

Trabajo de Fin de Grado

Desarrollo de un Gemelo Digital para estimar la Producción Primaria en el océano

TITULACIÓN: Grado en Ingeniería Informática

AUTOR: Imanol Benito Quintero Bermúdez

TUTORIZADO POR:
Dr. José Évora Gómez
Dr. Yeray Santana Falcón

Las Palmas de Gran Canaria, Julio 2025

SOLICITUD DE DEFENSA DE TRABAJO DE FIN DE TÍTULO

D./D^a Imanol Benito Quintero Bermúdez, autor del trabajo Desarrollo de un Gemelo Digital para estimar la Producción Primaria en el océano, correspondiente a la titulación Grado en Ingeniería Informática

SOLICITA

que se inicie el procedimiento de defensa del mismo, para lo que se adjunta la documentación requerida, haciendo constar que

☒ se autoriza / ☐ no se autoriza la grabación en audio de la exposición y turno de preguntas.

Asimismo, con respecto al registro de la propiedad intelectual/industrial del TFT, declara que:

☐ Se ha iniciado o hay intención de iniciarlo (defensa no pública).

☒ No está previsto.

Y para que así conste firma la presente. (fecha en firma electrónica)

El/La estudiante

Fdo. _____

A rellenar y firmar **obligatoriamente** por el/la/los/las tutores

En relación con la presente solicitud, se informa: (firmar donde corresponda)

Positivamente (en caso de detección de copia, esta firma quedará invalidada)

Fdo. _____

Negativamente (justificación en TFT05)

Fdo. _____

DIRECTORA DE LA ESCUELA DE INGENIERÍA INFORMÁTICA

Agradecimientos

Rehúso a comenzar el epílogo de estos últimos cuatro años de nuestras vidas sin recordar a todos los acompañantes de esta hermosa travesía. Gracias por todo. Por remar por mí y por forjar la persona que a día de hoy soy.

A mis compañeros de carrera. Cuesta creer que después de tantos años de cafés, debates y confesiones, nuestros caminos se bifurquen. No dejen de llevar consigo la ilusión y la humildad de ese septiembre de 2021 en el que nos conocimos.

A mis amigos. A Miguel, Raúl, Laura y Alejandro, por guiarme en este difícil camino. A Marcos y a Álvaro, por ser mis hermanos. A mi grupo de Telde. Sin ustedes no hubiera sido lo mismo.

A José Évora Gómez y a Yeray Santana Falcón, por ser una inspiración para mí, y por el pleno apoyo prestado. Representan lo que quiero ser algún día.

A Ainhoa, por ser calma en la tormenta. Por tu amor, y por mirarme con otros ojos. Esto acaba de empezar.

A abuela y abuelo, porque no dejo de pensar en ustedes y en sus lecciones. En los momentos, en las anécdotas, y en la unión familiar por la que lucharon. Espero que, estén donde estén, sientan orgullo de mí.

A Carla, por ser mi mano derecha. Por regresar a Canarias, de la mano de Léo y del pequeño que viene en camino. Nos esperan grandes tardes en casa.

A ti, mamá, porque solo tú sabes lo que hemos pasado. Dejaste todo por nosotros, y remaste contra marea. Tú no perdiste la esperanza en mí, en mis logros, y yo no quise decepcionar. Esta batalla es tuya.

Resumen

El presente Trabajo de Fin de Grado presenta el desarrollo de un gemelo digital basado en redes neuronales artificiales para estimar la producción primaria neta (NPP) en el océano. Esta variable es clave en el ciclo del carbono y en la salud de los ecosistemas marinos. Se recopilieron datos satelitales y oceanográficos de múltiples fuentes, aplicando modelos como Eppley-VGPM y Lee-AbPM para la obtención de la variable objetivo en la región macaronésia, como conjuntos de entrenamiento y verificación. La metodología sigue el marco CRISP-DM, implementando modelos predictivos validados con otros conjuntos de datos reconocidos por la comunidad científica. Este proyecto contribuye a una mejor comprensión del cambio climático y de los procesos tróficos fundamentales del océano.

Abstract

This Final Degree Project presents the development of a digital twin based on artificial neural networks to estimate net primary production (NPP) in the ocean—a key variable in the carbon cycle and in the health of marine ecosystems—. Satellite and oceanographic data from multiple sources were collected, applying models such as Eppley-VGPM and Lee-AbPM to obtain the target variable in the Macaronesian region. These data were then used as training and validation datasets. The methodology follows the CRISP-DM framework, and implements predictive models validated with other data sets recognized by the scientific community. This project contributes to a better understanding of climate change and fundamental trophic processes of the ocean.

Índice general

1. Introducción	1
1.1. El cambio climático antropogénico y la huella humana en los océanos desde la industrialización	1
1.2. El océano como regulador climático y sumidero de carbono	3
1.3. El concepto de producción primaria y el rol del fitoplancton en la red trófica	5
1.4. Modelos actuales de estimación y limitaciones	6
1.5. Motivación	7
1.6. Organización del documento	7
2. Estado actual y objetivos iniciales	8
2.1. Motivación y antecedentes	8
2.1.1. El concepto de producción primaria neta	8
2.1.2. Área de estudio y la dinámica de la Corriente de Canarias	9
2.1.3. Evaluación de las mediciones y modelos actuales para la estimación de la producción primaria neta	9
2.1.4. Modelos empleados en el proyecto	14
2.1.5. Gemelos digitales y estado actual de los modelos estimativos para la producción primaria	16
2.2. Objetivos	20
3. Aportaciones del trabajo	24
3.1. Principales aportaciones	24
3.2. Competencias específicas	26
3.3. Alineación con los objetivos de desarrollo sostenible	27
4. Desarrollo	29
4.1. Metodología	29
4.1.1. Marco de trabajo	29
4.1.2. Herramienta de gestión de procesos	30
4.1.3. Recursos y tecnologías empleadas	31
4.1.4. Bibliotecas requeridas	36
4.1.5. Edición de la memoria final	37
4.2. Fases de desarrollo	37
4.2.1. Metodología, etapas y flujo de trabajo	37

4.2.2. Implementación y desarrollo técnico por etapas	39
5. Resultados	52
5.1. Análisis del modelo explicativo del entorno oceánico	52
5.1.1. La biodiversidad oceánica macaronesia en las últimas dos décadas . .	53
5.1.2. Factor de inflación de la varianza	61
5.1.3. Regresión lineal múltiple	62
5.1.4. Interpretación de coeficientes del modelo explicativo	68
5.2. Resultados del gemelo digital para estimación de NPP	71
5.2.1. Reestructuración previa del conjunto de datos	72
5.2.2. Definición del modelo	73
5.2.3. Resultados del modelo	73
6. Conclusiones y trabajo futuro	81
6.1. Conclusiones	81
6.2. Trabajo futuro	82

Índice de Algoritmos

1.	Modelo de gemelo digital para la estimación de NPP en Picota	74
----	--	----

Índice de figuras

1.1. Evolución de anomalías térmicas globales (1880-2024)	2
1.2. Aceleración del aumento del nivel del mar (1993-presente)	3
2.1. Esquema de la circulación superficial de la cuenca canaria en el Atlántico nororiental subtropical	10
2.2. Satélite Vanguard 2 sobre la Tierra	12
2.3. Evaluación de sesgos en modelos de NPP	18
2.4. Concepto de gemelo digital según el modelo de Grieves	22
4.1. Proceso CRISP-DM	31
4.2. Tablero Trello para gestión del proyecto	32
4.3. Interfaz de Panoply	33
4.4. Mapa de concentración de clorofila-a	34
4.5. Entorno de Google Colab para el análisis de NPP	35
4.6. Almacenamiento de datos en Google Cloud Storage	35
4.7. Automatización de descargas con Selenium WebDriver	36
4.8. Proceso CRISP-DM adaptado	38
4.9. Matriz de correlación de variables ambientales	43
4.10. Análisis de cobertura de variables oceanográficas	44
4.11. Problema de valores NaN en interpolación lineal	45
4.12. Solución de interpolación por vecino más cercano	46
4.13. Interpolación con funciones de base radial (RBF)	48
4.14. Análisis de cobertura de variables oceanográficas	49
4.15. Análisis de cobertura de variables oceanográficas	50
5.1. Matriz de correlación de variables ambientales	54
5.2. Anomalías de clorofila-a (2003-2007 vs. 2019-2023)	55
5.3. Evolución temporal de clorofila-a (2003-2023)	56
5.4. Composición y evolución del fitoplancton (2002-2022)	57
5.5. Anomalías de temperatura superficial del mar (2003-2007 vs. 2019-2023)	58
5.6. Anomalías de productividad primaria neta (2003-2007 vs. 2019-2023)	60
5.7. Análisis de multicolinealidad mediante VIF	61
5.8. Matriz de correlación de Spearman entre variables clave	63
5.9. VIF por tamaño celular	64
5.10. Matriz de correlación de Spearman entre variables ambientales y biológicas	65

5.11. Diagnóstico de residuos del modelo de productividad primaria	66
5.12. Diagnóstico de residuos tras ajuste del modelo	68
5.13. Coeficientes del modelo explicativo de NPP	70
5.14. Evaluación de horizonte temporal en modelo de estimación	75
5.15. Ciclo anual estandarizado de variables clave	79

Índice de tablas

2.1.	Incertidumbres en cada variable de entrada por región y programa de estudio, basadas en diferencias entre datos satelitales, modelados y <i>in situ</i> . Los modelos de color del océano utilizaron 81 perturbaciones de los datos de entrada para cada medición de NPP, según estas incertidumbres específicas por región. Fuente: Saba et al., 2011 [60].	15
2.2.	Medidas estadísticas de las comparaciones entre la NPP <i>in situ</i> y la modelada (NPP _{AbPM} , NPP _{VGPM} y NPP _{CbPM}) utilizando datos OC-CCI. Leyenda de variables: N = número de observaciones; R^2 = coeficiente de determinación; σ_d = desviación estándar de las diferencias ($\text{mg C m}^{-2} \text{ día}^{-1}$); $RMSD$ = raíz del error cuadrático medio; $Bias$ = sesgo medio; $uRMSD$ = RMSD no sesgado; B^* , $uRMSD^*$ = métricas normalizadas; Similitud coseno = medida de similitud angular (1 = máxima similitud). Fuente: Wu et al., 2024 [76]. .	17
2.3.	RMSEP y r (coeficiente de correlación entre valores observados y predichos de producción primaria) para las redes neuronales que implementan retardos de 1, 2 y 3 meses entre producción primaria y las variables de entrada, para evaluar la predicción de cambios futuros en producción primaria. Fuente: Belgrano et al., 2001 [7].	21
3.1.	Grado de relación del TFT con los objetivos de desarrollo sostenible.	27
5.1.	Métricas de diagnóstico del modelo	67
5.2.	Coeficientes del modelo de regresión lineal múltiple para la productividad primaria neta transformada (NPP)	69
5.3.	Desempeño comparativo entre modelos clásicos de NPP y el gemelo digital propuesto	78
5.4.	Principales variables de entrada que contribuyen a la estimación de <code>bio_npp</code>	80

Capítulo 1

Introducción

La capacidad de aprender y cambiar es lo que nos hace humanos.

Alan Turing

1.1. El cambio climático antropogénico y la huella humana en los océanos desde la industrialización

Históricamente, el cambio climático ha sido un fenómeno observado en la propia naturaleza que afecta a los sistemas físicos y biológicos de todo el planeta. No obstante, a finales de la década de 1970, la comunidad científica sugiere que las perturbaciones provocadas por el ser humano podrían provocar un cambio abrupto en el sistema climático (Weart, 2010 [73]).

Desde entonces, múltiples líneas de evidencia han demostrado que los actuales efectos del cambio climático tienen origen antropogénico, es decir, causado por las actividades humanas. El aumento en la quema de combustibles fósiles como carbón, petróleo y gas, la deforestación masiva y la expansión de la agricultura intensiva han incrementado de forma exponencial la concentración de gases de efecto invernadero en la atmósfera, en especial el dióxido de carbono (CO_2), el metano (CH_4) y el óxido nitroso (N_2O). Cuando se habla de efecto invernadero, se hace referencia al fenómeno natural por el cual ciertos gases presentes en la atmósfera retienen parte del calor emitido por la superficie terrestre, evitando que se disipe completamente al espacio, y permitiendo que la Tierra mantenga una temperatura compatible con la vida. No obstante, el aumento desmesurado de estos gases debido a la actividad humana ha intensificado este efecto, provocando un calentamiento global acelerado.

La concentración atmosférica del CO_2 desde 1750, a partir de procesos relacionados con la Revolución Industrial, se ha incrementado un 31,0 % —una tasa de incremento sin precedentes en 20 mil años y la más alta en los últimos 420 mil años (Porrúa, 2001 [53])—. En el caso del

CH_4 , la concentración atmosférica ha crecido un 151,0 % en el mismo lapso, mientras que la del N_2O se ha incrementado un 17 %.

Como consecuencia de estos procesos, la temperatura media global de la superficie terrestre ha aumentado significativamente desde la era preindustrial. En 2024, se registró un incremento de aproximadamente $1,55^\circ C$ por encima del promedio de 1850–1900, convirtiéndose en el año más cálido desde que se tienen registros (ilustración 1.1). Los diez años más cálidos registrados han ocurrido desde 2015, reflejando una tendencia clara de calentamiento global (NASA, 2024 [48]). Este aumento de temperatura no tiene precedentes en los últimos mil años y es atribuible principalmente a las actividades humanas, como la quema de combustibles fósiles y la deforestación (Rosenzweig et al., 2013 [57]).

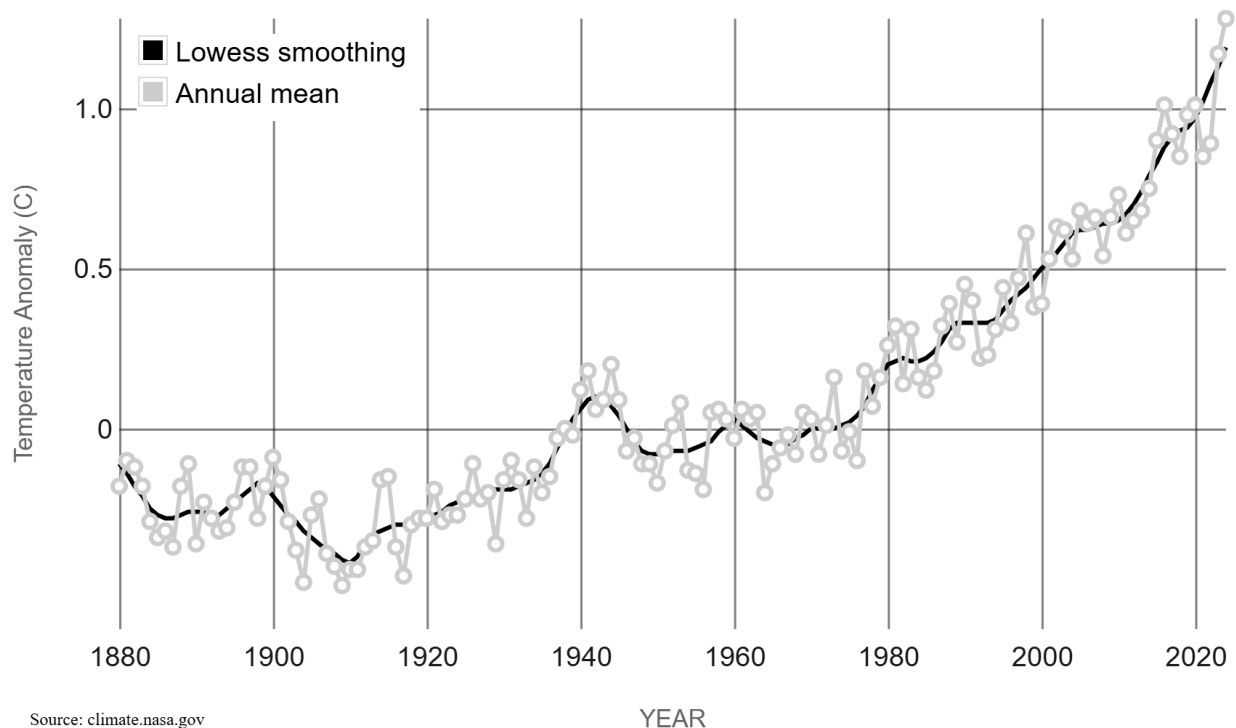


Ilustración 1.1: Suavizado *Lowess* de las anomalías de temperatura global anual (en $^\circ C$) desde 1880 hasta 2024. Los datos muestran el incremento acelerado post-1960, alcanzando anomalías superiores a $+0.5^\circ C$ en el siglo XXI. **Fuente:** NASA/GISS, 2024. **Datos:** NASA's Goddard Institute for Space Studies (GISS).

Los informes del *Intergovernmental Panel on Climate Change* (IPCC) han evidenciado el aumento del nivel del mar, la pérdida de masa de los glaciares, la reducción de la cobertura de nieve y hielo marino en el Ártico, y la alteración de patrones meteorológicos a escala global. Los eventos extremos, como las olas de calor, tormentas, sequías e inundaciones, han ganado en intensidad y frecuencia (Rosenzweig et al., 2013 [57]). Por ejemplo, se ha podido comprender gracias a distintas mediciones satelitales cómo el nivel global del mar ha aumentado más de 101 milímetros desde que comenzaron las mediciones en 1992 (ilustración 1.2), lo que ha incrementado las inundaciones costeras en algunos lugares. A medida que el agua del océano

se calienta, se expande y ocupa más espacio. El calor adicional en el aire y el océano también está derritiendo las capas de hielo y los glaciares, lo que añade agua dulce al océano y eleva aún más el nivel del mar.



Ilustración 1.2: Tendencia del nivel del mar global (en *mm*) desde 1993 hasta la actualidad, mostrando una aceleración en la última década. Los datos indican un aumento de 100 *mm* desde 1993, con una tasa media de 3.4 *mm/año* (2005-2020). **Fuente:** NASA's Goddard Space Flight Center. **Datos:** observaciones satelitales del nivel del mar.

1.2. El océano como regulador climático y sumidero de carbono

Los mares y océanos, que representan el 71 % de la superficie de la Tierra, 360 millones de km^2 y el 97 % de los recursos hídricos terrestres, desempeñan un papel fundamental en la regulación del clima global mediante la absorción de CO_2 y el transporte de calor. Las variaciones en el clima y el entorno, definidas por la interacción entre la atmósfera y el océano, no pueden explicarse sin los procesos acaecidos en el medio marino. Actúan como un amortiguador climático al absorber alrededor del 90 % del exceso de calor generado por las emisiones de gases de efecto invernadero.

Además, capturan aproximadamente el 25 % del CO_2 emitido por actividades humanas, lo que contribuye a mitigar el calentamiento atmosférico. Este proceso de absorción de CO_2

se realiza mediante dos mecanismos principales: la bomba física, donde el CO_2 se disuelve en aguas frías y se transporta a las profundidades oceánicas, y la bomba biológica, en la que el fitoplancton utiliza el CO_2 para la fotosíntesis, contribuyendo al secuestro de carbono .

En concreto, los océanos capturan aproximadamente el 90 % del exceso de calor generado por las emisiones de CO_2 a la atmósfera, actuando como un amortiguador térmico del sistema climático terrestre. Asimismo, capturan gran parte del CO_2 de origen antropogénico mediante procesos de eliminación y retención de este compuesto, disolviéndolo en el agua para formar ácido carbónico y permitiendo almacenar mucho más dióxido que otros gases disueltos, secuestrándolo durante décadas o más. La retención de CO_2 desempeña un papel crucial en el ciclo global del carbono al almacenar hasta 50 veces más carbono inorgánico que la atmósfera preindustrial, y ya absorbe aproximadamente el 25 % de las emisiones antropogénicas de CO_2 .

Si bien esta sucesión de acontecimientos favorece a mitigar parcialmente la acumulación de CO_2 en la atmósfera, dos siglos absorbiendo el 26 % de las emisiones antropogénicas (equivalente a 40 mil millones de toneladas de CO_2 anuales) han llevado consigo un proceso de **acidificación oceánica y desoxigenación**, lo que supone un cambio radical en la química acuática con diversos efectos sobre la vida marina. La superficie del océano absorbe de forma natural el CO_2 atmosférico para lograr el equilibrio entre el aire y el mar, y la fotosíntesis del fitoplancton o las macroalgas captura más CO_2 disuelto de la capa superior del océano en la biomasa vegetal. Cuando esta biomasa se consume o se descompone, el CO_2 capturado se libera disuelto en la columna de agua.

La acidificación oceánica es el resultado de la absorción de CO_2 atmosférico por parte de los océanos, formando ácido carbónico y disminuyendo el pH del agua de mar. El pH promedio del océano desde el período industrial ha disminuido en 0.1 unidades, lo que representa un aumento del 30 % en la acidez. Este cambio afecta a organismos calcificadores como corales, moluscos y ciertos tipos de plancton, esenciales para la cadena alimentaria marina.

Paralelamente, la desoxigenación oceánica se refiere a la disminución del oxígeno disuelto en el agua marina, impulsada principalmente por el calentamiento global. El aumento de la temperatura reduce la solubilidad del oxígeno y provoca una mayor estratificación del agua, limitando la mezcla de capas y el intercambio de gases con la atmósfera. Desde la década de 1950, los océanos han perdido aproximadamente un 2 % de su oxígeno. Este descenso en los niveles de oxígeno afecta especialmente a las zonas superiores del océano, donde habita la mayoría de las especies marinas, induciendo estrés en los organismos, alterando los ciclos de nutrientes y modificando la distribución de las especies.

Además, estudios recientes indican que el cambio climático está provocando una **reorganización de las comunidades de fitoplancton**, con una reducción en la equitatividad de las especies, lo que podría aumentar la inestabilidad en la estructura de la comunidad y presentar desafíos significativos para la productividad de toda la red alimentaria marina.

1.3. El concepto de producción primaria y el rol del fitoplancton en la red trófica

El fitoplancton está constituido por plantas suspendidas en el océano, gran parte de ellas unicelulares. Se trata de un organismo fotoautótrofo que capta la luz para convertir el carbono inorgánico en orgánico, y a su vez suministra este carbono orgánico a diversos heterótrofos, organismos que obtienen su energía únicamente de la respiración de materia orgánica. Los heterótrofos del océano abierto incluyen bacterias, así como zooplancton (animales flotantes) unicelulares y multicelulares más complejos, necton (organismos nadadores, como peces y mamíferos marinos) y bentos (la comunidad de organismos del fondo marino).

Precisamente, con el cambio climático, esta familia de organismos, el fitoplancton, base de la cadena trófica marina, está experimentando variaciones en su diversidad y distribución. Es probable que el cambio climático antropogénico impulse la pérdida de biodiversidad y productividad primaria debido a una menor mezcla vertical y una disminución de nutrientes a la zona eufótica y, por ende, disminuya la estabilidad de los ecosistemas, afectando así tanto el funcionamiento como la estructura de los ecosistemas marinos. El calentamiento de los océanos y las alteraciones en el suministro de nutrientes mediante cambios en la circulación o la estratificación, combinados con factores de estrés adicionales como la acidificación y la desoxigenación de los océanos, probablemente forzarán la reorganización de las comunidades de fitoplancton. Predecir los cambios futuros en los ecosistemas marinos se convertirá en un desafío. Sin embargo, dado que se predice que los océanos futuros serán entre 2 y 4°C más cálidos, más ácidos y con una concentración de oxígeno reducida, las especies deberán adaptarse, migrar a regiones con condiciones análogas o enfrentar la extinción. Es probable que los cambios resultantes previstos en la biodiversidad afecten el funcionamiento y los procesos fundamentales de los ecosistemas, como la producción de biomasa y el mantenimiento de la calidad del agua, así como toda la estructura del ecosistema marino, con consecuencias en la capacidad del océano para la producción de alimentos y la regulación climática.

Como base de la red trófica marina, el fitoplancton desempeña un papel fundamental en la determinación de la productividad de todo el ecosistema marino. Cada grupo específico de fitoplancton también desempeña un papel clave en el funcionamiento biogeoquímico del océano. Es el caso de la fijación del nitrógeno atmosférico, en la que actúan diazótrofos, o el ciclo del sílice, donde toman partida las diatomeas. Además, la estructura de tamaño de la comunidad afecta las interacciones tróficas, la productividad de la red trófica y el potencial de secuestro de carbono.

Las alteraciones en las condiciones oceánicas están provocando cambios en la composición y distribución del fitoplancton. Se observa un desplazamiento hacia latitudes más altas de ciertas especies de fitoplancton en respuesta al calentamiento de las aguas. La disminución en la equitatividad de las especies puede aumentar la inestabilidad en la estructura de la comunidad fitoplanctónica, afectando la productividad y la resiliencia del ecosistema marino.

Estos cambios no solo afectan la biodiversidad marina, sino que también tienen implicaciones en procesos clave como la bomba biológica de carbono, que depende en gran medida de la actividad del fitoplancton para el secuestro de carbono en el océano profundo. Los

cambios en la composición y productividad del fitoplancton pueden afectar la eficiencia de este proceso, con implicaciones significativas para el ciclo global del carbono y el clima.

Uno de los procesos clave en este contexto es la producción primaria oceánica. La productividad oceánica se refiere principalmente a la producción de materia orgánica por parte del fitoplancton. Por lo tanto, se puede definir como el proceso mediante el cual los organismos vivos sintetizan formas inorgánicas de carbono en compuestos orgánicos simples.

El cambio climático afecta la producción primaria oceánica a través de varios mecanismos. El aumento de la temperatura superficial del océano intensifica la estratificación, reduciendo la mezcla vertical y, por ende, el suministro de nutrientes esenciales a la zona eufótica, donde reside la mayoría del fitoplancton. Además, la acidificación oceánica disminuye el pH del agua, afectando negativamente a organismos calcificantes como los coccolitóforos, lo que puede alterar la composición y funcionalidad de las comunidades fitoplanctónicas. La desoxigenación y la estratificación también contribuyen a la disminución de oxígeno en ciertas zonas oceánicas, creando ambientes menos habitables para diversas especies marinas.

1.4. Modelos actuales de estimación y limitaciones

La estimación de la producción primaria neta (NPP) en el océano se ha abordado históricamente mediante tres enfoques principales: observaciones satelitales, mediciones directas in situ y modelos biogeoquímicos computacionales. Cada uno de estos métodos aporta una perspectiva distinta, pero presenta limitaciones importantes que afectan la precisión y representatividad de las estimaciones obtenidas.

Las mediciones in situ, aunque son consideradas el estándar de referencia, presentan una cobertura espacial y temporal muy limitada debido a los altos costos y dificultades logísticas que implican las campañas oceanográficas de recogida de datos.

En contraposición, surgen las técnicas de teledetección satelital, que han permitido avances significativos gracias a su capacidad para monitorizar grandes extensiones oceánicas de forma continua. Sin embargo, estos sensores únicamente captan información de la superficie oceánica (aproximadamente los primeros 20 metros de profundidad), lo que resulta en un impedimento en la evaluación de procesos fotosintéticos más profundos y condiciona la representatividad vertical de las mediciones. También se ven afectadas las mediciones por otras limitaciones como la nubosidad persistente y la baja calidad de los datos en latitudes altas.

Los modelos biogeoquímicos como VGPM (*Vertically Generalized Production Model*) o CbPM (*Carbon-based Production Model*) emplean parametrizaciones que combinan datos de clorofila-a, irradiancia y temperatura superficial para estimar la NPP (Behrenfeld et al., 1997 [6]; Behrenfeld et al., 2005 [5]). Aunque ampliamente utilizados, estos modelos presentan incertidumbres asociadas a su calibración y a la elección de variables representativas. Algunos modelos tratan de superar estas limitaciones integrando otros factores. Sin embargo, continúan dependiendo de supuestos empíricos que pueden no reflejar adecuadamente la complejidad del entorno marino.

Dado el carácter multiescala, no lineal e interdisciplinario de los procesos oceánicos, la integración de datos multifuente con capacidad de aprendizaje y generalización se ha vuelto esencial. En este contexto, emergen nuevas aproximaciones como los gemelos digitales, que integran modelos dinámicos basados en datos con capacidades de simulación en tiempo real, y las redes neuronales artificiales, capaces de capturar patrones complejos en conjuntos de datos satelitales y oceanográficos. Estas tecnologías ofrecen mejoras sustanciales en la estimaciones de la productividad primaria, ya que permiten adaptar los modelos a regiones concretas y mejorar la resolución temporal y espacial en las estimaciones, dependiendo de la adecuada selección de variables, validación con datos empíricos y una arquitectura robusta que garantice la interpretabilidad y fiabilidad del modelo (Chattopadhyay et al., 2024 [12]; Cutolo et al., 2024 [19]).

1.5. Motivación

El estudio de la producción primaria neta en los océanos es clave en la comprensión del ciclo global del carbono y en la evaluación de la salud de los ecosistemas marinos. Ante el incremento acelerado del cambio climático de origen antropogénico y sus efectos en la biodiversidad marina, surge la necesidad de desarrollar herramientas predictivas más precisas, adaptadas a contextos regionales concretos. En este sentido, el empleo de gemelos digitales y modelos basados en redes neuronales artificiales ofrece un enfoque innovador para estimar variables oceánicas complejas a partir de datos satelitales y oceanográficos. La elección de la región macaronesia, caracterizada por una alta productividad biológica y una dinámica oceanográfica particular, aporta un valor añadido al estudio. Este Trabajo de Fin de Grado se enmarca en la creciente demanda de soluciones tecnológicas aplicadas a la sostenibilidad ambiental, y pretende contribuir a la mejora en la estimación de la producción primaria neta mediante el desarrollo de un gemelo digital entrenado con modelos validados científicamente.

1.6. Organización del documento

La presente memoria se estructura en siete capítulos. En el primer capítulo, se introduce el contexto global del cambio climático y el papel del océano en la regulación climática, describiendo además los conceptos clave relacionados con la producción primaria y los modelos existentes para su estimación. El segundo capítulo expone el estado actual del conocimiento y los objetivos del trabajo, profundizando en los antecedentes científicos y tecnológicos. En el tercer capítulo, se detallan las principales aportaciones del trabajo, las competencias adquiridas y su alineamiento con los Objetivos de Desarrollo Sostenible. El cuarto capítulo describe la metodología empleada, los recursos utilizados y las distintas fases del desarrollo del proyecto. A continuación, el quinto capítulo presenta los resultados obtenidos. El sexto capítulo recoge las conclusiones alcanzadas y plantea posibles líneas de trabajo futuro. Finalmente, el séptimo capítulo incluye anexos y otros contenidos complementarios que enriquecen el desarrollo del proyecto.

Capítulo 2

Estado actual y objetivos iniciales

En este segundo capítulo, se expone el origen y la motivación del proyecto, se introducen los conceptos clave, y se describe el contexto oceanográfico en el que se desarrolla el gemelo digital. Asimismo, se analizan los principales enfoques y modelos existentes para la medición y estimación de la productividad primaria, identificando sus limitaciones y debilidades, con el objetivo de señalar aquellas áreas en las que el gemelo digital propuesto puede aportar mejoras o innovaciones.

2.1. Motivación y antecedentes

2.1.1. El concepto de producción primaria neta

La producción primaria constituye la base de la red alimentaria en los océanos, y es la responsable de controlar la cantidad de material y energía disponible para la biosfera en su conjunto. También constituye el punto de entrada de nutrientes inorgánicos y carbono a la red alimentaria. Por tanto, es el flujo del ecosistema que se ve más directamente influenciado por los cambios antropogénicos en los ciclos biogeoquímicos globales (Buitenhuis et al., 2013 [10]).

Como descripción de este fenómeno, la mayor parte del carbono en la Tierra se encuentra en formas oxidadas inorgánicas, como el CO_2 . El carbono inorgánico debe reducirse químicamente para formar las moléculas orgánicas que son los componentes básicos de la vida y el mecanismo por el cual se almacena la energía en los organismos vivos. La reducción del carbono inorgánico requiere una inversión de energía, que puede provenir de la luz o de la energía almacenada en algunos compuestos inorgánicos reducidos. Los autótrofos son organismos capaces de fijar el carbono inorgánico. Los fotoautótrofos utilizan la energía de la luz para fijar el carbono, mientras que los quimioautótrofos utilizan la energía liberada a través de la oxidación de sustratos inorgánicos reducidos para fijar el carbono en compuestos orgánicos.

La producción primaria es de magnitud similar en las biosferas terrestre y marina, de

orden de $100 \text{ Pg C año}^{-1}$ para la producción bruta y 50 Pg C año^{-1} para la producción neta (Longhurst et al., 1995 [38]). El fitoplancton constituye solo alrededor del 0,05 % de la biomasa terrestre, pero puede fijar carbono equivalente a las plantas terrestres a través de procesos de producción primaria.

2.1.2. Área de estudio y la dinámica de la Corriente de Canarias

El sistema de afloramiento de la Corriente de Canarias (*Canary Current*, CanC) es una de las cuatro principales regiones de afloramiento de borde oriental del planeta (Carr et al., 2003 [11]), que se caracteriza por una alta productividad biológica debido al ascenso de aguas profundas ricas en nutrientes hacia la zona fótica (Chavez et al., 2009 [13]). Este sistema se extiende dentro de la cuenca canaria, cuyos límites están definidos aproximadamente entre los 10° norte y los 40° norte en el océano Atlántico nororiental (Aristegui, 2009 [3]). No obstante, aunque el gran ecosistema marino de la CanC, en su sentido más amplio, se extiende desde el extremo noroeste de la península Ibérica hasta el sur de Senegal, en este trabajo la región CanC hace referencia al área que se extiende desde el norte de Marruecos hasta Senegal (ilustración 2.1).

La dinámica del afloramiento de la CanC está impulsada principalmente por los vientos alisios, cuya intensidad y dirección varían estacionalmente, afectando la magnitud del afloramiento y, por ende, la disponibilidad de nutrientes en la zona eufótica. Además, la interacción entre la topografía costera y las corrientes oceánicas genera estructuras mesoescalares, como filamentos y remolinos, que facilitan el transporte horizontal de nutrientes y biomasa desde la costa hacia el océano abierto (Santana Falcón et al., 2020 [62]). Las masas de agua que predominan en la región también influyen en la productividad. Al norte del Cabo Blanco (21° norte), las Aguas Centrales del Atlántico Norte (NACW) son las principales, mientras que al sur predominan las Aguas Centrales del Atlántico Sur (SACW), que son más ricas en nutrientes. Otro factor de relevancia es la deposición de polvo sahariano, que aporta nutrientes esenciales como hierro y fósforo, potenciando la productividad en áreas donde estos elementos son limitantes. El transporte de aerosoles minerales desde el desierto del Sahara y el Sahel convierte al CanC en una de las regiones con las tasas de deposición de polvo más altas del mundo (Mahowald et al., 2005 [40]).

Todos estos factores físicos y químicos, en su conjunto, configuran un contexto y un entorno altamente dinámico y productivo, y cuya comprensión es esencial para la estimación precisa de la producción primaria neta (NPP).

2.1.3. Evaluación de las mediciones y modelos actuales para la estimación de la producción primaria neta

Las características físicas del océano, como las corrientes marinas o la radiación fotosintéticamente activa, influyen en la distribución de los organismos y las tasas de actividad biológica. Muchas características y procesos físicos, biológicos y químicos toman parte en estos procesos, así como en la productividad primaria oceánica, la red trófica oceánica o la

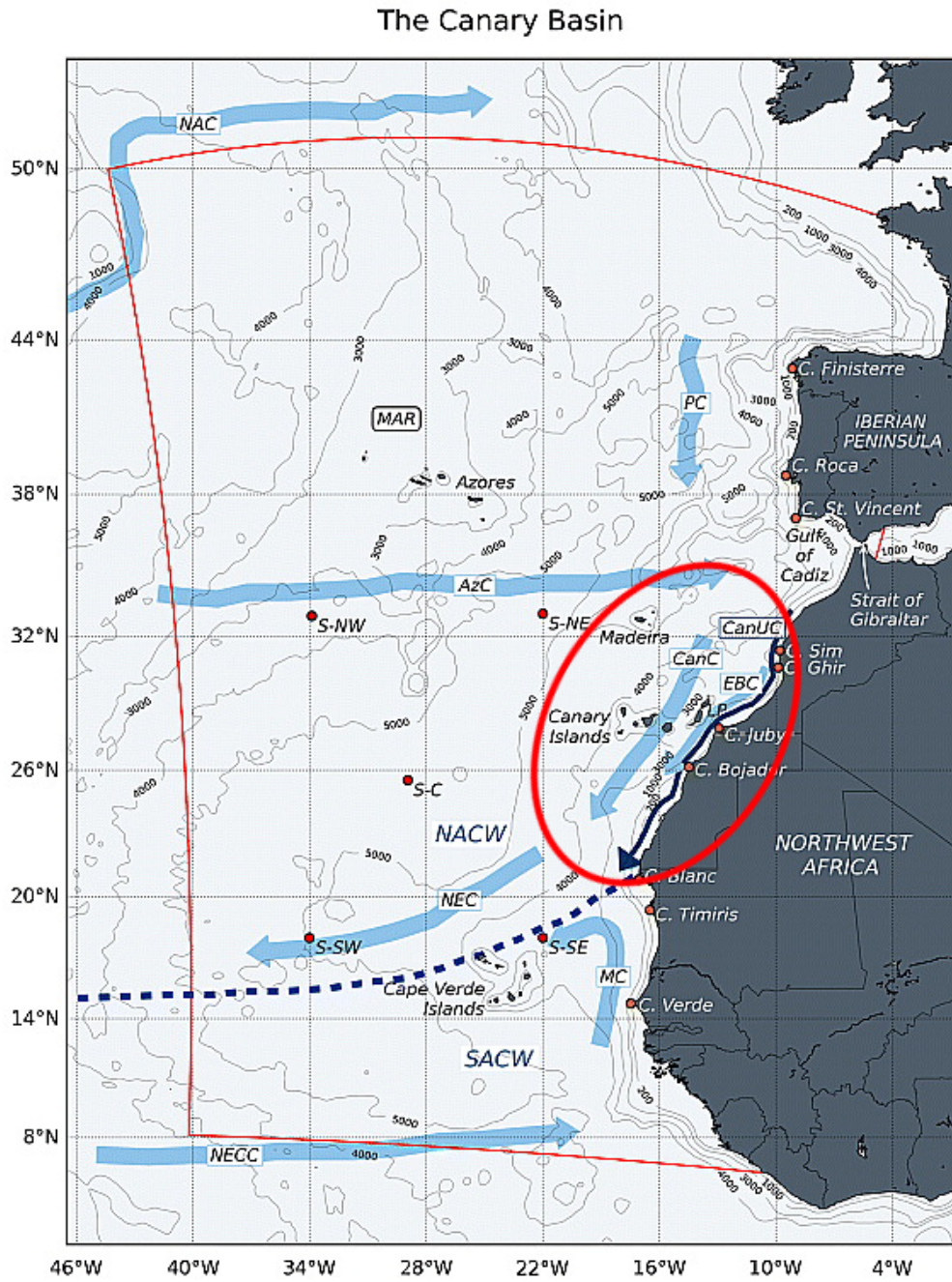


Ilustración 2.1: Esquema de la topología y circulación superficial de la cuenca canaria en el Atlántico nororiental subtropical. Véase CanC, la corriente de Canarias. Los contornos en negro marcan las isóbatas a 200, 1000, 3000, 4000 y 5000 metros. **Fuente:** Mason et al., 2011 [42].

bomba biológica, y su papel en el cambio del ciclo global del carbono (Lermusiaux, 2006 [37]). La mayoría de procesos oceánicos importantes son interdisciplinarios, multiescala y no lineales (Hsieh et al., 1998 [29]). A menudo, esta variabilidad no puede muestrearse ni modelarse de forma sostenida y sustancial, a escalas y precisiones suficientes para una representación definitiva (Chattopadhyay et al., 2024 [12]).

Los datos marinos in situ son necesarios para estudios realistas, pero su cobertura temporal y espacial es limitada (Johannessen et al., 2000 [30]). Los modelos oceánicos son aproximados debido a simplificaciones prácticas, representaciones o parametrizaciones inexactas e implementaciones numéricas (Buitenhuis et al., 2013 [10]). Esto genera incertidumbres o diferencias entre los valores reales, que son desconocidos, y los valores medidos o modelados de los campos y propiedades físicas, biológicas y acústicas (Saba et al., 2011 [60]).

Una predicción integral debe cuantificar la incertidumbre de las estimaciones. Esto permite un uso adecuado de estas estimaciones para la aplicación científica. En una predicción con un modelo que se integra en el tiempo y en el espacio, los errores en los datos iniciales, las condiciones de contorno y las imperfecciones del modelo afectan a la precisión. Las incertidumbres predichas contienen entonces los efectos integrados del error inicial y de los errores introducidos continuamente durante la integración del modelo.

Hasta la fecha, se han acumulado miles de mediciones in situ para el NPP en el océano global, basadas en muestreos discretos desde buques e incubaciones en botellas. Los trabajos pioneros de Nielsen, en 1951 (Nielsen, 1951 [50]) propusieron la técnica del trazador de carbono 14 (^{14}C). No obstante, la cobertura espacial de estos datos es notablemente limitada, introduciendo posibles artefactos y limitando la cobertura espacial y temporal del océano global, debido a las dificultades logísticas y económicas asociadas a los muestreos desde buques (Longhurst et al., 1995 [38]). Las campañas oceanográficas suelen concentrarse en regiones costeras o rutas navegables, dejando vastas áreas del océano abierto (p. ej., Pacífico Sur o Índico subtropical) sin datos directos. Además, las incubaciones en botellas —aunque útiles— pueden alterar las condiciones naturales, generando artefactos metodológicos. Esta escasez de datos introduce incertidumbres en la extrapolación de resultados a escala global y dificulta la captura de variabilidad estacional o eventos transitorios (p. ej., florecimientos algales). Todos estos datos representan solo una pequeña parte del conocimiento, en comparación con el gran tamaño del océano. La producción primaria del fitoplancton es un proceso complejo, y la escasez de datos limita la precisión de su parametrización en modelos globales (Zhang et al., 2025 [80]).

En contraposición, se encuentran los satélites de observación terrestre, que proporcionan una forma eficaz de explorar el océano, la atmósfera y el ciclo del carbono de la Tierra, transmitiendo información valiosa de manera constante o en tiempo real desde el espacio. El *Vanguard 2* (ilustración 2.2), diseñado por el Laboratorio de Investigación Naval de los Estados Unidos, NRL (en inglés, *United States Naval Research Laboratory*), lanzado en 1959, es considerado el primer satélite diseñado específicamente para la observación de la Tierra para medir la distribución de la cobertura de nubes. Desde entonces, se han lanzado numerosos satélites y sensores remotos de observación de la Tierra, entre los que destacan para la observación de los procesos oceánicos el Espectrorradiómetro de Imágenes de Resolución Moderada (en inglés, *Moderate-Resolution Imaging Spectroradiometer*, MODIS), Sentinel (parte

del programa europeo Copernicus), Sensor de Color del Océano de Vista Amplia (en inglés, *Sea-Viewing Wide Field-of-View Sensor*, SeaWiFS) y el Radiómetro de Imágenes por Infrarrojos Visibles (en inglés, *Visible Infrared Imaging Radiometer Suite*, VIIRS). Las mediciones de reflectancia del océano se pueden emplear, por ejemplo, para detectar la concentración de clorofila-a, indicador esencial de la biomasa fitoplanctónica.

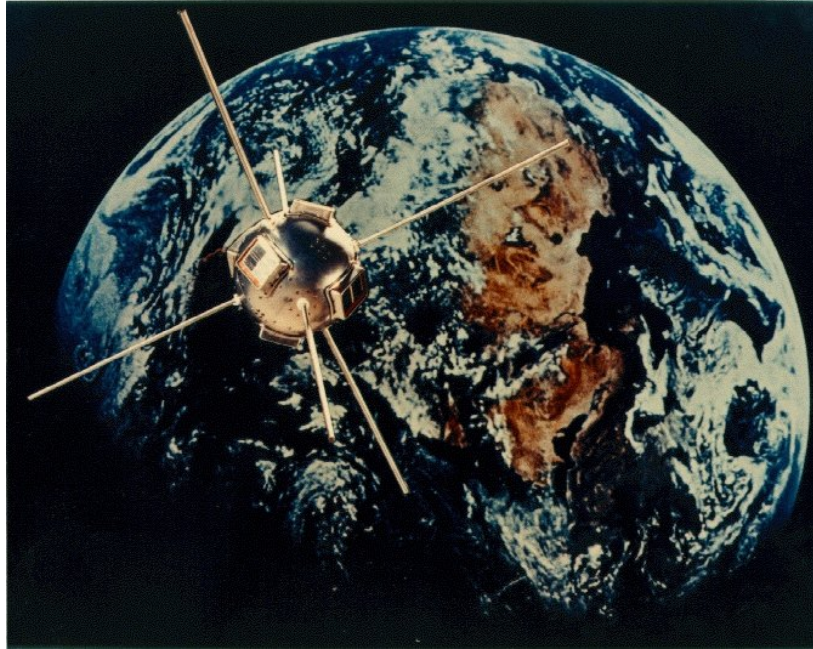


Ilustración 2.2: El satélite *Vanguard 2* en órbita sobre la Tierra. Fue uno de los primeros satélites diseñados para estudiar la atmósfera y la superficie terrestre. **Fuente:** *National Aeronautics and Space Administration* (NASA).

En las últimas décadas, el establecimiento de redes satelitales de observación oceánica han experimentado un auge, y han permitido realizar estimaciones razonables de los patrones de la NPP oceánica a gran escala. Se han logrado importantes avances en la resolución espacial, espectral y temporal de los datos de teledetección, y gracias a las mejoras en la tecnología de teledetección, han transformado drásticamente la forma en que se observa la Tierra, permitiendo un monitoreo global y de gran rapidez. Con la disminución de los costes de los instrumentos y el lanzamiento de satélites, la disponibilidad de datos de observación de la Tierra es mayor, y proporcionan información consistente, repetitiva y útil para la detección de tendencias a largo plazo.

La distribución global de la NPP se ha estimado utilizando observaciones satelitales con modelos como *Vertically Generalized Production Model*, VGPM (Behrenfeld et al., 1997 [6]), el *Absorption-based Productivity Model*, AbPM (Silsbe et al., 2016 [65]) y el *Carbon-based Productivity Model*, CbPM (Behrenfeld et al., 2005 [5]; Westberry et al., 2008 [74]), entre otros. Sin embargo, la teledetección satelital presenta limitaciones inherentes, como la falta de resolución vertical, ya que los sensores solo captan señales superficiales (únicamente los primeros metros de la columna de agua, hasta unos 20), lo que impide caracterizar procesos críticos como la producción primaria en la zona eufótica profunda; también existen lagunas de

cobertura en latitudes altas debido a los altos ángulos cenitales solares, que reducen la calidad de los datos espectrales. Estas restricciones introducen incertidumbres en la estimación de la NPP, especialmente en regiones donde la productividad está acoplada a procesos profundos o bajo condiciones de luminosidad variable (Hill et al., 2013 [27]). Con todo, en comparación con las mediciones in situ, la teledetección por satélite ofrece ventajas significativas debido a su amplia cobertura global y capacidades de observación a largo plazo (Platt et Sathyendranath, 1988 [52]; Yang et al., 2013 [77]; Lenton et al., 2024 [36]).

El monitoreo del NPP global del océano ha sido el objetivo principal de los estudios de teledetección satelital del color del océano desde el lanzamiento del sensor SeaWiFS en 1997. SeaWiFS fue el primer sensor satelital de color del océano con capacidad de calibración lunar, lo que permitió mantener una estabilidad radiométrica a largo plazo con una precisión relativa del 1% y una estabilidad radiométrica de 0,1% en las radiancias de la parte superior de la atmósfera (Eppley et al., 2012 [21]). Estas mediciones permiten estimar la concentración de clorofila en la capa superior del océano, utilizada para evaluar las reservas y la productividad del fitoplancton en toda la zona fótica. Los cambios interanuales en la abundancia de fitoplancton son pequeños en relación con las reservas totales, lo que requiere mediciones precisas a largo plazo para su detección. El legado de SeaWiFS fue continuado por el MODIS-Aqua, lanzado en 2002, que mejoró la resolución espectral (36 bandas) y la corrección atmosférica, clave para regiones costeras complejas. Posteriormente, el VIIRS, en 2011, introdujo una mayor resolución espacial (750 metros en bandas visibles) y capacidades de imagen en condiciones de baja luminosidad, extendiendo la observación a latitudes polares. Juntos, estos sensores han construido un registro sin precedentes de más de 25 años de datos homogeneizados, esencial para estudiar tendencias en la productividad fitoplanctónica y su respuesta al cambio climático.

Hasta la fecha, se han desarrollado muchos modelos para estimar la producción primaria a partir de mediciones satelitales (Platt et al., 1988 [52]; Morel, 1991 [45]; Behrenfeld et al., 1997 [6]; Westberry et al., 2008 [74]; Friedrichs et al., 2009 [25]; Lee et al., 2011 [35]; Saba et al., 2011 [60]), y se han publicado múltiples productos de estimación del NPP mediante observaciones satelitales, como los ofrecidos por la Universidad Estatal de Oregón [51]. Estos productos tienen un valor significativo para comprender las variaciones en las capacidades de fijación de carbono en el océano global, particularmente en un contexto de cambio climático como el actual (Westberry et al., 2023 [75]). En general, estos modelos se pueden agrupar en dos categorías según el producto satelital utilizado. Los primeros son los modelos basados en biomasa, que se basan en 1) la concentración de clorofila-a ($Chl-a$) (Platt y Sathyendranath, 1988 [52]; Longhurst et al., 1995 [38]; Behrenfeld y Falkowski, 1997 [6]; Brewin et al., 2021 [9]) o en 2) las concentraciones de carbono del fitoplancton (Behrenfeld et al., 2005 [5]; Westberry et al., 2008 [74]). La segunda categoría son los modelos independientes de la biomasa que, en cambio, se basan en el coeficiente de absorción del fitoplancton (aph) (Lee et al., 1996 [34]; Marra et al., 2003 [41]; Hirawake et al., 2011 [28]; Lee et al., 2011 [35]; Barnes et al., 2014 [4]).

Desde la perspectiva técnica y científica, algunos de los modelos de estimación de NPP ampliamente utilizados se han basado en estimaciones de clorofila-a desde el espacio, como el modelo VGPM (Behrenfeld et al., 1997 [6]), el CAFE (Silsbe et al., 2016 [66]) y el modelo de productividad basado en carbono, CbPM (Behrenfeld et al., 2005 [5]; Westberry et al.,

2008 [74]), siendo VGPM (Behrenfeld et al., 1997 [6]) el más popular, en parte debido a su simplicidad y facilidad de uso.

2.1.4. Modelos empleados en el proyecto

Para este trabajo, se ha planteado esencialmente el empleo de dos tipos de conjuntos de datos para la estimación de la NPP, basados en observaciones satelitales: los derivados del modelo Eppley-VGPM (Eppley et al., 1985 [22]) y los obtenidos a partir del modelo AbPM (Silsbe et al., 2016 [65]).

Para los modelos VGPM, la producción primaria integrada dentro de la zona eufótica se estima mediante la siguiente ecuación:

$$NPP_{VGPM} = 0,66125 \times P_{opt}^B \times \frac{PAR_{día}}{PAR_{día} + 4,1} \times Z_{eu} \times Chl - a \times DL$$

donde P_{opt}^B (medido en $mg\ C\ (mg\ Chl)^{-1}\ h^{-1}$) es la tasa máxima de fijación de carbono de la columna de agua normalizada por la clorofila-a ($Chl - a$), $PAR_{día}$ es la radiación fotosintética activa diaria disponible (medida en $mol\ fotones\ m^{-2}\ d^{-1}$), Z_{eu} (en m) es la profundidad de la zona eufótica y DL es la duración del día (en horas). $Chl - a$, $PAR_{día}$ y Z_{eu} están disponibles o se derivan directamente de mediciones del color del océano (Wu et al., 2024 [76]). P_{opt}^B se originó como una función polinómica de la temperatura superficial del mar (SST) (Behrenfeld et Falkowski, 1997 [6]).

Para el modelo Eppley-VGPM actualizado (derivado de VGPM) y empleado en el proyecto, P_{opt}^B se modeló como una función exponencial de la temperatura (Morel et al., 1991 [45]), obtenida en base a la función de crecimiento dependiente de la temperatura presentada por Eppley en 1972. La razón para seleccionar Eppley-VGPM se basa en su mejor rendimiento en comparación con el VGPM original, como se ha demostrado en diversos estudios previos (Friedland et al., 2012 [24]; Zhang et al., 2018 [79]). Asimismo, los productos globales de NPP basados en Eppley-VGPM están disponibles para su descarga en el sitio web de los desarrolladores del modelo [51] y ha sido aplicado sobre datos satelitales provenientes de tres sensores distintos, empezando por SeaWiFS, MODIS-Aqua y VIIRS, lo que ha permitido generar un conjunto de datos continuo y armonizado con una cobertura temporal mensual desde 2002 hasta fechas muy recientes. De estos, se han considerado datos de resolución mensual, e interpolando los valores de NPP a una malla regular para su uso como conjunto de entrenamiento de una red neuronal (un área espacial equivalente a un cuadrado de 480 puntos en ambas dimensiones, a lo largo de su longitud y latitud).

No obstante, se ha sugerido que los modelos que usan solo datos de teledetección, como estos últimos, generalmente subestiman el NPP. Diversas evaluaciones han señalado limitaciones en su precisión. Saba (Saba et al., 2011 [60]) señala que modelos derivados de perfiles de NPP resueltos en profundidad a partir de señales de superficie, como VGPM, tienden a subestimar la NPP en ciertas regiones oceánicas, especialmente en áreas con alta productividad o condiciones ambientales complejas. Por ejemplo, en la estación de series temporales

del océano de Hawái (HOT; 22,75° norte, 158° oeste) cuando se empleó la concentración de $Chl - a$ (tabla 2.1), variable biológica derivada de las radiancias satelitales y detectada remotamente como sustituta de la masa fitoplanctónica (Westberry et al., 2008 [74]). Además, diversos estudios comparativos han demostrado que VGPM puede no capturar adecuadamente la variabilidad temporal y espacial de la NPP, lo que sugiere la necesidad de modelos más avanzados que consideren factores adicionales, como la absorción óptica del fitoplancton y la fluorescencia natural, para mejorar las estimaciones de la productividad primaria en el océano.

Región	$Chl - a \pm$	SST $\pm$	PAR $\pm$	MLD $\pm$
BATS (Océano Atlántico Noroeste)	35 %	1°C	20 %	40 %
NABE (Océano Atlántico Nororiental)	50 %	1°C	20 %	40 %
NEA (Océano Atlántico Nororiental)	50 %	1°C	20 %	20 %
Mar Negro	50 %	1°C	20 %	40 %
Mar Mediterráneo	65 %	1°C	20 %	40 %
Mar Arábigo	50 %	1°C	20 %	40 %
HOT (Océano Pacífico Norte)	35 %	1°C	20 %	40 %
Mar de Ross	65 %	1°C	20 %	60 %
WAP (Océano Antártico)	65 %	1°C	20 %	60 %
APFZ (Océano Antártico)	65 %	1°C	20 %	40 %

Tabla 2.1: Incertidumbres en cada variable de entrada por región y programa de estudio, basadas en diferencias entre datos satelitales, modelados y *in situ*. Los modelos de color del océano utilizaron 81 perturbaciones de los datos de entrada para cada medición de NPP, según estas incertidumbres específicas por región. **Fuente:** Saba et al., 2011 [60].

Ante estas limitaciones, han surgido modelos más avanzados como el modelo CAFE (*Carbon, Absorption, and Fluorescence Euphotic-resolving*, Silsbe et al., 2016 [66]) o el modelo AbPM (*Absorption-based Productivity Model*, Lee et al., 1996 [34]), que aprovechan parámetros ópticos como la absorción del fitoplancton o la fluorescencia natural para mejorar las estimaciones de producción primaria. Estos modelos muestran una sensibilidad mayor a la variabilidad fisiológica del fitoplancton y tienden a ofrecer mejores resultados que los algoritmos tradicionales.

Los modelos AbPM estiman la NPP a partir de la radiación solar absorbida por el fitoplancton y su eficiencia de conversión en carbono orgánico, y su ecuación fundamental se expresa como:

$$NPP_{AbPM} = E_{abs} \times \phi$$

donde E_{abs} representa la radiación solar absorbida por los pigmentos fotosintéticos del fitoplancton, mientras que ϕ es el rendimiento cuántico de la fotosíntesis del fitoplancton (medido en mol C (mol fotones)⁻¹), que representa la eficiencia de conversión de energía lumínica en carbono fijado (Kiefer y Mitchell, 1983; Antoine y Morel, 1996; Lee et al., 1996).

Para modelos AbPM que incluyen variación vertical, E_{abs} se calcula integrando la absorción espectral del fitoplancton a lo largo de la zona eufótica:

$$E_{abs}(z) = \int_{400}^{700} a_{ph}(\lambda) \times E_{día}(z, \lambda) d\lambda$$

donde $a_{ph}(\lambda)$ es el coeficiente de absorción del fitoplancton a la longitud de onda, que se estima a partir $a_{ph}(443)$ utilizando el modelo empírico presentado por Lee (Lee et al., 1999), asumiendo un perfil constante en la zona eufótica, en pasos de 5 nm entre 400 y 700 nm, y $E_{día}(z, \lambda)$ es la irradiancia espectral diaria a profundidad z (en mol fotones $m^{-2} d^{-1}$), que se puede calcular a partir de la irradiancia superficial $E_{día}(0-, \lambda)$ y el coeficiente de atenuación difusa $K_d(\lambda)$.

La variación vertical de ϕ se modela como:

$$\phi(z) = \phi_m \times \frac{K_\phi}{K_\phi + PAR_{día}(z)} \times \exp(-\nu \times PAR_{día}(z))$$

donde ϕ_m es el rendimiento cuántico máximo de la fotosíntesis, K_ϕ es un parámetro del modelo que describe la reducción de ϕ bajo mayor radiación, y ν es un parámetro para la fotoinhibición (Lee et al., 1996 [34]; Lee et al., 2011 [35]). Los valores de ϕ_m , K_ϕ y ν se tomaron como 0,06 mol C (mol fotones) $^{-1}$ (Morel, 1991 [45]; Lee et al., 2011 [35]), 10,0 mol fotones $m^{-2} d^{-1}$ y 0,01 mol fotones $m^{-2} d^{-1}$ (Lee et al., 2011 [35]), respectivamente, y se mantuvieron constantes en este estudio. La integración de la ecuación sobre la profundidad de la zona eufótica proporciona entonces la NPP de la columna de agua (Wu et al., 2024 [76]).

Por tanto, para tareas como el entrenamiento de redes neuronales que estimen NPP, es preferible emplear conjuntos de datos como Eppley-VGPM por su amplia cobertura y continuidad. Para validación, resulta adecuado utilizar conjuntos como AbPM, por su mayor fidelidad con observaciones regionales y fisiológicas del fitoplancton. Esta combinación mejora la capacidad predictiva del modelo y garantiza su coherencia con observaciones reales. El tabla 2.2 proporciona medidas estadísticas para la comparación entre los modelos VGPM, AbPM y CbPM.

2.1.5. Gemelos digitales y estado actual de los modelos estimativos para la producción primaria

El avance en la modelización de variables oceánicas, como la NPP, ha dado un salto cualitativo con la adopción de enfoques híbridos que combinan observaciones satelitales, mediciones *in situ*, modelos biogeoquímicos y, más recientemente, gemelos digitales. Estos últimos representan una revolución metodológica al permitir la simulación dinámica y en tiempo cuasi-real de los ecosistemas marinos, integrando flujos de datos multiescala y técnicas de inteligencia artificial.

Estación	Modelo	N	R^2	σ_d	RMSD	Bias	uRMSD	B^*	$uRMSD^*$	$uRMSD^*$	Sim. coseno
HOT	AbPM	188	0.33	72.4	0.10	0.00	0.10	-0.02	-0.87	0.87	0.90
HOT	VGPM	188	0.22	48.6	0.21	-0.19	0.10	-1.64	-0.87	1.86	0.90
HOT	CbPM	188	0.23	92.1	0.32	-0.28	0.16	-2.42	1.40	2.80	0.89
BATS	AbPM	226	0.25	117.0	0.21	0.07	0.20	0.31	-0.89	0.94	0.87
BATS	VGPM	226	0.19	153.4	0.28	-0.10	0.26	-0.46	-1.14	1.23	0.84
BATS	CbPM	226	0.07	105.1	0.71	-0.40	0.58	-1.78	2.57	3.13	0.71
CARIACO	AbPM	187	0.54	761.5	0.19	0.02	0.19	0.08	0.83	0.83	0.87
CARIACO	VGPM	187	0.52	845.7	0.20	-0.06	0.19	-0.28	0.86	0.90	0.85
CARIACO	CbPM	187	0.40	611.8	0.21	0.11	0.18	0.48	-0.81	0.94	0.86
Total	AbPM	601	0.67	546.9	0.18	0.03	0.17	0.13	-0.66	0.67	0.87
Total	VGPM	601	0.64	588.5	0.24	-0.11	0.21	-0.44	0.79	0.90	0.84
Total	CbPM	601	0.59	633.9	0.48	-0.20	0.44	-0.75	1.68	1.84	0.82

Tabla 2.2: Medidas estadísticas de las comparaciones entre la NPP *in situ* y la modelada (NPP_{AbPM}, NPP_{VGPM} y NPP_{CbPM}) utilizando datos OC-CCI. **Leyenda de variables:** N = número de observaciones; R^2 = coeficiente de determinación; σ_d = desviación estándar de las diferencias (mg C m⁻² día⁻¹); $RMSD$ = raíz del error cuadrático medio; $Bias$ = sesgo medio; $uRMSD$ = RMSD no sesgado; B^* , $uRMSD^*$ = métricas normalizadas; Similitud coseno = medida de similitud angular (1 = máxima similitud). **Fuente:** Wu et al., 2024 [76].

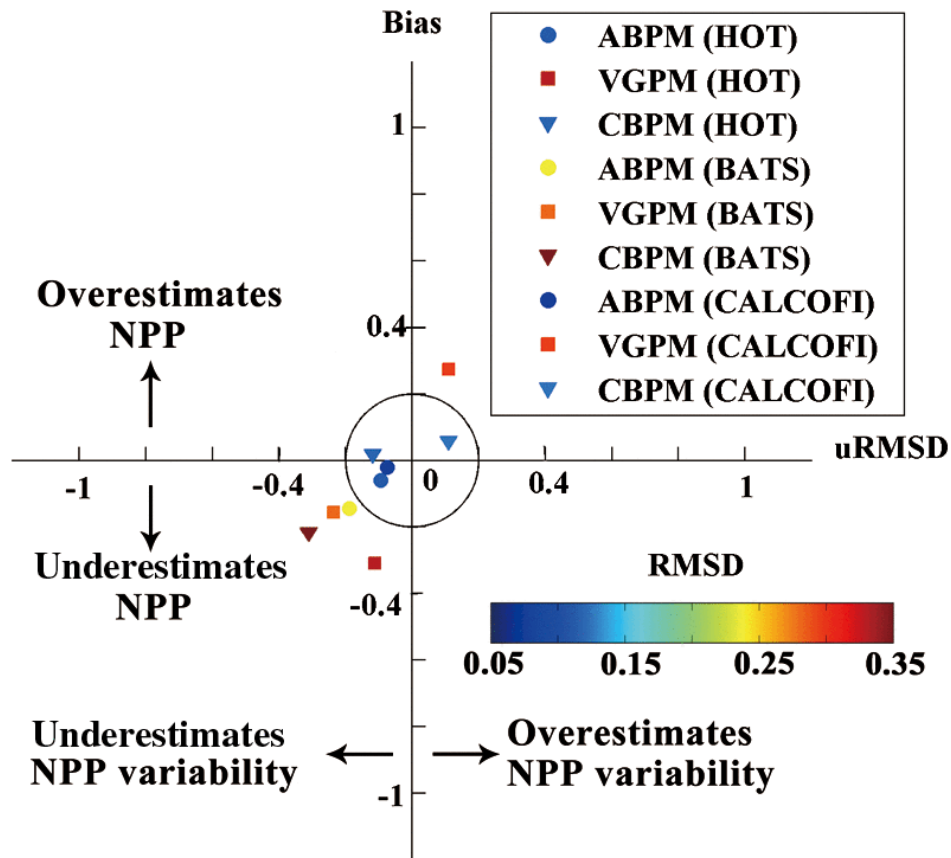


Ilustración 2.3: Diagramas objetivo comparando el rendimiento de los modelos AbPM, VGPM y CbPM en las estaciones HOT, BATS y CALCOFI. (A) El círculo sólido representa la desviación estándar promedio de los datos observados (σ_d). (B) La distancia desde el origen al símbolo de cada modelo indica el error cuadrático medio total (RMSD). Colores: azul (subestimación), rojo (sobreestimación) de la variabilidad de NPP. **Fuente:** adaptado de Ma et al., 2014 [39].

Conceptualmente, un gemelo digital es una réplica virtual de un sistema físico (en este caso, el océano) que se alimenta de datos observacionales continuos para actualizar sus estados y predecir escenarios futuros. Su aplicación en el estudio de los océanos y variables como la NPP proporciona importantes ventajas, incluyendo la integración en tiempo real de datos satelitales (MODIS, VIIRS), mediciones in situ (BGC-Argo, boyas) y modelos biogeoquímicos (NPZ, VGPM, AbPM) en un sistema unificado, la capacidad para simular y evaluar los impactos de perturbaciones ambientales como el calentamiento global y la acidificación oceánica, y una notable eficiencia computacional que permite análisis y pronósticos operacionales con menor costo energético y temporal.

El modelado basado en datos en el ámbito de las ciencias terrestres y atmosféricas se ha centrado tradicionalmente en la predicción de la dinámica atmosférica global, con la predicción meteorológica como referencia, pero recientemente se ha explorado su aplicación al NPP oceánico. La principal ventaja de estos enfoques reside en su eficiencia computacional, que permite ejecutar un gran número de ensayos, facilitando una asimilación de datos fluida y eficiente, lo que conduce a mejores estimaciones de fenómenos extremos. En este contexto, las últimas investigaciones apuntan al modelo XGBoost_CbPM entrenado con medidas BGC-Argo y datos de MODIS-Aqua que alcanza un $R^2 = 0,87$ y un error medio porcentual absoluto del 12,5 %, superando en exactitud a productos tradicionales de NPP base mapa en series temporales como BATS y HOT (Zhang et al., 2025[80]). Así, estudios recientes demuestran que los modelos basados en datos pueden alcanzar una precisión comparable —y a menudo superior— a la de los modelos numéricos de vanguardia, con un coste computacional significativamente reducido. Este avance resulta muy prometedor para la incorporación de modelos basados en datos y aprendizaje profundo en los sistemas operativos de pronóstico de productividad primaria oceánica.

La introducción de métodos empíricos o estadísticos en la meteorología y la oceanografía puede clasificarse, a grandes rasgos, en cuatro etapas distintas: 1) la regresión y el análisis de correlación de Galton (1885) para hallar una relación lineal entre pares de variables x y z ; 2) técnicas multivariantes (ACP), introducidas por Pearson para hallar los patrones correlacionados en un conjunto de variables $x_1, \dots, x_n$. Se introdujo en la física oceanográfica a mediados de la década de 1970; 3) el análisis de correlación canónica (ACC) de Hotelling, para relacionar linealmente un conjunto de variables $x_1, \dots, x_n$ con otro conjunto de variables $z_1, \dots, z_m$ y 4) el método de redes neuronales, que apenas comienza a aparecer en meteorología y oceanografía, y tiene un origen biológico-psicológico, ya que se desarrolló a partir de investigaciones sobre la función cerebral humana. El motivo por el que se escoge el cerebro como fundamento de estos razonamientos es que es mucho más robusto y tolerante a fallos que una computadora. Esta robustez es resultado de la estructura computacional masivamente paralela del cerebro, que surge de la interconexión de las neuronas por una red. Es la estructura computacional paralela subyacente en este sistema la que permite que la red neuronal, con velocidades de reloj neuronales típicas de unos pocos milisegundos, aproximadamente una millonésima parte de la de una computadora, supere a la computadora en visión, control motor y muchas otras tareas.

Fascinados por estas características, los científicos comenzaron a estudiar el procesamiento lógico en redes neuronales en la década de 1940 (McCulloch et al., 1943 [43]). A finales

de la década de 1950 y principios de la de 1960, F. Rosenblatt y B. Widrow investigaron de forma independiente el aprendizaje y la capacidad computacional de las redes neuronales. En la segunda mitad de la década de 1980, el redescubrimiento del algoritmo de retropropagación (Rumelhart et al., 1986 [58]) resolvió con éxito el problema de encontrar los pesos en un modelo con capas ocultas. Desde entonces, la investigación sobre redes neuronales se popularizó rápidamente en inteligencia artificial, robótica y muchos otros campos (Crick, 1989 [18]). Se descubrió que las redes neuronales superaban a los modelos lineales de Box-Jenkins (Box et al., 1970 [8]) en la predicción de series temporales con memoria corta y plazos de entrega más largos (Tang et al., 1991 [70]).

En los últimos años, se han desarrollado modelos estimativos avanzados que incorporan inteligencia artificial y aprendizaje profundo para modelar y predecir dinámicas oceánicas complejas. Un ejemplo destacado es *OceanNet*, un gemelo basado en operadores neuronales para predecir la altura de la superficie del mar en regiones específicas, como la Corriente del Golfo. *OceanNet* emplea un esquema para mitigar el crecimiento de errores autorregresivos y mejorar la estabilidad a lo largo del tiempo. Además, utiliza un regularizador espectral para contrarrestar el sesgo espectral en escalas más pequeñas. Este modelo ha demostrado un rendimiento competitivo en la predicción de la altura de la superficie del mar, superando a modelos oceánicos dinámicos no acoplados de última generación, y reduciendo el costo computacional significativamente (Chattopadhyay et al., 2024 [12]). También destacan estudios como el de Belgrano (Belgrano et al., 2001 [7]), donde se aplicaron redes neuronales artificiales para predecir la productividad primaria a partir de series temporales de datos ambientales, demostrando una mejora en la precisión respecto a métodos tradicionales (tabla 2.3). Investigaciones más recientes han empleado algoritmos de aprendizaje automático, como Random Forest (Yu et al., 2019 [78]), para escalar observaciones de campo y estimar la productividad primaria bruta a nivel global.

La integración de estos enfoques podría proporcionar una visión más completa y precisa de la producción primaria neta y la dinámica del fitoplancton, esencial para comprender y predecir los impactos del cambio climático en los ecosistemas marinos.

2.2. Objetivos

El objetivo principal de este Trabajo de Fin de Grado es el desarrollo de un gemelo digital para la estimación de la NPP en el océano mediante el uso de redes neuronales artificiales, que son aceptadas hoy en día como una tecnología alternativa que ofrece una forma de abordar problemas complejos y de difícil definición. El fin de este gemelo digital es estimar la cantidad de partículas biogénicas en función de diversas variables ambientales, contribuyendo a una estimación precisa y dinámica de la producción primaria en distintos escenarios ambientales.

Uno de los objetivos específicos para este proyecto transita por la identificación y la recopilación de datos relevantes para el gemelo digital, obtenidos de múltiples fuentes, incluyendo mediciones in situ siempre que sea posible, datos de laboratorio y datos de satélite. Estos datos incluirán variables clave como la temperatura del agua, la concentración de clorofila-a,

Retardo	RMSEP	r	N° variables	Variables de entrada	Método RMSEP
0	0.6548	0.866	Sin restricción	$Chl - a$ 15-30 m, PO_4^{3-} 0-10 m, PO_4^{3-} 15-30 m, NO_3^- 0-10 m, Temperatura 15-30 m, Salinidad 0-10 m, Precipitación, Vientos SO y NO PO_4^{3-} 15-30 m, NO_3^- 0-10 m, NAO, Densidad 0-10 m, Precipitación, Vientos SE y NO	Conjunto prueba
1	0.6413	0.870	Sin restricción	Todas las 18 variables	Conjunto prueba
2	0.6439	0.871	Fijas	$Chl - a$ 0-10 m, PO_4^{3-} 0-10 m, NO_3^- 0-10 m, NO_3^- 15-30 m, Salinidad 0-10 m, Densidad 0-10 m, Precipitación, Vientos NE, NO, SO y SE	Conjunto prueba
3	0.6529	0.868	Sin restricción		Ajuste red

Leyenda: RMSEP: Raíz del Error Cuadrático Medio de Predicción; r : coeficiente de correlación de Pearson; N° variables: “Fijas” = todas las variables incluidas; “Sin restricción” = selección automática de variables; Método RMSEP: “Conjunto prueba” = evaluación con datos independientes; “Ajuste red” = evaluación durante el entrenamiento.

Tabla 2.3: RMSEP y r (coeficiente de correlación entre valores observados y predichos de producción primaria) para las redes neuronales que implementan retardos de 1, 2 y 3 meses entre producción primaria y las variables de entrada, para evaluar la predicción de cambios futuros en producción primaria. **Fuente:** Belgrano et al., 2001 [7].

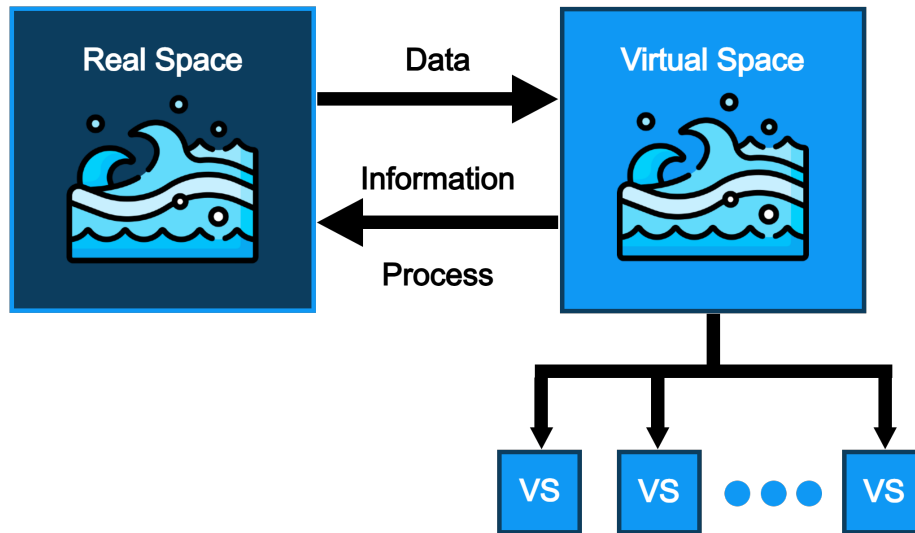


Ilustración 2.4: Representación esquemática del modelo de espacios reflejados (*Mirrored Spaces Model*) propuesto por Michael Grieves, considerado el precursor conceptual de los gemelos digitales. **Elementos clave:** (1) **Espacio real:** entidad física (ej: sistema oceanográfico), (2) **Espacio virtual (VS):** réplica digital sincronizada, (3) **Flujo bidireccional:** intercambio de datos en tiempo real (flechas). **Fuente:** adaptado de Singh et al., 2021 [67], basado en el trabajo original de Grieves (2002).

los tipos de fitoplancton y sus densidades en la región del océano escogida, el CO_2 atmosférico, la radiación fotosintéticamente activa, las corrientes marinas, y la localización y tiempo de las mediciones, que influyen notablemente en la producción primaria en el océano.

Posteriormente, se desarrollará y entrenará un modelo de estimación de la productividad primaria en el océano basado en redes neuronales artificiales, capaz de correlacionar las variables ambientales y físicas recopiladas para estimar la cantidad de partículas biogénicas por volumen oceánico. Se evaluarán adicionalmente diferentes arquitecturas de redes neuronales para optimizar la precisión de las predicciones.

Otro de los objetivos clave será la implementación de un gemelo digital para la predicción de la producción primaria. El gemelo digital utilizará el modelo de redes neuronales anteriormente mencionado para simular la producción primaria en el océano bajo distintos escenarios ambientales. Esto permitirá realizar predicciones dinámicas y visualizar los resultados en gráficos interactivos.

A continuación, se llevará a cabo una validación del modelo mediante la comparación de sus resultados con datos reales, o, en su defecto, con datos lo más representativos posible del entorno real. Esta validación rigurosa se basará en la confrontación de las estimaciones obtenidas con datos procedentes de campañas oceanográficas y observaciones satelitales. Este proceso permitirá ajustar el modelo, mejorando así su precisión y sensibilidad.

Más adelante, se buscará la visualización de los resultados de manera dinámica, mediante el desarrollo de una interfaz gráfica que permita mostrar las predicciones de producción

primaria de forma dinámica e interactiva, facilitando el análisis de las variaciones en función de las condiciones ambientales.

Por último, se atenderá a la evaluación del rendimiento del gemelo digital en distintos escenarios ambientales, probando su funcionamiento en diferentes regiones oceánicas y bajo condiciones ambientales variables, evaluando su capacidad para generalizar las predicciones y su aplicabilidad en escenarios de cambio climático.

Esta serie de objetivos, en su conjunto, permitirá crear una herramienta útil para la estimación de la producción primaria en el océano, contribuyendo a una mejor gestión de los recursos marinos y al entendimiento de las fluctuaciones en los ecosistemas oceánicos.

Capítulo 3

Aportaciones del trabajo

El presente capítulo se encamina a destacar las contribuciones más relevantes de este Trabajo de Fin de Grado, tanto desde una perspectiva técnica como científica. A lo largo del desarrollo del proyecto, se han abordado desafíos significativos en el ámbito de la oceanografía computacional, integrando tecnologías avanzadas de inteligencia artificial y análisis de datos para la estimación de la producción primaria neta (NPP) en el océano. Este capítulo se estructura en tres secciones principales. En primer lugar, se presentan las aportaciones clave del trabajo; en segundo lugar, se detallan las competencias específicas del Grado en Ingeniería Informática que se han desarrollado durante el proyecto; finalmente, se analiza la alineación del trabajo con los Objetivos de Desarrollo Sostenible (ODS). Se trata de reflejar el rigor metodológico y técnico del proyecto, además de su potencial impacto en la comprensión de los ecosistemas marinos y su relación con el cambio climático.

3.1. Principales aportaciones

El presente Trabajo de Fin de Grado presenta varias aportaciones significativas en los ámbitos técnico, científico y socioeconómico. La primera de las aportaciones es el desarrollo de un gemelo digital innovador, basado en el empleo de redes neuronales artificiales, para estimar la producción primaria neta (NPP) en el océano.

El modelo plantea el uso de datos satelitales aplicados en diferentes algoritmos y modelos biogeoquímicos provenientes de múltiples fuentes, aplicando técnicas de aprendizaje automático para alcanzar la precisión de las estimaciones existentes. La implementación de este gemelo digital representa una aportación en la integración de tecnologías de inteligencia artificial en el estudio de los ecosistemas marinos, cada vez más presentes y generalizadas.

Por otra parte, el modelo resultante desarrollado busca superar las limitaciones que presentan los métodos tradicionales, como Eppley-VGPM y AbPM, al incorporar variables adicionales y técnicas de interpolación avanzadas (como las funciones de base radial, RBF). Se busca una estimación más precisa y dinámica de la producción primaria, especialmente en regiones con alta variabilidad ambiental, como la Corriente de Canarias.

Asimismo, uno de los aportes elementales del trabajo es contribuir en parte a una mejora comprensión del ciclo del carbono y de los procesos tróficos fundamentales, a fin de evaluar y obtener una comprensión adicional de los impactos del cambio climático en los ecosistemas marinos.

Desde una perspectiva técnica propia de la Ingeniería Informática, la gestión del proyecto se ha basado en la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), complementada con herramientas de gestión ágil como los tableros Kanban de Trello. Para el desarrollo e implementación del proyecto, se han utilizado tecnologías avanzadas como Python, junto con bibliotecas de procesamiento y visualización de datos (**matplotlib**, **seaborn**) y software especializado de terceros (NASA Panoply). Todo ello ha permitido integrar enfoques informáticos y metodologías propias del área en el contexto de la investigación oceanográfica.

Además, una de las principales contribuciones del presente Trabajo de Fin de Grado es la incorporación de un lenguaje específico de dominio (DSL) para el modelado de gemelos digitales, mediante la herramienta Picota (Monentia, 2025 [44]). Este enfoque permite describir de forma estructurada y comprensible los elementos del sistema oceánico a estudiar (variables físicas, químicas y biológicas) y facilita su conexión con modelos de inferencia entrenados a partir de datos reales.

La formalización mediante DSL aporta múltiples ventajas en términos de legibilidad, mantenibilidad y reproducibilidad del gemelo digital. Además, se ha planteado un primer análisis sobre si el uso de estos lenguajes contribuye a facilitar la tarea a usuarios no expertos en programación, aumentando así la accesibilidad y escalabilidad de este tipo de herramientas para comunidades científicas o equipos multidisciplinares. Se trata de una aportación que abre las puertas a futuras investigaciones centradas en la evaluación empírica del impacto del uso de DSL en la eficiencia del desarrollo de gemelos digitales, así como en su precisión predictiva o su facilidad de uso en entornos colaborativos.

En esta misma línea, el impacto científico del proyecto busca que los resultados obtenidos pueden ser utilizados por la comunidad científica para validar hipótesis sobre la dinámica del fitoplancton y su relación con el clima. Económicamente, la herramienta desarrollada puede apoyar la gestión sostenible de recursos marinos, como la pesca y la acuicultura, al proporcionar datos precisos sobre la productividad oceánica. En el ámbito educativo, el proyecto sirve como ejemplo de cómo las tecnologías de la información pueden aplicarse a problemas ambientales complejos, fomentando la interdisciplinariedad.

El trabajo se alinea con varios Objetivos de Desarrollo Sostenible (ODS), en especial con el ODS 14 (vida submarina), al promover la conservación de los océanos y el ODS 15 (vida de ecosistemas terrestres), al analizar indirectamente los efectos del cambio climático en ecosistemas terrestres a través del estudio de las interacciones del océano, el transporte de nutrientes y la regulación del ciclo del carbono.

Este Trabajo de Fin de Grado no solo busca contribuir en los avances del campo de la oceanografía computacional, sino que también ofrece herramientas y conocimientos valiosos para abordar desafíos globales como el cambio climático y la sostenibilidad de los ecosistemas marinos. Su repercusión se extiende a la academia, la industria y la sociedad, destacando el

potencial de la inteligencia artificial para impulsar la investigación científica.

3.2. Competencias específicas

En esta sección, se detallan las competencias específicas del Grado de Ingeniería Informática que se han desarrollado de forma directa a través del presente Trabajo de Fin de Grado, a partir de la documentación del título dispuesta en las fuentes de la universidad:

- ✓ La competencia específica 1 (CI1, *capacidad para diseñar, desarrollar, seleccionar y evaluar aplicaciones y sistemas informáticos, asegurando su fiabilidad, seguridad y calidad, conforme a principios éticos y a la legislación y normativa vigente*) se aborda mediante el diseño y desarrollo del sistema de análisis y estimación de datos oceanográficos. Se garantiza la fiabilidad de los resultados mediante validaciones y documentación del flujo de trabajo.
- ✓ La competencia específica 6 (CI6, *conocimiento y aplicación de los procedimientos algorítmicos básicos de las tecnologías informáticas para diseñar soluciones a problemas, analizando la idoneidad y complejidad de los algoritmos propuestos*) llega mediante los requerimientos de selección y aplicación de algoritmos de aprendizaje automático y su adaptación al contexto de los datos seleccionados, considerando su rendimiento y su complejidad computacional.
- ✓ La competencia específica 7 (CI7, *conocimiento, diseño y utilización de forma eficiente de los tipos y estructuras de datos más adecuados a la resolución de un problema*) referencia al procesamiento de grandes volúmenes de datos oceanográficos, haciéndose un uso eficiente de estructuras de datos concretas como arrays multidimensionales, *dataframes*, o estructuras jerárquicas, en formato NetCDF.
- ✓ La competencia específica 8 (CI8, *capacidad para analizar, diseñar, construir y mantener aplicaciones de forma robusta, segura y eficiente, eligiendo el paradigma y los lenguajes de programación más adecuados*) se aproxima a las implementaciones realizadas con Python y la selección de bibliotecas especializadas, priorizando eficiencia en el código y la trazabilidad en los resultados.
- ✓ La competencia específica 13 (CI13, *conocimiento y aplicación de las herramientas necesarias para el almacenamiento, procesamiento y acceso a los sistemas de información, incluidos los basados en web*) hace referencia al trabajo realizado con herramientas de almacenamiento y procesamiento de datos oceanográficos (p. ej., mediante NetCDF y NASA Panoply) y bibliotecas de acceso y tratamiento de conjuntos, como `xarray`, `pandas` y `matplotlib`.
- ✓ La competencia específica 15 (CI15, *conocimiento y aplicación de los principios fundamentales y técnicas básicas de los sistemas inteligentes y su aplicación práctica*) resulta esencial en el uso de técnicas de *machine learning*, como clasificación y clustering, aplicadas a la interpretación de datos ambientales y su análisis exploratorio.

- ✓ La competencia específica TFG (*ejercicio original a realizar individualmente y presentar y defender ante un tribunal universitario, consistente en un proyecto en el ámbito de las tecnologías específicas de la Ingeniería en Informática de naturaleza profesional en el que se sintetizan e integren las competencias adquiridas en las enseñanzas*) representa la culminación del proceso formativo. Se han integrado conocimientos técnicos en distintas áreas de la Ingeniería Informática, como la programación, el análisis de datos, la inteligencia artificial, y la gestión de proyectos, para abordar un problema de carácter científico en el ámbito oceanográfico.

3.3. Alineación con los objetivos de desarrollo sostenible

Los objetivos de desarrollo sostenible (ODS) son un conjunto de 17 medidas adoptadas para la protección del planeta, la lucha contra la pobreza y la erradicación de esta, con el fin de construir un mundo más próspero, justo y sostenible para las generaciones futuras. Los objetivos de desarrollo sostenible, atendiendo a su grado de relación con el presente Trabajo de Fin de Grado, están representadas en la siguiente tabla 3.1:

ODS	Grado de relación			
	0 No procede	1 Bajo	2 Medio	3 Alto
1 Fin de la Pobreza	X			
2 Hambre cero	X			
3 Salud y Bienestar	X			
4 Educación de calidad	X			
5 Igualdad de género	X			
6 Agua limpia y saneamiento	X			
7 Energía Asequible y no contaminante	X			
8 Trabajo decente y crecimiento económico	X			
9 Industria, Innovación e Infraestructuras	X			
10 Reducción de las desigualdades	X			
11 Ciudades y comunidades sostenibles	X			
12 Producción y consumo sostenibles	X			
13 Acción por el clima	X			
14 Vida submarina				X
15 Vida de ecosistemas terrestres				X
16 Paz, justicia e instituciones sólidas	X			
17 Alianzas para lograr objetivos	X			

Tabla 3.1: Grado de relación del TFT con los objetivos de desarrollo sostenible.

El objetivo de desarrollo sostenible central en este trabajo acontece en la vida submarina. El objetivo principal es desarrollar un gemelo digital basado en inteligencia artificial para

estimar la NPP en ecosistemas marinos, con especial foco en la región macaronesia, contribuyendo a la conservación y gestión sostenible de los océanos al proporcionar herramientas precisas para monitorear la salud de los ecosistemas fitoplanctónicos, que se sustenta como la base de la cadena trófica marina. Por ende, el proyecto está centrado en el entorno oceánico, y se emplean datos biogeoquímicos y modelos para simular procesos esenciales en los ecosistemas marinos, lo que contribuye a una mejor conservación, gestión sostenible de los océanos y comprensión de la biodiversidad marina, avanzando en la comprensión de la dinámica del fitoplancton y su papel en el secuestro de carbono. Esto permite evaluar impactos ambientales, como el calentamiento global y la acidificación oceánica, y diseñar estrategias para proteger la biodiversidad marina. La herramienta desarrollada facilita el acceso a información científica rigurosa, esencial para políticas de conservación y gestión pesquera sostenible.

Por otra parte, el objetivo de vida de ecosistemas terrestres fue seleccionado pues se analizar indirectamente los efectos del cambio climático en ecosistemas terrestres a través del estudio de interacciones océano-atmósfera, como el transporte de nutrientes y la regulación del ciclo del carbono. Aunque el trabajo se enfoca en ecosistemas marinos, los resultados tienen implicaciones indirectas para ecosistemas terrestres. Por ejemplo, la productividad fitoplanctónica influye en la generación de nubes y la deposición de nutrientes en zonas costeras, afectando a humedales y manglares. Además, el gemelo digital podría adaptarse en el futuro para estudiar sistemas terrestres con dinámicas similares, como lagos o bosques, ampliando su alcance científico.

Capítulo 4

Desarrollo

En este cuarto capítulo, se describe de manera detallada la metodología y las fases de desarrollo que han guiado la realización del proyecto. El trabajo se ha basado en el marco de trabajo CRISP-DM (*Cross-Industry Standard Process for Data Mining*), una metodología ampliamente reconocida en proyectos de ciencia de datos, que garantiza un enfoque estructurado y eficiente. Además, se han empleado herramientas de gestión ágil, como *Trello*, para organizar y supervisar las tareas.

El capítulo se divide en dos secciones principales. La primera de ellas explica los recursos tecnológicos y las herramientas utilizadas, desde lenguajes de programación como Python hasta bibliotecas especializadas para el análisis de datos. En la segunda, se detallan las fases de desarrollo, incluyendo la recopilación y procesamiento de datos, que culminarán en la implementación de un modelo explicativo y estimativo. El enfoque metodológico permitió abordar de manera sistemática y ordenada los objetivos planteados, asegurando de esta modo la reproducibilidad y validez científica del trabajo.

Mediante la lectura de este apartado, se documenta el trabajo técnico realizado y se evidencia la aplicación de prácticas innovadoras en el manejo de datos oceanográficos y la integración de técnicas de inteligencia artificial en un contexto interdisciplinario.

4.1. Metodología

4.1.1. Marco de trabajo

En el período anterior al comienzo del proyecto, se mantuvo un breve intervalo de discusión en torno a qué metodología seguir, que brindase los requerimientos organizativos y garantizara la eficacia para alcanzar los objetivos propuestos, garantizando la calidad del trabajo y la adaptabilidad frente a posibles imprevistos, cambios o refactorizaciones, todo ello dentro de un enfoque alineado a los principios de la ciencia de datos.

Finalmente, el marco de trabajo empleado en el proyecto se ha basado en CRISP-DM (*Cross-Industry Standard Process for Data Mining*) (ilustración 4.1), siendo una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos, puesto en marcha bajo el financiamiento del Programa de Investigación y Desarrollo en Tecnologías de Información de la Unión Europea (*European Strategic Programme for Research in Information Technology*, ESPRIT) en 1997. Actualmente, IBM es la principal empresa que promueve el uso de esta metodología [23].

CRISP-DM consta de seis etapas fundamentales, descritas genéricamente a continuación:

1. **Comprensión del problema o negocio**, en la que se evalúa la situación de para obtener una visión general de los recursos disponibles y necesarios. Determinar el objetivo de la minería de datos es uno de los aspectos más importantes de esta fase. Primero, se debe explicar el tipo de minería de datos (p. ej., la clasificación) y los criterios de éxito (como la precisión). Se debe crear un plan de proyecto obligatorio.
2. **Comprensión de los datos**, en la que se recopilan, exploran y describen los datos de las fuentes, y se verifica su calidad.
3. **Preparación de los datos**, donde la selección de datos debe realizarse definiendo los criterios de inclusión y exclusión. La mala calidad de los datos puede solucionarse con la limpieza y acondicionamiento de los mismos, mediante eliminación de *outliers* e imputación de valores faltantes. En función del modelo utilizado, definido en la primera fase, se deben construir los atributos derivados. Para todos estos pasos, existen diferentes métodos que dependen del modelo.
4. **Modelado**, que consiste en seleccionar la técnica de modelado, construir el caso de prueba y el modelo. Se pueden utilizar todas las técnicas de minería de datos. En general, la elección depende del problema de negocio y de los datos. Lo más importante es cómo explicar la elección. Para construir el modelo, se deben establecer parámetros específicos. Para evaluar el modelo, es conveniente evaluarlo con respecto a los criterios de evaluación y seleccionar los mejores.
5. **Evaluación del modelo**, en la que los resultados se comparan con los objetivos de negocio definidos. Por lo tanto, es necesario interpretarlos y definir acciones futuras. Además, es necesario revisar el proceso en general.
6. **Implementación del modelo**, que puede tratarse de un informe final o de un componente de software. La guía del usuario describe que la fase de implementación consiste en la planificación, la supervisión y el mantenimiento de la implementación.

4.1.2. Herramienta de gestión de procesos

Para la organización y el desarrollo de las tareas, se manejaron los servicios ofrecidos por *Trello*, un software de gestión de proyectos en línea para el trabajo de forma colaborativa basado en la metodología *Kanban*, que exige a los usuarios utilizar tarjetas de trabajo en un



Ilustración 4.1: Diagrama del proceso *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*), que describe las fases estándar en proyectos de minería de datos. **Fuente:** IBM *SPSS Modeler Documentation*.

tablero, llevando de este modo una línea de producción (*Product Backlog*) de tareas con sus estados correspondientes (ilustración 4.2).

Esta herramienta ha sido fundamental para facilitar la organización y la colaboración entre las partes involucradas. Su elección responde al objetivo de integrar herramientas propias de la Ingeniería Informática en el campo de la Oceanografía. En el ámbito científico, no es habitual el uso de herramientas de gestión de proyectos, a diferencia de lo que ocurre en informática, donde su uso resulta imprescindible debido a la complejidad y cantidad de tareas que se manejan. Por esta razón, y como parte de una prueba de integración de nuevas metodologías, se optó por implementar este software.

4.1.3. Recursos y tecnologías empleadas

Respecto a los recursos empleados, el desarrollo del trabajo ha requerido de la integración de diversas herramientas y tecnologías para facilitar las técnicas de adquisición, procesamiento, análisis, visualización y entendimiento de los datos, así como la documentación y la colaboración con los tutores.

La mayor parte de datos oceanográficos y de ámbito esencialmente científico utilizados se encontraron en formato NetCDF (*Network Common Data Form*), que se define como un formato estándar de abstracción para almacenar y recuperar datos multidimensionales. NetCDF se distribuye como una biblioteca de software que proporciona un formato independiente de la máquina para representar datos científicos. En conjunto, la abstracción, la biblioteca y el

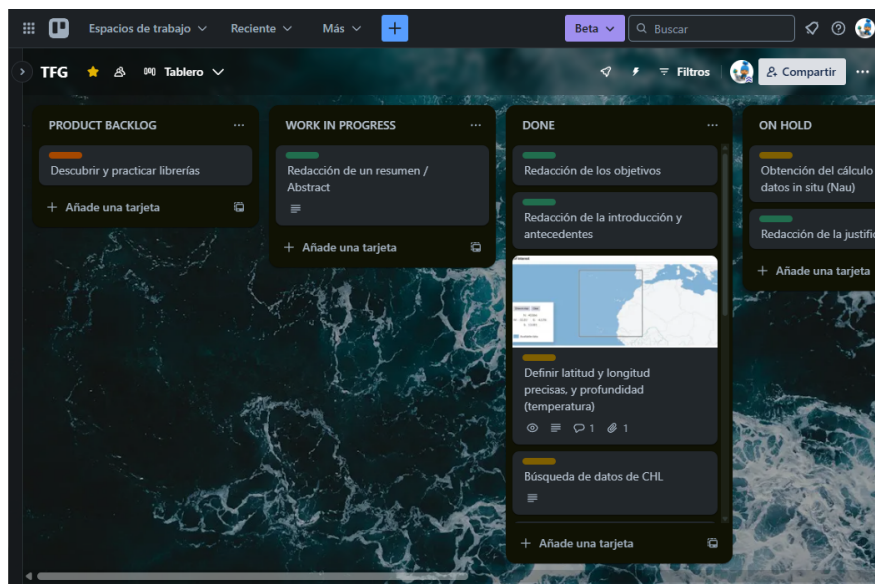


Ilustración 4.2: Tablero de *Trello* utilizado para la organización y seguimiento de tareas en el proyecto. Muestra las columnas *Product Backlog*, *Work in Progress*, *Done*, *On Hold* y *Up Next*, con tarjetas que incluyen actividades como la redacción del resumen, búsqueda de datos de clorofila-a (CHL) y definición de parámetros geográficos. **Fuente:** elaboración propia mediante *Trello*.

formato de datos facilitan la creación, el acceso y el intercambio de información científica. Resulta útil para admitir objetos que contienen diferentes tipos de datos, en un entorno de red heterogéneo, y para el desarrollo de software que no depende de formatos específicos. El análisis abarca la abstracción e interfaz de datos de NetCDF, así como dimensiones, variables y atributos, y extensas bibliotecas y utilidades asociadas a NetCDF, como `ncdump` y `cdo` [56].

`ncdump` es una utilidad en línea de comandos que convierte los datos NetCDF en formato de texto legible, usando el lenguaje CDL (*Common Data form Language*) [72], desarrollada por el *Unidata Program Center*. AtaÑe principalmente al proyecto la opción `ncdump -h`, que permite observar solo información del encabezado, sus dimensiones, variables y atributos, sin datos, a fin de verificar la integridad y el contenido de los archivos de esta extensión mediante herramientas poderosas.

Por otro lado, `cdo` (*Climate Data Operators*) es un conjunto de más de 600 operadores en línea de comandos, desarrollados por el *Max Planck Institute for Meteorology*, para el procesamiento estándar de datos de modelos climáticos y de modelos de predicción numérica del tiempo, soportando el formato NetCDF. Los operadores incluyen funciones estadísticas y aritméticas sencillas, herramientas de selección y submuestreo de datos, regrado, transformaciones espectrales, e interpolación espacial, lo que facilita notablemente el procesamiento de grandes volúmenes de datos científicos [63].

Adicionalmente, y de manera puntual, se han utilizado las herramientas del *HDF-EOS Tools and Information Center*, en particular el programa `h4tonccf_nc4`, que permite convertir archivos en formato HDF (*Hierarchical Data Format*), un formato de archivo de datos

para ayudar a los usuarios en el almacenamiento y manipulación de datos científicos [46], al formato común NetCDF [26], permitiendo así su posterior análisis con herramientas como `ncdump`, `cdo` o bibliotecas científicas en Python.

Todas estas utilidades han sido fundamentales en el análisis de los datos, proporcionando funcionalidades rápida y eficazmente mediante línea de comandos.

Asimismo, la alternativa de software empleada para la visualización preliminar y el estudio ágil de los datos es Panoply (ilustración 4.3), desarrollada por la *NASA Goddard GISS*. Se trata de una aplicación multiplataforma compatible con la mayoría de ordenadores de escritorio, y se precisa de un JDK compatible con Java 11 o una versión posterior. El programa se encarga de trazar matrices georreferenciadas y de otros tipos a partir de formatos NetCDF, incluido HDF y otros conjuntos de datos, crear gráficos de contorno a color de dimensiones de latitud y longitud, y dibujar datos globales en un mapa mundial o regional utilizando más de 200 proyecciones cartográficas (ilustración 4.4). También permite acceder a conjunto de datos remotos almacenados en un sitio web o en un bucket, y guardar todas las creaciones como imágenes de mapa de bits GIF, JPEG, PNG o TIFF en el disco, o como archivos gráficos PDF, PostScript o SVG [49].

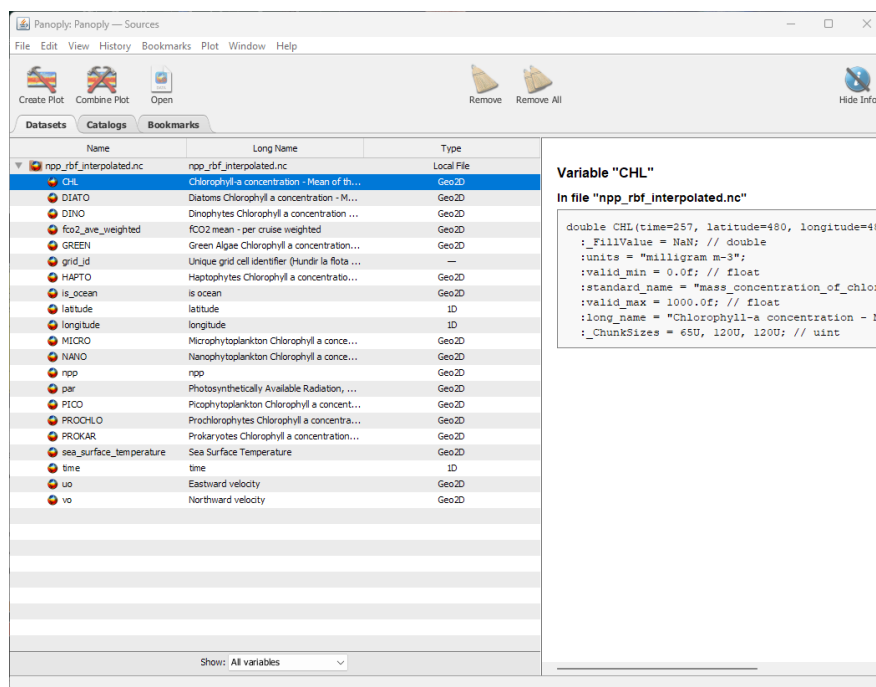


Ilustración 4.3: Interfaz del software *Panoply* mostrando la lista de variables disponibles en un archivo NetCDF, incluyendo concentraciones de clorofila-a y otros parámetros oceanográficos. **Fuente:** elaboración propia mediante *NASA Panoply NetCDF Data Viewer*.

El principal editor de código utilizado durante el desarrollo del proyecto ha sido Visual Studio Code, que se presenta como el IDE (*Integrated Development Environment*) estándar y más usado por otros desarrolladores, ya que ofrece versatilidad y una amplia gama de extensiones y soporte para múltiples lenguajes de programación. Ocasionalmente, se ha hecho uso de PyCharm para la creación de scripts sencillos para la extracción de datos.

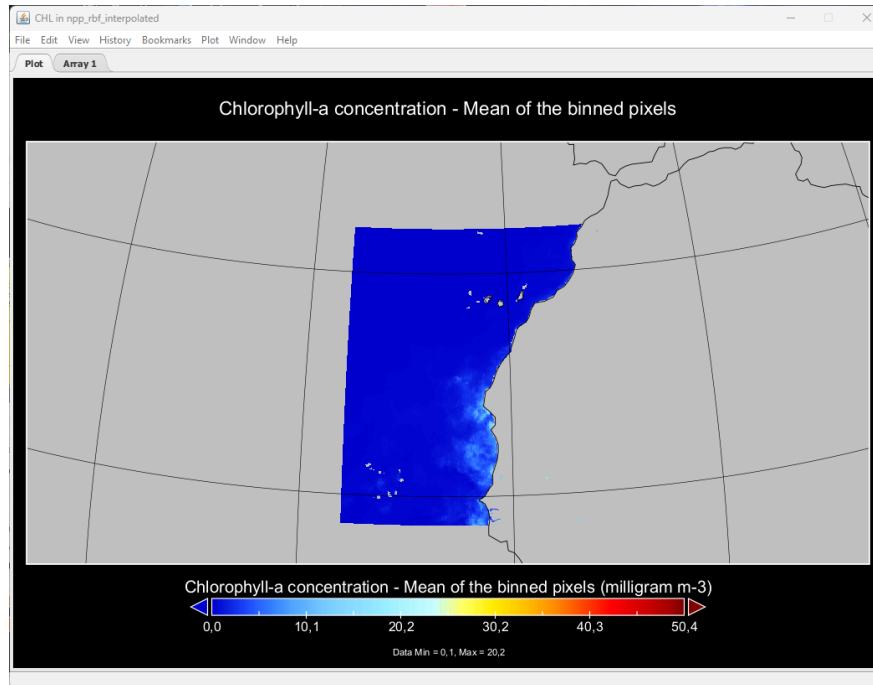


Ilustración 4.4: Mapa generado en *Panoply* que muestra la concentración media de clorofila-a (en mg m^{-3}) basada en datos binados. Los valores oscilan entre 0,1 y 20,2 mg m^{-3} . **Fuente:** elaboración propia mediante *NASA Panoply NetCDF Data Viewer*.

El uso de cuadernos de Jupyter Notebook (con la extensión .ipynb) ha permitido combinar código en lenguaje de programación Python con *Data Visualization* y texto explicativo en un solo documento, facilitando la documentación y reproducibilidad de los análisis.

Eventualmente, se sirvió un cuaderno de Google Colab (ilustración 4.5), con el que se facilitó la colaboración y el acceso compartido a los avances del proyecto de los integrantes del mismo, permitiendo la ejecución de las celdas de código de los cuaderno en la nube, sin la necesidad de una configuración local.

De igual manera, por medio de otro servicio de Google, *Google Cloud Storage* (GCS), se utilizó un espacio para almacenar y compartir los conjuntos de datos de forma eficiente, mediante la creación de un *bucket* (ilustración 4.6). GCS permite el acceso a los datos directamente desde la nube, facilitando su descarga y procesamiento sin necesidad de almacenamiento local, lo que resultó útil para el análisis de un gran volumen de datos como el planteado.

El lenguaje de programación escogido fue Python, debido a su amplia aplicación en estudios científicos y la disponibilidad de bibliotecas especializadas. Se considera uno de los lenguajes de programación más populares para la ciencia y el análisis de datos. Gracias a la comunidad de desarrolladores y de código abierto, se han desarrollado numerosas bibliotecas útiles para la computación científica y el aprendizaje automático [55].

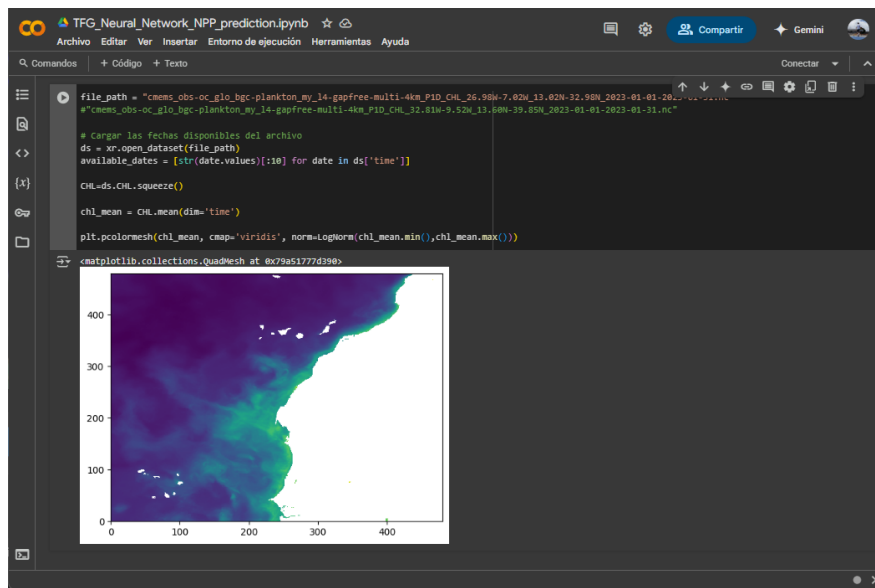


Ilustración 4.5: Entorno de trabajo en *Google Colab* utilizado para el procesamiento de datos de clorofila-a (CHL)). Muestra la carga de archivos NetCDF, manipulación con *xarray*, y visualización mediante *matplotlib*. **Fuente:** elaboración mediante *Google Colab*.

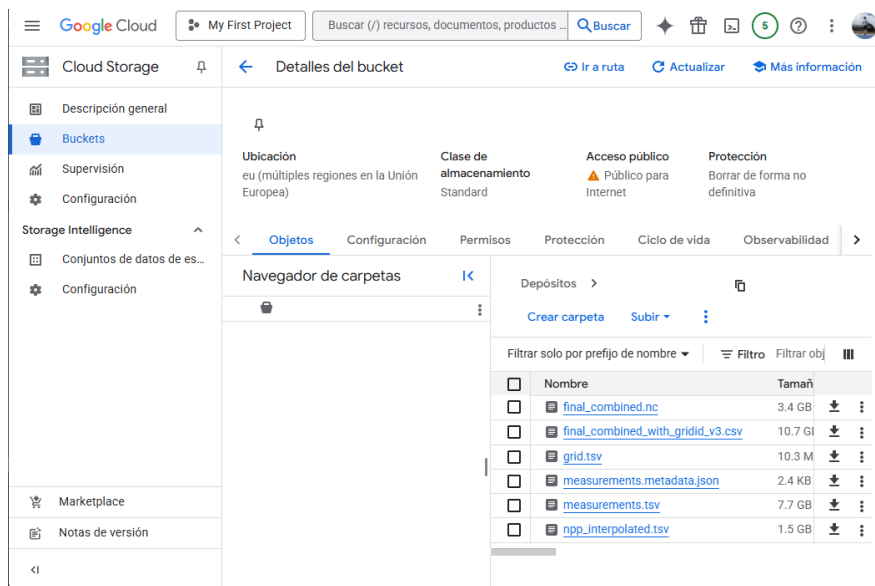


Ilustración 4.6: Interfaz de *Google Cloud Storage* mostrando la organización de datos en buckets, incluyendo archivos NetCDF y CSV utilizados para el análisis de productividad primaria neta (NPP). Destaca la visualización de metadatos, permisos de acceso y tamaños de archivos (ej: `final_combined.nc` con 3.4 GB). **Fuente:** elaboración mediante *Google Cloud Platform*.

4.1.4. Bibliotecas requeridas

Las bibliotecas empleadas para la modificación y la extracción de datos fueron **xarray**, **numpy**, y **pandas**. La primera de ellas es utilizada para trabajar con estructuras de datos para matrices dimensionales. En cambio, **numpy** y **pandas** proporcionan soluciones para operaciones numéricas y manipulación de las estructuras de datos, respectivamente. Además, **netCDF4** es la librería que ofrece la principal interfaz utilizada, siendo también utilizada [31].

Para los conjuntos de datos de clorofila, los tipos de fitoplancton y las corrientes marinas, se ha hecho uso de **copernicusmarine**, una herramienta de fácil manejo que interopera con *Copernicus Marine Data Store* con la intención de recuperar metadatos, conjuntos de datos completos o subconjuntos, para cualquier tipo de producto, sean modelos numéricos, observaciones satelitales o mediciones de datos in situ [14]. Esta librería ofrece la posibilidad de descargar perfiles únicamente con el identificador del conjunto, el lapso de tiempo y las coordenadas de latitud y longitud deseadas.

Para la descarga directa de grandes conjuntos de datos individualizados o separados, se ha encontrado solución en **selenium** y **webdriver_manager** (ilustración 4.7), para la automatización de la descarga de datos desde sitios web que requieren interacción dinámica. Tal es el caso de los datos de la variable radiación fotosintéticamente activa, del que se recuperaron enlaces para las descargas de los datasets desde *NASA Earthdata*. Para optimizar el proceso, se llevó a cabo una mecanización mediante un script.

```
# Configuración
USER =          # usuario NASA
PASS =          # passwd
DOWNLOAD_DIR = r"C:\Users\Imano\Desktop\4o CARRERA\TFG\Datos\Irradiancia\monthly_data" # carpeta de descarga
URLS_FILE = r"C:\Users\Imano\Desktop\4o CARRERA\TFG\Datos\Irradiancia\enlaces.txt" # archivo con URLs

# configuración del driver
chrome_options = webdriver.ChromeOptions()
prefs = {
    "download.default_directory": DOWNLOAD_DIR,
    "download.prompt_for_download": False,
    "download.directory_upgrade": True,
    "safebrowsing.enabled": True
}
chrome_options.add_experimental_option("prefs", prefs)

driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()), options=chrome_options)

driver.get("https://urs.earthdata.nasa.gov")
WebDriverWait(driver, 10).until(EC.presence_of_element_located((By.ID, "username"))).send_keys(USER)
driver.find_element(By.ID, "password").send_keys(PASS)
driver.find_element(By.NAME, "commit").click()
time.sleep(3)

with open(URLS_FILE, "r") as f:
    urls = [line.strip() for line in f.readlines() if line.strip()]

for url in urls:
    try:
        filename = url.split("/")[-1]
        if os.path.exists(os.path.join(DOWNLOAD_DIR, filename)):
            print(f"✓ {filename} ya existe")
            continue

        print(f"Descargando {filename}...")
        driver.get(url)
```

Ilustración 4.7: Script de *Python* utilizando *Selenium WebDriver* para automatizar la descarga de datos de radiación fotosintéticamente activa desde el portal *NASA Earthdata*. Incluye la configuración del navegador, inicio de sesión mediante credenciales, y gestión de descargas desde URLs almacenadas en un archivo. **Fuente:** elaboración propia.

cftime tuvo una utilidad clave para el manejo y gestión de fechas y tiempos de calendarios

no estándar de archivos NetCDF.

Adicionalmente, a fin de visualizar los datos, poder graficar las observaciones y variables de los conjuntos, se usó `matplotlib` y `seaborn`. La visualización es esencial para comprender la dinámica de los datos, obtener información sobre su estructura, e ideas sobre las transformaciones adecuadas para aplicar al conjunto de datos. `matplotlib` es fácilmente compatible con NumPy, Pandas y otras bibliotecas relevantes, mientras `seaborn` se sitúa como una biblioteca de visualización gráfica basada en las configuraciones principales de la primera, que ofrece a los usuarios acceso a algunos de los procesos de visualización de datos más comunes, con ciertas necesidades de visualización, como la asignación de color a una variable [64].

Cabe destacar `scipy`, una librería para realizar interpolaciones espaciales. Mediante `RBF Interpolator`, se implementó un método de interpolación basado en funciones de base radial (RBF, *Radial Bases Functions*). Esta herramienta ofrece ventajas significativas, y se trata de una interpolación suave por definición, aplicable en campos escalares y vectoriales [68].

Para el desarrollo del modelo explicativo, la librería `scikit-learn` (`sklearn`) se ha consolidado como una herramienta fundamental, al ofrecer un amplio conjunto de algoritmos de aprendizaje supervisado y no supervisado, así como utilidades para la selección de características, evaluación de modelos y procesamiento de datos. En particular, se ha empleado para construir modelos interpretables que permitan analizar la importancia de las variables y su relación con la variable objetivo, contribuyendo así a una mejor comprensión del sistema estudiado y a una preparación más precisa de los datos para las fases posteriores del proyecto.

4.1.5. Edición de la memoria final

Por último, para las fases finales y la redacción de la memoria de este Trabajo de Fin de Grado, se dispuso de LaTeX como sistema de composición de textos, pues es ampliamente utilizado en la creación de documentos del ámbito científico y técnico. La plataforma de edición seleccionada fue Overleaf, pues ofrece un servicio en línea esencial para la edición colaborativa de documentos en LaTeX.

4.2. Fases de desarrollo

4.2.1. Metodología, etapas y flujo de trabajo

Tal como se señaló previamente, el marco de trabajo empleado en el proyecto se ha basado en la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), una de las más empleadas para el desarrollo de proyectos de minería de datos.

El proceso se dividió en seis etapas diferenciadas (ilustración 4.8) que permitieron una secuencia de fases no estricta gracias al carácter flexible y personalizable de CRISP-DM. En este trabajo, se avanzó y retrocedió entre las distintas fases de forma iterativa, realizando cambios y refactorizaciones siempre que fue necesario:

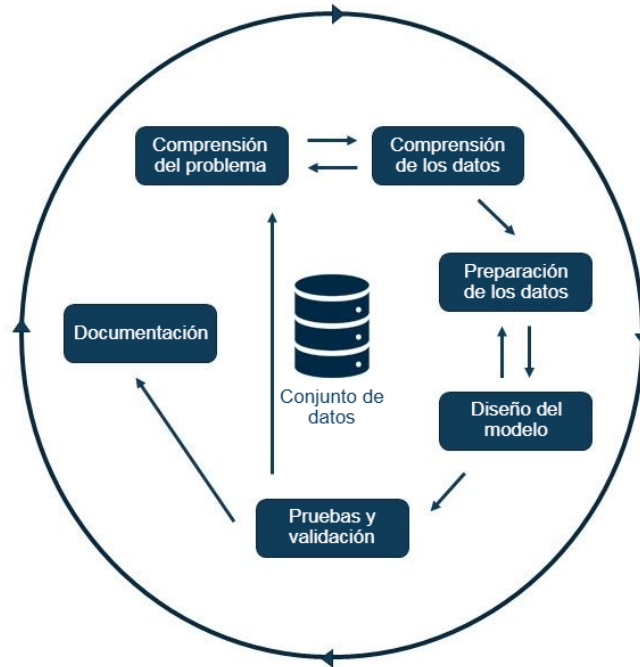


Ilustración 4.8: Diagrama del proceso *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) adaptado, mostrando las fases clave: **Comprensión del problema** (documentación, datos y validación) y **Comprensión de los datos** (preparación, modelado y evaluación). **Fuente:** adaptado de *IBM SPSS Modeler Documentation*.

Las etapas específicas fueron:

1. **Comprensión del problema**, en la que se evaluaron las variables que influyen en la variabilidad de la producción primaria neta, lo que permitió obtener una visión más detallada del fenómeno. Para ello, se priorizó la revisión de artículos científicos y explicaciones relacionadas con los procesos tróficos implicados.
2. **Comprensión de los datos**, donde se recopilaron y exploraron datos procedentes de distintas fuentes, se verificó su calidad y se analizó su relevancia en relación con el fenómeno principal. Asimismo, se definió un plan de proyecto que incluyó los pasos a seguir, las fuentes de datos definitivas y, finalmente, el procesamiento necesario de la información recopilada.
3. **Preparación de los datos**, donde se seleccionaron y descargaron datos relevantes sobre clorofila, fitoplancton, temperatura superficial, corrientes marinas, radiación fotosintéticamente activa y producción primaria. Se definieron y analizaron los criterios y resoluciones temporales y espaciales de cada conjunto de datos. Posteriormente, se llevó a cabo un proceso de acondicionamiento que incluyó la eliminación de *outliers* y datos espurios, la aplicación de técnicas de interpolación y extrapolación, y la redefinición de mallas espaciales con el apoyo de Python y sus librerías especializadas. Además, se construyó un nuevo atributo denominado `tile`, definido como una combinación alfanumérica de las coordenadas de longitud y latitud correspondientes a la malla regional. Paralelamente, se inició la documentación detallada de los pasos realizados utilizando

LaTeX, con el objetivo de facilitar la fase final del proyecto. Finalmente, se elaboró un modelo explicativo con el propósito de describir las características del entorno y las relaciones entre las variables, lo cual permitió orientar mejor el proceso de preparación de los datos.

4. **Diseño del modelo**, en el que se seleccionó la técnica de modelado más adecuada en función del problema y las características de los datos disponibles, a través de la herramienta de modelado de gemelos digitales Picota. Se construyó un modelo explicativo para la interpretación de variables e interrelaciones entre perfiles, y se definió el modelo y los parámetros necesarios para optimizar el rendimiento del sujeto del modelo. La elección metodológica se justificó teniendo en cuenta tanto los objetivos del análisis como la capacidad explicativa del modelo sobre el fenómeno observado.
5. **Pruebas y validación**, donde se evaluaron los resultados obtenidos frente a los objetivos definidos inicialmente, interpretando el comportamiento del modelo y su ajuste a la realidad del fenómeno estudiado. Esta fase incluyó la validación con métricas, la identificación de posibles desviaciones y la propuesta de mejoras o ajustes futuros. Además, se revisó el proceso completo para asegurar su coherencia y calidad metodológica.
6. **Documentación** de todo el proceso realizado, bajo la estructura de la presente memoria del Trabajo de Fin de Grado. Se expone en la misma la problemática que se pretende estudiar, los fundamentos teóricos de la misma, la solución desarrollada, los resultados de la misma, y las metodologías desplegadas.

4.2.2. Implementación y desarrollo técnico por etapas

Después de concretar cada uno de los aspectos relevantes del tema, los objetivos y los contenidos del presente Trabajo de Fin de Grado, se procedió a una fase de comprensión profunda del problema. Para ello, se revisaron múltiples artículos que abordaban la producción primaria neta oceánica como eje principal, prestando especial atención a las metodologías de medición y a las incertidumbres asociadas a las mediciones in situ. Esta revisión de distintas fuentes bibliográficas permitió identificar las limitaciones de las mediciones tradicionales y la necesidad de enfoques complementarios para estimar la NPP de manera más precisa.

Asimismo, se exploró el uso de redes neuronales en la estimación de la NPP, con el objetivo de comprender el contexto actual y las ventajas que ofrecen estos modelos. Estudios como el de Belgrano (Belgrano et al., 2001 [7]) aplicaron redes neuronales artificiales para predecir la productividad primaria a partir de series temporales de datos ambientales, demostrando una mejora en la precisión respecto a métodos tradicionales. Asimismo, investigaciones más recientes han empleado algoritmos de aprendizaje automático, como Random Forest (Yu et al., 2019 [78]), para escalar observaciones de campo y estimar la producción primaria bruta a nivel global. Estos enfoques proporcionan una base sólida para el desarrollo de modelos de estimación de NPP precisos y adaptados a las condiciones de la región de estudio.

La etapa de comprensión de los datos estuvo estrechamente ligada a la fase posterior de preparación. Se seleccionaron datos y perfiles provenientes de diversas fuentes y bases de

datos, incluyendo modelos y campañas oceanográficas. Estos conjuntos de datos abarcaban diferentes variables oceanográficas relevantes para la estimación de la producción primaria, como la temperatura, la concentración de clorofila-a, los tipos de fitoplancton y la radiación fotosintéticamente activa. La integración de estos datos requirió un análisis de la calidad, resolución temporal y espacial, y cobertura geográfica de la información de la que se disponía.

En ocasiones, fue necesario realizar un procesamiento previo de los datos para permitir comparaciones lógicas entre los distintos conjuntos. Esto incluyó la armonización de las resoluciones temporales y espaciales, la interpolación de datos faltantes y la definición de nuevas mallas para evitar descompensaciones entre los datasets. Además, se aplicaron técnicas de normalización y estandarización para garantizar la coherencia entre las diferentes variables y facilitar su uso en modelos de estimación de la NPP.

A continuación, se implementó un modelo explicativo con el objetivo de obtener una comprensión más profunda de las características del entorno, tanto biológicas como ambientales, y analizar las relaciones causales entre los distintos perfiles y variables. Este modelo permitió identificar patrones y tendencias, así como explorar las posibles causas subyacentes mediante técnicas de interpretabilidad y análisis causal. Aspectos como las correlaciones, la multicolinealidad y los coeficientes asociados resultaron fundamentales para guiar la preparación del conjunto de datos, favoreciendo una selección de variables más coherente, robusta y alineada con los objetivos del estudio.

Durante la fase de preparación de los datos, se seleccionaron y descargaron los conjuntos de datos definitivos para la implementación de la red neuronal artificial. Se llevó a cabo un procesamiento y estandarización de los datos mediante acciones técnicas específicas, con el objetivo de convertirlos en formatos adecuados para su cohesión y uso en la red neuronal.

El ciclo general de procesamiento de los conjuntos de datos fue el siguiente:

1. **Descarga y extracción del conjunto de datos** objetivo a través de las distintas fuentes.
2. **Conversión del fichero resultante a formato NetCDF** si fuera necesario. Muchos ficheros se encontraban separados y en otros formatos, como HDF.
3. **Selección de la resolución espacial y temporal** adecuada, abarcando desde el año 2002 hasta diciembre de 2023, en el rango de longitudes 27° oeste a 7° oeste, y latitudes 13° norte a 33° norte.
4. **Definición de una nueva malla** con una resolución espacial de 480 mediciones por lado.
5. **Redefinición de las variables y atributos**, según fuera necesario.
6. **Interpolación y extrapolación** de los datos faltantes tras la definición de la nueva malla, según fuera necesario.
7. **Comprobación y validación** de la salida a través de herramientas de visualización como Panoply.

Las variables extraídas y los conjuntos de datos empleados en este estudio provienen de diversas fuentes reconocidas por su calidad y cobertura global:

- ✓ **Los perfiles de producción primaria neta** se obtuvieron del banco de datos de productividad oceánica de la Universidad Estatal de Oregon [51]. En particular, se utilizaron estimaciones del modelo Eppley-VGPM, que calcula la NPP diaria máxima en una columna de agua basada en la temperatura y la concentración de clorofila-a. Los datos provienen de tres sensores satelitales distintos: MODIS (2002–2012), SeaWiFS (1997–2002) y VIIRS (desde 2012).
- ✓ **Los perfiles de clorofila-a y tipos de fitoplancton** se obtuvieron a través de productos satelitales del programa Copernicus, específicamente del producto identificado como `cmems_obs-oc_glo_bgc-plankton_my_l4-multi-4km_P1M` [16]. Este producto proporciona todas las variables biogeoquímicas del océano de interés, con una resolución espacial de 4 kilómetros y temporal mensual. La extracción de datos se realizó utilizando la librería de Python `copernicusmarine`, que facilita el acceso y descarga de datos marinos de Copernicus.
- ✓ **Los perfiles de temperatura** fueron conseguidos desde el producto satelital de Copernicus, identificado como `SST_GLO_SST_L4_REP_OBSERVATIONS_010_011` [15]. Estos datos fueron posteriormente procesados para ajustarse a los requerimientos de resolución espacial y temporal del estudio.
- ✓ **Los perfiles de radiación fotosintéticamente activa** se encontraron en las bases de datos de NASA Earthdata, bajo el nombre de *Solar Irradiance Data (Monthly)* [47]. Se desarrolló un script personalizado para automatizar la descarga y consolidación de los archivos necesarios.
- ✓ **Los perfiles de dióxido de carbono** se obtuvieron desde el sitio web de SOCAT (*The Surface Ocean CO₂ Atlas*) [69], un proyecto de síntesis que compila observaciones de fCO_2 (fugacidad de dióxido de carbono) de la superficie del océano, realizadas por la comunidad internacional de investigación sobre el carbono marino. Estos datos se recopilan a lo largo de transectos oceánicos, proporcionando información detallada sobre la variabilidad espacial y temporal del fCO_2 en la superficie marina.
- ✓ **Los perfiles de corrientes marinas** fueron extraídos del programa Copernicus, bajo el identificador `cmems_mod_glo_phy_my_0.083deg_P1M-m` [17]. Este producto ofrece componentes de velocidad de las corrientes oceánicas hacia el norte y el este, con una resolución espacial de 0.083° y temporal mensual.

La recolección de todos los datos comentados implicó un procesamiento sistemático, que incluyó la automatización en la normalización, optimización y mejora de la información. Para abordar estos desafíos, y tras un uso continuado de cuadernos Jupyter, se diseñó e implementó una biblioteca propia con funciones fundamentales orientadas a la integración y tratamiento de datos provenientes de múltiples fuentes, estructurada según procesos ETL (por sus siglas en inglés, *Extract, Transform, Load*) (Kimball et al., 2004 [32]).

El proceso ETL se estructura en tres fases principales:

1. **Extracción**, donde se recopilan datos de diversas fuentes, que pueden incluir bases de datos, archivos planos, APIs o sistemas *legacy*. Esta fase se encarga de obtener la información en su formato original sin alteraciones.
2. **Transformación**, en la cual se aplican una serie de operaciones para limpiar, validar, convertir y estandarizar los datos según reglas de negocio. La transformación es esencial para garantizar que la información sea coherente y compatible, facilitando su análisis posterior.
3. **Carga**, en la que, una vez transformados, los datos se integran en un sistema de destino, que puede ser un almacén de datos, un *data lake* u otro repositorio especializado. Esto permite centralizar la información y hacerla accesible para análisis, reportes y toma de decisiones.

Se modularizaron además las herramientas utilizadas en cada una de estas etapas, con el fin de mejorar la eficiencia y mantener la flexibilidad del sistema. Esta arquitectura modular permitió resolver de forma consistente los problemas recurrentes identificados en distintos conjuntos de datos y también facilitó la reutilización y extensión del proceso ETL en futuros proyectos. Todo el código y las utilidades desarrolladas han sido centralizadas en un repositorio de GitHub, lo cual puede ser de gran utilidad como referencia o para el empleo en nuevas implementaciones (repositorio [54]).

Tras la extracción y consolidación de todos los conjuntos de datos, se procedió a un análisis de correlación entre las variables consideradas en el estudio, tomando como variable objetivo la producción primaria neta oceánica (**bio.npp**).

La ilustración 4.9 muestra una matriz de correlación de Pearson, ordenada de forma descendiente según el grado de correlación con la NPP. Esta visualización permitió identificar qué variables están más estrechamente relacionadas con la productividad primaria oceánica, con el fin de contar o prescindir de algunas de ellas.

Posteriormente se examinó la integridad de los datos, verificando la existencia de valores ausentes y la posible necesidad de aplicar métodos de interpolación o extrapolación (ilustración 4.10):

Los resultados del análisis de cobertura de los datos revelaron heterogeneidad en la cobertura de las variables. **env.par** (variable de radiación fotosintéticamente activa) mostró la mayor completitud ($\sim 100\%$), seguida de **env.sst** y corrientes oceánicas. **bio.chl** presenta datos casi completos en su totalidad, y la variable objetivo, **bio.npp**, contiene una importante nulidad de sus datos. En contraste, **env.fco2** ($\sim 60\%$ de datos faltantes) y **bio.phyto.diato** para las diatomeas ($\sim 40\%$) presentaron las mayores carencias, lo que sugiere la necesidad de aplicar técnicas de imputación para estas variables.

La distribución por filas indicó que cerca del 60% de los registros tenían una completitud superior al 90% , mientras que menos de un 10% mostraba valores inferiores al 50% . Este patrón, unido a la gran cantidad de gaps en la variable objetivo **bio.npp**, que debería encontrarse en su completitud, respalda la viabilidad de emplear métodos de interpolación para los casos con gaps moderados, mientras que los datos con alta fragmentación (**env.fco2**) podrían requerir enfoques alternativos, como modelos basados en variables correlacionadas.

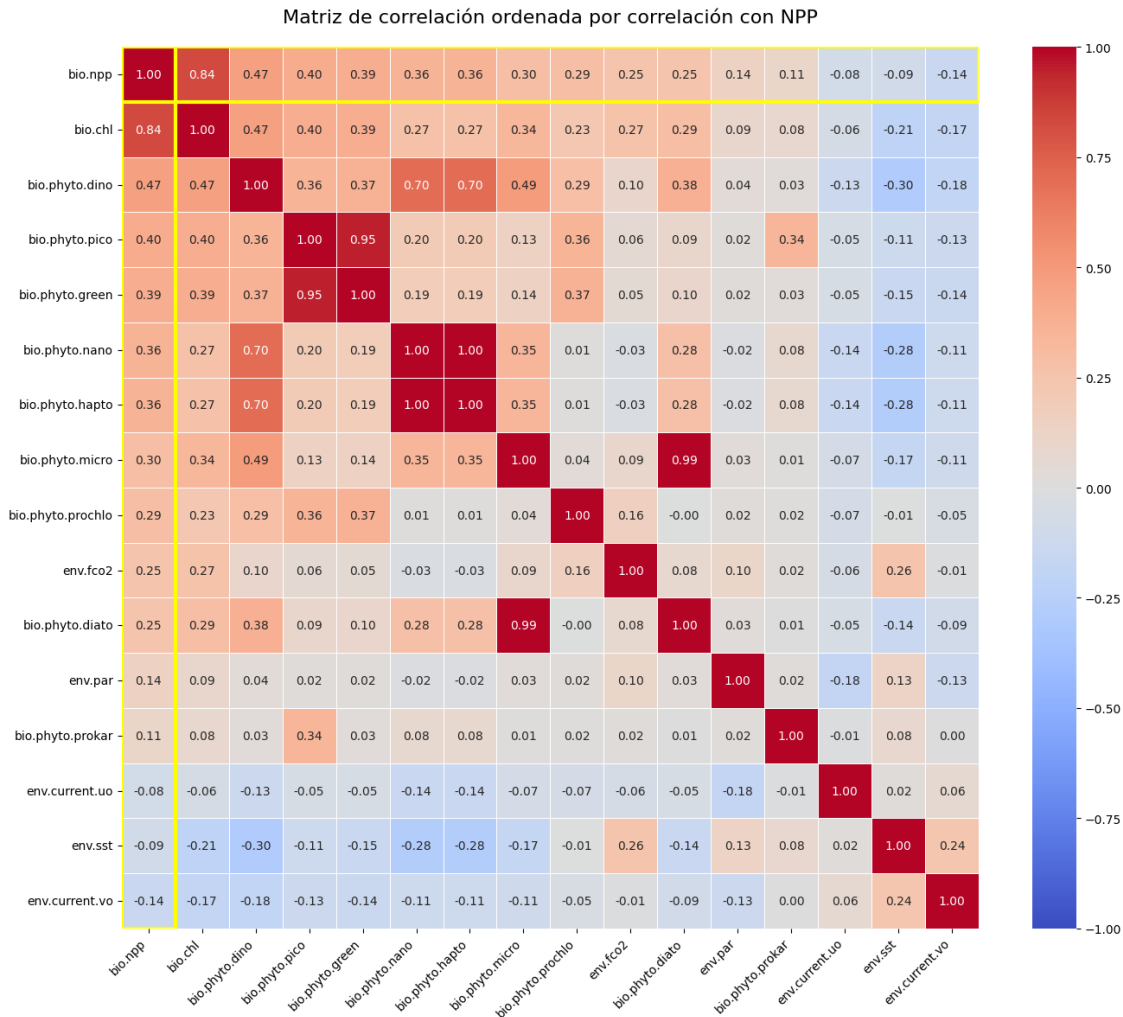


Ilustración 4.9: Matriz de correlación ordenada por su relación con la productividad primaria neta (NPP), mostrando las interacciones entre variables biogeoquímicas y ambientales. Destacan: **(1)** Alta correlación NPP-CHL ($r = 0,84$), **(2)** Relaciones fuertes entre grupos fitoplanctónicos (ej. picoplancton-nanoplancton $r = 0,95$), y **(3)** Correlaciones negativas con SST (temperatura superficial). Los valores en rojo indican anticorrelaciones significativas ($r < -0,1$). **Fuente:** elaboración con datos propios.

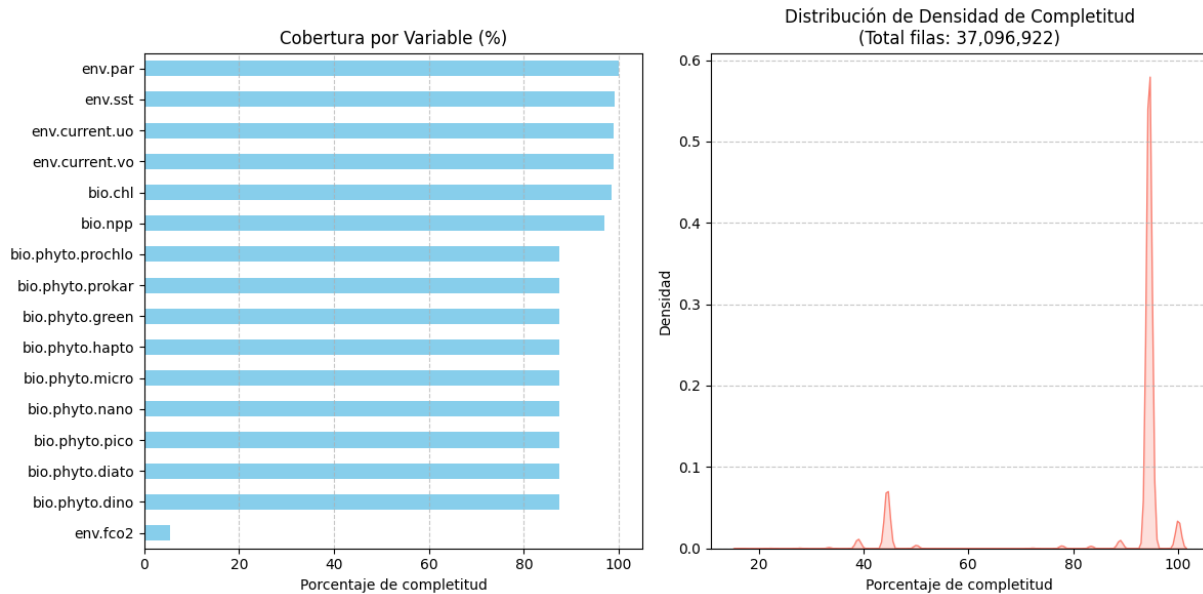


Ilustración 4.10: Evaluación de la cobertura de datos para 16 variables oceanográficas (2002-2023): **(A)** Porcentaje de datos disponibles por variable (PAR, SST, UO, VO, etc.), **(B)** Distribución de la cobertura por registros individuales. Destacan gaps en **bio.npp** ($\sim 97,02\%$), **bio.chl** ($\sim 98,50\%$), fitoplancton ($\sim 87,45\%$) y **env.fco2** ($\sim 5,51\%$), frente a **env.par** ($\sim 100\%$). **Fuente:** elaboración con datos propios.

Mediante el empleo de cuadernos de Jupyter, el trabajo procedimental resultó en un enfoque trivial. La etapa más profunda fue la interpolación de datos faltantes. En ocasiones, al definir la nueva malla, interpolar los datos mediante una interpolación lineal, y visualizar los perfiles en la herramienta Panoply, resultaba evidente que varias mediciones por unidad de longitud y latitud, eminentemente en los extremos de la región que se pretendía ahondar, presentaban datos nulos (vistos como *NaN*).

El problema reside en que la interpolación lineal en grillas latitud/longitud requiere que al menos dos puntos vecinos tengan datos válidos (es decir, no nulos o *NaN*) para poder calcular el valor interpolado. Si el punto que se intenta interpolar está rodeado por valores nulos (incluso parcialmente rodeado), entonces no se puede calcular una línea entre esos puntos y se conserva como *NaN*.

Para solucionar esta cuestión, se procedió al replanteamiento del problema, y se blindó la interpolación con una operación de vecino más cercano (*nearest*).

La interpolación por el vecino más cercano es esencialmente más veloz y sencilla, pues busca el punto válido más cercano (no *NaN*) y le asigna su valor. No necesita ningún tipo de continuidad o varios vecinos, solo uno. Por eso puede rellenar huecos donde la interpolación lineal falla.

A pesar de su velocidad y simplicidad, debido a una técnica de remuestreo más rápida y sencilla, con cálculos mínimos, tiene el inconveniente de ser una operación muy capciosa, pues se asume que asimila menos los valores reales y sus tendencias, y puede producir artefactos

Y Axis: latitude (degrees_north)		-27,0	-26,95825	-26,91649	-26,87474	-26,83299
	13,0	NaN	NaN	NaN	NaN	NaN
	13,04175365	NaN	NaN	NaN	NaN	NaN
	13,08350731	NaN	NaN	NaN	NaN	NaN
	13,12526096	NaN	NaN	476,0	478,1	480,2
	13,16701461	NaN	NaN	477,4	481,8	486,2
	13,20876827	NaN	NaN	478,8	485,5	492,2
	13,25052192	NaN	NaN	480,2	489,2	498,1
	13,29227557	NaN	NaN	491,4	498,0	504,6
	13,33402923	NaN	NaN	506,4	508,9	511,3
	13,37578288	NaN	NaN	521,5	519,7	517,9
	13,41753653	NaN	NaN	536,5	530,5	524,6
	13,45929019	NaN	NaN	547,3	538,7	530,1
	13,50104384	NaN	NaN	556,5	545,8	535,1
	13,54279749	NaN	NaN	565,7	552,9	540,2
	13,58455115	NaN	NaN	574,9	560,1	545,2
	13,6263048	NaN	NaN	573,8	559,7	545,6
	13,66805846	NaN	NaN	568,8	556,6	544,3
	13,70981211	NaN	NaN	563,8	553,4	543,0
	13,75156576	NaN	NaN	558,9	550,2	541,6
	13,79331942	NaN	NaN	561,7	554,0	546,4
	13,83507307	NaN	NaN	567,3	560,4	553,5
	13,87682672	NaN	NaN	572,9	566,7	560,5
	13,91858038	NaN	NaN	578,6	573,1	567,5

Ilustración 4.11: Visualización en *Panoply* de los resultados de interpolación lineal, mostrando valores *NaN* (no numéricos) en las coordenadas de latitud y longitud. Este artefacto surge al procesar datos con *xarray* cuando los puntos de grilla no cubren áreas específicas del dominio espacial. **Fuente:** elaboración con datos propios mediante *NASA Panoply*.

Y Axis: latitude (degrees_north)	X Axis: longitude					
	-27,0	-26,95825	-26,91649	-26,87474	-26,83299	
13,0	476,0	476,0	476,0	478,1	480,2	
13,04175365	476,0	476,0	476,0	478,1	480,2	
13,08350731	476,0	476,0	476,0	478,1	480,2	
13,12526096	476,0	476,0	476,0	478,1	480,2	
13,16701461	477,4	477,4	477,4	481,8	486,2	
13,20876827	478,8	478,8	478,8	485,5	492,2	
13,25052192	480,2	480,2	480,2	489,2	498,1	
13,29227557	491,4	491,4	491,4	498,0	504,6	
13,33402923	506,4	506,4	506,4	508,9	511,3	
13,37578288	521,5	521,5	521,5	519,7	517,9	
13,41753653	536,5	536,5	536,5	530,5	524,6	
13,45929019	547,3	547,3	547,3	538,7	530,1	
13,50104384	556,5	556,5	556,5	545,8	535,1	
13,54279749	565,7	565,7	565,7	552,9	540,2	
13,58455115	574,9	574,9	574,9	560,1	545,2	
13,6263048	573,8	573,8	573,8	559,7	545,6	
13,66805846	568,8	568,8	568,8	556,6	544,3	
13,70981211	563,8	563,8	563,8	553,4	543,0	
13,75156576	558,9	558,9	558,9	550,2	541,6	
13,79331942	561,7	561,7	561,7	554,0	546,4	
13,83507307	567,3	567,3	567,3	560,4	553,5	
13,87682672	572,9	572,9	572,9	566,7	560,5	
13,91858038	578,6	578,6	578,6	573,1	567,5	

Ilustración 4.12: Resultado en *Panoply* tras aplicar interpolación por *vecino más cercano*, eliminando los valores *NaN* visibles en las coordenadas de latitud y longitud. Comparado con la interpolación lineal, este método preserva los valores originales de los píxeles adyacentes cuando existen huecos en los datos. **Nota:** La interpolación lineal original (Fig. 4.11) generaba huecos en zonas costeras, mientras que este método mantiene continuidad espacial. **Fuente:** elaboración propia mediante *NASA Panoply*.

de bloques. Dado que replica píxeles sin interpolación, la grilla resultante puede mostrar patrones de píxeles cuadrados visibles, lo cual puede no ser deseable en ciertos contextos.

Con el fin de solucionar este problema, se descubrió y optó por el empleo de la herramienta `RBFInterpolator` de la librería `scipy`, que implementa una interpolación basada en funciones de base radial (*Radial Basis Functions*, RBF). Se trata de una técnica que permite estimar valores faltantes en una malla irregular, considerando todos los puntos válidos disponibles, o una muestra de ellos, y ajustando una función suave que varía con la distancia a dichos puntos.

Es un tipo de interpolación que se caracteriza por su capacidad de generar superficies continuas y suaves, incluso en zonas donde los datos son escasos o están rodeados de valores nulos. A diferencia de los métodos anteriores lineales y de cercanía, este enfoque no requiere que los puntos estén alineados en una grilla estructurada, y resulta adecuado para reconstruir patrones espaciales complejos [61].

Se hizo uso del núcleo *thin plate spline* con un parámetro de suavizado `1e-2`, que permite mitigar oscilaciones excesivas en la superficie interpolada. Para evitar los elevados costes computacionales derivados de esta operación trasladada a toda una cobertura espacial y temporal, se estableció un límite máximo de 5000 puntos válidos por iteración, seleccionados aleatoriamente si se superaba esta cifra. La interpolación se aplicó individualmente en cada paso temporal del conjunto de datos, asegurando además que las zonas terrestres (identificadas mediante la variable `is_ocean`) permanecieran como valores nulos.

Este procedimiento produjo resultados visualmente más coherentes en comparación con las anteriores técnicas, eliminando huecos costeros sin generar artefactos de replicación abrupta.

Como puede observarse en los resultados generados con *NASA Panoply* (ilustración 4.13), las interpolaciones RBF lograron capturar la tendencia de los valores existentes incluso en los extremos del dataset, a diferencia de lo ocurrido con el método del vecino más cercano, que mostró limitaciones en estas zonas.

En un nuevo análisis de la integridad y cobertura de los datos, se verifica que buena parte de las variables principales del conjunto se encuentran completas en su totalidad (ilustración 4.14):

Para la producción del modelo explicativo, se trataron los datos con la ayuda de la librería `sklearn`, que se utilizó para el preprocesamiento, escalado de variables y la transformación Box-Cox de la variable objetivo. La librería `statsmodels` se empleó para el ajuste del modelo de regresión lineal múltiple mediante mínimos cuadrados ordinarios (OLS), permitiendo además obtener métricas estadísticas detalladas del ajuste, como los coeficientes estimados, errores estándar, valores *p* y medidas de bondad del modelo. También se emplearon `matplotlib` y `seaborn` para la visualización de resultados.

Cabe destacar que, en múltiples ocasiones, fue ineludible utilizar la herramienta `Simple Imputer` de la biblioteca `sklearn`, la cual permitió reemplazar datos ausentes —representados como valores nulos— por estimaciones o valores derivados. Esta acción resultó fundamental para mantener la integridad del conjunto de datos y garantizar el correcto funcionamiento de

Y Axis: latitude (degrees_north)		-27,0	-26,95825	-26,91649	-26,87474	-26,83299
	13,0	462,9	461,2	459,4	457,8	456,3
	13,04175365	468,7	466,8	464,9	463,1	461,4
	13,08350731	474,7	472,5	470,4	468,5	466,7
	13,12526096	481,0	478,6	476,0	478,1	480,2
	13,16701461	487,6	484,9	477,4	481,8	486,2
	13,20876827	494,7	491,8	478,8	485,5	492,2
	13,25052192	502,4	499,2	480,2	489,2	498,1
	13,29227557	510,6	507,5	491,4	498,0	504,6
	13,33402923	519,4	516,6	506,4	508,9	511,3
	13,37578288	528,4	525,9	521,5	519,7	517,9
	13,41753653	537,0	534,7	536,5	530,5	524,6
	13,45929019	544,9	542,5	547,3	538,7	530,1
	13,50104384	552,1	549,5	556,5	545,8	535,1
	13,54279749	558,6	555,6	565,7	552,9	540,2
	13,58455115	564,5	561,3	574,9	560,1	545,2
	13,6263048	569,9	566,5	573,8	559,7	545,6
	13,66805846	574,8	571,4	568,8	556,6	544,3
	13,70981211	579,3	575,7	563,8	553,4	543,0
	13,75156576	583,1	579,5	558,9	550,2	541,6
	13,79331942	586,3	582,6	561,7	554,0	546,4
	13,83507307	588,8	585,1	567,3	560,4	553,5
	13,87682672	590,5	586,7	572,9	566,7	560,5
	13,91858038	591,2	587,5	578,6	573,1	567,5

Ilustración 4.13: Visualización del campo interpolado mediante RBF con núcleo *thin plate spline*. La interpolación ofrece una transición suave entre zonas con datos válidos y aquellas originalmente vacías, sin generar patrones de bloque ni pérdidas abruptas. **Fuente:** elaboración propia mediante *NASA Panoply*.

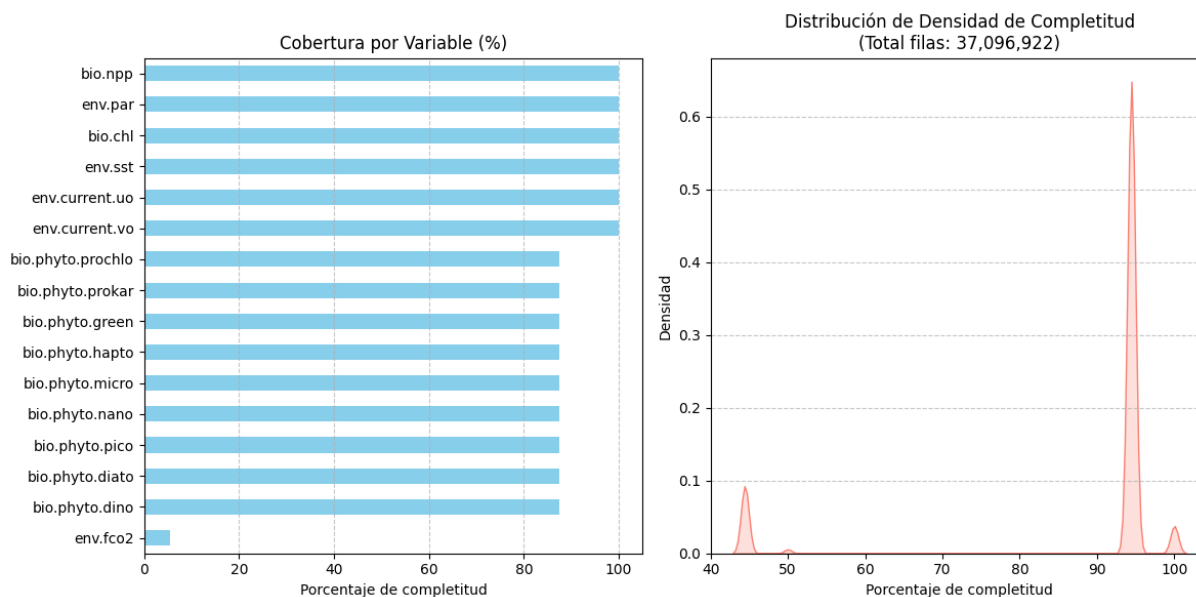


Ilustración 4.14: Evaluación de la cobertura de datos para 16 variables oceanográficas (2002-2023): **(A)** Porcentaje de datos disponibles por variable (NPP, PAR, SST, CHL, etc.), **(B)** Distribución de la cobertura por registros individuales. Destacan los gaps en fitoplancton ($\sim 87.45\%$) y `env.fc02` ($\sim 5.51\%$), frente a la completitud de `bio.npp`, `env.par`, `env.sst`, `bio.chl` y corrientes oceánicas ($\sim 100\%$). **Fuente:** elaboración con datos propios.

los algoritmos de análisis y aprendizaje automático, ya que, en muchos casos, no era posible trabajar directamente con valores faltantes.

Para el desarrollo del modelo en el proyecto Picota, fue necesario aplanar el conjunto de datos espaciales con el objetivo de enriquecer la representación de cada perfil individual. En lugar de considerar únicamente los datos propios de cada tesela (unidad espacial), se optó por incorporar también la información de sus vecinas adyacentes —específicamente, las teselas superior, inferior, izquierda y derecha—. Esta estrategia permite al modelo captar mejor las correlaciones espaciales y mejorar así su capacidad estimativa.

Cada tesela representa a un sujeto o entidad y se describe mediante un único archivo TSV que contiene sus registros temporales. Dado que inicialmente se trabajaba con una malla de 480 x 480 puntos, esto implicaba la generación y el procesamiento de más de 230.000 registros por marco temporal, lo que suponía una carga computacional muy elevada tanto en almacenamiento como en tiempo de procesamiento.

Para mitigar este problema, se decidió reducir la resolución de la malla a 240 x 240 puntos. Esta simplificación permitió disminuir drásticamente el coste computacional asociado al procesamiento de estos datos, sin comprometer de forma significativa la calidad o utilidad de los datos para el entrenamiento del modelo.

La nueva reproyección del conjunto de datos volvió a presentar desafíos similares a los encontrados previamente. Concretamente, muchas de las teselas correspondientes a zonas oceánicas no contaban con datos completos de las variables principales, como la producción

primaria y la concentración de clorofila-a. Esta ausencia de información dificultaba el uso directo de estos puntos en el modelo.

En consecuencia, fue necesario aplicar nuevamente las técnicas de interpolación espacial para estimar los valores faltantes, al menos para las variables fundamentales. Esta etapa resultó crucial para garantizar la coherencia del conjunto de datos y permitir su utilización en las siguientes fases del procesamiento y entrenamiento del modelo, sin introducir sesgos significativos debidos a la falta de datos en regiones clave (ilustración 4.15).

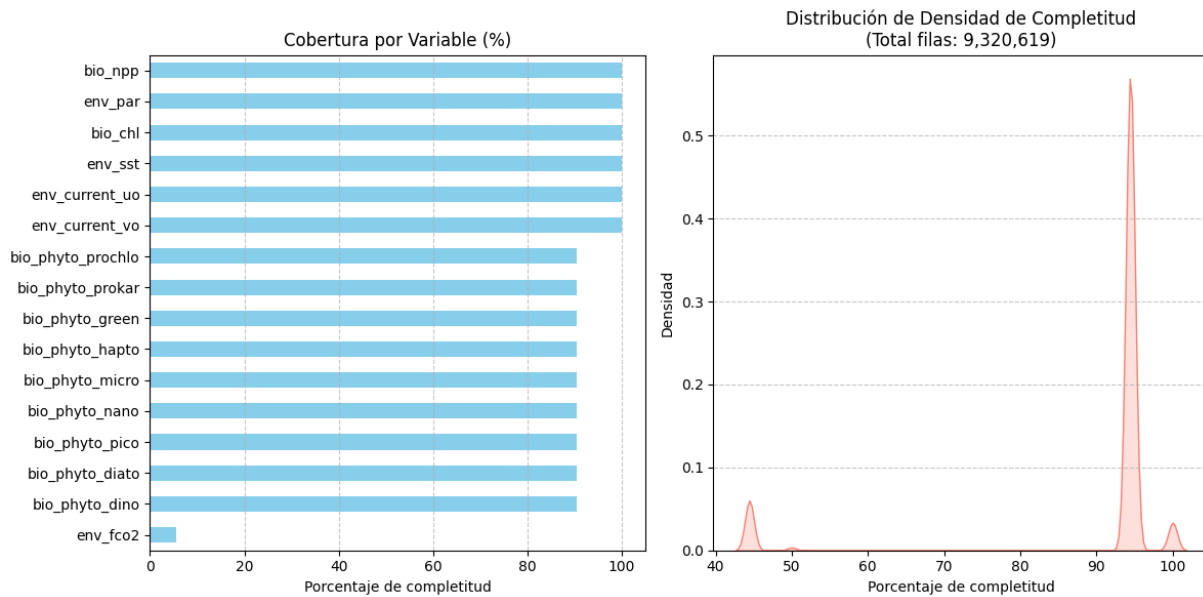


Ilustración 4.15: Evaluación de la cobertura de datos para 16 variables oceanográficas (2002-2023): **(A)** Porcentaje de datos disponibles por variable (NPP, PAR, SST, CHL, etc.), **(B)** Distribución de la cobertura por registros individuales. Destacan los gaps en fitoplancton ($\sim 90.54\%$) y `env.fco2` ($\sim 5.51\%$), frente a la completitud de `bio.npp`, `bio.chl`, `env.par`, `env.sst` y las corrientes oceánicas ($\sim 100\%$). **Fuente:** elaboración con datos propios.

Para el modelado y despliegue de los gemelos digitales, se ha empleado la plataforma Picota, diseñada específicamente para la construcción de *Digital Twins* a partir de datos temporales, descripciones estructurales y modelos inferenciales.

Picota opera mediante tres fases integradas:

1. Modelado descriptivo con Picota DSL, en el que se define la estructura del sistema (Realidad, Sujetos y Variables) mediante un lenguaje específico de dominio (DSL), que permite describir declarativamente el entorno y las propiedades observables o inferidas de cada sujeto.
2. Entrenamiento de modelos de inferencia con Picota DT, donde, a partir de datos históricos, se entrenan las redes neuronales para estimar valores actuales o futuros de las variables marcadas para inferencia.
3. Despliegue operativo con Picota DT una vez entrenados los modelos, que se agrupan en

un gemelo digital funcional, que opere como un servicio web, capaz de simular, predecir y adaptarse en tiempo real.

En este trabajo, Picota se ha empleado para modelar la NPP como variable de salida, a partir de múltiples entradas como clorofila-a, temperatura superficial del mar, radiación fotosintéticamente activa, fugacidad de CO_2 o corrientes oceánicas, entre otras. El modelado estructural permitió organizar las variables por teselas espaciales, y establecer relaciones espaciales entre cada celda y sus vecinas (izquierda, derecha, arriba, abajo), enriqueciendo así el contexto de entrada del modelo.

Además, se diseñaron sujetos para representar las distintas unidades espaciales del océano, y se ha configurado la resolución temporal de inferencia (mensual) en cada caso. Gracias al entorno visual y al flujo de trabajo modular de Picota, ha sido posible construir, entrenar y desplegar de manera eficiente el gemelo digital, integrando datos reales en el proceso de entrenamiento y validación.

Capítulo 5

Resultados

Este capítulo presenta los principales hallazgos obtenidos a partir del análisis de datos oceanográficos y la aplicación del modelo desarrollado para la estimación de la producción primaria neta (NPP) en la región macaronesia. Se estructura en dos bloques diferenciados. Por un lado, se analiza el comportamiento de variables oceanográficas clave en las últimas dos décadas y su relación con la biodiversidad marina, y las propias interrelaciones entre perfiles de datos; por otro, se muestran los resultados del gemelo digital diseñado para estimar la NPP a partir de dichas variables.

Los resultados se exponen tanto desde una perspectiva descriptiva como predictiva, incorporando métricas, comparativas, y evaluaciones del rendimiento del modelo y su utilidad como herramienta de simulación ecológica, lo que permite validar el enfoque metodológico, e identificar patrones espaciales y temporales relevantes para la gestión marina sostenible en el contexto macaronesio.

5.1. Análisis del modelo explicativo del entorno oceánico

Con el objetivo de contextualizar los resultados del modelo estimativo, se ha llevado a cabo un análisis exploratorio y explicativo de los principales factores oceanográficos que influyen en la NPP de la región macaronesia. Esta sección se centra en el estudio de variables ambientales y biológicas de resolución temporal y espacial —como la temperatura superficial del mar, la concentración de clorofila-a y los diferentes grupos fitoplanctónicos— que permiten identificar patrones de variabilidad interanual y estacional, e interrelaciones entre los distintos perfiles.

El análisis proporciona una base sólida sobre la evolución del ecosistema oceánico en las últimas décadas, y aporta información valiosa para interpretar el comportamiento del modelo desarrollado. Se busca así establecer relaciones causales o correlacionales entre los cambios observados y la NPP, facilitando una comprensión más profunda del entorno marino que sustenta el funcionamiento del gemelo digital.

5.1.1. La biodiversidad oceánica macaronesia en las últimas dos décadas

La región macaronesia, ubicada en el Atlántico nororiental, constituye un enclave singular en términos de biodiversidad marina, caracterizado por una elevada productividad biológica derivada de fenómenos de afloramiento, vientos alisios estacionales y la influencia de la Corriente de Canarias, como se nombró en los apartados anteriores.

Durante las dos últimas décadas, diversos estudios han evidenciado cambios significativos en las condiciones oceanográficas de la región y en la composición de sus comunidades fitoplanctónicas. Con el objetivo de analizar y modelar esta evolución, en el presente trabajo se han recopilado y procesado datos de múltiples variables clave para la estimación de la NPP. Estas variables, además de alimentar el modelo basado en redes neuronales artificiales, permiten evaluar tendencias y correlaciones que influyen directamente en la productividad marina.

Los principales perfiles y variables considerados en este estudio son:

- ✓ **Concentración de clorofila-a ($Chl - a$)**. Es el indicador de biomasa fitoplanctónica, obtenido mediante teledetección. Altamente correlacionada con la actividad fotosintética en la zona eufótica.
- ✓ **Composición del fitoplancton**. Se han incluido estimaciones de los principales grupos fitoplanctónicos, cuya distribución ha cambiado en respuesta al calentamiento oceánico, la estratificación y la acidificación del agua.
- ✓ **Radiación fotosintéticamente activa (PAR)**. Representa la energía disponible para la fotosíntesis (400–700 nm). Su disponibilidad es esencial para determinar la eficiencia fotosintética del fitoplancton.
- ✓ **Temperatura superficial del mar (SST)**. Es el factor clave que regula la estratificación de la columna de agua y, por tanto, la disponibilidad de nutrientes. Influye directamente en la fisiología del fitoplancton.
- ✓ **Fugacidad de CO_2 (fCO_2)**. Es la variable que indica la presión parcial de dióxido de carbono en la superficie oceánica, relevante tanto para el secuestro de carbono como para los procesos fotosintéticos.
- ✓ **Corrientes marinas**. Determinan el transporte de nutrientes, organismos planctónicos y calor. Su variabilidad puede inducir importantes redistribuciones de biomasa fitoplanctónica.
- ✓ **Producción primaria neta (NPP)**. Es la variable objetivo del presente trabajo. Se refiere a la fracción del carbono fijado por el fitoplancton que queda disponible tras descontar la respiración autótrofa.

A la hora de tomar en consideración únicamente los perfiles de mayor impacto en la variable objetivo NPP en base a estos conjuntos, se construyó una matriz de correlación de Pearson entre todos los conjuntos recopilados (ilustración 5.1). Este análisis ha permitido

identificar las relaciones más significativas entre los factores físicos y biológicos, y cómo estos condicionan directamente la variabilidad observada en la NPP, estableciendo al mismo tiempo validez teórica a las observaciones realizadas:

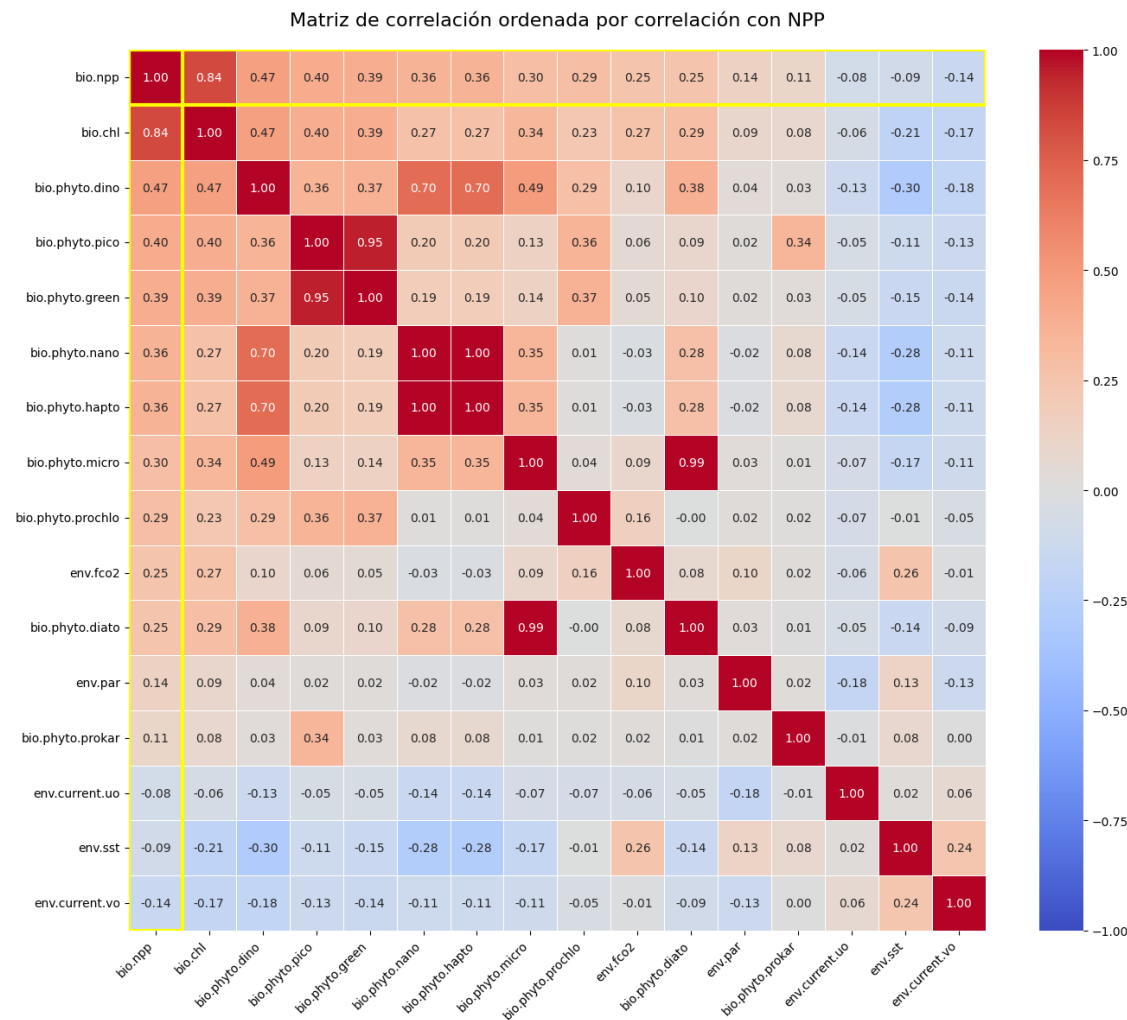


Ilustración 5.1: Matriz de correlación ordenada por su relación con la productividad primaria neta (NPP), mostrando las interacciones entre variables biogeoquímicas y ambientales. **Fuente:** análisis con datos propios procesados.

Para evaluar los cambios en las variables clave del ecosistema marino, se han generado visualizaciones comparativas entre dos períodos de referencia: 2003-2007 y 2019-2023. Esta selección temporal permite identificar tendencias decenales vinculadas al cambio climático o a fenómenos oceanográficos de gran escala al abarcar ciclos biológicos completos, y analizar impactos en la productividad primaria neta, mediante múltiples variables como la clorofila-a (CHL), directamente relacionada con la biomasa fitoplanctónica. Los resultados se presentan mediante mapas de anomalías, que destacan desviaciones espaciales y temporales respecto al período base, facilitando la interpretación de patrones regionales.

Para la clorofila-a, la disposición de los datos sobre la región revelan cambios significativos

en la distribución espacial y la intensidad de esta variable clave para la producción primaria en la Macaronesia.

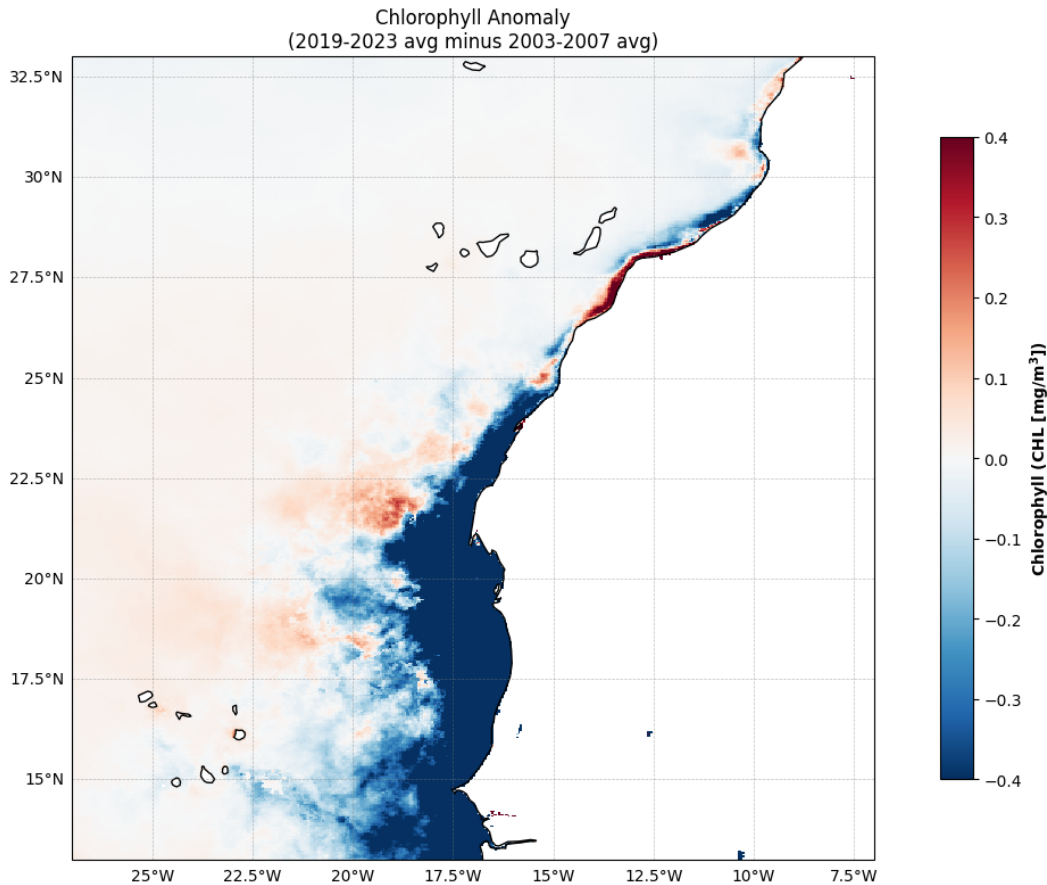


Ilustración 5.2: Diferencias en concentración de clorofila-a (mg/m^3) entre los periodos 2003-2007 (línea base) y 2019-2023. Valores positivos (rojo) indican aumento de productividad, negativos (azul) disminución. **Fuente:** elaboración con datos propios procesados de Copernicus Marine.

El mapa de anomalías de clorofila-a (ilustración 5.2) revela cambios significativos en la productividad primaria del Atlántico noreste entre 2003-2007 y 2019-2023. Se observa un patrón espacial heterogéneo, con áreas de aumento en rojo en la concentración de clorofila, y que sugieren mayor productividad fitoplanctónica, y valores negativos en azul, que expresan una disminución en la concentración de clorofila, asociada a una reducción en la biomasa fitoplanctónica.

Los mayores aumentos de la concentraciones de clorofila se localizan principalmente en zonas costeras y áreas de afloramiento al norte de la región, donde el aporte de nutrientes desde aguas profundas puede haber favorecido el crecimiento de fitoplancton, posiblemente vinculados a cambios en las corrientes oceánicas o a un incremento en la disponibilidad de nutrientes en estas regiones. Las disminuciones notables se concentran en zonas costeras al sur de la región, donde el calentamiento de la superficie del mar, debido al aumento de las temperatura, y la estratificación de la columna de agua pueden haber limitado el

acceso a nutrientes esenciales. Coinciden con el declive observado en diatomeas, especie de fitoplancton especialmente dominante en aguas productivas, que dependen de aguas frías y ricas en nutrientes. La tendencia general (ilustración 5.3) muestra esa disminución en las concentraciones de clorofila.

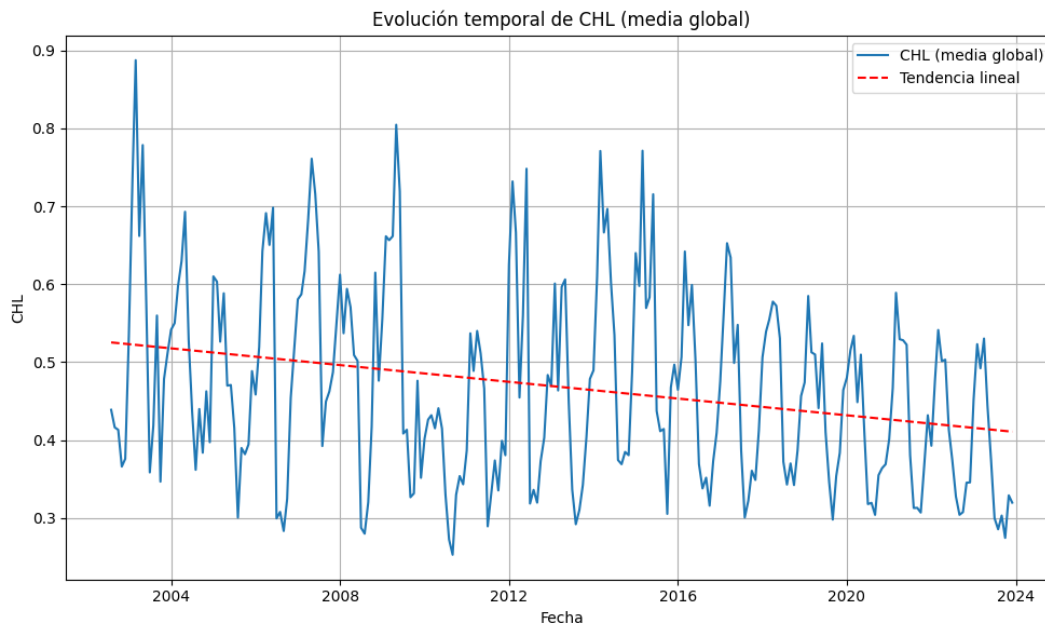


Ilustración 5.3: Tendencia temporal de la concentración media global de clorofila-a (CHL, mg/m^3) entre 2003 y 2023. La línea roja muestra la tendencia lineal, indicando una disminución significativa en las concentraciones del 13 % durante el período establecido. **Fuente:** elaboración con datos propios procesados de Copernicus Marine.

Además de identificar relaciones estadísticas entre variables, se exploró la variabilidad temporal de la composición fitoplanctónica, ya que la proporción relativa de los diferentes grupos puede tener un impacto directo en la eficiencia de la productividad primaria y la transferencia de energía en la red trófica. Esta visualización permite detectar patrones estacionales y cambios estructurales en la comunidad fitoplanctónica a lo largo de los años (ilustración 5.4).

Los grupos fitoplanctónicos incluidos fueron A) diatomeas, algas con paredes celulares de sílice, comunes en aguas frías y ricas en nutrientes, B) dinoflagelados, microorganismos móviles, algunos con capacidad fotosintética o mixotrófica, C) algas verdes o clorofitas, dominantes en ambientes de agua dulce pero también presentes en marinos, D) haptófitas, incluyen especies importantes como los cocolitóforos, relevantes en ciclos biogeoquímicos, y E) microfitoplancton, nanofitoplancton y picofitoplancton, clasificados por tamaño (20 a 200 μm , 2 a 20 μm y menos de 2 μm , respectivamente), con roles ecológicos distintos.

El análisis denota una predominancia de picofitoplancton y nanoplancton, lo que sugiere su alta adaptabilidad a condiciones variables y su papel clave en la producción primaria. Este patrón es característico de sistemas oligotróficos, es decir, pobres en nutrientes, como suele ser el caso de las aguas abiertas oceánicas. No obstante, en ciertas zonas costeras,

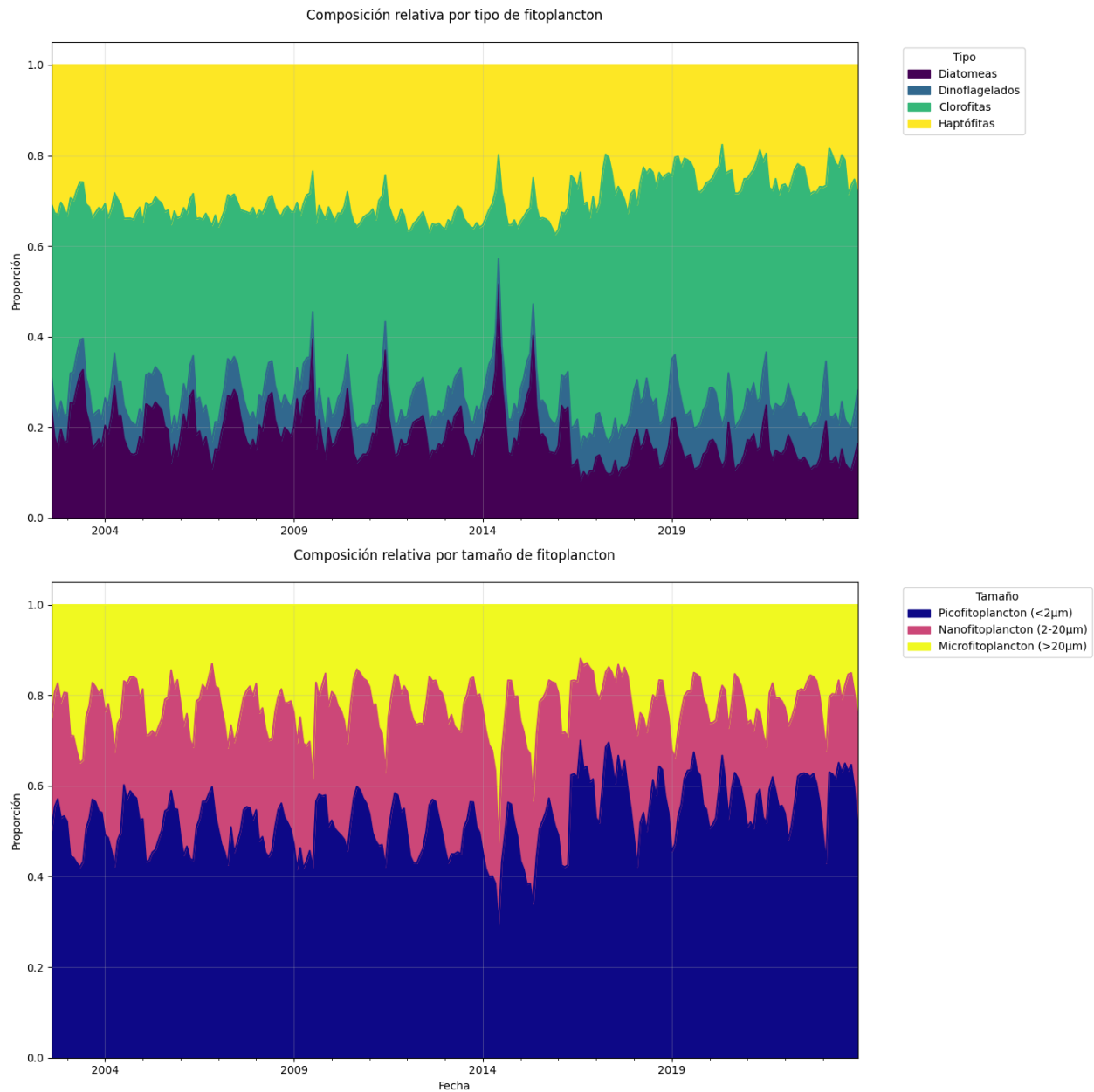


Ilustración 5.4: Dinámica temporal de la comunidad fitoplanctónica: **(A)** Composición por grupos funcionales (diatomeas, dinoflagelados, etc.), **(B)** Distribución por tamaño celular (pico-, nano- y microfitoplancton). **Tendencias:** (1) Aumento progresivo de picofitoplancton (2014: 15 % → 2019: 28 %), (2) Dominancia persistente de diatomeas (> 40 % en todos los periodos). **Fuente:** análisis con datos propios de Copernicus Marine.

procesos de afloramiento —donde las aguas profundas, frías y ricas en nutrientes ascienden a la superficie— pueden generar picos de productividad y favorecer temporalmente a grupos fitoplanctónicos como las diatomeas.

Aunque se observa variabilidad estacional, la estructura comunitaria mostró una notable estabilidad durante el período estudiado. Sin embargo, destaca un decrecimiento progresivo en la proporción de diatomeas en los últimos años, un patrón que podría asociarse al aumento de las temperaturas u otros factores de estrés ambiental. Estos cambios en la dominancia de grupos fitoplanctónicos (p. ej., aumento relativo del picoplancton) podrían responder a alteraciones en variables como la concentración de nutrientes, la irradiancia o la temperatura, con implicaciones para la eficiencia de la productividad primaria y el ciclo del carbono en el ecosistema.

Para estudiar la posible causa de la disminución de diatomeas en los últimos años de la década, se procede a estudiar la anomalía en los efectos de la temperatura superficial del océano en los períodos establecidos:

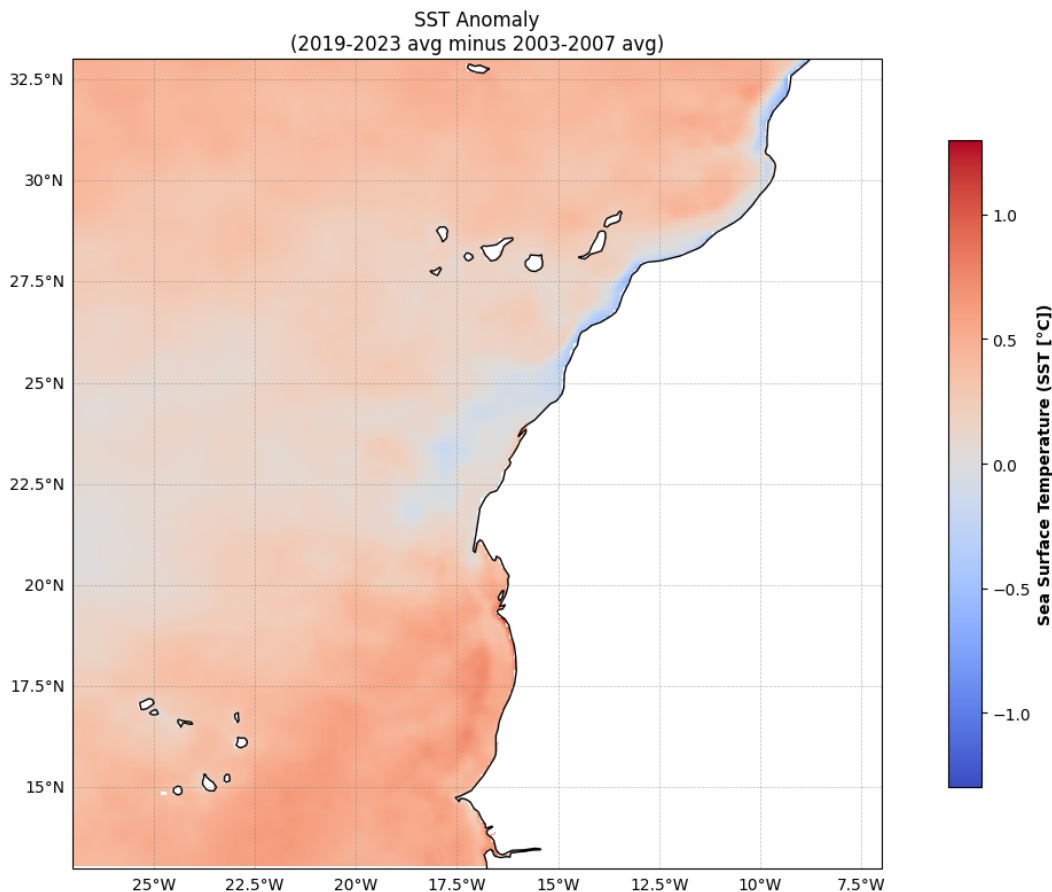


Ilustración 5.5: Cambios en temperatura superficial ($^{\circ}C$) entre 2003-2007 y 2019-2023. Patrón espacial muestra calentamiento en áreas oceánicas profundas ($\geq +1,5^{\circ}C$) y enfriamiento en zonas costeras ($\geq -0,5^{\circ}C$). **Escala:** $-1,0^{\circ}C$ a $+1,0^{\circ}C$ (intervalos de $0,2^{\circ}C$). **Fuente:** elaboración con datos propios procesados de Copernicus Marine.

La ilustración 5.5 revela la distribución espacial de la anomalía de SST en el Atlántico noreste. La escala de colores (el azul indica una disminución de temperaturas $< 0^{\circ}\text{C}$, y el rojo, un aumento de temperaturas $> 0^{\circ}\text{C}$) permite inferir un posible aumento de la temperatura superficial oceánica (SST), que coincide con tendencias globales de calentamiento oceánico.

Un aumento de SST podría reducir el habitat óptimo para diversos tipos de fitoplancton, como las diatomeas, que prosperan en aguas frías (en torno a 20°C o menos), favoreciendo a grupos más pequeños, como picoplancton o nanoplancton, adaptados a aguas más cálidas y estratificadas. Esto tiene impacto directo en la productividad primaria, pues las diatomeas son grandes exportadoras de carbono al fondo marino, y su declive podría reducir la eficiencia de la bomba biológica de carbono. Además, las redes tróficas pueden verse afectadas, pues un cambio de comunidades podría impactar a organismos que dependen de diatomeas (Krause et al., 2020 [33]).

La variable objetivo, NPP, comparada en los dos períodos, complementa el análisis de temperatura superficial del mar, la concentración de clorofila y los cambios en las comunidades fitoplanctónicas. Dado que se detectó un declive en las diatomeas, que son productores eficientes en aguas frías y ricas en nutrientes, y una preponderancia del picofitoplancton y nanofitoplancton, a pesar de tener menores tasas de exportación de carbono, se espera que el aumento de la temperatura superficial redujera la NPP en zonas que antaño se consideraban productivas. Además, un cambio hacia comunidades dominadas por picoplancton alterara los patrones espaciales de NPP.

La ilustración 5.6 permite inferir posibles tendencias temporales. Destacan principalmente los colores azules, que representan una reducción de la productividad primaria, lo que apoyaría la hipótesis de que el calentamiento reduce la NPP. Se sugieren como causas probables la estratificación inducida por el calentamiento, que reduce la mezcla vertical de nutrientes (p. ej., nitratos, fosfatos y silicatos), así como eventos extremos como olas de calor que afectan al océano. Adicionalmente, este descenso podría estar acompañado por cambios en la composición del fitoplancton como efecto de estas condiciones. Por otro lado, los tonos rojos (mayor índice de aumento de NPP) se encuentran en todas las aguas abiertas y en aguas de afloramiento costero al noroeste de África. Las posibles causas son el aporte continuo de nutrientes por la surgencia de aguas profundas, el dominio de tipos de fitoplancton adaptados a la variabilidad térmica y las corrientes locales, como la Corriente de Canarias, que redistribuye los nutrientes.

Las implicaciones ecológicas de esta reducción de la producción primaria neta podrían significar la disminución en la exportación de carbono a aguas profundas, debilitando la bomba biológica. Además, una reducción en la calidad nutricional del fitoplancton dominante —por ejemplo, predominio de picoplancton frente a diatomeas, que son más nutritivas— afectaría a los consumidores primarios, como el zooplancton. Los cambios en la fenología (es decir, en la sincronización estacional de los ciclos de vida y reproducción de los organismos) podrían además causar desacoplamientos temporales entre productores y consumidores, alterando las interacciones entre presas y depredadores y comprometiendo la eficiencia del flujo de energía en el ecosistema.

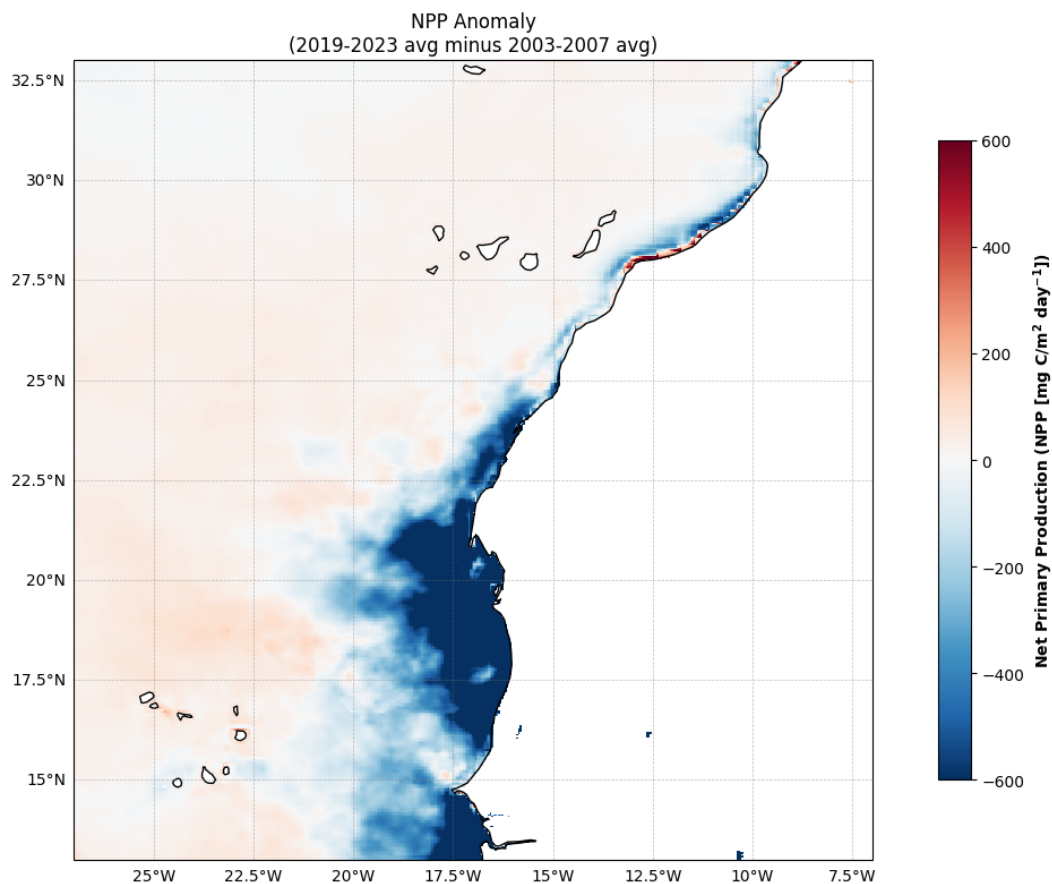


Ilustración 5.6: Variación en productividad primaria neta (NPP, $\text{mg C/m}^2/\text{day}$) entre períodos de referencia. Destacan zonas de disminución (hasta $-600 \text{ mg C/m}^2/\text{day}$) asociadas a afloramientos costeros. **Fuente:** elaboración con datos propios procesados de la Universidad Estatal de Oregón. **Método:** modelado con Eppley-VGPM.

5.1.2. Factor de inflación de la varianza

El factor de inflación de la varianza (VIF, por sus siglas en inglés *Variance Inflation Factor*) es una medida estadística utilizada para detectar la presencia de multicolinealidad entre las variables independientes de un modelo de regresión. La multicolinealidad ocurre cuando una variable explicativa puede predecirse linealmente a partir de las demás con un alto grado de precisión, lo que puede afectar negativamente la estabilidad y fiabilidad de los coeficientes estimados (Akinwande et al., 2015 [2]).

El VIF indica cuánto se incrementa la varianza de un coeficiente estimado debido a la colinealidad con otras variables del modelo. Un VIF cercano a 1 sugiere que no hay colinealidad significativa. Valores entre 5 y 10 indican una multicolinealidad moderada, que podría requerir atención, mientras que valores superiores a 10 se interpretan como evidencia de alta multicolinealidad, lo cual puede comprometer la interpretación del modelo y, por tanto, justificar la eliminación o combinación de variables.

Con el objetivo de evaluar este fenómeno, se realizó un análisis de multicolinealidad para todas las variables explicativas empleadas en el estudio de la producción primaria neta. A continuación, se presentan los resultados obtenidos para el VIF de cada una de las variables consideradas:

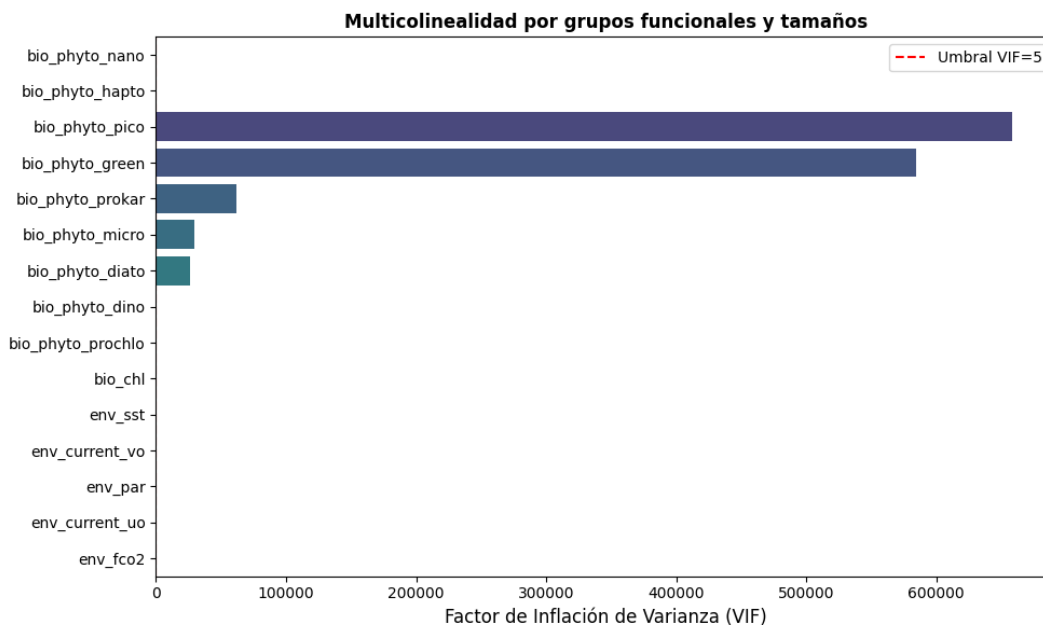


Ilustración 5.7: Factores de Inflación de Varianza (VIF) para variables ambientales y grupos fitoplanctónicos. **Umbral crítico:** $VIF > 5$ (línea roja), indicando multicolinealidad severa en bio_phyto ($VIF \approx 600,000$). **Patrones:** (1) Variables físicas (`env_sst`, `env_par`) con baja multicolinealidad ($VIF < 5$), (2) Grupos fitoplanctónicos interrelacionados (`bio_phyto`). **Método:** `sklearn` y `statsmodels` (Python). **Fuente:** elaboración con datos propios procesados.

En el análisis inicial de multicolinealidad (ilustración 5.7), se evaluó el conjunto completo

de variables relacionadas con grupos funcionales fitoplanctónicos y parámetros ambientales. Los resultados mostraron alta multicolinealidad, ya que muchas variables superaban ampliamente el umbral crítico de $VIF = 5$, con algunas alcanzando valores extremadamente altos (hasta 600,000), y variables problemáticas. principalmente entre los grupos fitoplanctónicos (como `bio_phyto_nano`, `bio_phyto_hapto`, `bio_phyto_pico`, etc.), que presentaban fuertes interdependencias al combinar clasificaciones por tipo funcional y por tamaño del fitoplancton.

Además, se analizaron las interdependencias mediante una matriz de correlación, lo que permitió confirmar las fuertes relaciones existentes entre las variables fitoplanctónicas (ilustración 5.8).

La situación de multicolinealidad extrema hace inviable la interpretación confiable de los coeficientes en un modelo de regresión, ya que las varianzas de los estimadores se inflan enormemente.

Para abordar estos problemas, se implementaron medidas de agrupación por tamaño, reorganizando el conjunto de datos y seleccionando únicamente las variables fitoplanctónicas clasificadas por tamaño, en lugar de por grupos funcionales. Estas categorías abarcan todos los grupos: **nano** para organismos entre 0.2 y 2 μm , **micro** entre 2 y 20 μm , y **pico** entre 20 y 200 μm , lo que permitió reducir la redundancia. Además, se aplicó selección de variables, en las que se eliminaron los perfiles más colineales, manteniendo solo aquellas con menor VIF y mayor relevancia ecológica. Asimismo, se transformaron los datos, para reducir las correlaciones lineales entre predictores.

El segundo gráfico muestra los resultados después de estas modificaciones (ilustración 5.9), con valores de VIF cercanos a 1 para todas las variables, indicando multicolinealidad mínima. La gráfica demuestra que, mediante reorganización y selección de variables, fue posible resolver los problemas de multicolinealidad identificados inicialmente, permitiendo ahora la construcción de modelos de regresión más estables e interpretables.

La estrategia adoptada mejoró las propiedades estadísticas del modelo, y mantuvo la información ecológica y ambientalmente relevante necesaria para el estudio y estimación de la producción primaria neta.

5.1.3. Regresión lineal múltiple

El modelo de regresión lineal múltiple es una técnica estadística comúnmente utilizada para modelar la relación entre una variable dependiente y múltiples variables independientes (Abuín, 2007 [1]). En este caso, se emplea para estimar la NPP a partir de diversas variables ambientales y biológicas.

Una vez ajustado el modelo, es esencial evaluar la calidad del ajuste mediante el análisis de los residuos. Los residuos, definidos como la diferencia entre los valores observados y los predichos, deben cumplir ciertos supuestos para que los resultados del modelo sean válidos y confiables. Entre los supuestos clave se encuentran la homocedasticidad, que establece que

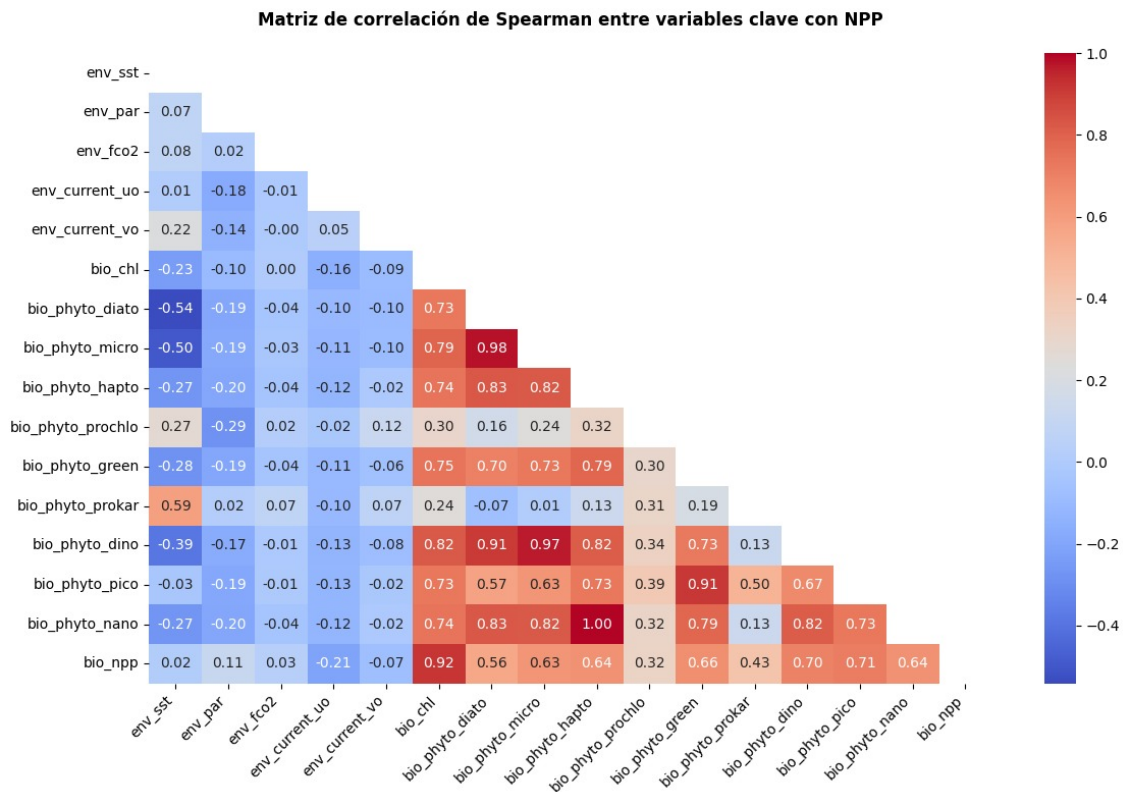


Ilustración 5.8: **Matriz de correlación de Spearman (ρ) entre variables ambientales y biológicas, incluyendo NPP.** Se analizaron relaciones monótonas entre: (a) Variables ambientales (SST, PAR, fCO_2 , corrientes) y (b) Biomasa fitoplanctónica (clorofila y grupos taxonómicos, NPP).

Patrones destacados: 1. Correlación moderada-negativa entre SST (`env_sst`) y microfitoplancton ($\rho = -0,50$), diatomeas ($\rho = -0,54$) y procariotas ($\rho = -0,59$); 2. Alta correlación positiva entre clorofila (`bio_chl`) y diatomeas ($\rho = 0,73$), microfitoplancton ($\rho = 0,79$) y dinoflagelados ($\rho = 0,82$); 3. Relaciones físicas débiles: corrientes (uo/vo) muestran $|\rho| < 0,18$ con variables biológicas.

Interpretación: valores en negrita destacan correlaciones fuertes ($|\rho| > 0,5$). **Método:** análisis con `pandas.DataFrame.corr(method='spearman')` (Python). **Fuente:** elaboración con datos propios procesados.

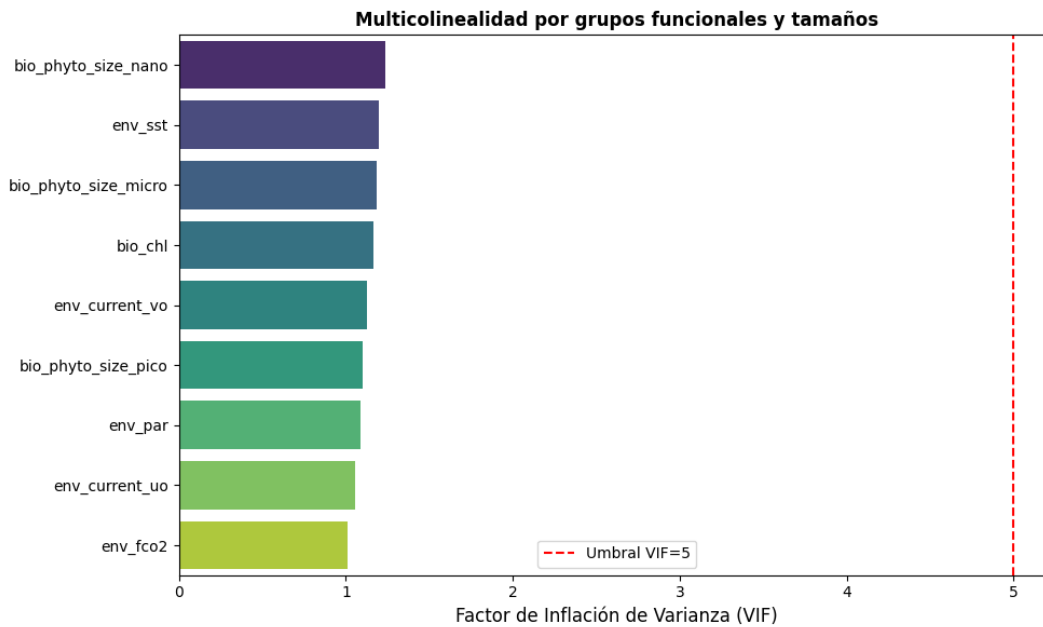


Ilustración 5.9: Factores de Inflación de Varianza (VIF) según tamaño celular del fitoplancton y variables ambientales. Se presentan los VIF calculados para cada categoría. Todos los conjuntos (nanofitoplancton, microfitoplancton, picofitoplancton, y variables ambientales y biológicas) presentan $VIF \approx 1$, lo que indica riesgo moderado de multicolinealidad (umbral de inspección ≈ 5). **Conclusión:** la reclasificación del fitoplancton por tamaño celular ha reducido eficazmente la multicolinealidad general, manteniéndola por debajo del umbral seguro ($VIF < 5$). **Método:** `VarianceInflationFactor` de `statsmodels` (Python). **Fuente:** elaboración con datos propios procesados.

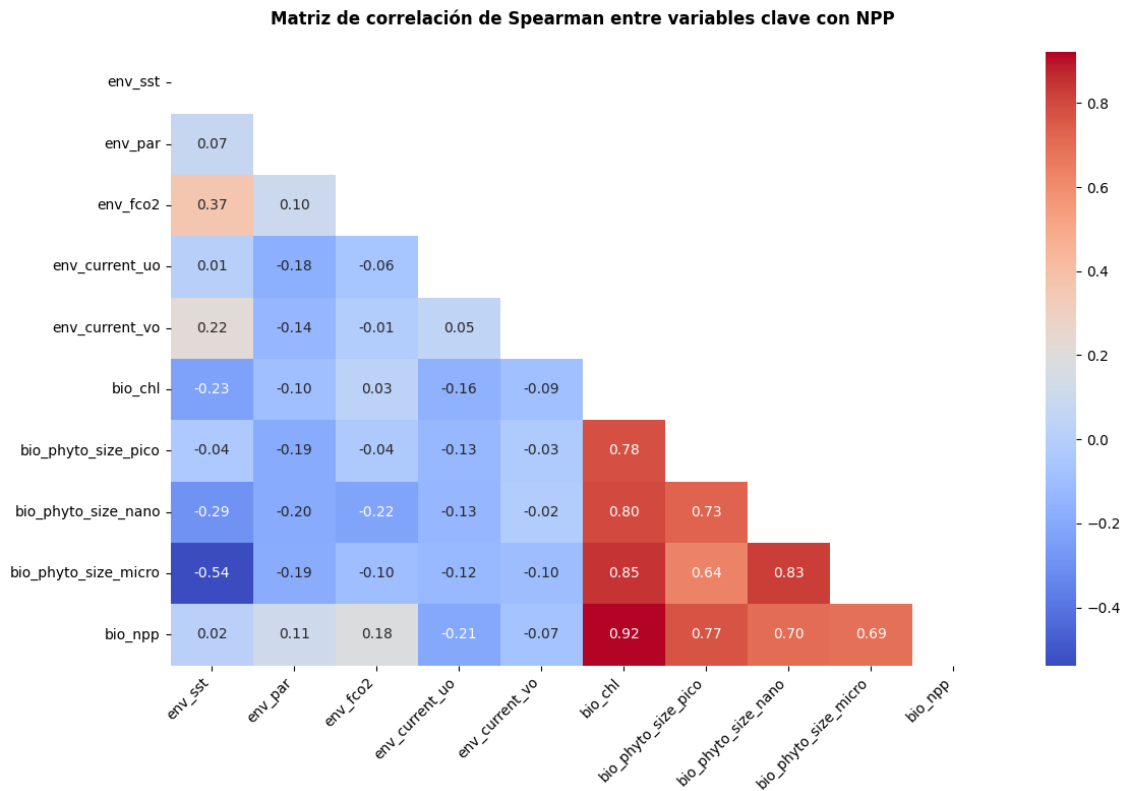


Ilustración 5.10: Matriz de correlación de Spearman (ρ) entre variables ambientales y biológicas y grupos fitoplanctónicos por tamaño, incluyendo NPP. Se empleó correlación de rango para identificar relaciones monótonas entre: (a) Variable física (SST, PAR, corrientes), biomasa (CHL) y productividad (NPP), (b) grupos por tamaño celular (pico, nano, micro).

Patrones destacados: 1. Anticorrelación leve entre *Chl - a* y SST ($\rho = -0,23$), indicando tendencia de menor biomasa en aguas más cálidas; 2. Baja correlación negativa entre PAR y corrientes ($\rho = -0,18 / -0,14$), reflejando relaciones físicas débiles; 3. Microfitoplancton muestra la mayor correlación positiva con clorofila ($\rho = 0,55$), sugiriendo mayor contribución a la biomasa total.

Interpretación de valores: coeficientes $-\rho-$ entre 0–0.10 se consideran bajos o débiles.

Método: cálculos realizados con `pandas.DataFrame.corr(method='spearman')` (Python).

Fuente: elaboración con datos propios procesados.

la variabilidad de los residuos es independiente de las variables explicativas, y la normalidad de los residuos.

Para evaluar estos supuestos, se presentan los resultados del diagnóstico de residuos en la ilustración 5.11. La gráfica de residuos frente a valores ajustados (izquierda) permite identificar si la varianza de los errores se mantiene constante, lo cual es crucial para la validez de las inferencias estadísticas. Por otro lado, el QQ-plot (derecha) proporciona una representación visual de la distribución de los residuos en comparación con una distribución normal teórica, lo que permite evaluar si los residuos siguen una distribución normal, como se espera en la regresión lineal.

Como se observa, la heterocedasticidad y la desviación de la normalidad sugieren que los supuestos clásicos no se cumplen completamente, lo que podría requerir ajustes en el modelo o considerar alternativas metodológicas.

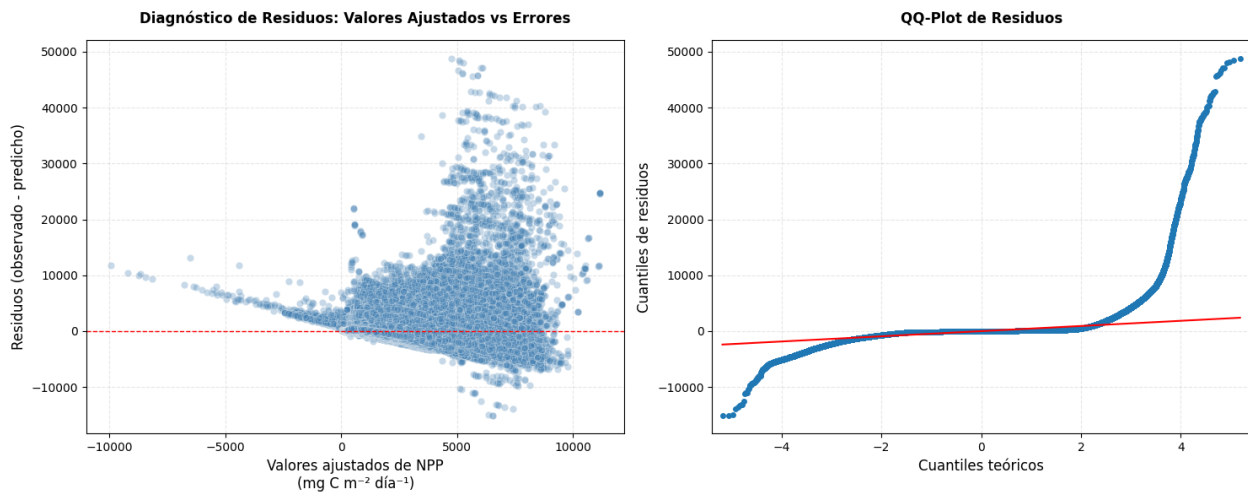


Ilustración 5.11: **Diagnóstico de residuos del modelo de estimación de NPP (*Net Primary Production*)**: *Izquierda*: gráfico de residuos frente a valores ajustados muestra heterocedasticidad: la varianza de los errores aumenta con la magnitud de los valores predichos. *Derecha*: QQ-plot de residuos frente a una distribución normal teórica revela asimetría y colas pesadas, lo que indica una distribución no normal de los errores.

Implicaciones: los supuestos clásicos de homocedasticidad y normalidad no se cumplen completamente, lo que sugiere necesidad de transformación de variables, robustecimiento del modelo o métodos alternativos de estimación. **Método:** `statsmodels` y `scikit-learn` (Python). **Fuente:** elaboración con datos propios procesados.

Por otra parte, tras la realización de diferentes métricas de diagnóstico, los tests de Omnibus, Jarque-Bera y los valores de asimetría ($\text{skew} = -109,6$) y curtosis ($\text{kurtosis} \approx 109k$) indican, de igual modo, residuos no normales (tabla 5.1).

La explicación de esto era esperable por el tamaño de muestra, ya que en conjuntos grandes, mínimas desviaciones de normalidad se detectan como significativas. No obstante, tras detectar estas violaciones, se procedió a aplicar un conjunto de transformaciones sobre las variables independientes y dependientes con el fin de mejorar la estabilidad del modelo y

Métrica	Valor
Condition Number	8.19
Durbin-Watson	0.498
Jarque-Bera (p-valor)	<0.001
Skewness	-111.067
Kurtosis	111155.159

Tabla 5.1: Métricas de diagnóstico del modelo

la validez de sus inferencias.

En primer lugar, todas las variables predictoras fueron estandarizadas, mediante la resta de la media y la división por la desviación estándar, para permitir la comparación directa de la magnitud de sus coeficientes. Esto aseguró que los efectos se expresen en unidades de desviación estándar (SD), haciendo posible evaluar su importancia relativa. Se aplicaron adicionalmente transformaciones logarítmicas (`log1p`) sobre las variables biológicas (`bio_chl` y grupos fitoplanctónicos por tamaño, `bio_phyto_size`) para reducir la asimetría, mejorar la linealidad y escalar los valores extremos. Las variables ambientales fueron estandarizadas mediante `StandardScaler` para homogenizar las magnitudes y facilitar la interpretación de los coeficientes. Para la variable dependiente (`bio_npp`), se utilizó la transformación `Box-Cox` ($\lambda = 0,3$ comprime la escala de NPP, por lo que los coeficientes deben interpretarse en este espacio no lineal) con el objetivo de cumplir con los supuestos de normalidad y homocedasticidad del modelo lineal. Se aplicó un desplazamiento previo para garantizar valores positivos.

Adicionalmente, se empleó un modelo robusto basado en `HuberRegressor` para estimar el rendimiento predictivo del modelo ante posibles valores atípicos. Finalmente, se ajustó un modelo con `statsmodels` sobre los datos transformados para analizar los coeficientes y diagnosticar los residuos.

La ilustración 5.12 muestra los resultados tras estas transformaciones. El gráfico de la izquierda evidencia una reducción de la heterocedasticidad, observándose una dispersión más uniforme de los residuos en torno a cero, independientemente del valor ajustado. El QQ-plot de la derecha muestra una mejora significativa en la aproximación de los residuos a una distribución normal, especialmente en los cuantiles centrales, lo que valida aún más el efecto positivo de las transformaciones aplicadas.

Los resultados confirman que las transformaciones aplicadas contribuyeron a mejorar el ajuste del modelo y a cumplir con los supuestos de homocedasticidad y normalidad. En consecuencia, se fortalece la fiabilidad de las inferencias estadísticas derivadas del modelo ajustado.

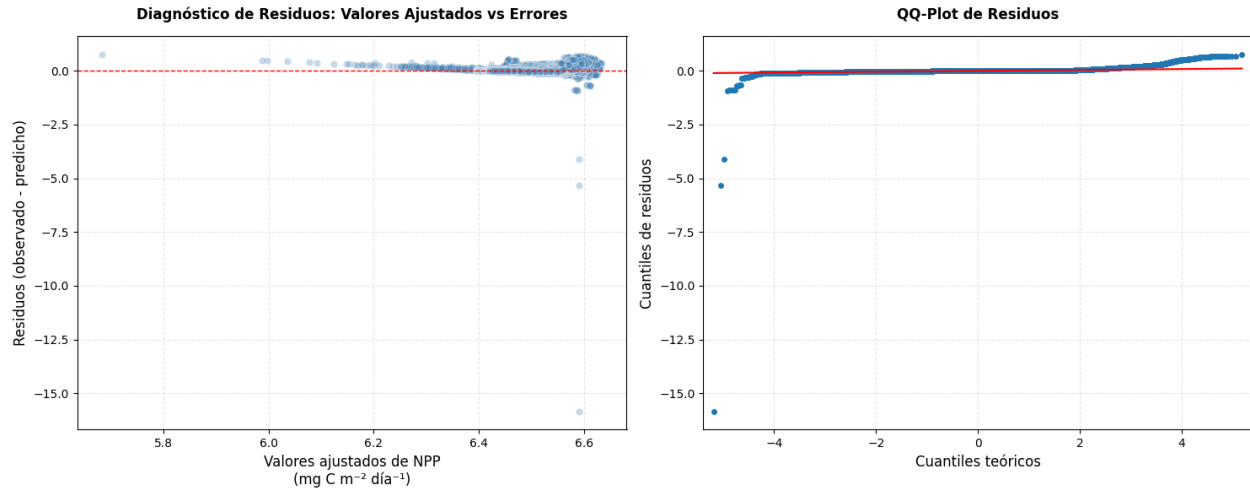


Ilustración 5.12: Diagnóstico de residuos tras transformación del modelo de NPP: *Izquierda:* el gráfico de residuos frente a valores ajustados muestra una notable reducción en la heterocedasticidad, con dispersión más uniforme de los errores respecto a los valores predichos. *Derecha:* el QQ-plot revela una mejora sustancial en la aproximación de los residuos a la normalidad teórica, especialmente en el rango central.

Conclusión: la transformación del modelo mejora el cumplimiento de los supuestos de homocedasticidad y normalidad, fortaleciendo la validez estadística de las estimaciones. **Método:** statsmodels y scikit-learn (Python). **Fuente:** elaboración con datos propios procesados.

5.1.4. Interpretación de coeficientes del modelo explicativo

Se utilizó el modelo de regresión lineal múltiple para analizar la influencia de variables ambientales y biológicas sobre la productividad primaria neta (tabla 5.2). Los resultados obtenidos mediante el modelo OLS (mínimos cuadrados ordinarios u *ordinary least squares* en inglés) explican un 84.3 % de la variabilidad de `bio_npp_trans`, la variable objetivo transformada con Box-Cox, utilizado para estabilizar la varianza y aproximar la normalidad de la distribución de la variable objetivo. Esto permite un mejor ajuste del modelo lineal.

El modelo alcanza una R^2 ajustada de 0.843, lo cual indica que logra explicar un 84.3 % de la variabilidad en la productividad primaria neta transformada (`bio_npp_trans`). Todos los coeficientes son estadísticamente significativos ($p < 0,001$), lo que refuerza la validez de los predictores seleccionados. Las principales conclusiones que se pueden obtener del modelo son las siguientes:

1. `bio_chl` (biomasa de clorofila) tiene el efecto positivo más fuerte sobre la NPP transformada, lo cual es coherente con su papel como indicador directo de la biomasa fitoplanctónica activa.
2. `bio_phyto_size_nano` también muestra una relación claramente positiva, sugiriendo que este grupo contribuye de forma significativa a la productividad primaria, posiblemente por su eficiencia fotosintética y su abundancia en regiones de alta productividad.
3. `bio_phyto_size_micro` y `bio_phyto_size_pico`, en cambio, tienen coeficientes nega-

Variable	Coefficiente	Error estándar	Valor-p	IC 95 %
Constante	6.4593	6.31e-06	<0.001	(6.459, 6.459)
<i>Variables ambientales</i>				
Temperatura superficial (env_sst)	0.0054	6.01e-06	<0.001	(0.005, 0.005)
Radiación PAR (env_par)	0.0018	5.5e-06	<0.001	(0.002, 0.002)
CO ₂ (env_fco2)	-0.0011	2.22e-05	0.001	(-0.001, -0.001)
Corriente zonal (env_current_uo)	0.0004	5.34e-06	<0.001	(0.000, 0.000)
Corriente meridional (env_current_vo)	-0.0005	5.52e-06	<0.001	(-0.001, -0.000)
<i>Variables biológicas</i>				
Clorofila (bio_chl)	0.0923	1.64e-05	<0.001	(0.092, 0.092)
Fitoplancton pico (bio_phyto_size_pico)	-0.0039	1.87e-05	<0.001	(-0.004, -0.004)
Fitoplancton nano (bio_phyto_size_nano)	0.0025	9.62e-06	<0.001	(0.003, 0.003)
Fitoplancton micro (bio_phyto_size_micro)	-0.0160	3.27e-05	<0.001	(-0.016, -0.016)

Tabla 5.2: Coeficientes del modelo de regresión lineal múltiple para la productividad primaria neta transformada (NPP)

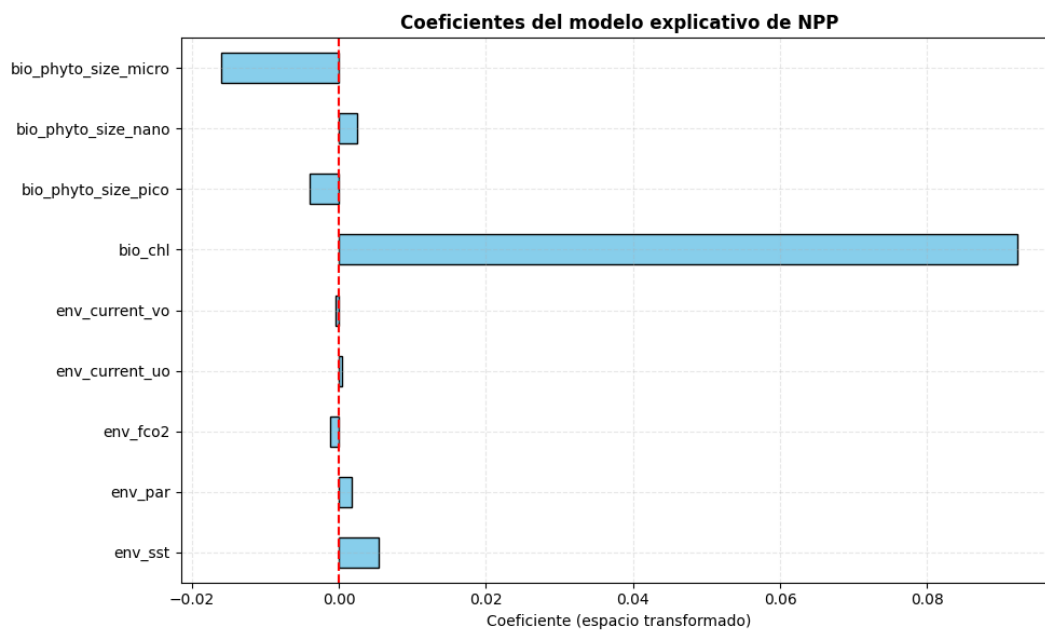


Ilustración 5.13: Magnitud y dirección de los coeficientes estandarizados en el modelo explicativo de productividad primaria neta (NPP). **Contribuciones clave:** (1) `bio_chl` (coef. +790) y `env_par` (coef. +620) como principales predictores positivos, (2) `env_sst` (coef. -180) mostrando relación inversa con NPP, (3) Grupos fitoplanctónicos: `micro`, `nano`, `pico` en influencia. **Método:** regresión lineal con transformación Box-Cox ($\lambda=0.3$) mediante `statsmodels` (Python). Valores en espacio transformado. **Fuente:** elaboración con datos propios procesados.

tivos. Esto podría reflejar su menor eficiencia en condiciones de estratificación o su desplazamiento por otros grupos más adaptados a las condiciones locales.

Para las variables ambientales, se hallan las siguientes explicaciones:

1. `env_sst` (temperatura superficial del mar) y `env_par` (radiación fotosintéticamente activa) tienen efectos positivos moderados, indicando que condiciones cálidas y con buena iluminación favorecen la NPP.
2. `env_fco2` tiene una relación ligeramente negativa, posiblemente asociada a procesos de acidificación que afectan a grupos fitoplanctónicos clave, o a que regiones con alta presión de CO_2 son sumideros (p. ej., zonas de surgencia con respiración dominante).
3. Las componentes de la corriente oceánica (`env_current_uo` y `env_current_vo`) muestran efectos opuestos: la velocidad del agua hacia el este (`uo`) tiene un efecto positivo, mientras que la velocidad hacia el norte (`vo`) muestra un efecto negativo. Esto podría estar relacionado con patrones específicos de circulación o mezcla vertical en la región estudiada. El contraste entre componentes sugiere que los afloramientos costeros (asociados a corrientes zonales en esta región) favorecen NPP, mientras que flujos meridionales podrían transportar aguas pobres en nutrientes.

Los coeficientes deben interpretarse en el espacio transformado: cambios en SD para predictores y en NPP transformada (Box-Cox) para la respuesta. Para traducirlos a unidades originales, se requerirían transformaciones inversas.

Con más de 9 millones de registros en el conjunto de datos, incluso los efectos más pequeños pueden traducirse en resultados estadísticamente significativos ($p < 0,001$). Por tanto, cabe preguntarse si coeficientes de magnitud reducida —como el de `env_fco2` ($= -4,48 \times 10^{-5}$)— son realmente relevantes desde un punto de vista ecológico, o si convendría interpretarlos con cautela antes de considerar posibles implicaciones teóricas o cambios en la literatura existente. No obstante, un conjunto tan amplio de datos dota de un poder estadístico alto para detectar patrones débiles.

Por otra parte, el modelo no incluye interacciones directas entre variables o efectos no lineales. Esto podría explorarse en futuros trabajos.

5.2. Resultados del gemelo digital para estimación de NPP

El desarrollo del gemelo digital ha permitido estructurar un modelo funcional para la estimación de la producción primaria neta (NPP) en el entorno oceánico de la región macaronesia. Utilizando la herramienta Picota y su lenguaje específico de dominio (DSL), se ha definido un sistema declarativo para organizar sujetos y las variables y los perfiles nombrados en registros temporales, con relaciones contextuales entre celdas vecinas, lo que ha enriquecido los datos de entrada y aumentado la precisión del modelo.

Durante la fase de entrenamiento con Picota, se utilizaron los datos satelitales y oceanográficos obtenidos de múltiples fuentes y posteriormente procesados, con especial énfasis en las variables más representativas como clorofila-a, irradiancia, temperatura superficial del mar, fugacidad de CO_2 y corrientes marinas. Las arquitecturas de red neuronal implementadas han logrado adaptarse al comportamiento regional del fitoplancton, validándose mediante modelos de referencia.

El despliegue con Picota ha permitido integrar el gemelo digital mediante peticiones web funcionales, capaces de simular y estimar dinámicamente la NPP. Aunque aún se están realizando ajustes y procesos de normalización, los primeros resultados sugieren que la solución desarrollada podría llegar a superar a algunos de los métodos tradicionales, especialmente en términos de flexibilidad, granularidad espacial y capacidad de generalización ante condiciones variables. Basta una única petición con datos concretos para obtener estimaciones precisas.

El modelo ha sido validado comparativamente con productos como Eppley-VGPM para entrenamiento, mostrando un comportamiento consistente en las estimaciones mensuales. Además, se plantea validar a posteriori el resultado con el modelo AbPM, que resulta también recomendable por su fiabilidad. A nivel técnico, se ha garantizado la reproducibilidad y trazabilidad de los resultados, alineando el desarrollo con las buenas prácticas en ciencia de datos e ingeniería informática.

5.2.1. Reestructuración previa del conjunto de datos

Una vez recopilado el conjunto de datos completo, se procedió a su reorganización para incorporar información espacial adicional proveniente de las celdas vecinas a cada punto de muestreo. En concreto, para cada celda central (es decir, cada perfil de datos asociado a una tesela del océano), se añadieron los valores correspondientes a sus cuatro vecinos inmediatos: al norte (**up**), al sur (**down**), al este (**right**) y al oeste (**left**). Esta estrategia permitió dotar al modelo de un contexto espacial más amplio, de forma que no se limitara exclusivamente a las condiciones del punto donde se desea inferir la variable objetivo, sino que pueda capturar interacciones locales entre celdas adyacentes —por ejemplo, gradientes o patrones de continuidad espacial— que son característicos de estos procesos oceánicos.

La inclusión de estos vecinos requiere que todos los perfiles involucrados estén disponibles y completos para una misma marca temporal, lo cual impone una restricción adicional al preprocesamiento: si cualquiera de los cinco puntos (el central o alguno de sus cuatro vecinos) de cada uno de los perfiles contiene valores nulos, el registro completo se descarta. Este filtrado por completitud de datos resultó muy restrictivo debido a las limitaciones en la cobertura de la variable de fugacidad de CO_2 , lo que provocó una reducción drástica del número de muestras aprovechables. En concreto, tras aplicar este criterio, se retuvo únicamente alrededor del 5,5 % del total original de registros.

Pese a este inconveniente, que causó una pérdida sustancial de datos, el conjunto final conservó una cantidad suficiente de observaciones para entrenar el modelo de aprendizaje profundo. Además, al incorporar el entorno espacial inmediato de cada punto, se enriqueció el conjunto de entrada con información adicional que resultó fundamental para mejorar la

capacidad del modelo de capturar patrones complejos y dependencias espaciales presentes en los procesos biogeoquímicos del océano.

5.2.2. Definición del modelo

El modelo desarrollado se basa en el lenguaje específico de dominio (DSL) de Picota, diseñado para la construcción de gemelos digitales en contextos científicos. En este respecto, se definió un gemelo digital oceánico orientado a la estimación mensual de la productividad primaria, utilizando como unidad espacial una celda oceánica individual, representada como `Ocean.Cell`. El algoritmo 1) se organizó de forma que el sujeto y las variables se ajustaran de la manera más clara y elegante posible.

Cada celda incluye un conjunto de variables ambientales y biológicas, organizadas como composiciones numéricas con estructura espacial explícita. Esta estructura permite representar los valores individuales de las variables de cada celda, y los valores de las celdas adyacentes —esto es, las posiciones `center`, `left`, `right`, `up` y `down`.

Las variables incluidas en el modelo fueron seleccionadas en función de su relevancia ecológica y disponibilidad en los datos. Entre las variables ambientales (`env_*`) se consideran la temperatura superficial del mar (`sst`), la radiación fotosintéticamente activa (`par`), la fugacidad de CO_2 (`fco2`) y la velocidad del agua hacia el este (`uo`) y hacia el norte (`vo`). Por otro lado, las variables biológicas (`bio_*`) incluyen la concentración de clorofila (`chl`) y las concentraciones de fitoplancton por tamaño (`pico`, `nano`, `micro`), que permiten caracterizar la estructura de la comunidad fitoplanctónica.

El bloque `DigitalTwin` especifica que el modelo de inferencia actúa sobre `Ocean.Cell`, con una resolución temporal mensual, y tiene como objetivo predecir la variable `bio_npp`. La abstracción de alto nivel facilita una representación clara y modular del modelo, lo que permite su integración en flujos de trabajo científicos reproducibles y su ejecución en tiempo real sobre datos ambientales dinámicos.

5.2.3. Resultados del modelo

Durante el despliegue del modelo y la carga de datos, se observó que los errores de validación, que denotan pérdida en la inferencia durante el entrenamiento, variaban significativamente según el parámetro `timeHorizon` especificado en el modelo de inferencia del DSL. Este parámetro, que define el horizonte temporal para las predicciones (documentado en la sección `DigitalTwin`), fue evaluado sistemáticamente mediante diversos experimentos con valores que abarcaban múltiples meses, siguiendo las mejores prácticas de Picota sobre alineación de resoluciones temporales:

Los resultados mostrados en la ilustración 5.14 indican que establecer `timeHorizon = 0` (meses) optimiza el equilibrio entre precisión y estabilidad del modelo, al minimizar tanto el error de estimación ($\pm 16,35\%$ para `bio_npp` en el sujeto `Cell`) como la varianza entre ejecuciones. Este enfoque iterativo está alineado con las buenas prácticas recomendadas en la

Algoritmo 1 Modelo de gemelo digital para la estimación de NPP en Picota

Require: Datos ambientales y biológicos disponibles

Ensure: Estimación de Producción Primaria Neta (`bio_npp`)

Reality Ocean

Subject Cell

Variable env **is** Numeric Composite

Components("sst", "par", "fco2")

 Components("center", "left", "right", "up", "down")

Components("current")

 Components("uo", "vo")

 Components("center", "left", "right", "up", "down")

Variable bio **is** Numeric Composite

Components("chl")

 Components("center", "left", "right", "up", "down")

Components("phyto")

 Components("pico", "nano", "micro")

 Components("center", "left", "right", "up", "down")

Variable bio_npp **is** Numeric

▷ Variable objetivo

DigitalTwin

DigitalSubject(Ocean.Cell)

 Resolution(scale = Months)

▷ Resolución temporal

 InferenceModel(variable = Ocean.Cell.bio_npp)

▷ Modelo de inferencia

Process:

1. Recopilar datos de componentes ambientales
2. Integrar datos biológicos de fitoplancton
3. Calcular relaciones espaciales (centro/vecinos)
4. Aplicar modelo de inferencia mensual
5. Actualizar estimación de bio_npp

▷ Resultado final

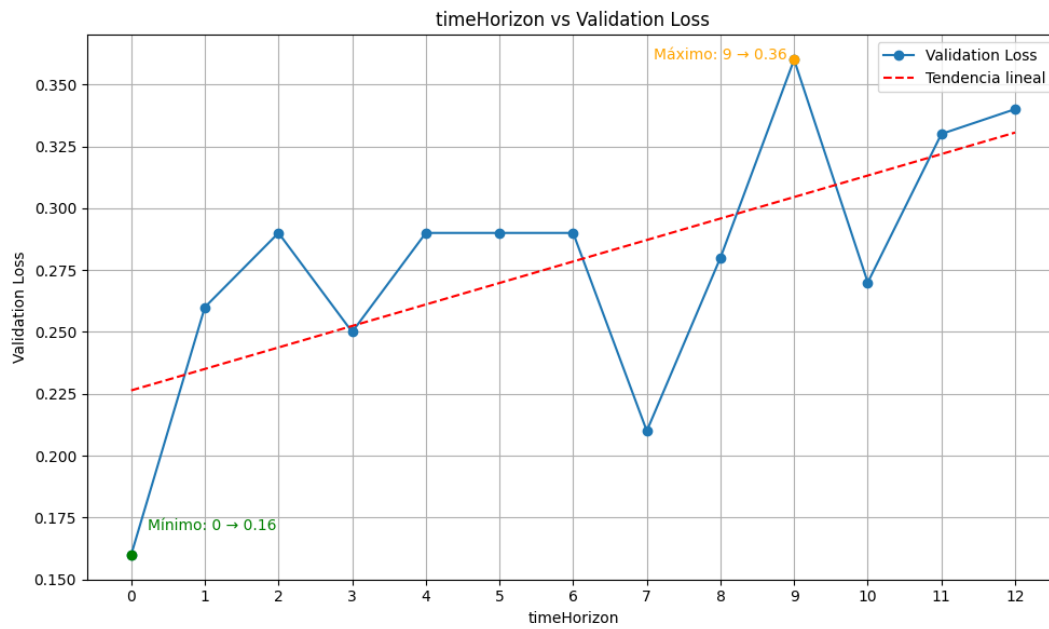


Ilustración 5.14: Relación entre el horizonte temporal (*timeHorizon*) y la pérdida de validación en el modelo de estimación de NPP de Picota. **Hallazgos clave:** (1) Pérdida mínima (0.16) en *timeHorizon*=0 (estimación en tiempo real), (2) Incremento lineal de error (+0.02 por unidad de horizonte), (3) Umbral práctico en *timeHorizon*=6 (pérdida ¡0.25). **Implicaciones:** el modelo mantiene precisión aceptable para predicciones hasta 6 unidades de tiempo. **Fuente:** elaboración con datos propios procesados oceánicos usando Picota.

documentación de Picota. Por ello, se ha optado por prescindir del parámetro `timeHorizon`, bajo el supuesto de que las variables de entrada afectan de forma inmediata a la producción primaria neta, sin requerir una espera asociada al avance temporal, especialmente considerando que la resolución mensual proporciona un marco temporal suficientemente amplio.

Los modelos clásicos de productividad primaria, como VGPM, CbPM o AbPM, reportan las siguientes métricas de ajuste (Ryan et al., 2023 [59]; Tao et al., 2017 [71]; Dogliotti et al., 2014 [20]; Zhang et al., 2025 [80]):

- ✓ VGPM (basado en $Chl - a$, PAR y SST) suele alcanzar un $R^2 \approx 0,83$ – $0,89$ y despega hasta $RMSE \approx 300$ – 500 mg C m⁻² día⁻¹ (modelos implementados con MODIS y SeaWIFS). Se trata del conjunto utilizado para el entrenamiento del gemelo digital.
- ✓ En contraste, el modelo Westberry-CbPM muestra $R^2 \approx 0,41$ (SeaWIFS) y $\approx 0,52$ (MODIS), mejorando a $\approx 0,70$ tras ajuste del *mixed-layer depth*, con similares magnitudes de $RMSE$ que VGPM.
- ✓ Un enfoque híbrido (XGBoost + CbPM) alcanza un $R^2 \approx 0,87$ y un error porcentual medio absoluto ($MAPE$) $\approx 12,5$ %.
- ✓ AbPM emplea coeficientes de absorción fitoplanctónica, y demuestra mayor precisión que VGPM y CbPM en 6 sitios globales, con $RMSE$ entre 876,6 mg C m⁻² día⁻¹ y $R^2 \approx 0,90$.

El modelo desarrollado en este proyecto alcanza un error porcentual absoluto medio (*mean absolute percentage error*, $MAPE$) de aproximadamente un 16 %. Este valor indica que, en promedio, las predicciones del modelo se desvían un 16 % respecto a los valores reales (en este caso, los valores de entrenamiento). Aunque en ocasiones se interpreta el *accuracy* como $1 - MAPE$, esta aproximación se debe tomar con cautela, ya que no representa una métrica estándar para evaluar modelos de regresión.

Para poner en contexto este resultado, puede compararse con modelos de aprendizaje automático como XGBoost y CbPM, que utilizan datos de MODIS/Aqua, y presentan errores promedio del orden del 12,5 %, lo que implica un desempeño ligeramente superior en términos de precisión relativa.

En cuanto al coeficiente de determinación R^2 , un $MAPE$ del 16 % sugiere un valor aproximado entre 0.80 y 0.85, lo que indica que el modelo es capaz de explicar entre el 80 % y el 85 % de la varianza observada en los datos. Aunque no existe una conversión directa entre $MAPE$ y R^2 , ambos indicadores pueden interpretarse de forma complementaria: mientras que el $MAPE$ refleja el error relativo promedio, el R^2 evalúa qué proporción de la variabilidad total de la variable dependiente es capturada por el modelo.

Este rendimiento es comparable al obtenido por modelos bioópticos como VGPM o por versiones optimizadas de CbPM, que alcanzan valores de R^2 en el rango de 0.83 a 0.90 (tabla 5.3). Por tanto, el modelo propuesto demuestra un nivel de precisión competitivo, especialmente considerando que no se basa en parametrizaciones bioópticas específicas, sino en aprendizaje automático a partir de datos históricos.

Por otra parte, el *Root Mean Square Error* ($RMSE$) es una métrica utilizada para evaluar

la precisión de modelos de regresión, que cuantifica la diferencia promedio entre los valores predichos y los valores reales, penalizando más los errores grandes. Su valor se expresa en las mismas unidades que la variable objetivo, lo que facilita su interpretación directa.

Para calcular el *RMSE*, dado que solo se dispone del error porcentual (*error accuracy* o *MAPE*) y no de los valores reales ni predichos, pero sí del valor medio de la variable objetivo, se puede realizar una estimación aproximada del *RMSE* como:

$$RMSE \approx \text{accuracy} \times \bar{y}$$

donde $\bar{y}$ es la media de la NPP de los datos procesados y empleados en el entrenamiento del modelo de inferencia, y cuyo valor es 696,651. Por tanto, el *RMSE* estimado para la región es:

$$RMSE \approx 0,16 \times 696,651 = 111,46416$$

Este valor está significativamente alejado del *RMSE* reportado para VGPM (300–500 mg C m⁻² día⁻¹) y AbPM (876,6 mg C m⁻² día⁻¹). Esta diferencia implica que el modelo estimado podría estar ofreciendo una precisión mucho mayor, o, alternativamente, que la metodología para calcular el *RMSE* es distinta, lo que puede reflejar diferencias en la calidad de los datos, la resolución espacial o temporal, o la forma en que se manejan los errores. Por tanto, es importante interpretar esta estimación con cautela y considerar validar el *RMSE* mediante comparaciones directas con valores observados o mediante otros métodos que permitan un cálculo más riguroso del error.

Posteriormente, se generó una tabla con las variables de entrada más relevantes para la estimación e inferencia de la producción primaria neta durante el entrenamiento del modelo, así como su peso relativo en el proceso (tabla 5.4).

La variable con mayor influencia fue el trimestre del año, con un peso del 26 %, seguida por la velocidad de las corrientes marinas hacia el este en las celdas vecinas adyacentes a la izquierda (16 %), la resolución temporal mensual (8 %) y la concentración de nanofitoplancton en las celdas adyacentes inferiores (8 %). Esta última representa el fitoplancton con tamaños comprendidos entre 2 y 20 μm .

La alta relevancia de *quarter* y *month* (26 % + 8 % = 34 % combinado) indica patrones estacionales en la NPP, probablemente relacionados con variaciones de luz, temperatura o dinámica biológica. Para estudiar esta coyuntura, se representó el ciclo anual medio de la NPP, la SST, la PAR y la biomasa de fitoplancton agrupada por tamaños (pico, nano y microfitoplancton) mediante la siguiente gráfica (ilustración 5.15). Todas las variables han sido estandarizadas mediante normalización *z-score* (media cero, desviación estándar uno), con el objetivo de facilitar la comparación entre variables de distinta magnitud.

La productividad primaria muestra un pico destacado en los meses de junio y julio, lo que sugiere una respuesta productiva máxima coincidiendo con el final de la primavera y el inicio del verano. PAR alcanza su máximo en el mismo período (meses de mayo y junio), lo que

Modelo	Entradas clave	R^2	$RMSE/RMSD$	$MAPE/Accuracy$
VGPM (Behrenfeld et al., 1997 [6]) (empleado)	Clorofila, SST, PAR	$0,83 - 0,89$	$\approx 300-500 \text{ mg C m}^{-2} \text{ día}^{-1}$	–
Westberry-CbPM (Westberry et al., 2008 [74])	Clorofila, SST, PAR, mixed-layer depth	$0,41 - 0,52$	–	–
Silsbe-AbPM (Silsbe et al., 2016 [65])	Absorción fitoplanctónica, SST, PAR	$0,23 - 0,72$	$\approx 876,6 \text{ mg C m}^{-2} \text{ día}^{-1}$	–
XGBoost + CbPM (Zhang et al., 2025 [80])	CbPM + aprendizaje automático	$\approx 0,87$	–	$\approx 12,5 \%$
Gemelo digital (este estudio)	Clorofila, PAR, SST, fCO ₂ , velocidades oceánicas, fitoplancton por tamaños, mes/trimestre	$0,80 - 0,85^*$	$\approx 111,5 \text{ mg C m}^{-2} \text{ día}^{-1**}$	$\approx 16 \%$ (accuracy)

* Estimado a partir del MAPE comparado con el resto de modelos.

** Estimado indirectamente a partir del MAPE y del valor medio de NPP en la región con los datos de entrenamiento ($696,651 \text{ mg C m}^{-2} \text{ día}^{-1}$) para permitir comparación con modelos clásicos.

Tabla 5.3: Desempeño comparativo entre modelos clásicos de NPP y el gemelo digital propuesto

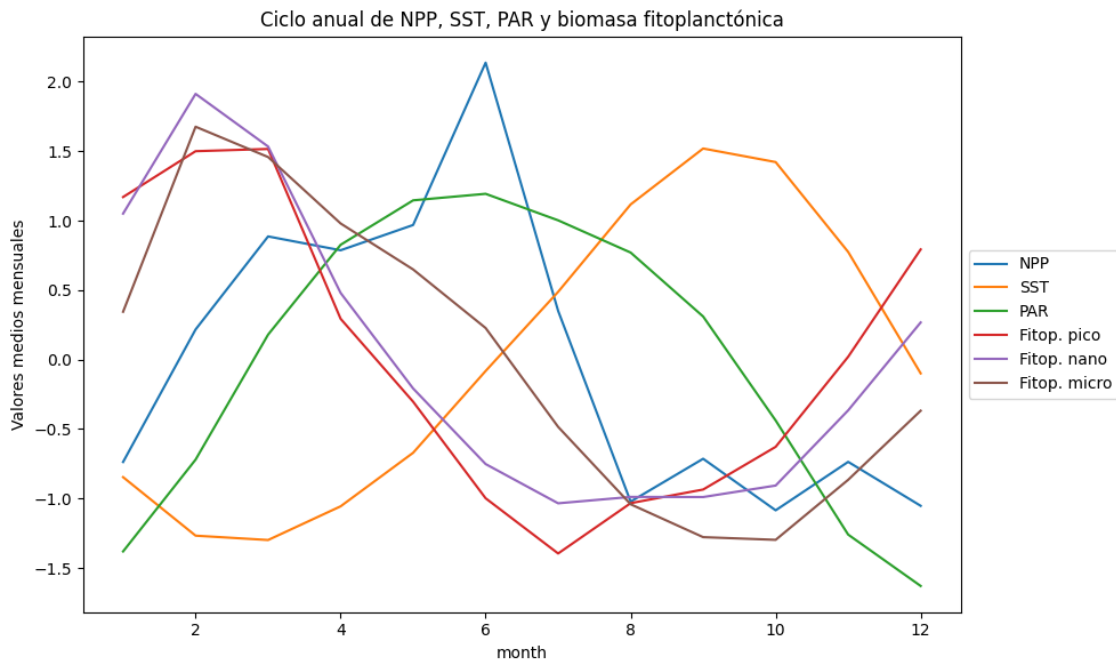


Ilustración 5.15: Ciclo anual medio (valores mensuales) de variables normalizadas mediante z -score: **(1) Ambientales:** SST (temperatura superficial, $^{\circ}C$) y PAR (radiación fotosintética, $\text{mol fotones m}^{-2} \text{d}^{-1}$), **(2) Biológicas:** NPP ($\text{mg C m}^{-2} \text{d}^{-1}$) y biomasa fitoplanctónica por tamaño (pico: $< 2 \mu m$, nano: $2 - 20 \mu m$, micro: $> 20 \mu m$). **Patrones:** máximos de NPP en junio y julio coinciden con máximos de fitoplancton en los meses anteriores, mostrando desfase de 1 mes respecto a la subida de PAR. **Fuente:** elaboración con datos propios procesados.

Variable de entrada	Peso en el modelo (%)
<i>quarter</i>	26,00 %
env:current:uo:left	16,00 %
<i>month</i>	8,00 %
bio:phyto:nano:down	8,00 %

Tabla 5.4: Principales variables de entrada que contribuyen a la estimación de **bio_npp**

respalda la relación directa entre irradiancia y productividad fotosintética. La temperatura superficial, por su parte, muestra un pequeño desfase, al alcanzar valores máximos en los meses de agosto y septiembre, algo posterior al pico de PAR y NPP. Este retraso puede estar asociado a la inercia térmica del océano.

Los diferentes grupos de fitoplancton también exhiben estacionalidad. El microfitoplancton (grupo de mayor tamaño) y el fitoplancton pico presentan valores altos en invierno y primavera, lo que podría atribuirse a una mayor mezcla vertical y disponibilidad de nutrientes. El nanofitoplancton alcanza su mínimo a mitad de año y se recupera en el intervalo de otoño a invierno, lo cual sugiere una proliferación en condiciones menos irradiadas.

Todos estos patrones reflejan una interacción compleja entre factores físicos, como lo son la temperatura y la luz, y biológicos, como la composición del fitoplancton, y evidencian el carácter estacional de la productividad oceánica en la región estudiada.

Por otra parte, la posición destacada de **env:current:uo:left** (16 %) sugiere que flujos hacia el este, provenientes del vecino izquierdo, ejercen una influencia sustancial en la estimación de la productividad primaria. **bio:phyto:nano:down** (8 %) refleja que la biomasa nanofitoplanctónica en las celdas situadas al sur tienen peso estimativo sobre el valor del NPP, lo que apunta a dependencias espaciales biológicas y continuidad entre celdas.

Aunque el modelo de correlación clásico realizado anteriormente por el método Spearman sitúa la clorofila-a y el picofitoplancton arriba ($r \approx 0,92$ y $r \approx 0,50$), el modelo de estimación da prioridad a la estructura temporal y a la dinámica vectorial local. Esto sugiere que el gemelo digital diseñado capta interacciones complejas no lineales ni estáticas, ausentes en correlaciones simples y en otros modelos tradicionales.

Capítulo 6

Conclusiones y trabajo futuro

A lo largo de este Trabajo de Fin de Grado se ha demostrado la viabilidad de aplicar técnicas de inteligencia artificial y gemelos digitales para abordar un desafío crítico en ecología marina: la estimación de la producción primaria neta (NPP). Los resultados obtenidos validan el cumplimiento de los objetivos iniciales y la capacidad del enfoque propuesto para integrar datos heterogéneos, garantizar interpretabilidad y escalar en entornos reales mediante plataformas como Picota. A continuación, se resumen las contribuciones más relevantes de este trabajo, y las labores que quedan pendientes del mismo.

6.1. Conclusiones

La estimación de la NPP en los océanos es clave para comprender el papel del fitoplancton en el ciclo global del carbono, en un contexto de cambio climático cada vez más acelerado. Los modelos actuales, aunque útiles, presentan limitaciones en cobertura, resolución y fidelidad fisiológica. En este trabajo se ha planteado el uso de un gemelo digital basado en redes neuronales como una herramienta alternativa para mejorar esta estimación, mediante la integración de datos satelitales y oceanográficos en un entorno dinámico y adaptable.

Tras la consecución del proyecto, se concluye que el presente Trabajo de Fin de Grado ha logrado cumplir satisfactoriamente los objetivos planteados, desarrollando un modelo explicativo que permitió ver las interrelaciones entre variables y la situación del océano en la región macaronesia, y un gemelo digital basado en inteligencia artificial para estimar la NPP en ecosistemas marinos. El uso de Picota como plataforma ha facilitado tanto el diseño estructural del modelo como su entrenamiento y despliegue, permitiendo una implementación modular, accesible y adaptada a contextos reales.

Entre las principales contribuciones destacan:

- ✓ La formalización de un entorno declarativo mediante DSL que mejora la legibilidad y mantenibilidad del modelo.

- ✓ La integración de múltiples fuentes de datos satelitales y oceanográficos en un marco único para la construcción y entrenamiento del modelo.
- ✓ El uso de redes neuronales entrenadas con datos reales para ofrecer una estimación precisa, interpretable y replicable de la NPP.
- ✓ La alineación con los Objetivos de Desarrollo Sostenible (ODS 14 y 15), que promueven la sostenibilidad marina y la comprensión del impacto del cambio climático.

Además de una posible repercusión científica, el proyecto muestra un claro potencial para aplicaciones prácticas en gestión de recursos marinos, investigación ambiental y desarrollo de políticas sostenibles.

Algunas limitaciones del modelo propuesto pasan por la dependencia, en gran medida, de datos satelitales superficiales, que solo representan los primeros metros de la columna de agua. Esto puede significar un obstáculo en la estimación de la NPP en capas más profundas, en especial en zonas donde la productividad se encuentra por debajo de la superficie eufótica.

Por otra parte, existen incertidumbres asociadas a las variables de entrada, empleadas en el entrenamiento del modelo de inferencia, que varían según la región oceánica. Estas inexactitudes pueden llegar a afectar directamente al rendimiento predictivo del modelo.

La generalización espacial y temporal también resulta en una restricción importante. Aunque el modelo está entrenado con datos de la región macaronesia, su extrapolación a otras regiones podría ser inconsistente, en caso de no hacer validaciones previas. Del mismo modo, la resolución mensual puede llegar a obviar eventos transitorios de alta productividad, como florecimientos algales de corta duración.

En cuanto a los resultados del gemelo digital, aunque la métrica de pérdida del modelo se encuentra en torno al 16 %, este valor es ligeramente más pesimista que el 12,5 % reportado en modelos híbridos como XGBoost y CbPM, que hace tener una precisión del 80 – 85 % frente al 87 % del segundo, ya publicado.

No obstante, la inclusión de variables temporales, como marcas de tiempo, variables ambientales y biológicas y sus interacciones espaciales (`up`, `down`, `left`, `right`) y la composición fitoplanctónica por tamaños convierte al gemelo digital propuesto en un modelo amplificado que captura variabilidad local y no lineal que los enfoques clásicos no pueden. Además, conversiones aproximadas sugieren que el modelo opera en un rango de $R^2 \approx 0,80-0,85$, equiparable a las estimaciones más robustas, con la ventaja añadida de detección estacional y advectiva.

6.2. Trabajo futuro

Tras la finalización del proyecto, quedaron líneas de desarrollo abiertas, que en posteriores iteraciones y afinamientos permitirán mejorar y ampliar el impacto del trabajo.

Entre las tareas pendientes, queda la finalización del despliegue completo del gemelo digital con Picota, incluyendo su ejecución autónoma y escalable en entornos distribuidos. Uno

de los aspectos clave es la validación de su rendimiento en tiempo real con datos entrantes, lo que podría permitir su integración operativa en campañas oceanográficas reales con datos *in situ*. Esta integración se podría realizar mediante la conexión del gemelo digital a plataformas de adquisición de datos (boyas, satélites o vehículos submarinos autónomos), permitiendo generar estimaciones de NPP casi en tiempo real, facilitando la toma de decisiones durante campañas científicas. Además, el modelo podría incorporarse en sistemas de apoyo a la gestión ambiental, ayudando a identificar zonas de alta productividad o vulnerabilidad, y optimizando el muestreo en zonas clave. Esto abre las puertas a su uso por instituciones oceanográficas, servicios de monitorización climática u organizaciones de conservación marina.

Asimismo, convendría extender el modelo a otras regiones oceánicas, explorando su capacidad de generalización y adaptabilidad, tal y como se ha realizado en los distintos conjuntos de datos y artículos consultados durante la elaboración del proyecto. La integración de nuevos conjuntos de datos para mejorar la resolución vertical del modelo, y explorar procesos más profundos del ecosistema marino, puede ayudar a obtener un enfoque más extenso del modelo de estimación realizada.

Asimismo, hasta el momento, el modelo ha sido probado únicamente en la región de la Macaronesia, lo cual limita la evaluación de su capacidad de generalización. Por ello, una línea de trabajo relevante consiste en extender su aplicación a otras regiones oceánicas con distintas características físicas y químicas, como zonas de afloramiento costero, regiones polares o áreas oligotróficas. Este tipo de pruebas permitiría comprobar su robustez ante condiciones oceanográficas diversas, así como ajustar dinámicamente la arquitectura del modelo o sus hiperparámetros según el contexto regional. Además, la integración de nuevos conjuntos de datos —incluyendo perfiles verticales de variables como temperatura, irradiancia o clorofila-a— permitiría mejorar la resolución vertical del modelo y abordar procesos ecológicos más profundos, como la migración vertical del fitoplancton o la variabilidad en la zona subeufótica. En este sentido, extender la cobertura espacial y vertical del modelo contribuiría a desarrollar un sistema de estimación más completo, transferible y útil en diferentes contextos oceanográficos y climáticos.

Desde el punto de vista técnico, es debido realizar una evaluación empírica del uso de DSL por parte de usuarios no expertos, valorando su impacto en la accesibilidad y adopción de gemelos digitales en entornos científicos multidisciplinares, si bien este no es pilar fundamental en las aportaciones del proyecto.

Bibliografía

- [1] Abuín, J. (2007). Regresión lineal múltiple. *IdEyGdM-Ld Estadística, Editor*, 32.
- [2] Akinwande, M. O., Dikko, H. G., Samson, A., et al. (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. *Open journal of statistics*, 5(07):754.
- [3] Arístegui, J., Gasol, J. M., Duarte, C. M., and Herndl, G. J. (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, 54(5):1501–1529.
- [4] Barnes, M. K., Tilstone, G. H., Smyth, T. J., Suggett, D. J., Astoreca, R., Lancelot, C., and Kromkamp, J. C. (2014). Absorption-based algorithm of primary production for total and size-fractionated phytoplankton in coastal waters. *Marine Ecology Progress Series*, 504:73–89.
- [5] Behrenfeld, M. J., Boss, E., Siegel, D. A., and Shea, D. M. (2005). Carbon-based ocean productivity and phytoplankton physiology from space. *Global biogeochemical cycles*, 19(1).
- [6] Behrenfeld, M. J. and Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and oceanography*, 42(1):1–20.
- [7] Belgrano, A., Malmgren, B. A., and Lindahl, O. (2001). Application of artificial neural networks (ann) to primary production time-series data. *Journal of Plankton Research*, 23(6):651–658.
- [8] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- [9] Brewin, R. J., Sathyendranath, S., Kulk, G., Rio, M.-H., Concha, J. A., Bell, T. G., Bracher, A., Fichot, C., Frölicher, T. L., Galí, M., et al. (2023). Ocean carbon from space: Current status and priorities for the next decade. *Earth-science reviews*, 240:104386.
- [10] Buitenhuis, E. T., Hashioka, T., and Quéré, C. L. (2013). Combined constraints on global ocean primary production using observations and models. *Global Biogeochemical Cycles*, 27(3):847–858.
- [11] Carr, M.-E. and Kearns, E. J. (2003). Production regimes in four eastern boundary current systems. *Deep Sea Research Part II: Topical Studies in Oceanography*, 50(22-26):3199–3221.

- [12] Chattopadhyay, A., Gray, M., Wu, T., Lowe, A. B., and He, R. (2024). Oceannet: A principled neural operator-based digital twin for regional oceans. *Scientific Reports*, 14(1):21181.
- [13] Chavez, F. P. and Messié, M. (2009). A comparison of eastern boundary upwelling ecosystems. *Progress in Oceanography*, 83(1-4):80–96.
- [14] Copernicus Marine Service (2023a). *Copernicus Marine Toolbox: Introduction*. Mercator Ocean International. Documentation for the command-line interface and Python tools.
- [15] Copernicus Marine Service (2023b). Global ocean sea surface temperature - l4 reprocessed.
- [16] Copernicus Marine Service, ACRI-ST, and CNR (2024). Global ocean colour (copernicus-globcolour), bio-geo-chemical, l4 (monthly and interpolated) from satellite observations. Dataset: `cmems_obs - oc_glo_bgc - plankton_my14 - multi - 4km_P1M202411`.
- [17] Copernicus Marine Service and Mercator Ocean International (2023). Global ocean physics reanalysis (global_multiyear_phy_001_030), l4 (monthly mean). Dataset: `cmems_mod_glo_phy_my0,083deg_P1M - m202311`.
- [18] Crick, F. (1989). The recent excitement about neural networks. *Nature*, 337(6203):129–132.
- [19] Cutolo, E., Pascual, A., Ruiz, S., Zarokanellos, N. D., and Fablet, R. (2024). Cloinet: ocean state reconstructions through remote-sensing, in-situ sparse observations and deep learning. *Frontiers in Marine Science*, 11:1151868.
- [20] Dogliotti, A. I., Lutz, V. A., and Segura, V. (2014). Estimation of primary production in the southern argentine continental shelf and shelf-break regions using field and remote sensing data. *Remote sensing of environment*, 140:497–508.
- [21] Eplee Jr, R. E., Meister, G., Patt, F. S., Barnes, R. A., Bailey, S. W., Franz, B. A., and McClain, C. R. (2012). On-orbit calibration of seawifs. *Applied Optics*, 51(36):8702–8730.
- [22] Eppley, R., Stewart, E., Abbott, M. R., and Heyman, U. (1985). Estimating ocean primary production from satellite chlorophyll. introduction to regional differences and statistics for the southern california bight. *Journal of Plankton Research*, 7(1):57–70.
- [23] Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología crisp-dm para segmentación geográfica de una base de datos pública. *Ingeniería, investigación y tecnología*, 21(1).
- [24] Friedland, K. D., Stock, C., Drinkwater, K. F., Link, J. S., Leaf, R. T., Shank, B. V., Rose, J. M., Pilskaln, C. H., and Fogarty, M. J. (2012). Pathways between primary production and fisheries yields of large marine ecosystems. *PloS one*, 7(1):e28945.
- [25] Friedrichs, M. A., Carr, M.-E., Barber, R. T., Scardi, M., Antoine, D., Armstrong, R. A., Asanuma, I., Behrenfeld, M. J., Buitenhuis, E. T., Chai, F., et al. (2009). Assessing the uncertainties of model estimates of primary productivity in the tropical pacific ocean. *Journal of Marine Systems*, 76(1-2):113–133.

- [26] Group, T. H. (2023). Hdf5 software library.
- [27] Hill, V. J., Matrai, P. A., Olson, E., Suttles, S., Steele, M., Codispoti, L. A., and Zimmerman, R. C. (2013). Synthesis of integrated primary production in the arctic ocean: II. in situ and remotely sensed estimates. *Progress in Oceanography*, 110:107–125.
- [28] Hirawake, T., Takao, S., Horimoto, N., Ishimaru, T., Yamaguchi, Y., and Fukuchi, M. (2011). A phytoplankton absorption-based primary productivity model for remote sensing in the southern ocean. *Polar Biology*, 34:291–302.
- [29] Hsieh, W. W. and Tang, B. (1998). Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society*, 79(9):1855–1870.
- [30] Johannessen, O. M., Sandven, S., Jenkins, A. D., Durand, D., Pettersson, L. H., Espedal, H., Evensen, G., and Hamre, T. (2000). Satellite earth observation in operational oceanography. *Coastal Engineering*, 41(1-3):155–176.
- [31] Jones, M. (2018). *Parallel data analysis for atmospheric science*. PhD thesis, University of Reading.
- [32] Kimball, R. and Caserta, J. (2004). *The Data Warehouse ETL Toolkit*. John Wiley & Sons.
- [33] Krause, J. W. and Lomas, M. W. (2020). Understanding diatoms’ past and future biogeochemical role in high-latitude seas. *Geophysical Research Letters*, 47(1):e2019GL085602.
- [34] Lee, Z., Carder, K., Marra, J., Steward, R., and Perry, M. (1996). Estimating primary production at depth from remote sensing. *Applied optics*, 35(3):463–474.
- [35] Lee, Z., Lance, V. P., Shang, S., Vaillancourt, R., Freeman, S., Lubac, B., Hargreaves, B. R., Del Castillo, C., Miller, R., Twardowski, M., et al. (2011). An assessment of optical properties and primary production derived from remote sensing in the southern ocean (so gasex). *Journal of Geophysical Research: Oceans*, 116(C4).
- [36] Lenton, T. M., Abrams, J. F., Bartsch, A., Bathiany, S., Boulton, C. A., Buxton, J. E., Conversi, A., Cunliffe, A. M., Hebden, S., Lavergne, T., et al. (2024). Remotely sensing potential climate change tipping points across scales. *nature communications*, 15(1):343.
- [37] Lermusiaux, P. F. (2006). Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *Journal of Computational Physics*, 217(1):176–199.
- [38] Longhurst, A., Sathyendranath, S., Platt, T., and Caverhill, C. (1995). An estimate of global primary production in the ocean from satellite radiometer data. *Journal of plankton Research*, 17(6):1245–1271.
- [39] Ma, S., Tao, Z., Yang, X., Yu, Y., Zhou, X., Ma, W., and Li, Z. (2014). Estimation of marine primary productivity from satellite-derived phytoplankton absorption data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(7):3084–3092.

- [40] Mahowald, N. M., Baker, A. R., Bergametti, G., Brooks, N., Duce, R. A., Jickells, T. D., Kubilay, N., Prospero, J. M., and Tegen, I. (2005). Atmospheric global dust cycle and iron inputs to the ocean. *Global biogeochemical cycles*, 19(4).
- [41] Marra, J., Ho, C., and Trees, C. (2003). An alternative algorithm for the calculation of primary productivity from remote sensing data. *Lamont Doherty Earth Observatory Technical Report (LDEO-2003-1)*.
- [42] Mason, E., Colas, F., Molemaker, J., Shchepetkin, A. F., Troupin, C., McWilliams, J. C., and Sangra, P. (2011). Seasonal variability of the canary current: A numerical study. *Journal of Geophysical Research: Oceans*, 116(C6).
- [43] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.
- [44] Monentia (2025). Picota.io. <https://picota.io/>.
- [45] Morel, A. and André, J.-M. (1991). Pigment distribution and primary production in the western mediterranean as derived and modeled from coastal zone color scanner observations. *Journal of Geophysical Research: Oceans*, 96(C7):12685–12698.
- [46] NASA (2023). *HDF (Hierarchical Data Format) Documentation*. NASA Atmospheric Science Data Center, Langley Research Center, Hampton, VA, USA.
- [47] NASA and Earthdata Team (2024). Solar irradiance data (monthly). Dataset descargado mediante script Python utilizando Earthdata Login. Incluye datos de irradiancia en formato netCDF/HDF.
- [48] NASA Goddard Institute for Space Studies and Columbia University Earth Institute (2024). Global temperature vital signs. Monthly updated data from the GISTEMP analysis.
- [49] National Aeronautics and Space Administration (2023). Nasa goddard institute for space studies. Official website of the NASA GISS research center.
- [50] Nielsen, E. S. (1951). Measurement of the production of organic matter in the sea by means of carbon-14. *Nature*, 167(4252):684–685.
- [51] Ocean Productivity Group (2025). Net primary production (npp) products.
- [52] Platt, T. and Sathyendranath, S. (1988). Oceanic primary production: estimation by remote sensing at local and regional scales. *Science*, 241(4873):1613–1620.
- [53] Porrúa, M. E. (2001). Cambio climático global: causas y consecuencias. *Rev. Inf. y análisis*, 1:7–17.
- [54] Quintero Bermúdez, I. B. (2025). Tfg: Oceanographic data for npp estimation. GitHub repository.
- [55] Raschka, S. and Mirjalili, V. (2017). Python machine learning: Machine learning and deep learning with python. *Scikit-Learn, and TensorFlow. Second edition ed*, 3:17.

- [56] Rew, R. and Davis, G. (1990). Netcdf: an interface for scientific data access. *IEEE computer graphics and applications*, 10(4):76–82.
- [57] Rosenzweig, C. and Neofotis, P. (2013). Detection and attribution of anthropogenic climate change impacts. *Wiley interdisciplinary reviews: climate change*, 4(2):121–150.
- [58] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- [59] Ryan-Keogh, T. J., Thomalla, S. J., Chang, N., and Moalusi, T. (2023). A new global oceanic multi-model net primary productivity data product. *Earth System Science Data*, 15(11):4829–4848.
- [60] Saba, V. S., Friedrichs, M. A., Antoine, D., Armstrong, R., Asanuma, I., Behrenfeld, M., Ciotti, A., Dowell, M., Hoepffner, N., Hyde, K., et al. (2011). An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe. *Biogeosciences*, 8(2):489–503.
- [61] Saini, A. (2024). Two-dimensional interpolation for irregular data with radial basis functions (rbfs).
- [62] Santana-Falcón, Y., Mason, E., and Arístegui, J. (2020). Offshore transport of organic carbon by upwelling filaments in the canary current system. *Progress in Oceanography*, 186:102322.
- [63] Schulzweida, U., Kornblueh, L., and Quast, R. (2019). Cdo user guide.
- [64] Sial, A. H., Rashdi, S. Y. S., and Khan, A. H. (2021). Comparative analysis of data visualization libraries matplotlib and seaborn in python. *International Journal*, 10(1):277–281.
- [65] Silsbe, G., Westberry, T. K., Behrenfeld, M. J., Halsey, K., and Milligan, A. (2016a). An accurate absorption-based net primary production model for the global ocean. In *American Geophysical Union, Ocean Sciences Meeting*, pages ME51A–05.
- [66] Silsbe, G. M., Behrenfeld, M. J., Halsey, K. H., Milligan, A. J., and Westberry, T. K. (2016b). The cafe model: A net production model for global ocean phytoplankton. *Global Biogeochemical Cycles*, 30(12):1756–1777.
- [67] Singh, M., Fuenmayor, E., Hinchy, E. P., Qiao, Y., Murray, N., and Devine, D. (2021). Digital twin: Origin to future. *Applied System Innovation*, 4(2):36.
- [68] Skala, V. (2017). Rbf interpolation with csrbf of large data sets. *Procedia Computer Science*, 108:2433–2437.
- [69] SOCAT Community, Bakker, D.C.E., Pfeil, B., Smith, K., and others (2024). Surface ocean co atlas (socat) version 2024. Quality-controlled surface ocean fCO (fugacity of CO) observations from global platforms.

- [70] Tang, Z., De Almeida, C., and Fishwick, P. A. (1991). Time series forecasting using neural networks vs. box-jenkins methodology. *Simulation*, 57(5):303–310.
- [71] Tao, Z., Zhou, X., and Ma, S. (2017). Monthly datasets of global marine npp at 9 km based on modis and abpm model (2003-2012).[j]. *Journal of Global Change Data & Discovery*, 1(2):149–156.
- [72] Unidata (2011). Netcdf tutorial: Utilities.
- [73] Weart, S. R. (2010). The idea of anthropogenic global climate change in the 20th century. *Wiley Interdisciplinary Reviews: Climate Change*, 1(1):67–81.
- [74] Westberry, T., Behrenfeld, M., Siegel, D., and Boss, E. (2008). Carbon-based primary productivity modeling with vertically resolved photoacclimation. *Global Biogeochemical Cycles*, 22(2).
- [75] Westberry, T. K., Behrenfeld, M. J., Milligan, A. J., and Doney, S. C. (2013). Retrospective satellite ocean color analysis of purposeful and natural ocean iron fertilization. *Deep Sea Research Part I: Oceanographic Research Papers*, 73:1–16.
- [76] Wu, J., Lee, Z., Goes, J., do Rosario Gomes, H., and Wei, J. (2024). Evaluation of three contrasting models in estimating primary production from ocean color remote sensing using long-term time-series data at oceanic and coastal sites. *Remote Sensing of Environment*, 302:113983.
- [77] Yang, G., Li, B., Ji, S., Gao, F., and Xu, Q. (2013). Ship detection from optical satellite images based on sea surface analysis. *IEEE Geoscience and Remote Sensing Letters*, 11(3):641–645.
- [78] Yu, B., Chen, F., and Chen, H. (2019). Npp estimation using random forest and impact feature variable importance analysis. *Journal of Spatial Science*, 64(1):173–192.
- [79] Zhang, S., Curchitser, E. N., Kang, D., Stock, C. A., and Dussin, R. (2018). Impacts of mesoscale eddies on the vertical nitrate flux in the gulf stream region. *Journal of Geophysical Research: Oceans*, 123(1):497–513.
- [80] Zhang, Y., He, X., Bai, Y., Wang, G., Li, T., Wang, D., Gong, F., and Zhu, Q. (2025). Satellite-estimation of the global ocean primary productivity via bgc-argo measurements. *Journal of Geophysical Research: Oceans*, 130(4):e2024JC021163.

Glosario

Digital Twin Gemelo digital. Representación virtual de un proceso, un sistema o un fenómeno del mundo real, que permite su simulación y análisis mediante datos y modelos computacionales. III, VII, 7, 8, 16, 19, 20, 22–25, 27, 28, 39, 50–52, 71–73, 80–83

AbPM *Absorption-based Productivity Model*. Modelo que estima la NPP utilizando la absorción espectral del fitoplancton y el rendimiento cuántico de la fotosíntesis. II, III, 12, 14–16, 18, 19, 24, 72, 76, 77

Afloramiento Proceso oceanográfico mediante el cual aguas profundas y frías, ricas en nutrientes, emergen a la superficie, estimulando la productividad biológica. 9, 53, 55, 58–60, 71, 83

CbPM *Carbon-based Productivity Model*. Modelo basado en observaciones satelitales que estima la NPP utilizando biomasa de carbono del fitoplancton en lugar de la concentración de clorofila-a. 6, 12, 13, 16, 18, 19, 76, 82

Clorofila-a Pigmento fotosintético presente en el fitoplancton, utilizado como aproximación de su biomasa mediante la teledetección. VII, 6, 12–15, 20, 32–36, 38, 40, 41, 43, 50–56, 59, 63, 65, 68, 72, 73, 76, 80, 83

CRISP-DM *Cross-Industry Standard Process for Data Mining*. Metodología ampliamente utilizada para proyectos de minería de datos, que estructura el proceso en fases, como la comprensión del negocio o del problema, comprensión de los datos, preparación, modelado, evaluación y despliegue. II, III, VII, 25, 29–31, 37, 38

Fitoplancton Microorganismos fotosintéticos acuáticos que constituyen la base de la cadena trófica marina. VII, 4–6, 9, 11, 13, 15, 16, 20, 22, 25, 28, 36, 38, 40, 41, 44, 49, 50, 53, 55, 56, 59, 62, 64, 72, 73, 77, 79–81, 83

Interpolación Técnica para estimar valores desconocidos entre puntos conocidos. En este trabajo se utilizan métodos como interpolación lineal, por vecinos cercanos o con funciones de base radial (RBF). VII, 24, 32, 37, 38, 40, 42, 44–48, 50

Macaronesia Región geográfica del Atlántico norte que incluye los archipiélagos de Canarias, Azores, Madeira, y Cabo Verde. II, III, 7, 52, 53, 55, 71, 81–83

- Multicolinealidad** Fenómeno estadístico en el que dos o más variables explicativas de un modelo de regresión están altamente correlacionadas linealmente entre sí. Esta situación puede distorsionar la estimación de los coeficientes individuales, aumentar la varianza e impedir una interpretación clara del modelo. VII, 40, 61, 62, 64
- NPP** Producción primaria neta, en inglés, *Net Primary Production*. Cantidad de carbono fijado por el fitoplancton mediante la fotosíntesis, restando la respiración autotrófica. Es un indicador clave de la productividad biológica oceánica. II, III, VI–IX, 4–9, 11–16, 18–20, 22–24, 35, 38–43, 49, 51–56, 58–63, 65–77, 79–83
- PAR** Radiación fotosintéticamente activa, en inglés, *Photosynthetically Active Radiation*. Fracción del espectro solar útil para la fotosíntesis (400–700 nm), esencial para estimar la NPP. 6, 38, 40, 41, 51, 53, 58, 63, 65, 71–73, 76, 77, 79, 80, 83
- Red neuronal artificial** Modelo computacional inspirado en el cerebro humano, utilizado para detectar patrones y realizar predicciones a partir de grandes volúmenes de datos. II, III, 7, 16, 19, 20, 22, 39, 50, 53, 82
- SST** Temperatura superficial del mar, en inglés, *Sea Surface Temperature*. Temperatura medida en la capa superior del océano, importante en modelos oceanográficos y climáticos. VII, 6, 14, 20, 38, 40, 41, 43, 51–53, 55, 58, 59, 63, 65, 71–73, 76, 77, 79, 80, 83
- VGPM** *Vertically Generalized Production Model*. Modelo empírico para estimar la NPP basado en datos satelitales de clorofila-a, irradiancia y temperatura superficial del mar. II, III, 6, 12–16, 18, 19, 24, 41, 72, 76, 77
- VIF** *Variance Inflation Factor*. Indicador de multicolinealidad entre variables explicativas en un modelo de regresión. VII, 61, 62, 64
- Zona eufótica** Capa superficial del océano donde penetra la luz solar suficiente para permitir la fotosíntesis. 5, 6, 9, 12, 14, 16, 53