, with a positive and highly significant effect across all taxonomic and ecological groups considered. Two categorical factors, thermoregulatory strategy and primary food type, showed significant main effects as well as significant interactions with body mass. This indicates that the relationship between body mass and space use is not uniform across groups.

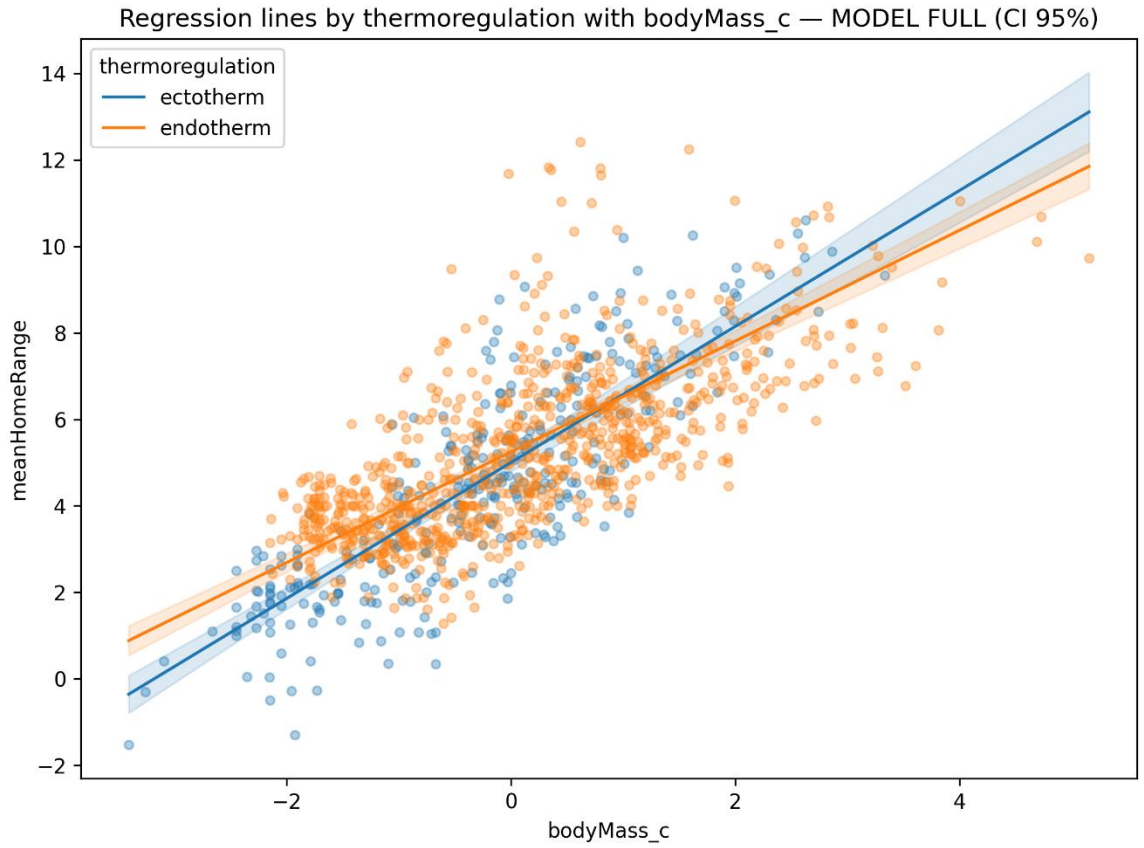
A similar pattern emerges for the food type predictor (Figure 3a): carnivores consistently show larger home ranges than herbivores, along with a steeper scaling with body mass. This interaction suggests that trophic ecology modulates the allometric relationship between body mass and space use.

As shown in Figure 3b ectotherms exhibit a steeper relationship between home-range size and body mass compared to endotherms, suggesting that increases in body mass translate into proportionally greater spatial requirements in ectothermic species. Post-hoc analyses on the estimated marginal means (Holm correction) confirm significant differences between the two groups, particularly at intermediate and high body-mass values (Figure 4).

Overall, the ANCOVA results suggest that body mass is the principal determinant influencing home-range size. However, the exact relationship between the two depends strongly on the species' thermal and trophic ecology.

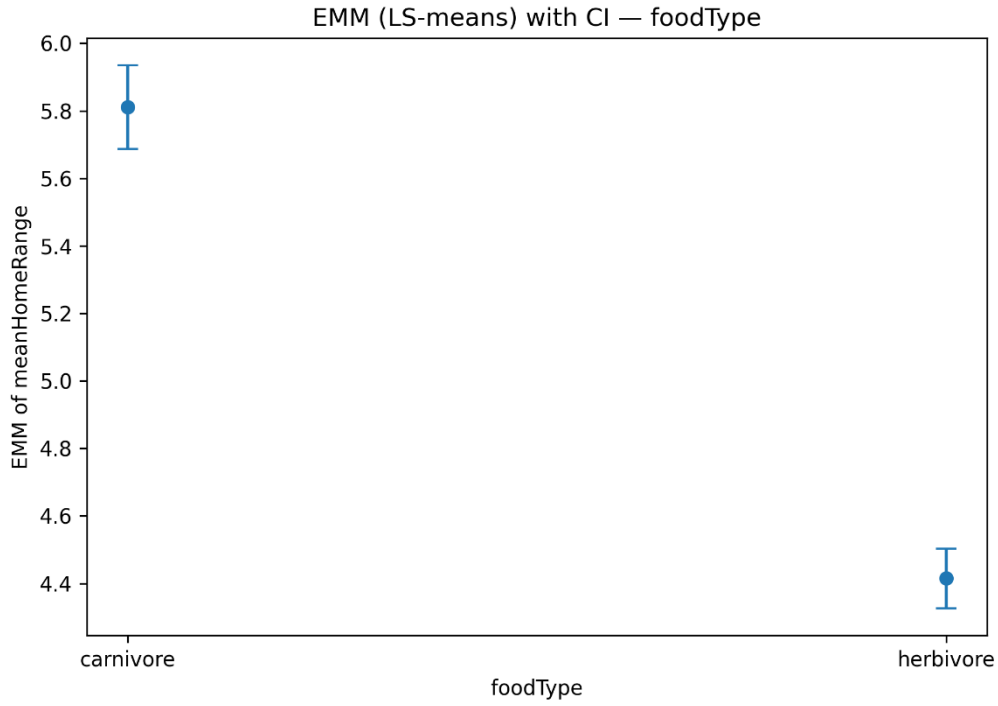


A

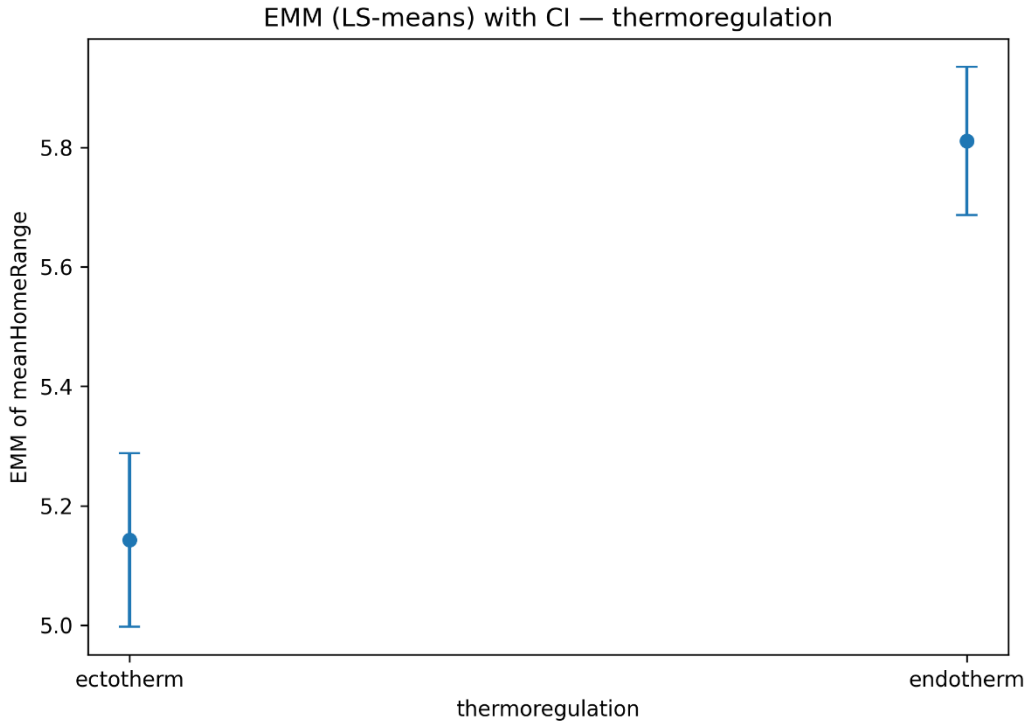


B

Figure 3. The plots show the relationships between body mass and home-range size in the selected models, highlighting how these relationships vary according to food type (Figure 3A) and thermoregulatory system (Figure 3B). In both cases, home-range size increases with body mass, but systematic differences emerge between groups.



A



B

Figure 4: Estimated marginal means (EMMs) for categorical predictors: A) foodType: carnivores vs herbivores; B) thermoregulation: ectotherms vs endotherms.

WF 4 - Application of machine learning models for ecological analysis.

Introduction

Workflow 4 (WF4) aims to evaluate a range of machine learning algorithms that can be adapted to different ecological problems. The goal is to identify the most suitable model for analysing the influence of predictor variables on a target variable. Designed with a pragmatic approach, the workflow is applicable across a wide spectrum of climate-change-related scientific domains, particularly regression analyses. Currently, WF4 supports regression modelling on any dataset of biotic and abiotic variables (characterised by a numerical variables), identifying key relationships based on user goals, available data and parameter settings.

The proposed case study showcases WF4's core capabilities. Using the selected dataset, it examines on the relationship between Primary Productivity (PP) and biotic variables such as chlorophyll concentration (Chl), as well as abiotic variables including sampling month (month), latitude (decimalLatitude), longitude (decimalLongitude), distance from the coast (distanceFromCoast, irradiance (Irr), day length (dayLength), and sea surface temperature (SST).

Dataset

The dataset used for this case study was derived from field data available on the Ocean Productivity website. We selected measurements taken at the sea surface (depth = 0 m). We created the dataset by adding the distance of each monitoring site from the coast, calculated using the geographic coordinates provided in the original data. Further details on the original data types can be found at: https://orca.science.oregonstate.edu/field.data.c14.online.php.

Methods

WF4 simultaneously evaluates five machine-learning regression algorithms, as eXtreme Gradient Boosting (XGBoost), Multiple Linear Regression, Neural Networks, Random Forests and Support Vector Machines, to identify the algorithm that best captures the relationships under study. In this case study, the goal was to estimate how PP (settled in this study as target variable in the "Parameter data" file) is influenced by the available biotic and abiotic variables (the predictors present in the input table "Training data" file).

In this case study, model selection is based on the Mean Absolute Error (MAE) by default. However, this can be configured in the 'Parameter data' file, and the best model can be chosen using the Root Mean Squared Error (RMSE) or the coefficient of determination (R^2).

Models are trained using cross-validation and then tested on a dataset that has been held back. Before training begins, WF4 can optionally apply data transformation pipelines. In this study, we train the model using data pre-processed through normalization and principal component analysis (PCA). The original dataset is automatically split into training and test

sets according to the specifications provided by the user in the 'Parameter data' file. The trained algorithm is then applied to the test set to evaluate its ability to generalise to new data.

Hyperparameters are defined as ranges or candidate values in the “Parameter data” file. WF4 explores these and selects the combination that yields the best performance for the chosen evaluation metric.

Finally, WF4 runs a SHAP (SHapley Additive exPlanations) analysis on the best-performing model to identify which variables most strongly influence model behavior.

Results

The Table 1 report R^2, MAE and RMSE for the training and testing phases for the algorithms, using both raw and transformed data. In this case study, normalization and PCA were applied to the data (Full results are provided in the 'model_parameters_description' file in the folder of each model). Using the chosen approach, eXtreme Gradient Boosting (XGBoost) achieves the lowest MAE in the training phase when trained on raw data (Table 1A). In the testing phase, however, the Support Vector Machine (SVM) is the best-performing algorithm when trained on raw data (Table1B).

Table 1. Tables reporting the performance results of the applied models. A) eXtreme Gradient Boosting; B) Multi Linear Regression; C) Neural Network; D) Random Forest and E) Support Vector Machine.

A

eXtreme Gradient Boosting						
	Training			Test		
Data transformation type	R-Square	MAE	RMSE	R-Square	MAE	RMSE
Raw data	0.9683	8.0390	16.4432	0.5818	17.3640	28.4081
Normalization	0.9690	8.0735	16.4619	0.5727	17.4307	28.9317
PCA	0.9655	8.5887	17.2835	0.5122	16.9785	30.2742

B

Multi Linear Regression						
	Training			Test		
Data transformation type	R-Square	MAE	RMSE	R-Square	MAE	RMSE
Raw data	0.4548	25.2567	42.5687	0.3978	21.2411	32.5453
Normalization	0.4548	25.2567	42.5687	0.3978	21.2411	32.5453
PCA	0.4537	25.2972	42.6103	0.3960	21.2445	32.5310

C

Neural Network						
	Training			Test		
Data transformation type	R-Square	MAE	RMSE	R-Square	MAE	RMSE
Raw data	0.3534	27.6501	46.3942	0.3444	23.5397	33.6402
Normalization	0.6484	21.9393	34.2835	0.5005	20.3747	30.7061
PCA	0.5746	22.9001	37.6052	0.4873	20.7923	32.0389

D

Random Forest						
	Training			Test		
Data transformation type	R-Square	MAE	RMSE	R-Square	MAE	RMSE
Raw data	0.9243	10.1476	18.2846	0.6046	18.6820	28.1342
Normalization	0.9244	10.1397	18.2786	0.6042	18.7120	28.1504
PCA	0.9221	10.9779	19.3436	0.5895	16.8323	27.0728

E

Support Vector Machine						
	Training			Test		
Data transformation type	R-Square	MAE	RMSE	R-Square	MAE	RMSE
Raw data	0.6481	17.9481	35.7210	0.7049	14.0892	22.6519
Normalization	0.6481	17.9481	35.7211	0.7049	14.0892	22.6519
PCA	0.6418	18.5455	36.1163	0.6721	14.6350	24.4148

A comparative analysis of the eXtreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) models reveals notable differences in their generalisation capabilities. While XGBoost performs excellently on the training set ($R^2 = 0.9683$), it shows a marked decline on the test set ($R^2 = 0.5818$), accompanied by substantial increases in MAE and RMSE. This behaviour indicates overfitting, whereby the model learns the characteristics of the training set too specifically, thereby compromising its ability to predict unseen data.

In contrast, the SVM model shows more balanced values in the training and testing phases, with R^2 increasing from 0.6481 to 0.7049, and errors reducing on the validation set. This consistency across the two phases suggests better generalisation and lower sensitivity to noise in the data. While XGBoost remains potentially more powerful, the current results show that SVM provides more reliable performance on the test set under the considered experimental conditions.

The plots below (Figure 1) show the relationship between the observed and predicted values for the best-performing algorithms in each training and testing phase for Extreme Gradient Boosting (Figure1a) and Support Vector Machine (Figure 1b).

A

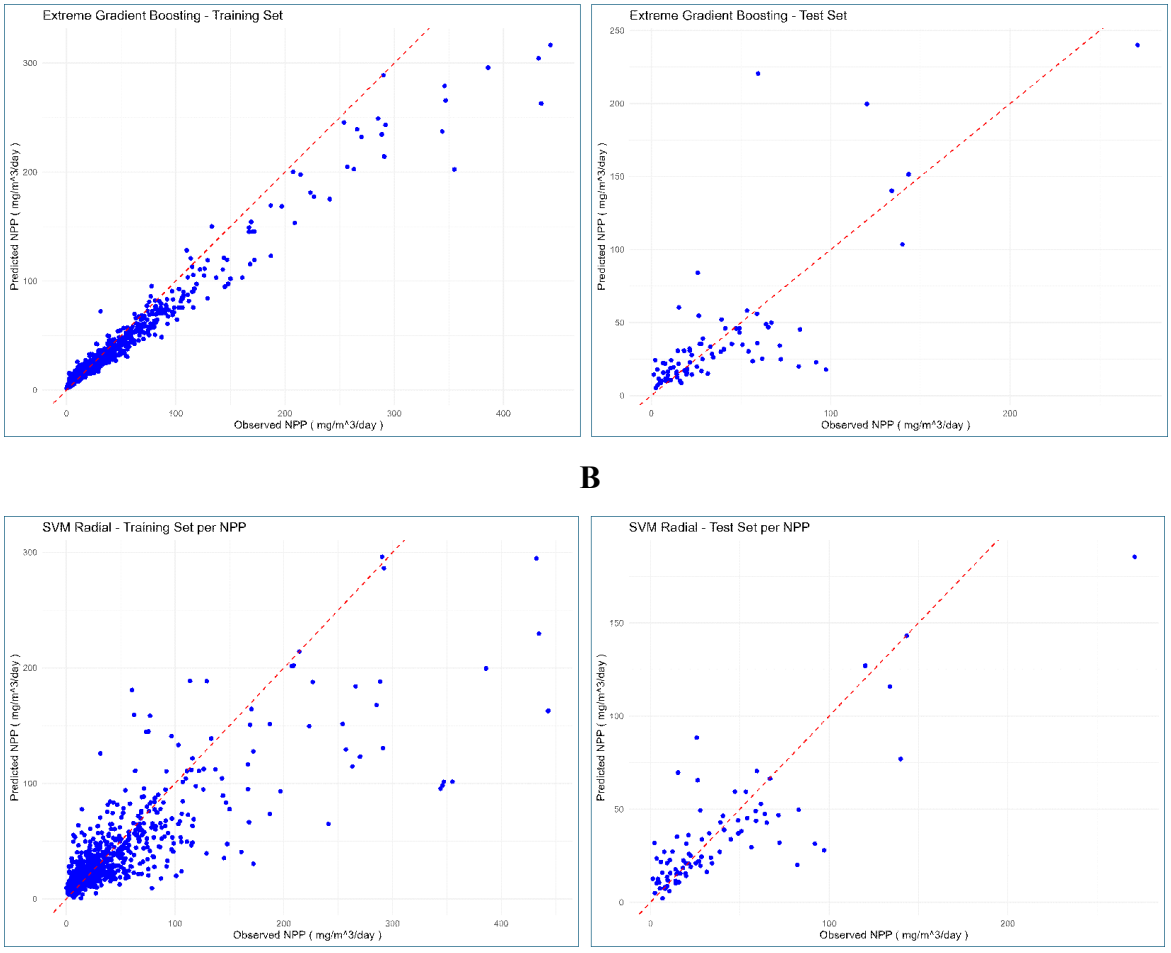


Figure 1. Scatter plots comparing observed and predicted values from the models. The blue points represent the model predictions relative to the actual data, while the red dashed line indicates the ideal 1:1 correspondence. A) Scatter plots of eXtreme Gradient Boosting; B) Scatter Plots of Support Vector Machine

This study aims to develop a machine-learning algorithm capable of making predictions based on new biotic and abiotic data, in order to evaluate the contribution of each variable to the target outcome (PP). Therefore, when choosing the best algorithm, we prioritise performance during the testing phase. Accordingly, we selected the Support Vector Machine (SVM) model, which, although it shows lower performance during training, generalises better to unseen data, that is, observations not used during training.

The SHAP analysis applied to the Support Vector Machine (SVM) model indicates that the variables with the greatest influence on the PP predictions are Chl (chlorophyll), Irr (irradiance) and SST (sea surface temperature). The SHAP bar plot for the SVM, which supports this finding, is provided among the outputs under the name 'shap_importance_bar_plot'.

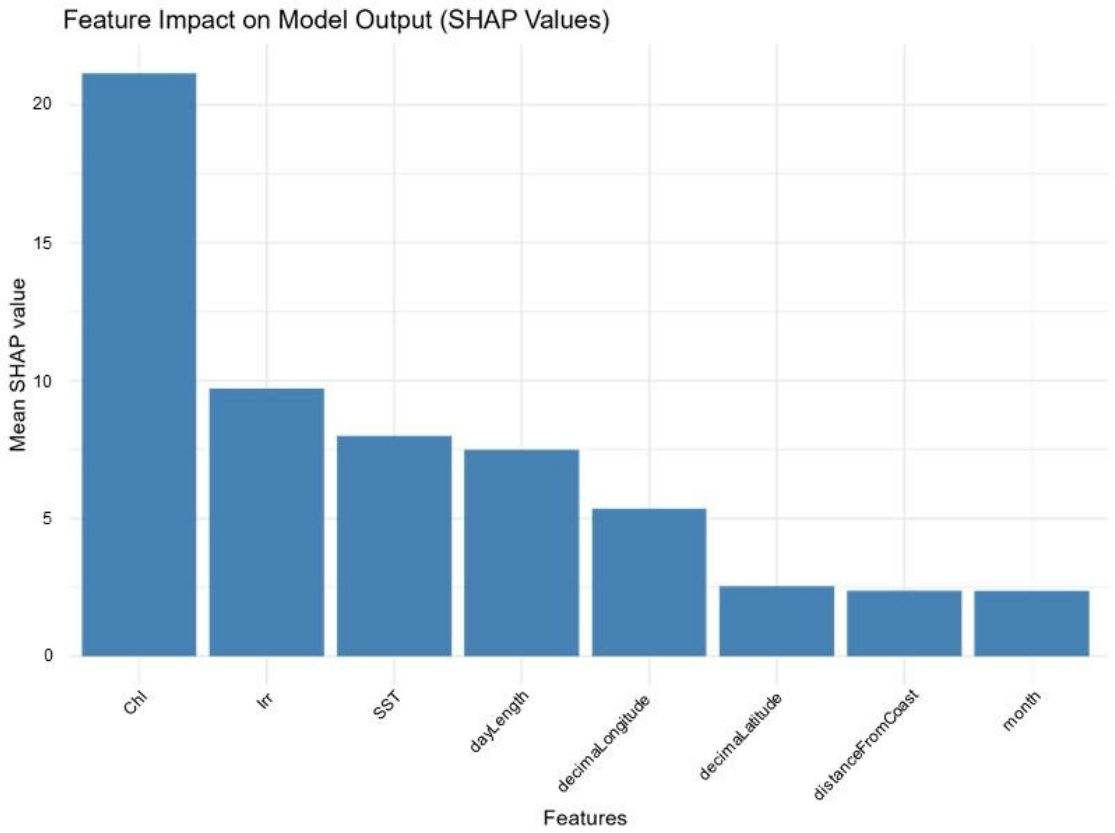


Figure 2. Relative importance of the variables used in the model. The bar chart shows the estimated contribution of each predictor to the model's performance.