



EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO



INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
MAESTRÍA EN SISTEMAS COMPUTACIONALES

**MODELO BIOGEOQUÍMICO PARA DETERMINAR LA DISTRIBUCIÓN DE LA
BIOMASA DE FITOPLANCTON (CHLA-A) MEDIANTE EL USO DE REDES
NEURONALES ARTIFICIALES.**

TESIS

QUE PARA OBTENER EL GRADO DE
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:

ING. JUAN MANUEL TALAMANTES MURILLO

DIRECTORES DE TESIS:

MC. JORGE ENRIQUE LUNA TAYLOR
DRA. ELVIA DENISSE SÁNCHEZ PÉREZ

LA PAZ, BAJA CALIFORNIA SUR, MÉXICO, NOVIEMBRE 2025.



Educación
Secretaría de Educación Pública



TECNOLÓGICO
NACIONAL DE MÉXICO



Instituto Tecnológico de La Paz
División de Estudios de Posgrado e Investigación

La Paz, B.C.S. **30/OCTUBRE/2025**
DEPI / 122 /2025

ASUNTO: Autorización de impresión

C. JUAN MANUEL TALAMANTES MURILLO,
ESTUDIANTE DE LA MAESTRÍA EN
SISTEMAS COMPUTACIONALES,
P R E S E N T E .

Con base en el dictamen de aprobación emitido por el Comité Tutorial de la Tesis denominada: **"MODELO BIOGEOQUÍMICO PARA DETERMINAR LA DISTRIBUCIÓN DE LA BIOMASA DE FITOPLANCTON (CHLA-A) MEDIANTE EL USO DE REDES NEURONALES ARTIFICIALES"**, mediante la opción de tesis (Proyectos de Investigación), entregado por usted para su análisis, le informamos que se **AUTORIZA** la impresión.

ATENTAMENTE
Excelencia en Educación Tecnológica

JUDITH GUADALUPE MARTÍNEZ TIRADO,
JEFA DE LA DIV. DE ESTUDIOS DE POSGRADO E INV.

c.c.p. Depto. de Servicios Escolares
c.c.p. Archivo.

JGMT/icl*



2025
Año de
La Mujer
Indígena

Boulevard Forjadores de B.C.S. #4720 Col. 8 de octubre 1ra. Sección
C.P. 23080 La Paz B.C.S. Tel. 612 121 0424
e-mail: depi_paz@tecnm.mx lapaz.tecnm.mx





DICTAMEN DEL COMITÉ TUTORIAL

La Paz, B.C.S., **29/OCTUBRE/ 2025**

**JUDITH GUADALUPE MARTÍNEZ TIRADO,
JEFA DE LA DIVISIÓN DE ESTUDIOS DE
POSGRADO E INVESTIGACIÓN,
P R E S E N T E.**

Por medio del presente, enviamos a usted dictamen del Comité Tutorial de tesis para la obtención del grado de Maestro, con los siguientes datos generales:

No. de Control M23310001	Nombre Juan Manuel Talamantes Murillo
Maestría en:	SISTEMAS COMPUTACIONALES
Título de la tesis: MODELO BIOGEOQUÍMICO PARA DETERMINAR LA DISTRIBUCIÓN DE LA BIOMASA DE FITOPLANCTON (CHLA-A) MEDIANTE EL USO DE REDES NEURONALES ARTIFICIALES	
DICTAMEN: Se autoriza el trabajo de investigación, en virtud de que realizó las correcciones correspondientes conforme a las observaciones planteadas por este Comité Tutorial.	

**Atentamente,
El Comité Tutorial**


DR. ISRAEL MARCOS SANTILLÁN MÉNDEZ


DR. MARCO ANTONIO CASTRO LIERA


DRA. ELVIA DENISSE SÁNCHEZ PÉREZ


M.C. JORGE ENRIQUE LUNA TAYLOR

c.c.p. Coordinador de la Maestría.
c.c.p. Departamento de Servicios Escolares.
c.c.p. Estudiante.

ITLP-DEPI-RTT-08

Rev.1



2025
Año de
**La Mujer
Indígena**

Boulevard Forjadores de B.C.S. #4720 Col. 8 de octubre 1ra. Sección
C.P. 23080 La Paz B.C.S. Tel. 612 121 0424
e-mail: depi_paz@tecnm.mx lapaz.tecnm.mx



Dedicatoria

A mi esposa, Selma Elizabeth Chaparro García, por su incondicional apoyo y confianza a lo largo de este camino.

Y una dedicatoria muy especial para nuestro hermoso hijo, Lionel Iban Talamantes Chaparro, quien quizás aún es pequeñito para comprenderlo, pero es la luz que ilumina mis días.

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mi esposa y a mi hijo, por su amor, paciencia y comprensión durante todo este proceso.

Mi especial gratitud al M.S.C. Jorge Enrique Luna Taylor, director de esta tesis, por su invaluable acompañamiento, su paciencia y su guía constante. Aprecio profundamente cada conversación, cada consejo y cada momento compartido. Más allá de lo académico, me llevo a un gran amigo.

También especial agradecimiento a la Dra. Elvia Denisse Sánchez Pérez, directora de tesis, por su compromiso, su confianza en mi trabajo, y por impulsarme siempre a dar lo mejor de mí. Su orientación y respaldo han sido clave en la culminación de este proyecto.

Deseo agradecer al Dr. Marco Antonio Castro Liera, por su apoyo genuino desde el primer día, siempre acompañado de una actitud amable y positiva; al Dr. Israel Marcos Santillán Méndez, por sus valiosas recomendaciones y consejos oportunos; y a la M.S.C. Iliana Castro Liera, por acompañarme en cada etapa de este trayecto, siempre con la mejor disposición.

Agradezco al Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCyT) por la beca otorgada, la cual fue de gran ayuda para la realización de este trabajo.

Finalmente, extendiendo mi agradecimiento al Instituto Tecnológico de La Paz (ITLP) y al Centro Interdisciplinario de Ciencias Marinas (CICIMAR) por facilitar y hacer posible el desarrollo de esta investigación.

Resumen

El fitoplancton es un microorganismo marino que aporta más del 50 % del oxígeno que hay en nuestro planeta. Es también el eslabón primario en las cadenas tróficas marinas, que constituye el alimento principal del zooplancton, derivando en el sustento de otras especies. Desde hace décadas, la actividad antropogénica ha generado que se viertan grandes volúmenes de contaminantes en los mares, alterando la abundancia y distribución de estos organismos. Sin embargo, aún se desconoce en qué medida se relacionan estas alteraciones con el fenómeno del cambio climático. Con el objetivo de aportar herramientas para abordar esta problemática, en este documento se presenta el desarrollo de un modelo computacional basado en redes neuronales artificiales, para predecir la concentración de biomasa del fitoplancton (*Chla-a*) a partir de datos de variables biogeoquímicas. Las pruebas del modelo se realizaron con datos obtenidos de la plataforma *Copernicus Marine Service*, para la zona subtropical del océano Pacífico que colinda con Perú. Los experimentos muestran una precisión del modelo del 98.5 %, con un coeficiente de determinación (R^2) de 0.9994.

Adicionalmente, se utilizó el modelo desarrollado para analizar escenarios de cambio climático, con el fin de predecir su potencial impacto en la distribución de la biomasa del fitoplancton. Se emplearon proyecciones de incrementos simulados de temperatura permitiendo explorar posibles cambios espaciales y temporales en la abundancia del fitoplancton.

Abstract

Phytoplankton is a marine microorganism that contributes more than 50 % of the oxygen present on our planet. It also represents the primary link in marine food chains, serving as the main food source for zooplankton and ultimately supporting other species. For decades, anthropogenic activity has led to the discharge of large volumes of pollutants into the oceans, altering the abundance and distribution of these organisms. However, the extent to which these alterations are related to the phenomenon of climate change remains unknown. In order to provide tools to address this issue, this document presents the development of a computational model based on artificial neural networks to predict phytoplankton biomass concentration (*Chl-a*) from physicochemical variable data. The model was tested using data obtained from the *Copernicus Marine Service* platform, focusing on the subtropical region of the Pacific Ocean off the coast of Peru. The experiments show a model accuracy of 98.5 %, with a coefficient of determination (R^2) of 0.9994.

Additionally, the developed model was used to analyze climate change scenarios in order to predict its potential impact on phytoplankton biomass distribution. Simulated projections of temperature increases were employed, allowing the exploration of possible spatial and temporal changes in phytoplankton abundance.

Índice general

Dedicatoria	I
Agradecimientos	II
Resumen	III
Abstract	IV
1. Introducción	1
1.1. Descripción del problema	2
1.2. Antecedentes	3
1.3. Hipótesis	5
1.4. Objetivos	5
1.4.1. Objetivo general	5
1.4.2. Objetivos específicos	6
1.5. Justificación	6
1.6. Alcance y limitaciones	6
2. Marco teórico	7
2.1. Inteligencia Artificial	7
2.2. Redes Neuronales	8
2.3. Biogeoquímica	9
2.4. Fitoplancton	9
2.5. Biomasa	10
2.6. Fotosíntesis	10
2.7. Clorofila (Chla-a)	11
2.8. Salinidad	12

2.9.	Temperatura	12
2.10.	Nutrientes	13
2.11.	Radiación solar	14
2.12.	Velocidad del mar hacia el Norte y Este	15
2.13.	Concentración de oxígeno disuelto en el mar	15
2.14.	Densidad del espesor de la capa mixta del océano	15
2.15.	Altura de la superficie del mar	16
2.16.	Producción primaria neta de la biomasa de carbono	16
2.17.	Concentración de biomasa de carbono de fitoplancton	17
2.18.	pH	18
2.19.	Dióxido de carbono	18
3.	Desarrollo de modelo para predicción de la concentración y distribución del fito- plancton	19
3.1.	Desarrollo del modelo basado en datos in situ	19
3.1.1.	Algoritmo 1: Integración de información de plataformas distintas	20
3.1.2.	Implementación del modelo para la predicción de fitoplancton (<i>Chla-a</i>)	22
3.1.3.	Entrenamiento del modelo	23
3.2.	Desarrollo del modelo basado en datos satelitales	25
3.2.1.	Relevancia de cada variable	27
3.2.2.	Preprocesado de datos	28
3.2.3.	Algoritmo 3: Conversión de datos NetCDF a CSV	28
3.2.4.	Algoritmo 4: Combinación de variables biogeoquímicas y físicas	29
3.2.5.	Incorporación de la variable irradiancia	30
4.	Experimentos y resultados de la aplicación de los modelos	31
4.1.	Experimentos con el modelo basado en datos in situ	31
4.2.	Experimentos con el modelo basado en datos satelitales	34
4.2.1.	Resultados de las pruebas con las primeras ocho variables de acuerdo el ranking obtenido del PCA	36
4.2.2.	Resultados de las pruebas con las primeras diez variables de acuerdo el ranking obtenido del PCA	38

4.2.3.	Resultados de las pruebas utilizando todas las variables	40
5.	Aplicación de modelo para predecir la biomasa del fitoplancton (Chla-a) derivado del efecto del cambio climático	45
5.1.	Selección de variables para el análisis del efecto del cambio climático sobre el fitoplancton	45
5.2.	Análisis de la tendencia de la temperatura promedio anual y generación de datos hipotéticos	48
5.3.	Generación y aplicación del modelo para la predicción bajo diferentes escenarios de cambio climático	49
5.4.	Resultados	49
6.	Conclusiones	53
Literatura Citada		59

1. Introducción

El fitoplancton, organismos autótrofos microscópicos presentes en ecosistemas costeros y marinos, constituyen un componente fundamental para el sostenimiento de la vida en el planeta. Estos organismos son responsables de aproximadamente la mitad de la producción global de oxígeno mediante la fotosíntesis, un proceso mediado por pigmentos fotosintéticos como la clorofila, la cual actúa como catalizador en la conversión de energía solar en energía química (Behrenfeld and Falkowski, 1997). La clorofila-a (chl-a) ha sido históricamente usada como un indicador cuantitativo de la biomasa fitoplanctónica en los ecosistemas acuáticos dada su correlación con la actividad fotosintética y abundancia celular (Rizzuto et al., 2020).

No obstante, el equilibrio de estas comunidades depende críticamente de factores biogeoquímicos como la disponibilidad de nutrientes, estratificación térmica, regímenes de luz y turbulencia, cuyos desajustes pueden alterar su productividad primaria y desencadenar proliferaciones algales nocivas (Tiwari et al., 2022; Chai et al., 2020). Para monitorear estas dinámicas, los sensores satelitales de color oceánico (*MODIS-Aqua*, *Sentinel-3*, *VIIRS*) han complementado los métodos *in situ*, proporcionando cobertura sinóptica y continua de parámetros biogeoquímicos mediante el análisis de la reflectancia espectral ($R_{rs}(\lambda)$) en múltiples longitudes de onda. Esta capacidad permite derivar no solo la concentración de chl-a, sino también proxies de material orgánico disuelto cromóforo (CDOM), partículas suspendidas (SPM) y nutrientes traza como nitritos y hierro (Hu et al., 2012; Kostadinov et al., 2010; Sun et al., 2022; Adhikary et al., 2022), ofreciendo una visión integral de la estructura de la comunidad.

Estos avances han impulsado innovaciones en el modelado bio-óptico. Un ejemplo destacado es la aplicación de regresión por componentes principales optimizada a datos hiperespectrales de $R_{rs}(\lambda)$, técnica que permite reconstruir simultáneamente perfiles de pigmentos accesorios (fucoxantina, zeaxantina) vinculados a grupos funcionales específicos como diatomeas y cianobacterias (Vandermeulen et al., 2020; Kramer et al., 2022). Sin embargo, la naturaleza no lineal y multivariable de los procesos biogeoquímicos exige metodologías analíticas más sofisticadas. En este escenario, las redes neuronales artificiales (RNA) destacan por su capacidad para procesar grandes volúmenes de datos satelitales y modelar interacciones complejas entre variables bióticas y abióticas superando las limitaciones de enfoques lineales como la regresión por componentes principales. Por ejemplo, arquitecturas de *Deep Learning* han logrado mayor

precisión en la predicción del fitoplancton al combinar datos hiperspectrales de reflectancia remota ($R_{rs}(\lambda)$) con mediciones oceanográficas *in situ*, como perfiles verticales de nitrógeno y estratificación térmica (Bracher et al., 2009; Pahlevan et al., 2020). Estudios recientes, utilizando técnicas de aprendizaje supervisado (*Random Forest*, *Gradient Boosting*, etc.) aplicadas a datos de reanálisis de variables biogeoquímicas, han predicho niveles de fitoplancton a través del coeficiente de determinación (R^2) de hasta 0.96, proponiendo su uso para complementar mediciones *in situ* y monitorear su rol crítico en ecosistemas marinos (Adhikary et al., 2024).

La integración sinérgica de información satelital y datos de campo promete no solo mejorar la modelización de ecosistemas acuáticos, sino también facilitar la identificación de patrones espacio-temporales críticos del fitoplancton. El desarrollo de modelos predictivos basados en esta aproximación ofrecería un soporte científico sólido para diseñar estrategias de gestión costera adaptativa, como la predicción temprana de floraciones algales nocivas o la optimización de zonas de conservación marina.

Dado que el fitoplancton es la base de la cadena trófica, estimar su concentración en términos de clorofila-a, es fundamental para comprender su productividad oceánica. Los recientes avances en modelos de inteligencia artificial, como las redes neuronales, ofrecen una oportunidad sin precedentes para predecir su comportamiento a corto y mediano plazo utilizando la compleja interacción de las variables oceanográficas. Por lo que, se esperaría que un modelo de regresión basado en redes neuronales artificiales sea capaz de predecir con un alto grado de precisión la variabilidad de la clorofila-a, al utilizar un conjunto de variables oceanográficas clave como predictores, dada la influencia directa de estas variables en la abundancia del fitoplancton.

1.1. Descripción del problema

La mayoría de los estudios enfocados en la biomasa fitoplanctónica se han basado en el análisis de la variabilidad estacional de la Chla-a 'in-situ' a través de la recolección de datos obtenidos por CTD (Conductividad-Temperatura-Presión) y sondas multiparamétricas. Sin embargo, a nivel de variabilidad interanual aún son escasos los estudios, debido a la duración relativamente corta de las series temporales. Hoy en día, una herramienta versátil y eficaz de bajo costo es la teledetección o imágenes satelitales, la cual ha permitido la identificación de patrones sinópticos físicos y biológicos del océano a nivel global.

Con base a lo anterior, la teledetección y la inteligencia artificial por medio de las redes

neuronales artificiales están tomando un papel muy importante en el desarrollo de estos estudios, ya que con el correcto uso de estas herramientas permitiría a los investigadores, mediante el desarrollo de modelos biogeoquímicos, determinar la distribución de la biomasa del fitoplancton (Chla-a).

Hoy en día existe una creciente preocupación sobre el impacto del cambio climático en los ecosistemas marinos, particularmente en el fitoplancton, que constituye la base de las redes tróficas oceánicas y desempeña un papel crucial en el ciclo global del carbono. El aumento de la temperatura superficial del mar, la estratificación de las masas de agua y la alteración de los patrones de circulación oceánica podrían modificar significativamente la distribución y productividad del fitoplancton, con consecuencias potencialmente graves para la biodiversidad marina. En este contexto, resulta fundamental contar con modelos predictivos sofisticados que permitan anticipar estos cambios y comprender las posibles respuestas del fitoplancton ante escenarios climáticos futuros.

1.2. Antecedentes

Actualmente, existe poca información sobre modelos de predicción de la biomasa del fitoplancton. Uno de los trabajos que consideramos más relevantes es la publicación titulada '*Global marine phytoplankton dynamics analysis with machine learning and reanalyzed remote sensing*' (Adhikary et al., 2024), esta investigación tuvo como objetivo predecir la concentración de chla-a aplicando diversas técnicas de regresión con modelos supervisados de aprendizaje automatizado en 6 diferentes regiones de los mares del planeta (Fig. 1).

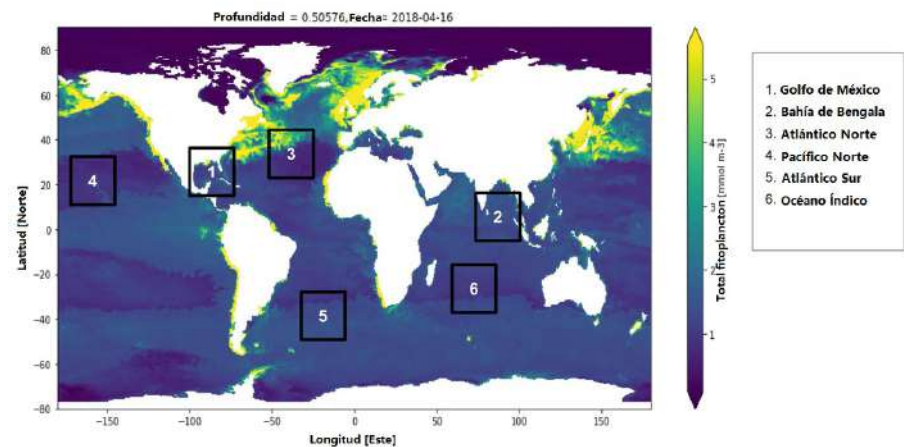


Figura 1. La región azulada indica la baja concentración de fitoplancton y la región amarilla la más alta concentración de fitoplancton (Tomado de Adhikary et al., 2024)

Los algoritmos de regresión utilizados en el experimento fueron *Random Forest*, *Bagging*, *Extra Trees* y *HGBR*. La investigación realiza comparaciones entre los resultados obtenidos en diferentes zonas específicas y el resto del mundo. Específicamente, evalúa cómo varían las concentraciones de fitoplancton y las propiedades biogeoquímicas en distintas regiones para entender las diferencias en las predicciones. Entrenaron modelos con datos de todas las áreas exceptuando una y luego probaron las predicciones en la región excluida. Esto permitió evaluar cómo las propiedades biogeoquímicas y físicas específicas de cada región afectan la capacidad del modelo para predecir la concentración de fitoplancton. El modelo más preciso fue el *Extra Trees*, con una media del error cuadrado de 0.0033 y con un puntaje de predicción de hasta 0.96 (Fig. 2).

Tabla 1 Desempeño de regresión de la estimación del fitoplancton en los mares globales					
Métricas		Random forest	Bagging	Extra trees	HGBR
Entrenamiento	Media del error cuadrado	0.000508	0.000771	3.02E-24	0.005344
	R ²	0.994511	0.991668	1.0	0.942291
	Max error	0.753581	0.885730	4.99E-10	1.574378
	Tiempo (s)	331.683	33.538	103.444	2.196
Prueba	Media del error cuadrado	0.003565	0.003962	0.003395	0.005837
	R ²	0.961293	0.956989	0.963142	0.936631
	Max error	1.396569	1.263875	1.077332	1.491765
	Tiempo (s)	2.211391	0.232537	2.426239	0.089830

Figura 2. Resultados del desempeño de los modelos en mares globales (Tomado de Adhikary et al., 2024)

En el caso del Golfo de México los resultados de los modelos de regresión mostraron un desempeño significativamente más bajo, indicando que a diferencia de las otras regiones analizadas, en el Golfo de México se tienen propiedades biogeoquímicas y físicas únicas que dificultan las predicciones con modelos globales. El modelo obtuvo una media del error cuadrado de 0.0417 y con un puntaje de predicción de hasta 0.59 (Fig. 3).

Tabla 6 Desempeño de regresión del Golfo de México vs el resto del mundo					
Métricas		Random forest	Bagging	Extra trees	HGBR
Entrenamiento	Media del error cuadrado	0.000382	0.000580	1.36E-30	0.004523
	R²	0.995809	0.993642	1.0	0.950445
	Max error	0.612501	0.662656	1.24E-14	1.729356
	Tiempo (s)	400.385	38.076	102.926	3.532
Prueba	Media del error cuadrado	0.053334	0.055418	0.041754	0.040747
	R²	0.479236	0.458879	0.592300	0.602135
	Max error	2.730470	2.643686	2.994876	2.540424
	Tiempo (s)	0.383338	0.044700	0.428533	0.043013

Figura 3. Resultados del desempeño de los modelos en el Golfo de México (Tomado de Adhikary et al., 2024)

El experimento concluye destacando que las diferencias en las propiedades biogeoquímicas y físicas en cada región afectan la precisión de los modelos. Esto significaría que existen regiones donde es necesario entrenar modelos específicos para mejorar las predicciones.

1.3. Hipótesis

Se esperaría que el desarrollo de modelos basados en redes neuronales artificiales, entrenados con variables biogeoquímicas obtenidas de imágenes satelitales, permitirá generar predicciones con alto nivel predictivo sobre la concentración de chl-a en el océano.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar un modelo para determinar la distribución espacio-temporal de la biomasa del fitoplancton (Chla-a) a través de redes neuronales artificiales, para aplicarlo como modelo de predicción en el Pacífico Tropical-Subtropical frente a Perú.

1.4.2. Objetivos específicos

- Desarrollar un modelo computacional a partir de redes neuronales artificiales alimentado de variables oceanográficas para el análisis de la concentración y distribución del fitoplancton.
- Aplicar el modelo computacional para predecir el efecto del cambio climático en la distribución de la biomasa del fitoplancton (Chla-a).

1.5. Justificación

Considerando los grandes esfuerzos que ha realizado el área de la ciencia e investigación para determinar a mayor detalle el comportamiento de la comunidad de fitoplancton, aún son muy incipientes los estudios que permitan pronosticar su dinámica, distribución y variabilidad, a corto, mediano y a largo plazo (Polovina et al., 2011).

Esta propuesta, mediante el diseño y entrenamiento de una red neuronal a partir de variables de entrada fisicoquímicas y biológicas, destaca la necesidad de modelos predictivos para evaluar la respuesta del fitoplancton ante escenarios climáticos futuros y en el potencial que ofrecen los enfoques basados en inteligencia artificial para integrar datos históricos (Brewin et al., 2015) reduciendo los tiempos de análisis frente a las técnicas tradicionales.

1.6. Alcance y limitaciones

El desarrollo de un modelo basado en variables oceanográficas utilizando redes neuronales artificiales permite analizar la distribución del fitoplancton a partir de la integración de grandes volúmenes de datos históricos y la relación entre variables físicas, químicas y biológicas, generando resultados predictores útiles para la toma de decisiones. Uno de los principales alcances de esta propuesta radica en su capacidad para aprovechar la información proveniente de imágenes satelitales, las cuales ofrecen una cobertura espacial y temporal amplia, permitiendo identificar patrones de distribución de la biomasa fitoplanctónica a diferentes escalas.

No obstante, este enfoque enfrenta limitaciones importantes. Una de las principales es la dependencia total de la disponibilidad y calidad de los datos históricos. Debido a la escasa cantidad de investigaciones previas específicamente orientadas a la predicción de la biomasa

fitoplanctónica mediante redes neuronales artificiales, no existe una base de datos estandarizada o completamente consolidada que pueda ser utilizada de manera directa. Esto obliga a la construcción de una base de datos propia, a partir de registros históricos obtenidos de diversas fuentes, principalmente imágenes satelitales. Esta tarea, además de demandante en términos de tiempo y recursos, requiere un cuidadoso proceso de selección y validación de las variables de entrada para asegurar que estén directamente relacionadas con los procesos biológicos y químicos fundamentales del fitoplancton, tales como la fotosíntesis y sus nutrientes.

2. Marco teórico

2.1. Inteligencia Artificial

La Inteligencia Artificial (IA), en el contexto de las ciencias de la computación, es una disciplina y un conjunto de capacidades cognoscitivas e intelectuales expresadas por sistemas informáticos o combinaciones de algoritmos cuyo propósito es la creación de máquinas que imiten la inteligencia humana para realizar tareas, y que pueden mejorar conforme recopilen información (McCarthy, 1996).

La habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana, la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano. Sin embargo, a diferencia de las personas, los dispositivos basados en IA no necesitan descansar y pueden analizar grandes volúmenes de información a la vez. Asimismo, la proporción de errores es significativamente menor en las máquinas que realizan las mismas tareas que sus contrapartes humanas (Rouhiainen, 2018) (Fig.4).

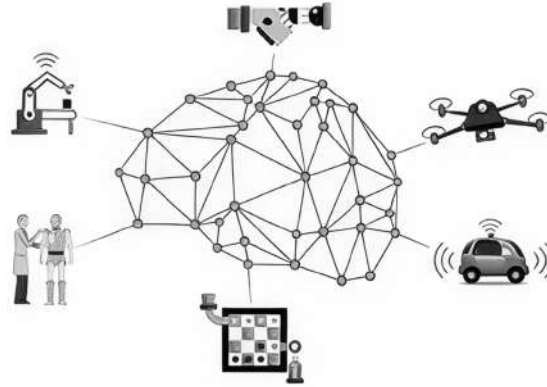


Figura 4. Ejemplos de aplicaciones de la Inteligencia Artificial (Tomado de Rouhiainen, 2018)

2.2. Redes Neuronales

Las redes neuronales artificiales, pilar de la inteligencia artificial, a partir del aprendizaje automatizado hace posible que las máquinas puedan ejecutar tareas o funciones que no son capaces de elaborar los sistemas convencionales.

Las redes neuronales artificiales son un modelo computacional que permite simular el comportamiento del cerebro humano, es decir, dotar a las máquinas de la capacidad de aprender de una manera similar a como lo hace nuestro cerebro. Una red neuronal artificial está formada por neuronas artificiales, que son unidades o nodos que reciben información del exterior o de otras neuronas, de manera similar a los impulsos nerviosos que reciben las neuronas del cerebro humano, las procesan y generan un valor de salida que alimenta a otras neuronas de la red o son la salida hacia el exterior de la red (Minsky, 1969) (Fig.5).

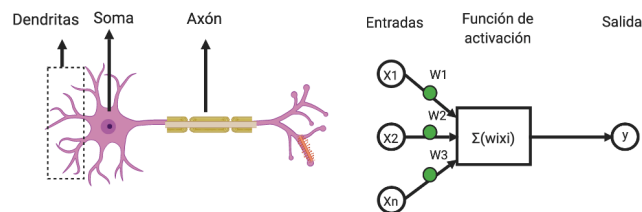


Figura 5. Comparación entre una neurona y un perceptrón o neurona artificial (Tomado de García, 2019)

2.3. Biogeoquímica

La biogeoquímica es la disciplina que estudia los procesos y reacciones químicas, físicas, geológicas y biológicas que gobiernan la composición del medio ambiente natural. En particular, la biogeoquímica es el estudio de los ciclos de los elementos químicos, tales como hidrógeno, el nitrógeno y el carbono, y su interacción e incorporación en los seres vivos (los Andes Colombia, 2023) (Fig.6).

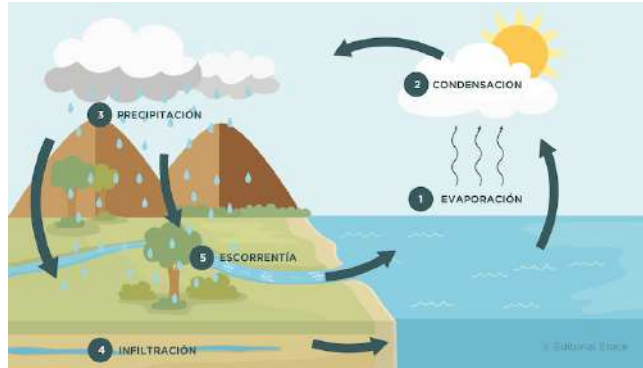


Figura 6. Representación del ciclo del hidrógeno (Tomado de Álvarez, 2024)

2.4. Fitoplancton

Fitoplancton, sustento vital de los seres vivos, es un microorganismo planctónico presente en los mares y capaz de generar más del 50 % del oxígeno en el planeta tierra por medio del proceso de la fotosíntesis. Considerado como la base de los ecosistemas acuáticos los cuales acaparan el 70 % del territorio del planeta. Además de la producción de oxígeno también, el fitoplancton, por el hecho de ser autótrofo puede, a partir de la materia inorgánica, proporcionar materia orgánica importante como alimento de otras especies como las proteínas, carbohidratos y grasas por consiguiente es el productor primario y eslabón basal en las cadenas tróficas marinas, constituye el alimento principal del zooplancton y el sustento de muchos peces, que son el objetivo de las principales pesquerías industriales a lo largo y ancho del planeta. Los principales grupos taxonómicos del fitoplancton son: Diatomeas y Dinoflagelados (DYNECO, 2011) (Fig.7).



Figura 7. Ejemplo de células del fitoplancton (Tomado de Han, 2023)

2.5. Biomasa

A la cantidad total de fitoplancton presente en una masa de agua en un momento determinado se le define como biomasa fitoplanctónica. Habitualmente se estima a través de la concentración de *chl-a*, pigmento fotosintético predominante en estos organismos expresados en miligramos por metro cúbico (mg/m^3), aun cuando la *chl-a* representa solo el 1 % del peso seco de una célula. Este indicador constituye una métrica fundamental para evaluar la productividad primaria de los ecosistemas marinos, su dinámica estacional e interanual, y su respuesta frente a factores ambientales y climáticos. La biomasa fitoplanctónica representa el peso total del fitoplancton por unidad de volumen o de área (Cullen, 1982).

2.6. Fotosíntesis

La fotosíntesis, también conocida como función clorofílica, es el proceso químico que tiene lugar en las plantas y otros organismos fotosintéticos como las algas y algunos tipos de bacterias, cuya finalidad es convertir la materia orgánica en inorgánica a partir del aprovechamiento de la luz del Sol. Dicho de otra manera, la fotosíntesis es el proceso por el cual la energía del Sol se convierte en energía química aprovechable para los organismos en la Tierra, liberando además en el proceso, el oxígeno indispensable para la vida en nuestro planeta (Bioenciclopedia, 2022) (Fig.8).

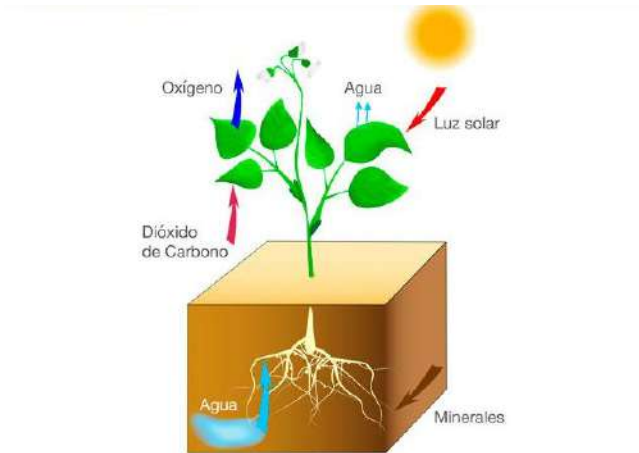


Figura 8. Representación del proceso de la fotosíntesis (Tomado de Bioenciclopedia, 2022)

2.7. Clorofila (Chla-a)

La chla-a es el pigmento verde que permite a las plantas (incluidas las algas) convertir la luz solar en compuestos orgánicos durante el proceso de fotosíntesis, permite la transferencia de energía entre las moléculas del compuesto y la oxidación química. Este compuesto es predominante en todas las algas, por lo que una alta concentración de chla-a en el agua (cerca a $1\text{mg}/\text{m}^3$) es un indicador de contaminación por nutrientes, porque existe un exceso de nutrientes que impulsa el crecimiento de las algas; por lo que la chla-a permite medir la cantidad de algas presentes en una zona específica y así establecer la cantidad del compuesto determinando si la cadena alimenticia se encuentra en equilibrio (Allinger, 1984) (Fig.9).

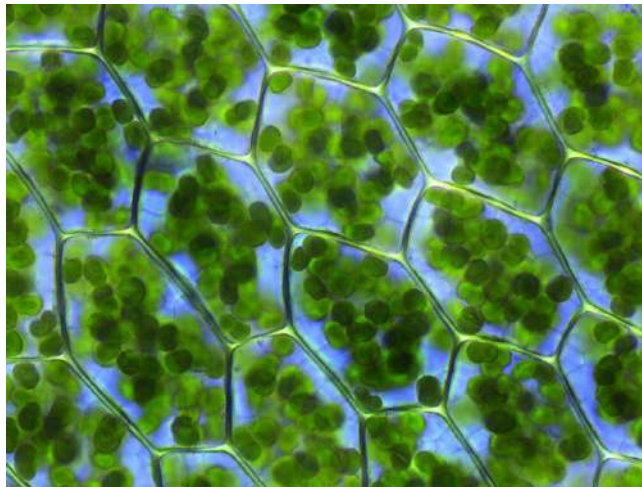


Figura 9. Representación de vista microscópica de los cloroplastos, los cuales contienen la chla-a (Tomado de Peters, 2017)

2.8. Salinidad

La salinidad de los océanos representa la cantidad de sales disueltas. La salinidad media de los océanos es de 35 g/l de agua. En los océanos, la salinidad varía entre 30 g/l en el Atlántico Norte y 40 g/l en el Mar Rojo.

La salinidad es un factor importante para determinar muchos aspectos de la química de las aguas marinas y de los procesos biológicos que tienen lugar en ellas, y es una variable de estado termodinámico que, junto con la temperatura y la presión, rige características físicas como la densidad y la capacidad calorífica del agua. La salinidad y la temperatura desempeñan un papel fundamental en la circulación oceánica mundial (Copernicus, 2023) (Fig.10).

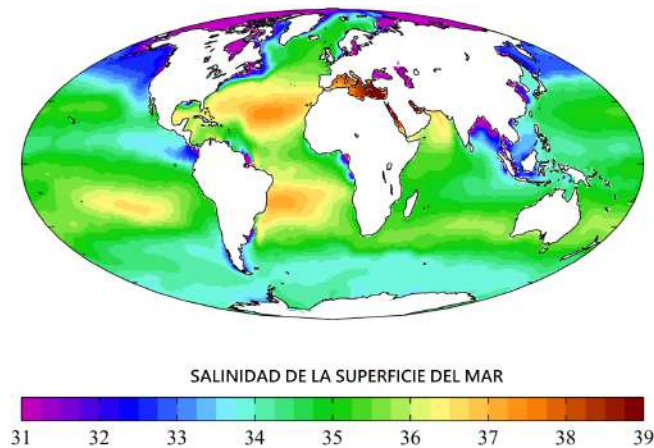


Figura 10. Salinidad de la superficie del mar (Tomado de NOAA-NESDIS-National, 2007)

2.9. Temperatura

La temperatura del mar es la medida de la calidez o frío del agua del mar en un lugar específico. Esta temperatura puede variar debido a factores como la latitud, la profundidad, la corriente marina y la estación del año. El mar recibe la energía proveniente de la radiación solar y debido a su elevada capacidad calorífica constituye el almacén de la mayoría del calor existente en nuestro planeta. Como consecuencia del proceso de rotación de la tierra alrededor del sol y las variaciones de su inclinación existen diferencias estacionales en la cantidad de radiación recibida por el mar, por lo cual la temperatura del agua tiende a ser más elevada en verano y más baja en invierno.

La capacidad calorífica del agua marina es relativamente alta en comparación con otros materiales, lo que significa que es un buen regulador de la temperatura en el planeta. Esto es

especialmente importante para el clima, ya que el agua marina puede retener y liberar calor en el curso de los ciclos climáticos (Good, 2020) (Fig.11).

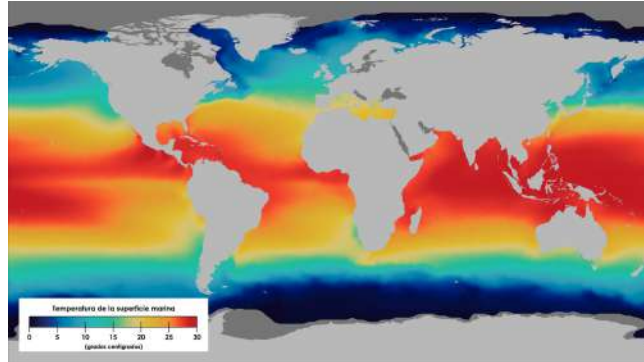


Figura 11. Mapa que muestra la temperatura de la superficie del mar (Tomado de NASA's-Earth-Observatory, 2017)

2.10. Nutrientes

Un nutriente es un compuesto que necesitan las células de un organismo para producir la energía empleada en las funciones de crecimiento, reparación y reproducción, metabolismo, entre otras. Se necesita de nutrientes y luz para realizar fotosíntesis en los océanos, en ellos se produce aproximadamente la mitad de la fotosíntesis del planeta, llevada a cabo principalmente por el fitoplancton. Los nutrientes se componen de nitrógeno, fósforo, sílice, manganeso, cobre y hierro; su concentración en las capas del mar depende de los procesos químicos y biológicos.

Cuando los organismos expulsan los desechos de su metabolismo o mueren, la materia orgánica queda en las capas superficiales y se hunde por gravedad hasta las profundidades, en donde por bacterias, algas y actividad química, se descompone (nitrificación), liberándose así los nutrientes, los cuales pueden ser utilizados nuevamente por otros vegetales para elaborar nueva sustancia orgánica por medio del proceso de fotosíntesis y así pasar a los animales (Jiménez, 2019) (Fig.12).

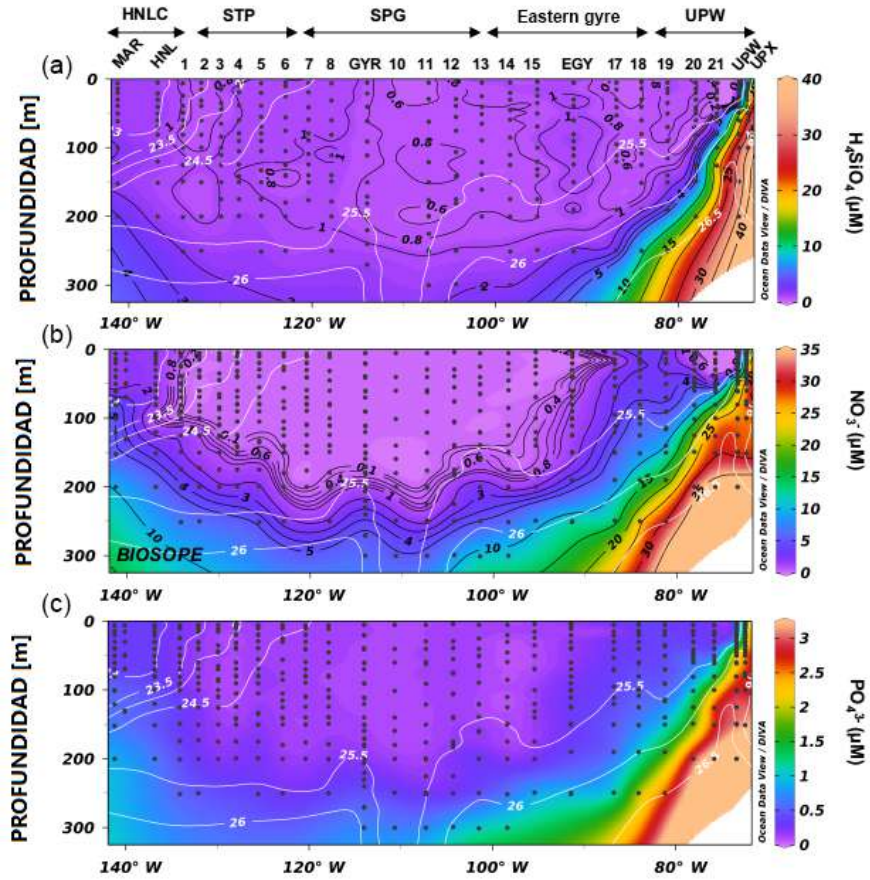


Figura 12. Distribución vertical de nutrientes del océano Pacífico (140°W a 80°W) mediante las distintas zonas de profundidad del océano desde la región HNLC (Alta concentración de nutrientes, baja de chla-a), la Provincia Subtropical (STP), el Giro del Pacífico Sur (SPG), el Giro Oriental (EGY) y la zona de surgencia costera (UPW) (Tomado de Copernicus, 2018).

2.11. Radiación solar

La radiación solar es uno de los principales factores ambientales que afectan la vida en nuestro planeta. Esta radiación controla el funcionamiento de los ecosistemas terrestres y acuáticos tanto a través del control de procesos fotobiológicos (fotosíntesis, fotoperíodo, fototropismos, etc.) como por medio de su acción sobre otros factores ambientales (temperatura, humedad, etc.) y ciclos naturales (ciclos diarios, anuales, hídricos, etc.) que finalmente inciden en la distribución de los organismos.

La radiación que llega a la Tierra abarca una amplia gama del espectro electromagnético y aproximadamente el 40 % de ella es la que conocemos como luz o radiación visible (Carrasco-Ríos, 2009).

2.12. Velocidad del mar hacia el Norte y Este

La velocidad de las corrientes superficiales del mar es un factor clave en la dispersión y distribución espacial del fitoplancton, ya que determina los patrones de movimiento horizontal en la capa fótica, donde estos organismos realizan la fotosíntesis. Si bien las corrientes superficiales facilitan la mezcla de nutrientes y evitan la acumulación excesiva de biomasa en un solo punto, también pueden representar un riesgo para el ecosistema fitoplanctónico cuando alcanzan niveles elevados de turbulencia o intensidad. En tales casos, la dispersión excesiva puede diluir las concentraciones locales de fitoplancton por debajo de los niveles necesarios para sostener una alta productividad primaria, o incluso arrastrar células fuera de la zona iluminada, afectando su capacidad fotosintética y, por ende, su supervivencia. (Fai PBA, 2023).

2.13. Concentración de oxígeno disuelto en el mar

El oxígeno disuelto en el agua es la cantidad de oxígeno presente en forma líquida en el agua. Se mide en miligramos por litro ($\text{mg}\cdot\text{L}^{-1}$) o partes por millón (ppm). El oxígeno disuelto resulta esencial para la supervivencia de la mayoría de las especies acuáticas, incluyendo a los peces, organismos invertebrados y plantas (Sonic, 2024) (Fig.13).

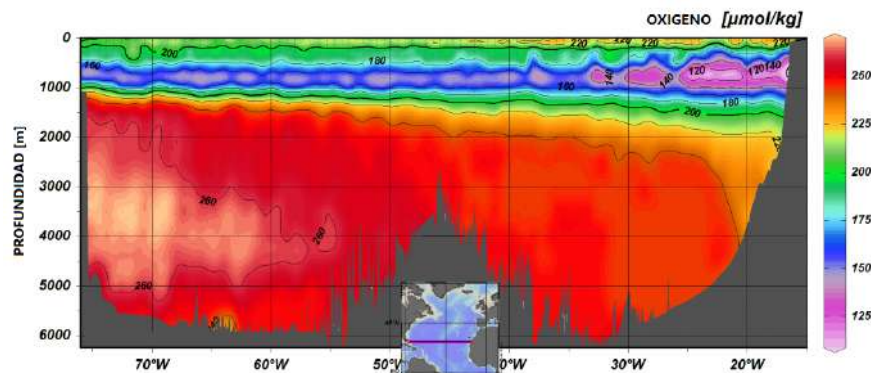


Figura 13. Representación de perfil de oxígeno disuelto de un transecto del océano Atlántico desde Florida hasta la costa de África (Tomado de Webb, 2023)

2.14. Densidad del espesor de la capa mixta del océano

La densidad del espesor de la capa mixta del océano es un parámetro importante, ya que al aumentar el espesor, la disponibilidad de luz disminuye, lo que reduce el metabolismo del

fitoplancton (Díaz, 2021) (Fig. 14).

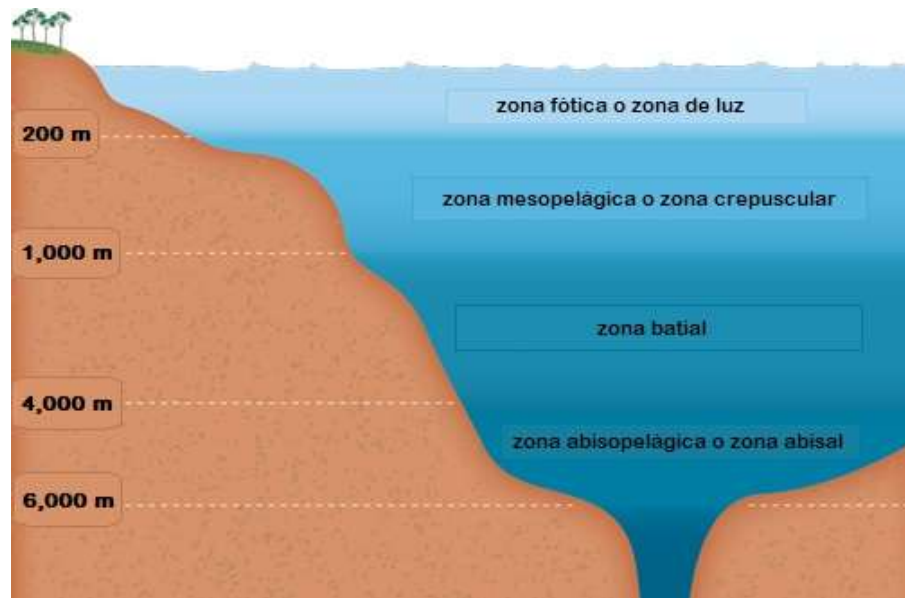


Figura 14. Representación del espesor de diversas capas del océano (Tomado de Ashlan Cous-teau, 2021)

2.15. Altura de la superficie del mar

La altura de la superficie del mar indica la diferencia en distancia entre la superficie media del mar y el elipsoide de referencia (Chen, 2021).

2.16. Producción primaria neta de la biomasa de carbono

La producción primaria neta es la tasa de fijación de carbono fotosintético menos la fracción de carbono usado por la respiración celular de los organismos. La producción de biomasa de carbono considera la radiación, nutrientes, micronutrientes, CO_2 y temperatura (Philip W. Boyd, 2014) (Fig.15).

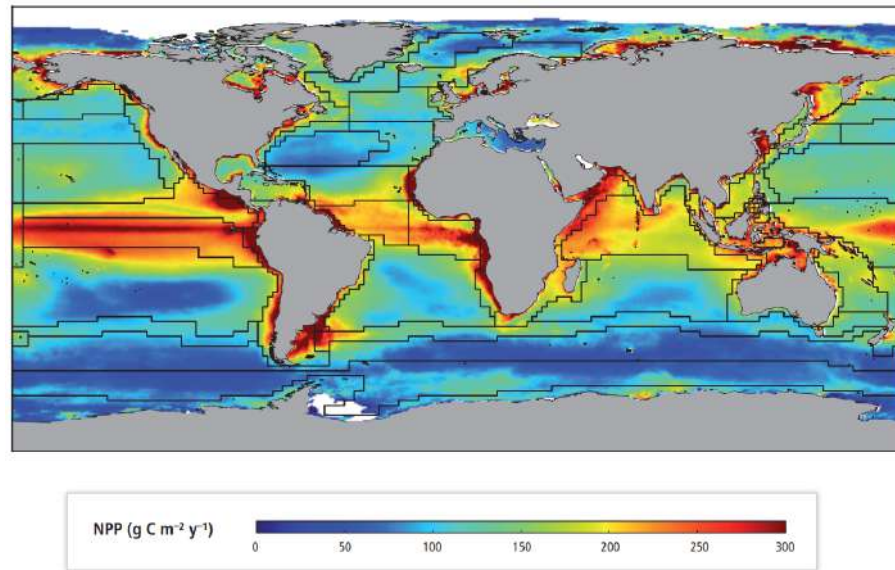


Figura 15. Mapa compuesto anual de tasas de nppv por área global (MODIS 2003-2012). Superpuesta hay una cuadrícula de finas líneas negras que representan 51 provincias biogeográficas oceánicas globales distintas (Tomado de Philip et al., 2014)

2.17. Concentración de biomasa de carbono de fitoplancton

La concentración de biomasa de carbono de fitoplancton es una medida de la cantidad de carbono dentro del fitoplancton. La distribución de la biomasa de carbono fitoplanctónica refleja los patrones globales de productividad primaria marina, y es clave para entender los flujos de carbono en el océano, el papel del fitoplancton en el ciclo global del carbono, y su contribución a la bomba biológica de carbono. (Kostadinov, 2016) (Fig.16).

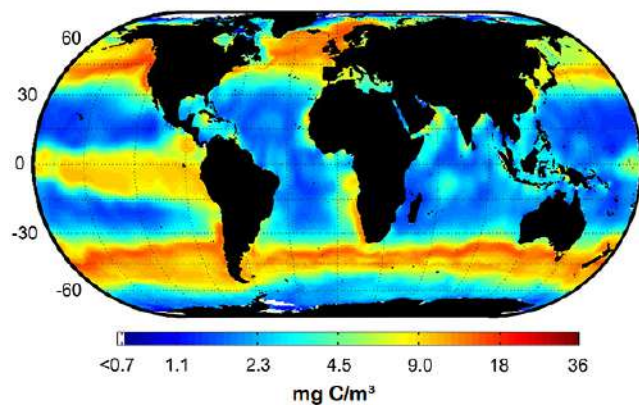


Figura 16. Mapa de distribución espacial de la biomasa de carbono de fitoplancton en la superficie oceánica. (Tomado de AQUA et al., 2023)

2.18. pH

El pH se define como la concentración de iones de hidrógeno en el agua. Su valor representa el grado de acidez o alcalinidad del mar. Los cambios en pH pueden alterar la concentración de otras sustancias en el agua modificando el nivel de toxicidad, por ejemplo: Una disminución en el pH puede aumentar la cantidad de mercurio soluble en el agua. Con una disminución del pH, el agua se hace más ácida y con un aumento de pH el agua se hace más básica.

Las aguas con crecimiento elevado de algas pueden mostrar un cambio diurno en el pH. Cuando las algas crecen y se reproducen usan CO_2 , esta reducción hace que el pH aumente. Por lo tanto, si las condiciones son favorables para el crecimiento de algas cuando la luz del sol entibia la temperatura, el agua será más alcalina. El pH máximo ocurre generalmente por la tarde y desciende por la noche. Puesto que el crecimiento de las algas está sujeto a la luz que penetra en las zonas, el pH puede variar en función de la profundidad en lagos, estuarios, agua de bahías y océanos (State-Water, 2019) (Fig.17).

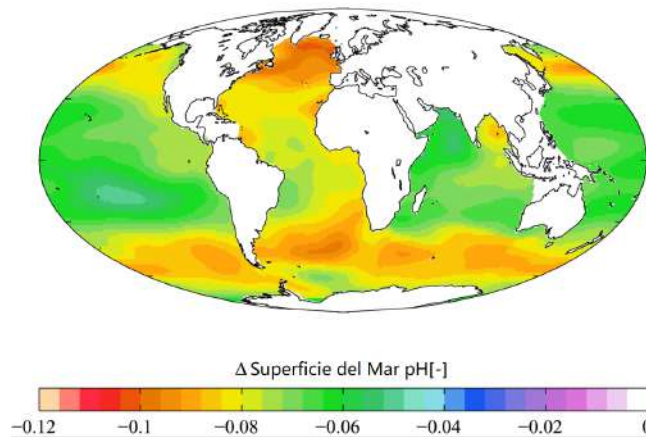


Figura 17. Imagen representativa de los niveles de acidez del océano (Tomado de University-Calgary, 2015)

2.19. Dióxido de carbono

El dióxido de carbono es un gas producido naturalmente a través de la respiración de animales y la degradación de la materia orgánica. Este gas se disuelve en el agua para formar ácido carbónico, el cual reduce el pH del agua. Los océanos son un sumidero de carbono que permite almacenar dióxido de carbono durante cientos de años. Sin embargo, cuando hay demasiado dióxido de carbono en el océano, puede haber efectos adversos en la vida marina (Energía-

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

Atómica, 2024) (Fig.18).

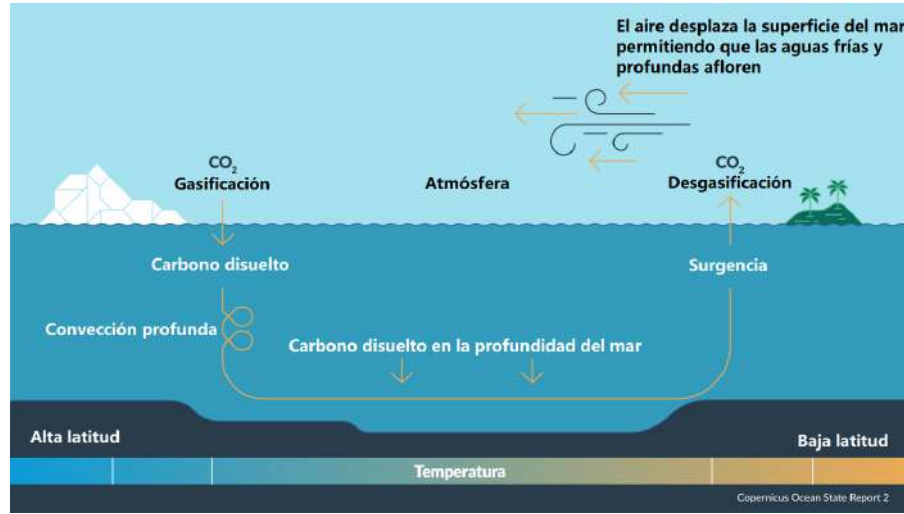


Figura 18. Representación del ciclo del carbono (Tomado de Copernicus-Services, 2024)

3. Desarrollo de modelo para predicción de la concentración y distribución del fitoplancton

Uno de los aspectos más relevantes a considerar en la generación de un modelo predictivo es la definición adecuada de las variables de entrada, ya que esta decisión influye de manera determinante en el nivel de precisión y desempeño del modelo. En el presente trabajo, se experimentó con dos fuentes principales de datos. La primera está conformada por mediciones *in situ*, obtenidas de la plataforma *World Ocean Database (WOD)*, complementadas con datos de irradiancia provenientes de la plataforma *POWER NASA*. La segunda fuente corresponde a un conjunto de datos satelitales, integrados a partir de información proporcionada por las plataformas *Copernicus Marine Service* y nuevamente *POWER NASA*.

3.1. Desarrollo del modelo basado en datos *in situ*

Para el desarrollo del modelo computacional predictivo, se emplearon datos de variables biogeoquímicas recolectados *in situ*, los cuales fueron obtenidos de la plataforma *World Ocean Database (WOD)* estableciendo un periodo de datos de descarga de 20 años (1999 - 2019). En particular, se consideraron aquellas variables, disponibles dentro de la plataforma, consideradas con mayor relevancia dentro de los procesos fisicoquímicos relacionados con el fitoplancton y los

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

procesos biogeoquímicos del océano: concentración de silicato, temperatura potencial del agua, salinidad, concentración de fosfato, concentración de nitrato y concentración de chl-a como variable objetivo.

Tabla 1: Propuesta inicial de variables obtenidas de WOD

Número	Descripción	Unidades
1	Silicatos	$mmol\ m^{-3}$
2	Temperatura potencial del agua de mar	$^{\circ}C$
3	Salinidad del agua de mar	psu
4	Fosfatos	$mmol\ m^{-3}$
5	Nitratos	$mmol\ m^{-3}$
6	Clorofila-a	$mmol\ m^{-3}$

Posteriormente, estos datos se complementaron con mediciones de irradiancia descargadas del portal *POWER NASA*, con el objetivo de enriquecer el conjunto de variables. No obstante, un desafío importante es la integración de información proveniente de plataformas distintas, cada una con su propio formato de almacenamiento, resolución espacial y temporal. Por un lado, los datos de WOD se componen principalmente de muestreos puntuales, realizados en campañas oceanográficas, con una cobertura espacial y temporal irregular y una fuerte dependencia de la disponibilidad logística. Por otro lado, los datos de irradiancia del portal *POWER NASA* se derivan de productos satelitales, que presentan una resolución más homogénea, lo que genera discrepancias cuando se busca alinear ambos conjuntos de datos.

Para resolver esta incompatibilidad, se diseñó un algoritmo de preprocesamiento capaz de realizar la interpolación espacial y temporal adecuada. Este algoritmo consideró la adecuación de coordenadas y la eliminación de valores atípicos, de manera que el conjunto final fuera consistente y utilizable como entrada para el modelo predictivo. Esta etapa fue muy importante para minimizar los errores sistemáticos de las fuentes de datos y asegurar el correcto conjunto de datos para el posterior entrenamiento del modelo.

3.1.1. Algoritmo 1: Integración de información de plataformas distintas

Una vez generado el conjunto de datos con la propuesta inicial de variables con datos *in situ* se implementó el Algoritmo 1 para incorporar la variable irradiancia. Derivado de la diferencia de resolución espacial que existe entre los datos descargados de la plataforma *World Ocean Database (WOD)* y los datos de irradiancia descargados de la plataforma *NASA POWER*, se

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

planteó una estrategia para realizar una combinación estableciendo un margen de $\pm 0.25^\circ$ en longitud y latitud para interpolar espacialmente cada registro.

La plataforma *NASA POWER* genera un archivo independiente por cada día de lecturas, lo que resulta en una amplia colección de archivos para cubrir el periodo de 20 años de los registros de WOD, por lo que se decidió limitar el conjunto de datos a un periodo de estudio de 5 años (1999 al 2005). Para esto, el algoritmo: (i) Carga y recorre secuencialmente todos los archivos de irradiancia, (ii) identifica y promedia las lecturas dentro del margen predefinido previamente, de tal manera que la fecha coincida en cada registro, y (iii) asigna el valor resultante a la base de datos. Con este enfoque se logra la integración de la irradiancia, superando las limitaciones por la disparidad en la resolución espacial de las fuentes de datos.

Algoritmo 1: Incorporación de la variable de irradiancia a la propuesta inicial de variables

Entrada: Archivo de variables (*input_file*)

Lista de archivos con datos de irradiancia (*irra_file_N*)

Cantidad de archivos de irradiancia (*contIrraFiles*)

Salida : Archivo de variables e irradiancia

```

1  output_file ← creación y apertura del archivo de salida
2  output_file.write('Fecha', 'Latitud', 'Longitud',..., 'Variables con irradiancia')
3  data_frame ← pd.read_csv(input_file, usecols=['Fecha', 'Latitud', 'Longitud', 'Variables'])
4  for registro in data_frame do
5      numIrraFiles = 1
6      contIrradiance = 0
7      sumIrradiance = 0
8      while numIrraFiles ≤ contIrraFiles do
9          for reg_irra in data_irra do
10             lat_irra ← reg_irra['latitude']
11             lon_irra ← reg_irra['longitude']
12             fecha ← reg_fecha['fecha']
13             irradiancia ← reg_irra['irradiancia']
14             coordRange = 0.25
15             numIrraFiles += 1
16         if contIrradiance > 0 then
17             output_file.write(variables+str(averageIrradiance))
18 return output_file

```

3.1.2. Implementación del modelo para la predicción de fitoplancton (*Chla-a*)

La implementación del diseño y entrenamiento del modelo se realizó en el lenguaje de programación Python, utilizando la biblioteca PyTorch. La arquitectura propuesta para la red neuronal se detalla en la Tabla 2. La capa de entrada consta de un número de neuronas igual a la cantidad de variables biogeoquímicas y físicas incluyendo la irradiancia, que son recibidas como entradas para cada muestra de entrenamiento. Por su parte, la capa de salida consta de una única neurona, que regresa el valor predicho de concentración de *chla-a*. El número de capas ocultas y neuronas por capa se determinó mediante experimentación iterativa, buscando

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

un equilibrio entre precisión predictiva y eficiencia computacional.

Tabla 2: Arquitectura de la red neuronal utilizada

Capa	Tipo	Neuronas	Activación
Entrada	Totalmente conectada	16	ReLU
Oculto	Totalmente conectada	1000	ReLU
Oculto	Totalmente conectada	1000	ReLU
Salida	Totalmente conectada	1	No Lineal

3.1.3. Entrenamiento del modelo

El proceso de entrenamiento se implementó a través del Algoritmo 2, que sigue un enfoque basado en lotes (*batch processing*), donde una vez definidos los parámetros de entrenamiento, el algoritmo ejecuta un ciclo principal que controla el número de épocas a realizar. Dentro de este ciclo, se activa el modo de entrenamiento, e inicia un ciclo interno donde el conjunto de datos es recorrido en lotes de 512 muestras. En este primer ciclo interno, el algoritmo realiza los cuatro pasos fundamentales del proceso de entrenamiento: 1) procesamiento de los lotes de muestra y obtención de las predicciones de la concentración de *chla-a*; 2) cálculo del error a partir de las predicciones obtenidas y los valores reales de concentración de *chla-a*; 3) pasos hacia atrás o *backward propagation* donde el error calculado se propaga hacia atrás a través de la red neuronal, desde la capa de salida hasta la capa de entrada; 4) actualización de pesos y bias utilizando los gradientes calculados para ajustar los pesos de las conexiones entre las neuronas.

Tras cada iteración, el algoritmo monitorea el desempeño del modelo. Para ello, calcula y acumula el error promedio sobre los datos de validación, evalúa si dicho error disminuye respecto a iteraciones anteriores. Si el error disminuye, el contador de paciencia se reinicia y almacena el modelo actual como el mejor obtenido. En caso contrario, el contador de paciencia se incrementa y si supera el umbral predefinido, finaliza el entrenamiento para evitar el sobre entrenamiento.

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

Algoritmo 2: Entrenamiento del modelo

Entrada: Conjunto de datos de entrenamiento

Salida : Mejor modelo generado

```
1 learning_rate  $\leftarrow$  0.001
2 loss_function  $\leftarrow$  MSE_Loss()
3 optimizer  $\leftarrow$  Adam(lr = learning_rate)
4 num_epochs  $\leftarrow$  1000
5 best_test_loss  $\leftarrow$  float('inf')
6 patience  $\leftarrow$  0
7 for epoch in range(1, num_epochs) do
8     model.train()
9     for (data_train, chla_train) in train_database do
10         chla_pred  $\leftarrow$  model(data_train)
11         loss_train  $\leftarrow$  loss_function(chla_pred, chla_train)
12         loss_train.backward()
13         optimizer.step()
14     model.eval()
15     total_loss_test  $\leftarrow$  0
16     for (data_test, chla_test) in test_database do
17         chal_pred  $\leftarrow$  model(data_test)
18         loss_test  $\leftarrow$  loss_function(chla_pred, chla_test)
19         total_loss_test  $\leftarrow$  total_loss_test + loss_test
20     current_test_loss  $\leftarrow$  total_loss_test/len(test_database)
21     if current_test_loss < best_test_loss then
22         patience  $\leftarrow$  0
23         best_test_loss  $\leftarrow$  current_test_loss
24         save(model, "model.pth")
25     else
26         patience  $\leftarrow$  patience + 1
27         if patience > 100 then
28             break
```

El modelo se entrenó con el 70 % del conjunto de datos, reservando el 30 % restante para su evaluación final. El entrenamiento finalizó en la época 686, utilizando un criterio de parada temprana con una paciencia de 100 épocas. Los parámetros empleados durante el entrenamiento del modelo se describen en la Tabla 3.

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

Tabla 3: Parámetros del entrenamiento del modelo

Parámetro	Valor
Número de muestras	12,961
Tipo de normalización	Standard Scaler
Tamaño de Lote	512
Tasa de aprendizaje	0.001
Función de pérdida	MSE Loss
Función de optimización	Adam
Número máximo de épocas	1000
Épocas de paciencia	100

3.2. Desarrollo del modelo basado en datos satelitales

Para el desarrollo del segundo modelo computacional se hizo uso de la plataforma *Copernicus Marine Service*, la cual es una plataforma internacional que pone a disposición del público una amplia variedad de datos marinos recopilados mediante observaciones realizadas por satélites que monitorean de manera continua las condiciones de los océanos, registrando diferentes variables de las cuales es posible conocer su estado con gran detalle incluyendo información específica sobre la fecha en la que fue tomado y la ubicación exacta donde se realizó la observación.

La definición de variables a considerar dentro del modelo es un factor importante para alcanzar índices aceptables de precisión en la predicción del modelo. Dado que la variable objetivo es el valor de concentración de chl-a, toda variable relacionada, directa o indirectamente, con el proceso de la fotosíntesis podría tener un gran impacto en los resultados del modelo. Es decir, se puede asumir que el peso de cada una de estas variables dentro de este proceso químico, aportará información relevante al modelo para lograr una mejor relación con la concentración de la chl-a.

Este estudio abarcó la región del Pacífico Tropical Oriental adyacente a la costa peruana (0.8-12 °S; 76°-87 °O; Figura 19), analizando el periodo comprendido entre el 31 de diciembre de 1999 y el 31 de diciembre de 2005; esta ventana temporal permite observar las condiciones posteriores al evento de El Niño (Wang and Weisberg, 2000). Para el desarrollo del modelo (entrenamiento y validación), se emplearon dos conjuntos de datos de acceso abierto obtenidos de la plataforma *Copernicus Marine Service* en formato NetCDF: 1) Variables biogeoquímicas provenientes del producto *Global Ocean Biogeochemistry Hindcast*: (<https://data.marine.>

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

copernicus.eu/product/GLOBAL_MULTIYEAR_BGC_001_029/description, Tabla 4); 2) Variables físicas obtenidas del *Global Ocean Ensemble Physics Reanalysis* (https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_ENS_001_031/description, Tabla 5). Ambos conjuntos presentan una resolución espacial de $1/4^\circ$ (4 km/píxel). Adicionalmente, se incorporaron datos de irradiancia superficial, obtenidos del proyecto *NASA Prediction of Worldwide Energy Resources* (<https://power.larc.nasa.gov/>). La variable objetivo seleccionada para el entrenamiento del modelo fue la concentración de chl-a (g/L), un proxy clave de la biomasa fitoplanctónica en estudios biogeoquímicos marinos.



Figura 19. Área de estudio de la investigación en el pacífico tropical-subtropical frente a Perú

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

Tabla 4: Variables biogeoquímicas obtenidas de la plataforma de Copernicus Marine Service

Número	Descripción	Unidades
1	Hierro disuelto	$mmol\ m^{-3}$
2	Nitratos	$mmol\ m^{-3}$
3	Producción neta de biomasa expresada como carbono	$mg\ m^{-3}$
4	Oxígeno disuelto	$mmol\ m^{-3}$
5	Potencial de hidrógeno	—
6	Concentración de fitoplancton expresada como carbono	$mmol\ m^{-3}$
7	Fosfatos	$mmol\ m^{-3}$
8	Silicatos	$mmol\ m^{-3}$
9	Presión parcial de dióxido de carbono	Pa

Tabla 5: Variables físicas obtenidas de la plataforma de Copernicus Marine Service

Número	Descripción	Unidades
1	Espesor de la capa mixta oceánica definido por sigma theta	m
2	Salinidad del agua de mar	psu
3	Temperatura potencial del agua de mar	C°
4	Velocidad del agua del mar hacia el este	m/s
5	Velocidad del agua del mar hacia el norte	m/s
6	Altura de la superficie del mar sobre el geoide	m

3.2.1. Relevancia de cada variable

La selección de variables se fundamentó en el rol de cada parámetro biogeoquímico y físico para modular la distribución, crecimiento y productividad fitoplanctónica, considerados parámetros clave en los ecosistemas marinos. Estas variables actúan como proxies del estado ecológico oceánico y se integran en mecanismos de retroalimentación con procesos globales, incluyendo el secuestro de carbono, la acidificación oceánica y los patrones de variabilidad climática. Por ejemplo, nutrientes como el nitrato, el hierro disuelto, el fosfato y el silicato son macronutrientes esenciales para la síntesis de biomoléculas en el fitoplancton (Moore TS et al., 2019), la biomasa de carbono y carbono fitoplanctónico como indicadores clave de la fijación de durante la fotosíntesis (Behrenfeld and Falkowski, 1997), el oxígeno disuelto es un indicador de balance entre la producción (fotosíntesis) y el consumo (respiración microbiana), con declives asociados al calentamiento oceánico (Lueker et al., 2000). El pH y la presión parcial de dióxido de carbono tienen un impacto significativo en la fisiología del fitoplancton y en la acidificación de los océanos (Doney et al., 2009)(Siegel et al., 2014). Por otro lado, las variables físicas como la temperatura regulan la tasa de crecimiento del fitoplancton (Van Heukelem and Thomas, 2001), la salinidad

3. DESARROLLO DE MODELO PARA PREDICCIÓN DE LA CONCENTRACIÓN Y DISTRIBUCIÓN

tiene un impacto significativo en la tasa de crecimiento del fitoplancton y en la composición de su comunidad, afectando procesos biológicos y ecológicos clave. Por otra parte, los forzantes atmosféricos como las corrientes impulsan la advección de nutrientes en sistemas de afloramiento, siendo un factor clave en la heterogeneidad espacial del fitoplancton. Mientras que el espesor de la capa mixta determina la disponibilidad de nutrientes en la zona eufótica (Soranno et al., 2017). La altura superficial del mar determinada por el promedio de la superficie del océano entre la marea alta y la marea baja, sus anomalías se correlacionan con eventos ENSO, mostrando una relación inversa con la productividad primaria durante fases cálidas de 'El Niño' (Bates et al., 2014) y la irradiancia es un factor determinante en la profundidad óptima para la fotosíntesis (Kirk, 1994).

3.2.2. Preprocesado de datos

El preprocesado de datos se realizó en tres etapas a partir del diseño e implementación de algoritmos secuenciales para exportar, validar y consolidar los datos en un único archivo estructurado. El conjunto de datos resultante consta de 17,140 registros, cada uno de los cuales incluye la fecha, latitud y longitud correspondientes a una lectura específica, junto con nueve variables biogeoquímicas, seis variables físicas y la irradiancia, utilizadas como variables de entrada en el modelo. Además, se incorporó la variable de concentración de chl-a como variable de salida. A continuación, se muestran los algoritmos utilizados:

3.2.3. Algoritmo 3: Conversión de datos NetCDF a CSV

Se desarrolló el Algoritmo 3 para extraer, validar y transformar los datos almacenados en formato NetCDF a un archivo estructurado en formato CSV.

Algoritmo 3: Conversión de documentos en formato .NC a formato .CSV

Entrada: Archivo en formato NetCDF con datos de las variablesRango de fechas de las lecturas (*DIAS*)Rango de latitudes de las lecturas (*LATS*)Rango de longitudes de las lecturas (*LONS*)**Salida :** Archivo estructurado con las variables en formato .CSV

```

1 data_in ← apertura del archivo NetCDF a través de la librería H5PY
2 data_out ← creación y apertura del archivo de salida en formato .CSV
3 data_out.write("Fecha, Latitud, Longitud, ...")

4 for dia in range(DIAS) do
5     for lat in range(LATS) do
6         for lon in range(LONS) do
7             // se extraen datos del archivo NetCDF:
8             variables ← data_in.get(variables)[dia, lat, lon]
9             // se validan datos:
10            if all([es_valido(variables)]) then
11                // se almacena un registro:
12                data_out.write(date, latitude, longitude, ...)

10 return data_out

```

3.2.4. Algoritmo 4: Combinación de variables biogeoquímicas y físicas

A partir de los archivos independientes en formato .CSV generados previamente, se implementaron en el Algoritmo 4 para realizar la mezcla espacio-temporal de registros mediante una operación 'innerjoin' basada en claves primarias compuestas (fecha, latitud, longitud).

Algoritmo 4: Unión de los conjuntos de datos de variables físicas y biogeoquímicas

Entrada: Archivo de variables biogeoquímicas (*bio_file*)
 Archivo de variables físicas (*phy_file*)
 Tamaño de bloque de mezcla (*chunk_size*)

Salida : Archivo con variables biogeoquímicas y físicas mezcladas

```

1  data_out ← creación y apertura del archivo de salida
2  data_out.write("Fecha, Latitud, Longitud, Variables mezcladas")
3  for chunk_bio in pd.read_csv(bio_file, chunksize=chunk_size, usecols=['Fecha', 'Latitud', 'Longitud',
    'Variables biogeoquímicas']) do
4      chunk_bio['Fecha'] ← convierte a formato yyyy-mm-dd
5      for chunk_phy in pd.read_csv(phy_file, chunksize=chunk_size, usecols=['Fecha', 'Latitud',
    'Longitud', 'Variables físicas']) do
6          chunk_phy['Fecha'] ← convierte a formato yyyy-mm-dd
7          merged_chunk ← pd.merge(chunk_bio, chunk_phy, on=['Fecha', 'Latitud', 'Longitud'],
    how='inner')
8          merged_chunk ← merged_chunk[['Fecha', 'Latitud', 'Longitud', 'Variables reordenadas']]
9          merged_chunk.to_csv(data_out, header = False, index = False)
10         delete merged_chunk
11         gc.collect()
12 return data_out

```

3.2.5. Incorporación de la variable irradiancia

Una vez integradas las variables biogeoquímicas y físicas, se implementó el Algoritmo 1 utilizado anteriormente para incorporar la variable irradiancia al conjunto de datos. A diferencia de los registros biogeoquímicos y físicos, que comparten una resolución espacial compatible, los datos de irradiancia descargados de la plataforma *NASA POWER* presentan una resolución espacial distinta a los datos de *Copernicus Marine Service*. Debido a esta discrepancia, no fue posible realizar una combinación basada en coincidencias exactas de fecha, latitud y longitud. En su lugar, se estableció un margen de $\pm 0.25^\circ$ en ambas direcciones, permitiendo asignar un valor de irradiancia a cada registro mediante interpolación espacial.

4. Experimentos y resultados de la aplicación de los modelos

4.1. Experimentos con el modelo basado en datos in situ

Para predecir la cantidad de biomasa del fitoplancton en términos de chl-a, en este trabajo se evaluaron diferentes conjuntos de datos que corresponden a variables oceanográficas que intervienen en los procesos fisicoquímicos del desarrollo del fitoplancton. En la Tabla 1 se representa el conjunto de datos descargado de la plataforma *World Ocean Database (WOD)*, considerado para el entrenamiento y prueba del modelo con las variables de propuesta inicial.

Los datos se descargaron sin establecer un espacio temporal definido, con el fin de tener la mayor cantidad de información posible. Posteriormente, se realizó un preprocesado donde se eliminan los registros con algún tipo de anomalía o no relevantes para la investigación como valores nulos, vacíos, duplicados, entre otros.

Por medio del desarrollo de un algoritmo se depuró el conjunto de datos para contemplar únicamente las lecturas ubicadas en la zona eufótica del mar, y así mantener la información bajo un mismo criterio de profundidad. El resultado de la depuración generó 2,872 registros, que abarcan datos desde el año de 1999 a la fecha. Como parte del preprocesado se promedió el valor de la concentración de chl-a con base en los campos de embarcación y la fecha de los registros. Posteriormente, se almacenó el conjunto de datos en formato .CSV.

Una vez depurado y normalizado el conjunto de datos, el modelo se entrenó durante 1,000 épocas, utilizando el 70 % de los registros para el entrenamiento y reservando el 30 % restante para las pruebas. Se estableció una tasa de aprendizaje de 0.001 y un criterio de parada temprana, el cual se reinicia cada vez que el error disminuye, en caso contrario el contador de paciencia aumenta hasta el límite de 100 épocas.

Los resultados obtenidos muestran que el modelo logra disminuir el error y converger de una manera muy rápida con los datos de entrenamiento, sin embargo, respecto a los datos de prueba, los resultados son irregulares, el modelo genera una reducida curva de aprendizaje inicial para posteriormente aumentar el error de forma pronunciada. A partir de estos resultados, se concluyó que las variables utilizadas, no son las adecuadas o que es necesario incluir más variables dentro del conjunto de datos (Fig 20).

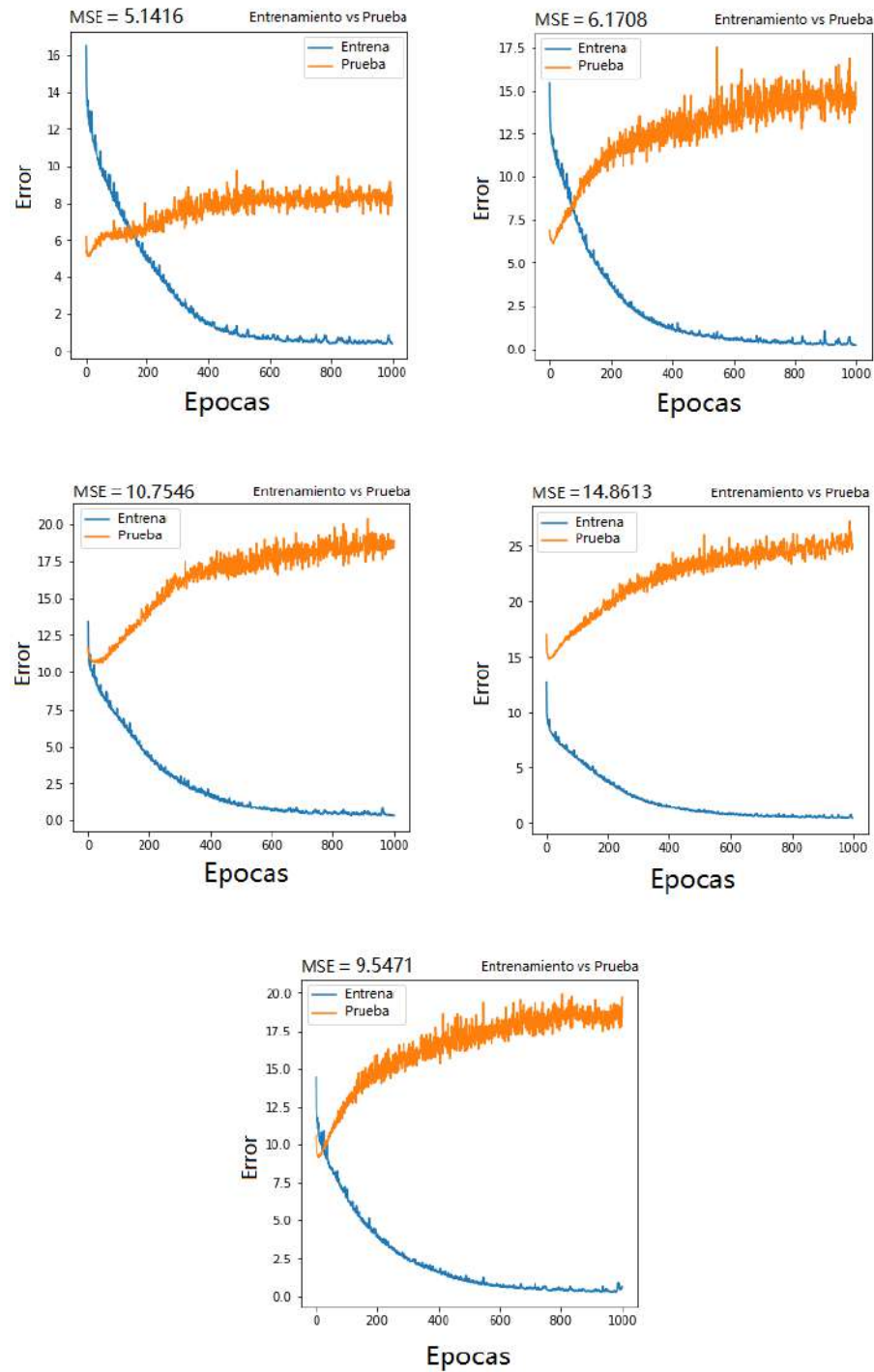


Figura 20. Resultado de cinco pruebas independientes de predicción de concentración de chla-a considerando las variables de salinidad, temperatura, fosfato, nitrato, silicato y chla-a.

El siguiente experimento consistió en agregar el dato de irradiancia a las seis variables de entrada del experimento anterior. Por medio del portal *Power NASA* se descendieron de los datos de irradiancia, bajo la definición 'radiación visible' por el hecho de ser el único de los tres

grupos de irradiancia disponibles que interviene dentro del proceso de fotosíntesis. Los datos se descargaron considerando el mismo periodo de tiempo que el conjunto de datos de las seis variables previas, y se unieron con estas en un solo archivo .CSV, basado en los campos de longitud, latitud y fecha de los registros.

El resultado de cinco pruebas independientes con el nuevo conjunto, muestra un comportamiento similar al experimento anterior en cuanto a los datos de entrenamiento. Con respecto a los datos de prueba, el modelo logra una disminución del error generando una curva ligeramente más pronunciada que en el experimento anterior, sin embargo, aun con niveles de precisión poco aceptables (Fig 21).

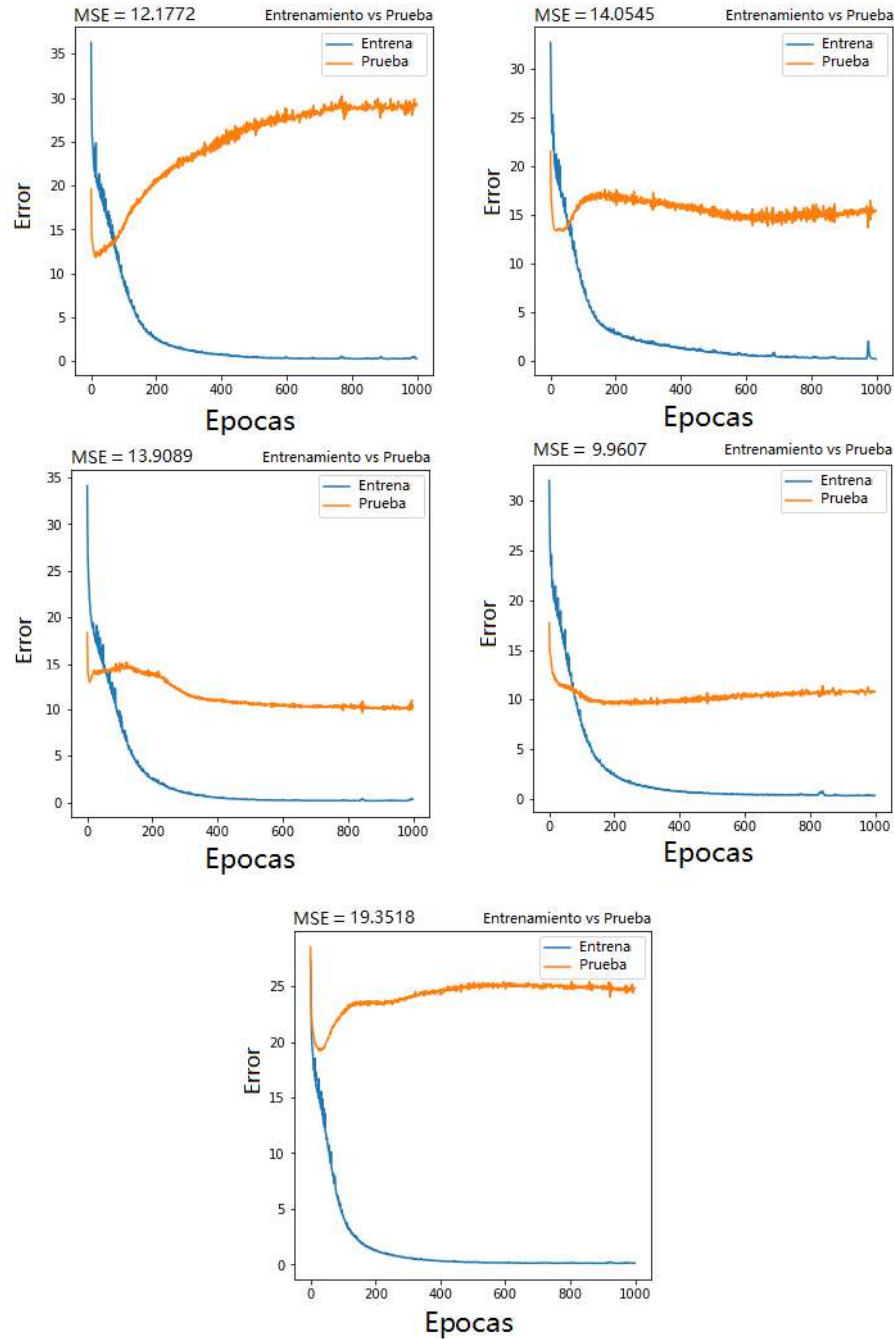


Figura 21. Resultado de cinco pruebas independientes considerando la variable de irradiancia.

4.2. Experimentos con el modelo basado en datos satelitales

World Ocean Database proporciona una gran cantidad de registros pero sin duda limitante es la poca diversidad de variables a disposición, por lo cual, se recurrió al uso de la plataforma *Copernicus Marine Service*, de donde se descargó un conjunto de datos bajo la etiqueta de

'producto biogeoquímico', contemplando 20 años de registros con 16 variables tanto físicas como químicas. Con el objetivo de identificar un subconjunto representativo de estas 16 variables fisicoquímicas para la predicción de la concentración del fitoplancton, se desarrolló un algoritmo basado en el Análisis de Componentes Principales (PCA, por sus siglas en inglés). La Tabla 6 muestra el ranking y puntaje de las variables obtenido por PCA.

Tabla 6: Puntaje y ranking asignado a cada variable por el PCA.

Número	Variable	Valor
1	EsteVelocidad	0.269
2	NorteVelocidad	0.243
3	Oxígeno	0.238
4	CapaOceano	0.202
5	Hierro	0.197
6	Silicato	0.195
7	Temperatura	0.180
8	AlturaSuperficie	0.178
9	BioCarbon	0.178
10	PhytoCarbon	0.172
11	PH	0.165
12	Salinidad	0.160
13	PresionDióxido	0.159
14	Fosfato	0.117
15	Nitrato	0.115

La Figura 22 presenta la proporción de varianza explicada por cada una de las 16 variables biogeoquímicas consideradas en el modelo, así como la varianza acumulada conforme se incorporan las variables sucesivamente. Se observa que las primeras variables concentran la mayor parte de la variabilidad de los datos, alcanzando más del 85 % de la varianza explicada con aproximadamente siete variables.

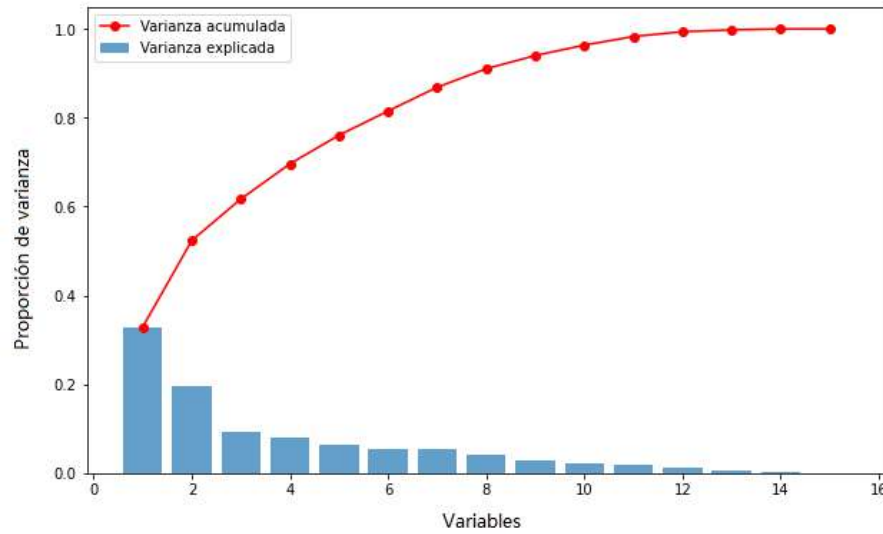


Figura 22. Medición de la dispersión o la cantidad de información (variabilidad) que aportan los datos.

A partir del ranking de la Tabla 6, se realizaron diferentes pruebas utilizando una cantidad diferente de variables.

4.2.1. Resultados de las pruebas con las primeras ocho variables de acuerdo al ranking obtenido del PCA

La Figura 23 muestra los resultados de tres experimentos independientes considerando únicamente las ocho primeras variables del ranking. En la primera columna de cada experimento muestra la tendencia de la media del error cuadrado (MSE) a lo largo de las épocas, tanto para el conjunto de entrenamiento como para el de prueba. En estas gráficas se observa cómo el error con los datos de entrenamiento disminuye de manera sostenida conforme avanza el proceso de entrenamiento, mientras que el error con los datos de prueba tiende a mantenerse después de un número determinado de épocas, alcanzando una Media del Error Cuadrado (MSE) de 0.0682, 0.0681 y 0.0678 respectivamente.

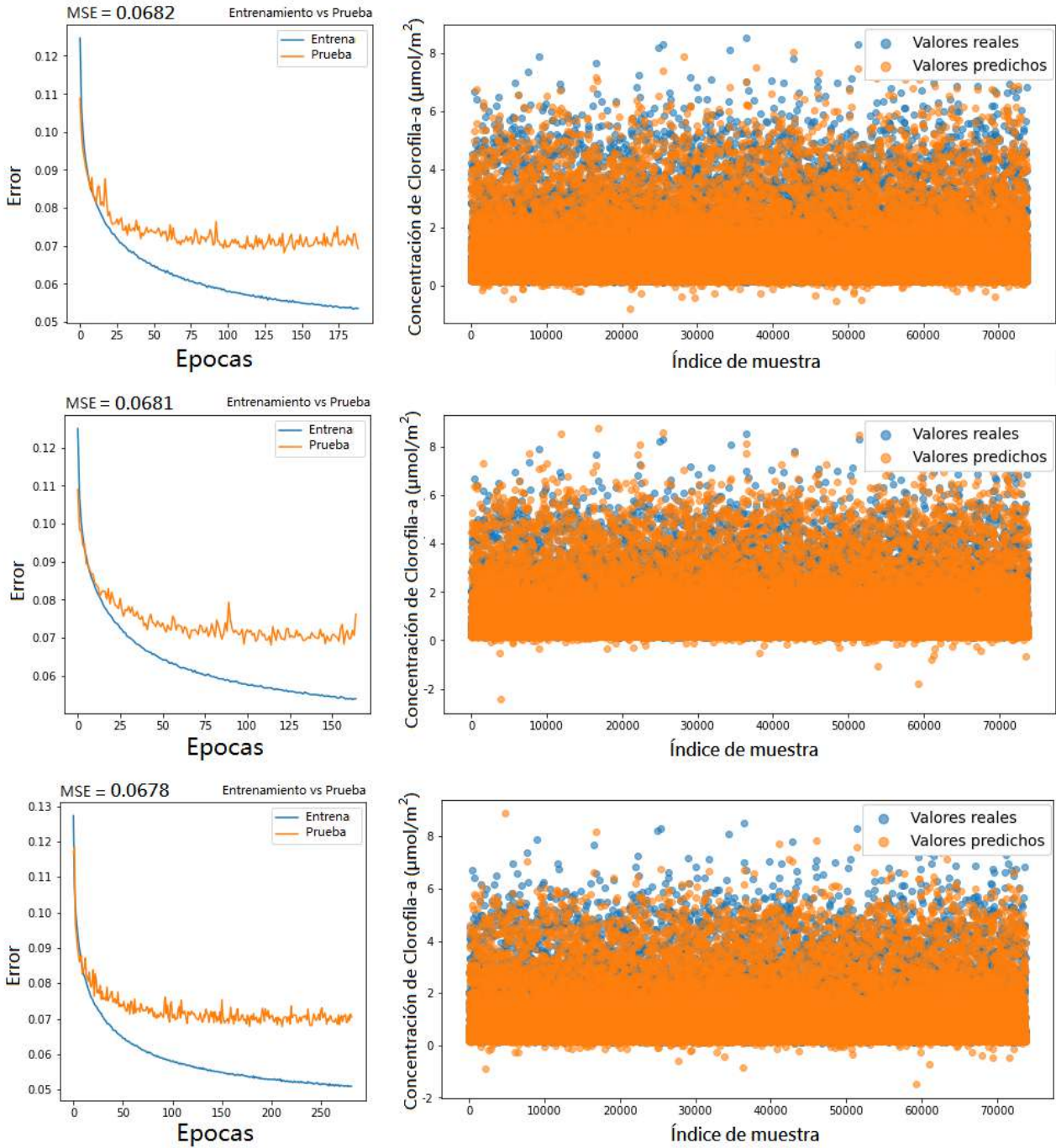


Figura 23. Prueba con las primeras ocho variables de acuerdo el ranking. Concentración de chla-a ($\mu\text{mol}/\text{m}^2$). Valores reales (puntos azules) y valores predichos (puntos naranjas).

La segunda columna de cada experimento corresponde a la comparación entre las concentraciones reales de chla-a y las predicciones generadas por el modelo sobre el conjunto de prueba. Cada punto azul representa un valor real de chla-a, mientras que cada punto naranja corresponde al valor predicho para la misma muestra. La proximidad de los puntos predichos a los valores

reales refleja la precisión alcanzada por el modelo, es decir, cuanto más cercanos se encuentran ambos conjuntos de puntos, mayor es la precisión de la predicción.

La Figura 24 muestra la correlación entre los valores reales y los valores predichos de concentración de chl-a por el modelo. La concentración de puntos en torno a la diagonal indican una mayor precisión de predicción del modelo. Con este modelo se obtuvo un coeficiente de determinación $R^2 = 0.87$.

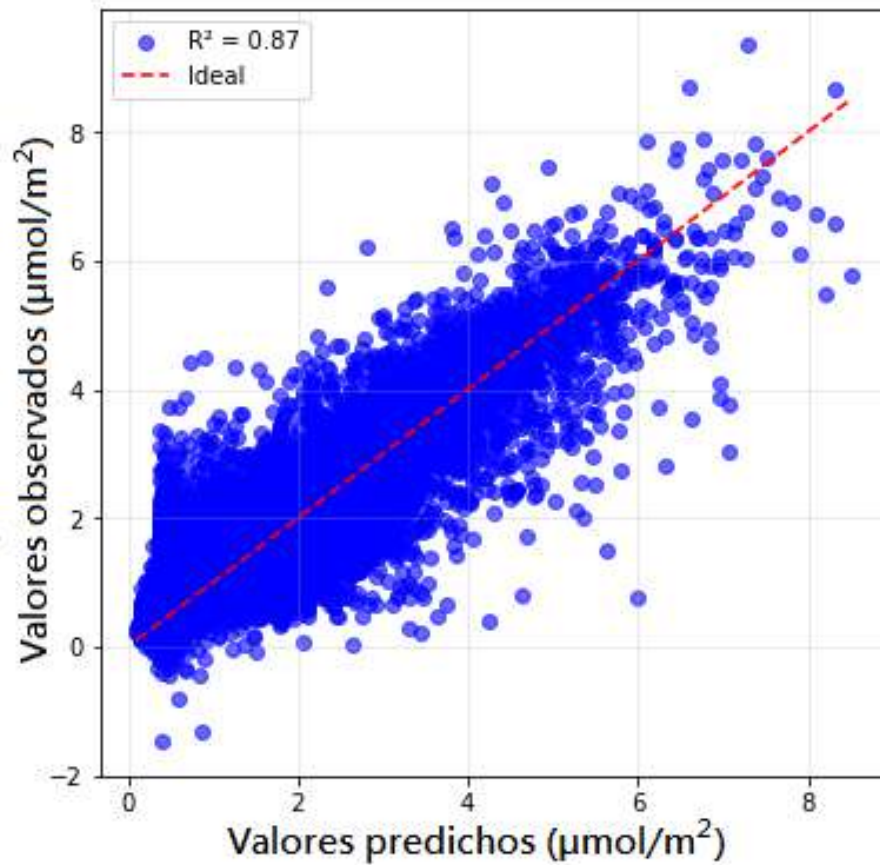


Figura 24. Correlación de los valores observados vs predichos de la concentración de chl-a ($\mu\text{mol}/\text{m}^2$). Coeficiente de determinación ($R^2 = 0.87$).

4.2.2. Resultados de las pruebas con las primeras diez variables de acuerdo el ranking obtenido del PCA

La Figura 25 muestra tres experimentos considerando las 10 primeras variables según el ranking del PCA. En estas pruebas se obtuvo una Media del Error Cuadrado (MSE) de entre 0.0022 y 0.0023. La Figura 26 muestra la gráfica de correlación entre los valores predichos y reales, con un coeficiente de determinación $R^2 = 0.95$.

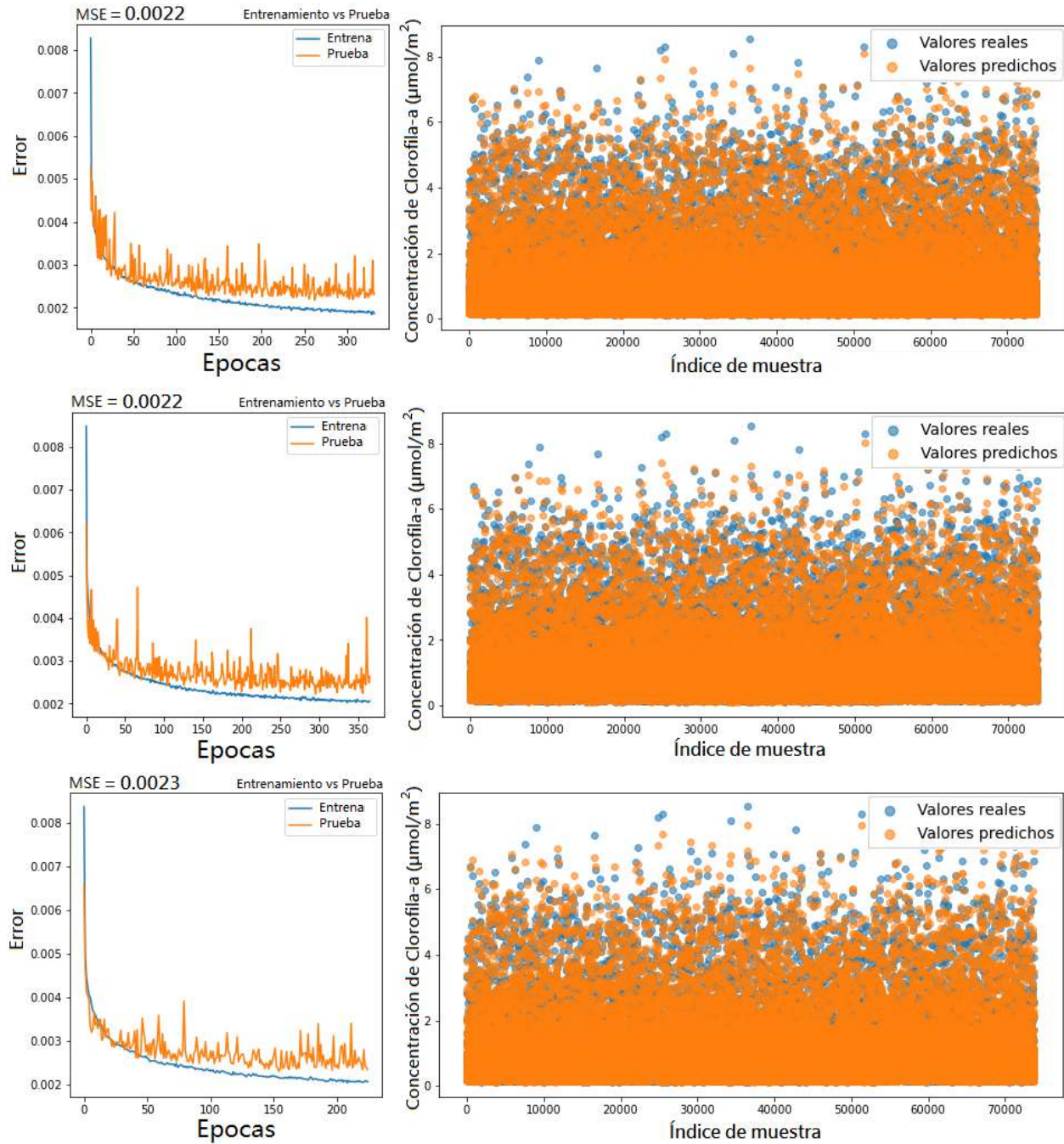


Figura 25. Prueba con las primeras diez variables de acuerdo el ranking. Concentración de chl-a (mol/m^2). Valores reales (puntos azules) y valores predichos (puntos naranja).

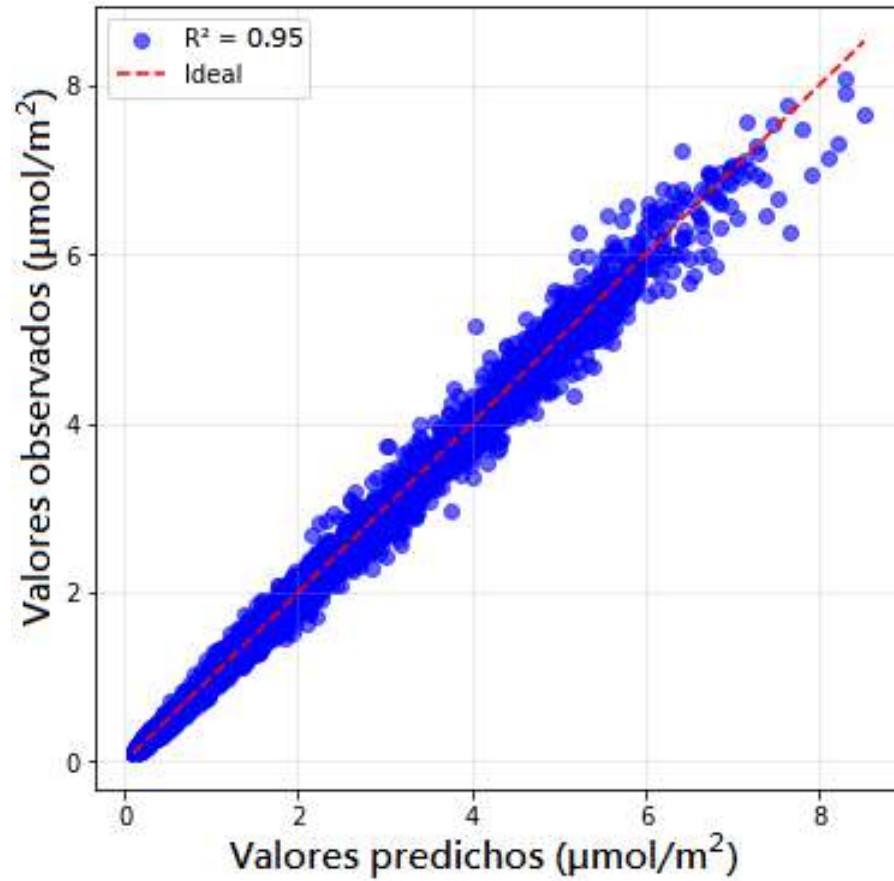


Figura 26. Correlación de los valores observados vs predichos de la concentración de chl-a (mol/m^2). Coeficiente de correlación ($R^2 = 0.95$).

4.2.3. Resultados de las pruebas utilizando todas las variables

Con base a los resultados previos, se generó la hipótesis que el modelo podría mejorar su desempeño a medida que más variables se incorporaran. Por lo que el siguiente experimento consistió en entrenar el modelo con las 16 variables.

Para evaluar el desempeño del modelo, se generaron predicciones de concentración de chl-a sobre el conjunto de prueba y se compararon con los valores de referencia. La Figura 27 muestra la dispersión entre los valores observados y predichos de chl-a, los cuales varían en un rango de 0 y 1 mol/m^2 , con un error máximo de 0.2875 mol/m^2 .

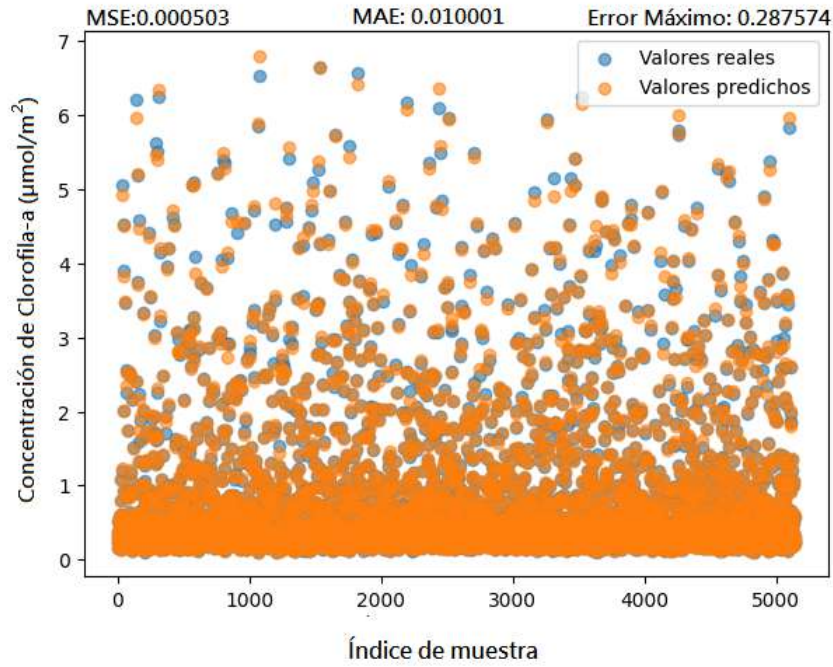


Figura 27. Concentración de chl-a ($\mu\text{mol}/\text{m}^2$). Valores observados (puntos azules) y valores predichos (puntos naranja).

La Figura 28 muestra la gráfica de correlación entre los valores reales de chl-a y los valores predichos por el modelo. Cada punto de la gráfica corresponde a una muestra de prueba, su posición en el eje X representa el valor de chl-a predicho, y su posición en el eje Y el valor real. Mientras más cercanos se encuentren los puntos a la diagonal, representada con la línea en color rojo, más precisa es la predicción del modelo. La Figura muestra el coeficiente de determinación R^2 de las pruebas del modelo, con un valor de 0.9994. Este coeficiente es un indicador que mide de forma cuantitativa la precisión de un modelo y la calidad de sus predicciones. Un valor cercano a cero indica que un modelo no tiene capacidad predictiva, mientras que un valor cercano a uno indica un alto nivel predictivo del mismo.

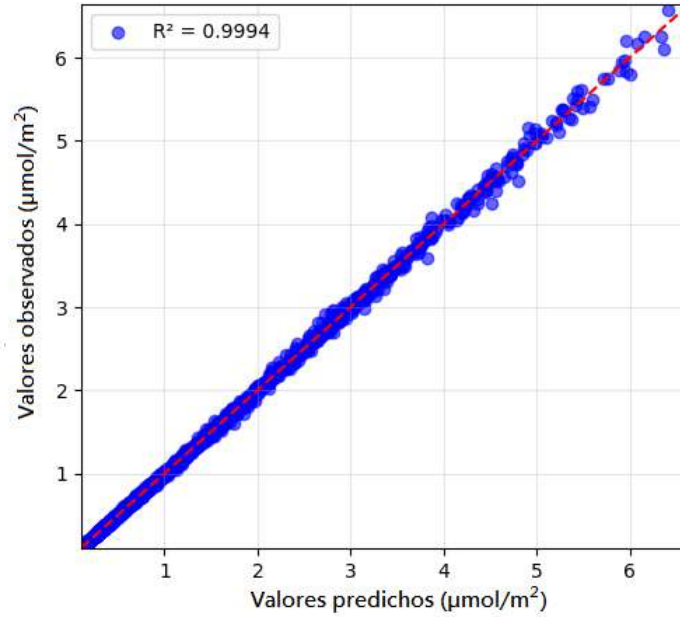


Figura 28. Correlación de los valores observados vs predichos de la concentración de chl-a ($\mu\text{mol}/\text{m}^2$). Coeficiente de correlación ($R^2 = 0.9994$).

El análisis de errores del modelo se realiza mediante mapas de calor, que permiten visualizar de manera clara las discrepancias entre los valores predichos y observados, así como identificar Falsos Positivos (FP) y/o Falsos Negativos (FN). Los valores de concentración de chl-a del conjunto de prueba se segmentaron en intervalos específicos, dividiendo los datos en dos grupos según su rango de concentración. El primer grupo incluye muestras de valores de chl-a de 0 a $1 \text{ mol}/\text{m}^2$, representando el 83 % del total de datos de prueba. Debido a la alta densidad de muestras en este intervalo, los valores se clasificaron en intervalos de $0.1 \text{ mol}/\text{m}^2$. En contraste, el segundo grupo abarca muestras con valores $> 1 \text{ mol}/\text{m}^2$, correspondiente al 17 % restante del conjunto de prueba, y se agrupan en intervalos de $0.5 \text{ mol}/\text{m}^2$.

El mapa de calor reveló un desempeño detallado del modelo al comparar los valores predichos con los observados. La Figura 29 representa el primer grupo, el eje X corresponde a los valores discretizados, usando intervalos de $0.1 \text{ mol}/\text{m}^2$, predichos por el modelo y el eje Y a los valores discretizados reales. Los valores en las celdas de la diagonal principal representan la cantidad de muestras con una predicción correcta. Por ejemplo, la celda con el valor 529 indica que 529 muestras con valor real de $0.1 \text{ mol}/\text{m}^2$, fueron predichas correctamente con ese valor. En el caso de la celda que se encuentra a un lado, con el valor 16, indica que 16 muestras con

valor real de 0.1 mol/m^2 se predijeron con un valor de 0.2. La celda por debajo, con valor 8, indica que 8 muestras con valor real de 0.2 mol/m^2 se predijeron con valor de 0.1. De esta forma, tomando como referencia el valor real de 0.1 mol/m^2 , el mapa muestra 529 predicciones Positivas Verdaderas (TP), 16 Falsos Negativos (FN) y 8 Falsos Positivos (FP). Estos valores corresponden a una precisión del modelo del 98.5 % (1) y una sensibilidad del 97.1 % (2) para este rango de valores.

$$\text{Precisión} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Sensibilidad} = \frac{TP}{TP + FN} \quad (2)$$

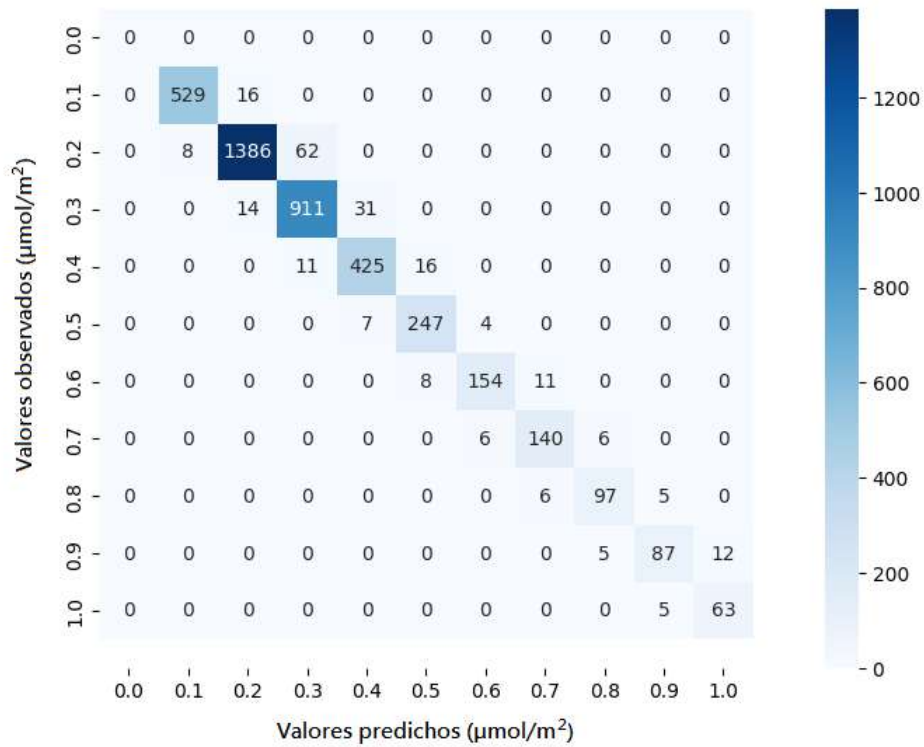


Figura 29. Grupo de muestras con valores de chla-a en el rango de 0 a 1 mol/m^2 discretizados en intervalos de 0.1. Entre más concentrado esté el color en la diagonal, mejor es el rendimiento del modelo.

La Figura 30 presenta el mapa de calor correspondiente al segundo grupo (valores $> 1 \text{ mol/m}^2$), donde se observa una marcada precisión predictiva. Los errores registrados (tanto FP como FN)

se localizaron exclusivamente en celdas adyacentes a la diagonal, lo que indica que las discrepancias no superaron los 0.5 mol/m^2 para ambos grupos.

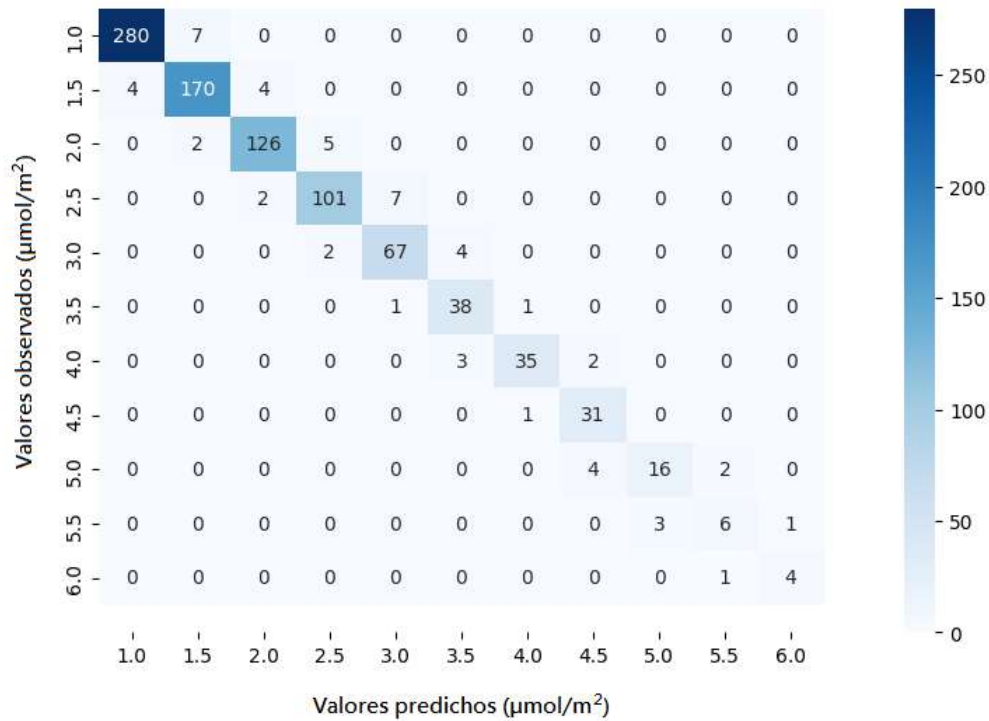


Figura 30. Grupo de muestras con valores de chla-a mayor a 1 mol/m^2 discretizados en intervalos de 0.5. Entre más concentrado esté el color en la diagonal, mejor es el rendimiento del modelo.

La estadística de las pruebas de predicción del modelo y de otros modelos del estado del arte presentados anteriormente se muestran en la Tabla 7. Las métricas reportadas son: Error Absoluto Medio (MAE) (3), Error Cuadrático Medio (MSE) (4), Coeficiente de Determinación (R^2) (5) y el Error Máximo. Como se puede observar, en las cuatro métricas consideradas, el modelo generado en este trabajo logra mejores resultados. En cuanto al error absoluto medio se obtuvo un valor de 0.0100 mol/m^2 superando por más del doble al mejor resultado de los otros métodos que reportan un error de 0.0255 mol/m^2 . Un error cuadrático medio de 0.0005 que mejora por más de seis veces al mejor resultado reportado de 0.0033 . Un coeficiente de determinación de 0.9994 contra 0.9631 del mejor resultado reportado. Finalmente, un error máximo del modelo para la predicción de cualquier muestra de 0.2875 mol/m^2 , que supera al mejor resultado de los otros métodos que reportan un error máximo de 1.0773 mol/m^2 .

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4)$$

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

Tabla 7: Resultados estadísticos de las pruebas de predicción de la concentración de chla-a del modelo propuesto y de los modelos tomados del estado del arte (Adhikary et al., 2024).

Métrica	Modelo Propuesto	Random Forest	Bagging	Extra Trees	HGBR
Error Absoluto Medio (MAE)	0.0100	0.0261	0.0279	0.0255	0.0395
Error Cuadrático Medio (MSE)	0.0005	0.0035	0.0039	0.0033	0.0058
Coefficiente Determinación (R^2)	0.9994	0.9612	0.9569	0.9631	0.9366
Error Máximo	0.2875	1.3965	1.2638	1.0773	1.4917

5. Aplicación de modelo para predecir la biomasa del fitoplancton (Chla-a) derivado del efecto del cambio climático

Con la finalidad de alcanzar el segundo objetivo de la investigación, se simularon condiciones oceanográficas futuras proyectadas bajo distintos escenarios de calentamiento global y se utilizó el modelo para estimar posibles alteraciones en los patrones de distribución y concentración del fitoplancton. Los resultados permiten visualizar tendencias potenciales en la biomasa fitoplanctónica asociadas al incremento de la temperatura del océano.

5.1. Selección de variables para el análisis del efecto del cambio climático sobre el fitoplancton

Como parte del proceso de preparación y depuración de los datos, se ajustó la lista de variables físicas y biogeoquímicas involucradas en el conjunto de datos, principalmente por

5. APLICACIÓN DE MODELO PARA PREDECIR LA BIOMASA DEL FITOPLANCTON (CHLA-A) D.

el hecho de hacer uso de dos variables que representan mediciones directas o derivadas de la biomasa del fitoplancton: El BioCarbon y PhytoCarbon.

El BioCarbon hace referencia al contenido de carbono asociado a la producción primaria neta del fitoplancton, es decir, la cantidad de carbono fijado por las células fitoplanctónicas a través de la fotosíntesis durante el proceso. Por otro lado, el PhytoCarbon corresponde a una estimación de la biomasa de carbono del propio fitoplancton presente en la columna de agua.

Ambas variables proporcionan datos que cuantifican directa o indirectamente la abundancia o productividad del fitoplancton, además, al incluir variables que son prácticamente equivalentes a la salida esperada, el modelo no reflejaría adecuadamente las relaciones entre las condiciones oceanográficas externas (como temperatura, salinidad, nutrientes, etc.) y el fitoplancton. La finalidad del modelo es aprender a estimar la biomasa fitoplanctónica a partir de variables ambientales (físicas y biogeoquímicas) independientes, y no a partir de mediciones que son manifestaciones directas del mismo fenómeno que se pretende modelar.

Por otro lado, y con el objetivo de seleccionar las variables dentro conjunto de datos con mayor relevancia para el modelo, se aplicó un procedimiento de evaluación de importancia de las variables basado en el desempeño predictivo dentro del entrenamiento del modelo.

Se entrenó el modelo con el conjunto completo de variables, excluyendo previamente al BioCarbon y PhytoCarbon, para evaluar el peso o correlación de cada variable con respecto al fitoplancton en términos de chla-a. En cada iteración, se excluyó temporalmente una de las variables del conjunto de entrada y se entrenó el modelo con las variables restantes. Una vez entrenado, se evaluó el desempeño del modelo calculando la Media del Error Cuadrático (MSE) sobre los datos de validación. Este proceso se repitió de forma independiente para cada una de las variables, registrando en cada caso el valor del MSE obtenido.

Se analizaron los resultados con base al principio de considerar únicamente las variables con mayor relevancia para la predicción del fitoplancton, acorde en el aumento significativo de MSE del modelo, dado que el aumento de MSE se relaciona con la importancia de la variable omitida en el ejercicio, y al no considerarla se estaría privando al modelo de información relevante. Se ordenaron las variables en función del MSE obtenido como se observa en la Tabla 8, priorizando aquellas cuya omisión produce el mayor aumento del error.

5. APLICACIÓN DE MODELO PARA PREDECIR LA BIOMASA DEL FITOPLANCTON (CHLA-A) D.

Tabla 8: Resultados del análisis para la selección de variables. Un alto MSE indica mayor dependencia del modelo respecto a la variable excluida.

Número	Variable	MSE
1	Hierro	0.0871
2	Silicato	0.0819
3	Temperatura	0.0743
4	Fosfato	0.0715
5	AlturaSuperficie	0.0715
6	pH	0.0665
7	PresionDióxido	0.0663
8	Nitrato	0.0655
9	Salinidad	0.0639
10	EsteVelocidad	0.0619
11	Oxígeno	0.0612
12	Irradiancia	0.0599
13	CapaOceano	0.0585
14	NorteVelocidad	0.0548

Con base en los resultados obtenidos, se seleccionaron únicamente 11 variables cuya omisión generaron un mayor incremento en el error. Con la intención de mantener la menor cantidad de variables posibles, sin afectar la precisión en los resultados, se descartaron las tres variables con menor impacto aparente: Irradiancia, CapaOceano y NorteVelocidad. Si bien estas variables podrían contener información útil, su bajo aporte en el desempeño del modelo sugiere una menor contribución en comparación con el resto del conjunto.

Tabla 9: Resultado de la selección de variables con mayor aporte en el desempeño del modelo.

Número	Variable	MSE
1	Hierro	0.0871
2	Silicato	0.0819
3	Temperatura	0.0743
4	Fosfato	0.0715
5	AlturaSuperficie	0.0715
6	pH	0.0665
7	PresionDióxido	0.0663
8	Nitrato	0.0655
9	Salinidad	0.0639
10	EsteVelocidad	0.0619
11	Oxígeno	0.0612

5.2. Análisis de la tendencia de la temperatura promedio anual y generación de datos hipotéticos

Con el objetivo de predecir el efecto del cambio climático sobre la concentración del fitoplancton, se analizó la variación de la temperatura del océano con base en los registros de esta variable del portal *Copernicus Marine Service*, del periodo de 1999 al 2019. Además de observarse un incremento sostenido de la temperatura promedio durante este periodo de tiempo, se observó un comportamiento de la temperatura inusual en el año 2015, donde la temperatura promedio se incrementó significativamente respecto a los demás años. Con base en estos datos, se decidió tomar el año 2015 como referencia para analizar diferentes escenarios hipotéticos a futuro (Fig 31).

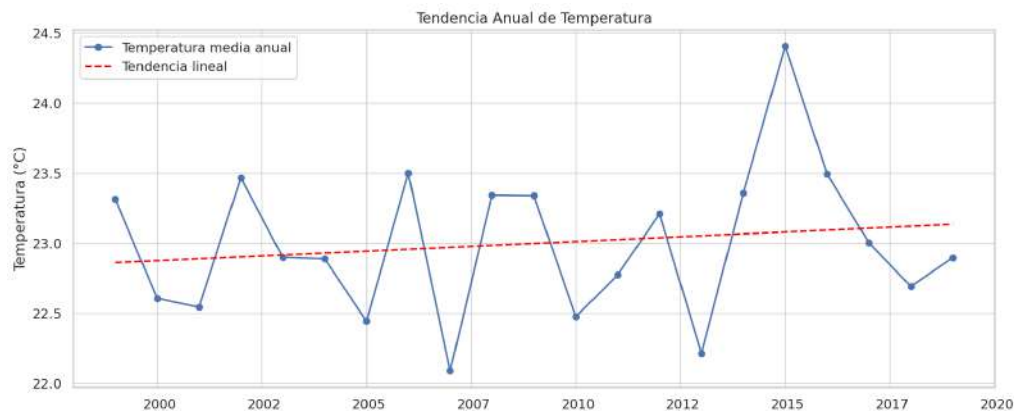


Figura 31. Tendencia anual de la temperatura promedio (1999–2019). La línea azul muestra la serie histórica con variabilidad anual, mientras que la línea roja indica la tendencia lineal.

Para la predicción de la biomasa del fitoplancton (chla-a) en torno a escenarios donde existen cambios constantes de temperatura por diversos factores, se simuló 50 años de datos futuros basados en el año 2015, previamente identificado como un año atípico según la tendencia. Este año 2015 se replicó 50 veces, con un incremento anual sostenido a partir del año 2027 hasta el año 2077.

Para esto, se obtuvo la tendencia de la temperatura a lo largo de 20 años (1999-2019) por medio de un modelo lineal aplicado a los datos históricos de temperatura promedio anual. Los resultados muestran un incremento de temperatura anual de $0.01367\text{ }^{\circ}\text{C}^{\circ}$. Con la intención de generar un aumento de temperatura hipotético dentro del nuevo conjunto de datos de proyección a 50 años futuros, se sumó este incremento anual a la variable temperatura.

5.3. Generación y aplicación del modelo para la predicción bajo diferentes escenarios de cambio climático

Una vez seleccionadas las variables del nuevo conjunto de datos, se entrenó y generó el modelo de predicción. El modelo fue implementado utilizando la biblioteca PyTorch, definiendo una arquitectura compuesta por una capa de entrada con 11 neuronas, dos capas ocultas de 1,000 neuronas, cada una con función de activación ReLU, y una capa de salida con una sola neurona para la predicción de la concentración de chla-a. Se normalizaron las variables mediante la función *StandardScaler* para asegurar que todas las entradas estuvieran en la misma escala. Los parámetros de normalización (la media y desviaciones estándar) fueron almacenados en archivos en formato .NPY independientes (*scaler.mean* y *scaler.scale*) para poder replicar el mismo preprocesamiento en la etapa de validación del modelo.

Una vez obtenido el modelo predictivo, se aplicó para predecir la biomasa de fitoplancton dentro del escenario hipotético generado a 50 años futuros. Para esto, las variables del conjunto de datos fueron organizadas en la misma forma que las utilizadas durante el entrenamiento del modelo, y se aplicó la normalización de los datos según los valores de media y desviación estándar anteriormente almacenados, garantizando así que las escalas de las variables se mantuvieran homogéneas respecto al modelo entrenado.

5.4. Resultados

Para aplicar el modelo predictivo bajo un escenario hipotético de cambio climático, se proyectaron los valores del conjunto de datos para representar distintas décadas futuras considerando un incremento progresivo únicamente en la variable de temperatura, mientras que el resto de las variables oceanográficas se mantuvieron constantes, con el objetivo de observar las posibles variaciones en la biomasa fitoplancton derivadas exclusivamente del aumento térmico. La Tabla 10 presenta un extracto representativo de los resultados obtenidos, donde se observa la variación entre los valores de chla-a reales y predichos.

5. APLICACIÓN DE MODELO PARA PREDECIR LA BIOMASA DEL FITOPLANCTON (CHLA-A) D.

Tabla 10: Representación de los resultados del modelo de los valores reales y predichos de chla-a.

Fecha	Lat	Lon	Temp 2015	Chla-a 2015	Temp hipot	Chla-a pred	Variación
2028-01-31	-12.00	-87.00	25.900	0.231	25.914	0.239	0.008
2028-01-31	-12.00	-86.75	25.919	0.228	25.933	0.235	0.007
2028-01-31	-12.00	-86.50	25.954	0.227	25.968	0.230	0.003
2028-01-31	-12.00	-86.25	25.963	0.226	25.976	0.224	0.002
2028-01-31	-12.00	-86.00	25.921	0.224	25.934	0.221	0.003
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
2048-11-30	-9.75	-80.00	19.823	1.587	20.110	1.522	0.065
2048-11-30	-9.75	-79.75	19.758	1.915	20.045	1.954	0.039
2048-11-30	-9.75	-79.50	19.684	2.322	19.971	2.421	0.099
2048-11-30	-9.75	-79.25	19.628	2.892	19.915	2.868	0.024
2048-11-30	-9.75	-79.00	19.609	3.652	19.896	3.748	0.096
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
2076-12-31	0.75	-81.50	21.473	0.371	22.143	0.345	0.026
2076-12-31	0.75	-81.25	21.510	0.560	22.180	0.434	0.125
2076-12-31	0.75	-81.00	21.575	0.863	22.245	0.809	0.053
2076-12-31	0.75	-80.75	21.635	1.122	22.305	1.090	0.032
2076-12-31	0.75	-80.50	21.673	1.173	22.343	0.767	0.407

La Figura 32 muestra un promedio de concentración de chla-a que oscila ligeramente entre 0.6674 y 0.6726 mol/m^3 a lo largo de las décadas, lo cual sugiere que bajo un escenario en el que sólo la temperatura varía, los cambios esperados en la biomasa fitoplanctónica serían mínimos.

5. APLICACIÓN DE MODELO PARA PREDECIR LA BIOMASA DEL FITOPLANCTON (CHLA-A) D.

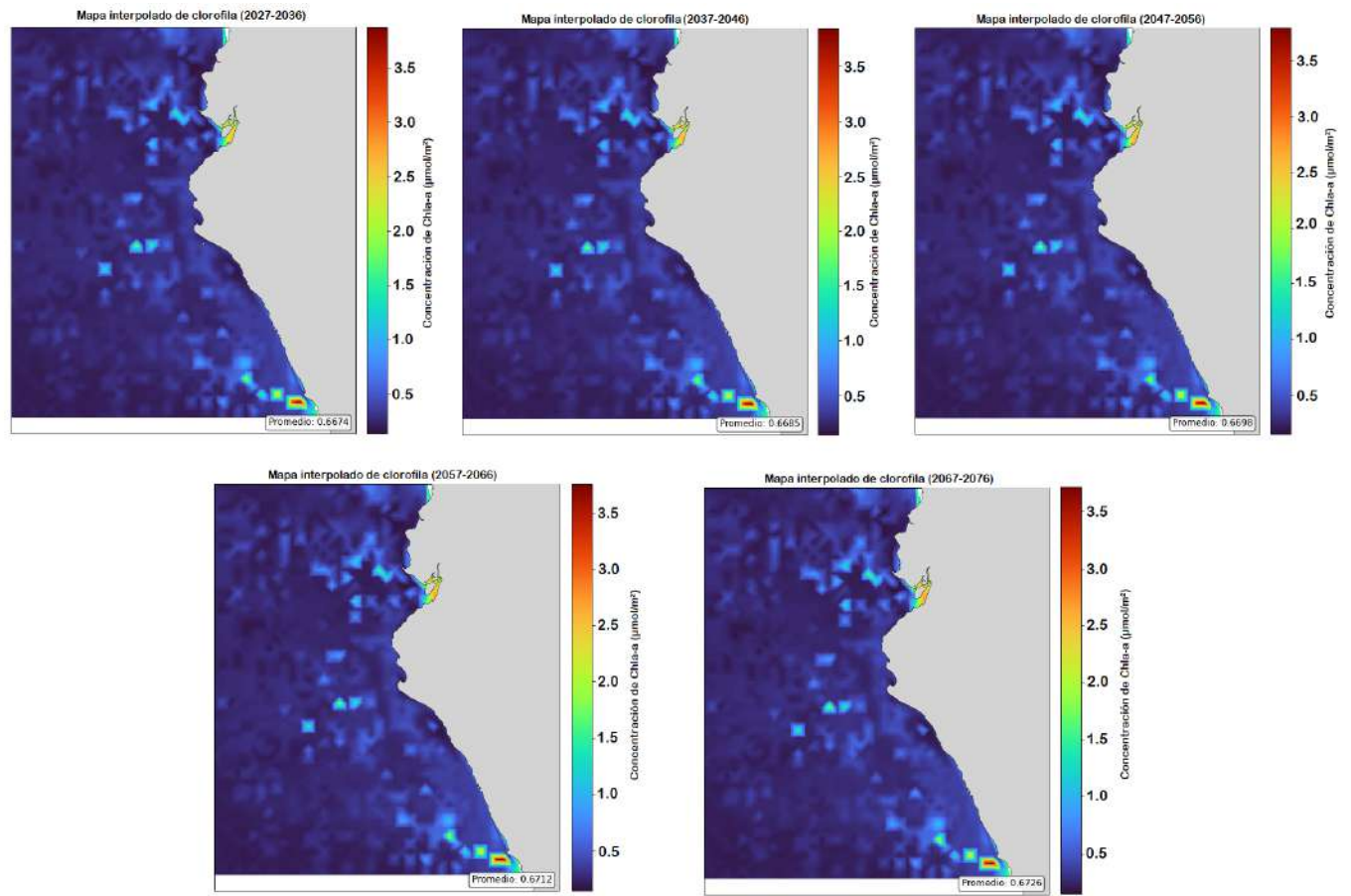


Figura 32. Representación de la distribución de la biomasa del fitoplancton en la zona tropical-subtropical del Pacífico que colinda con Perú.

La Figura 33 presenta la evolución del promedio anual de concentración de chla-a en el periodo comprendido entre los años 2027 y 2076. Este comportamiento sugiere una sensibilidad positiva del modelo ante el aumento gradual de la temperatura.

5. APLICACIÓN DE MODELO PARA PREDECIR LA BIOMASA DEL FITOPLANCTON (CHLA-A) D.

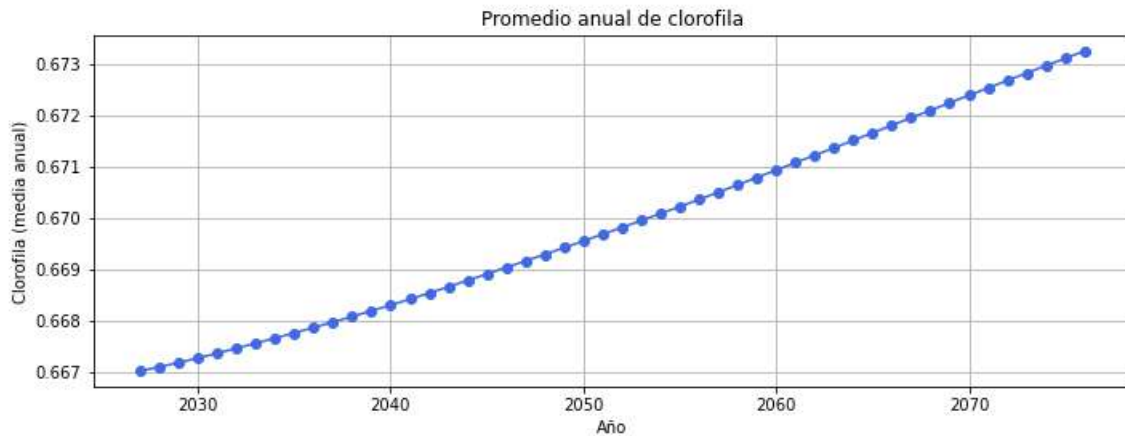


Figura 33. Representación de resultados de la prueba con un aumento de temperatura de 0.01367 C° anual

En los resultados se observa que el aumento de temperatura en 0.01367 C° por año, termina por ser favorable para la producción del fitoplancton.

Para poder determinar si el efecto del aumento de temperatura, asociado al cambio climático en algún punto tendría una repercusión negativa en el desarrollo y producción del fitoplancton, se realizaron tres pruebas en las cuales, dentro del actual conjunto de datos, se aplicó un incremento radical de temperatura. Si bien en el ejercicio anterior el aumento de temperatura fue con base en la tendencia anual del conjunto de datos de 20 años (1999-2019), para estas tres pruebas el aumento de temperatura anual fue de 0.1 C°, 0.2 C°, 0.3 C° respectivamente.

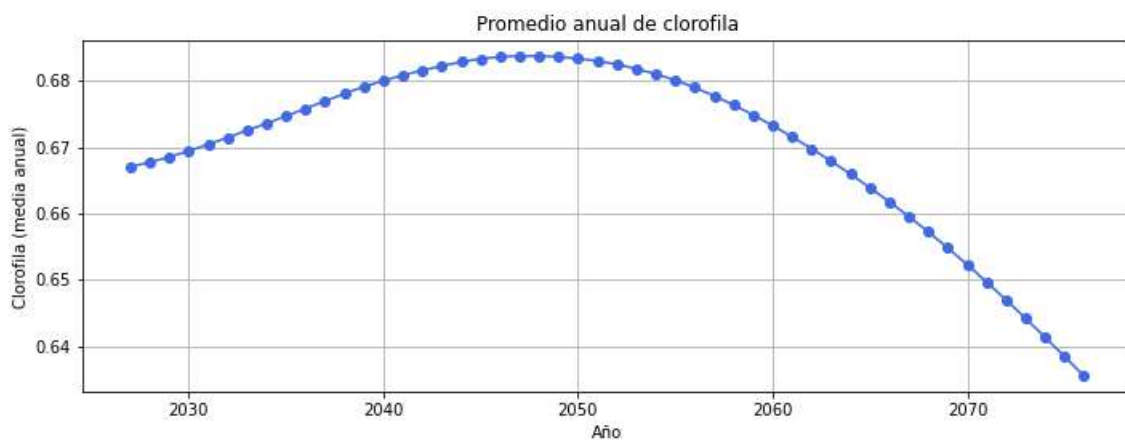


Figura 34. Primera prueba con un aumento de temperatura de 0.1 C° anual

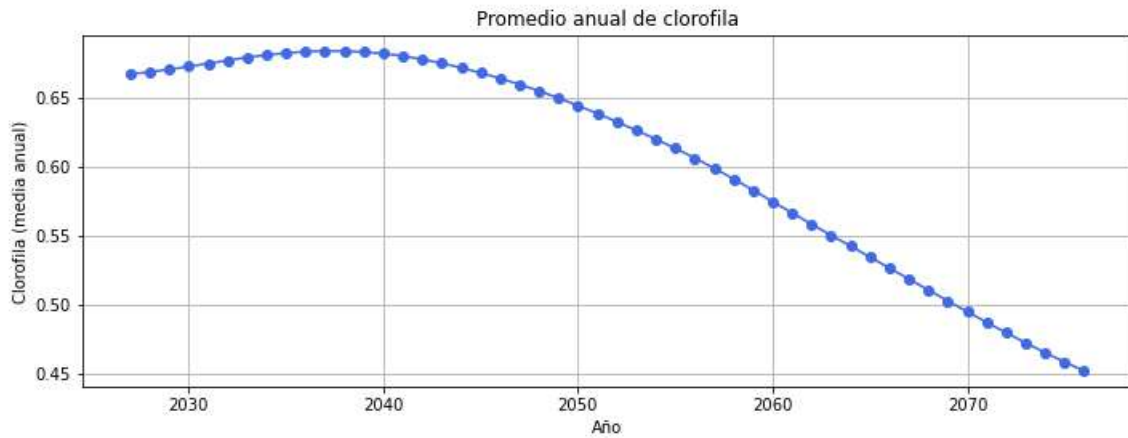


Figura 35. Segunda prueba con un aumento de temperatura de 0.2 C° anual

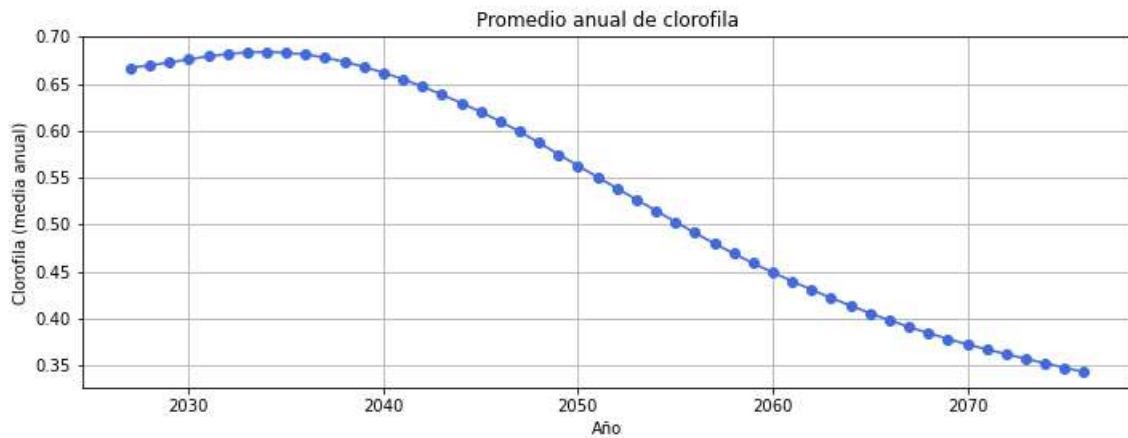


Figura 36. Tercera prueba con un aumento de temperatura de 0.3 C° anual

6. Conclusiones

El modelo propuesto mostró una precisión de 98.5 % con un coeficiente de determinación (R^2) de 0.9994. Confirmando la capacidad de las redes neuronales artificiales para capturar la complejidad no lineal de los procesos biogeoquímicos que determinan la concentración de chl-a-a los ecosistemas marinos.

El modelo permitió la evaluación de escenarios futuros asociados al cambio climático a partir de la simulación de incrementos hipotéticos en la temperatura superficial del mar, se evidenciaron posibles alteraciones en la distribución y concentración del fitoplancton, lo que resalta la vulnerabilidad de estos organismos ante los cambios en las condiciones oceanográficas.

Esta investigación permitió generar modelos predictivos de alta precisión a partir del uso de la inteligencia artificial.

Este estudio sienta las bases para futuras investigaciones enfocadas a la predicción de variables oceanográficas en los sistemas costero-marinos.

Literatura citada

- Adhikary, S., Tiwari, S. P., and Banerjee, S. (2022). Realtime oil spill detection by image processing of synthetic aperture radar data. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–5. IEEE.
- Adhikary, S., Tiwari, S. P., Banerjee, S., Dwivedi, A. D., and Rahman, S. M. (2024). Global marine phytoplankton dynamics analysis with machine learning and reanalyzed remote sensing. *PeerJ*, 12:e17361.
- Allinger (1984). Johnson and Lebel.
- AQUA (2023). MODIS, AQUA.
- Ashlan Cousteau, P. C. (2021). Vertical zones of the ocean’s water column. *Environmental Science*.
- Bates, N. R., Astor, Y. M., Church, M. J., Currie, K., Dore, J. E., González-Dávila, M., Lorenzoni, L., Muller-Karger, F., Olafsson, J., and Santana-Casiano, J. M. (2014). A time-series view of changing surface ocean chemistry due to ocean uptake of anthropogenic CO_2 and ocean acidification. *Oceanography*, 27(1):126–141.
- Behrenfeld, M. J. and Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and oceanography*, 42(1):1–20.
- Bioenciclopedia (2022). *Fotosíntesis: definición y fases*.
- Bracher, A., Vountas, M., Dinter, T., Burrows, J., Röttgers, R., and Peeken, I. (2009). Quantitative observation of cyanobacteria and diatoms from space using phytodoas on sciamachy data. *Biogeosciences*, 6(5):751–764.

- Brewin, R. J., Sathyendranath, S., Müller, D., Brockmann, C., Deschamps, P.-Y., Devred, E., Doerffer, R., Fomferra, N., Franz, B., Grant, M., et al. (2015). The ocean colour climate change initiative: Iii. a round-robin comparison on in-water bio-optical algorithms. *Remote Sensing of Environment*, 162:271–294.
- Carrasco-Ríos, L. (2009). *EFEECTO DE LA RADIACIÓN ULTRAVIOLETA-B EN PLANTAS*.
- Chai, F., Wang, Y., Xing, X., Yan, Y., Xue, H., Wells, M., and Boss, E. (2020). A limited effect of sub-tropical typhoons on phytoplankton dynamics. *Biogeosciences Discussions*, 2020:1–16.
- Chen, Jamet C, Z. Z.-H. Y. M. Z. P. D. W. T.-L. D. Y. D. (2021). Vertical distribution of subsurface phytoplankton layer in South China Sea using airborne lidar.
- Copernicus (2018). Biogeosciences.
- Copernicus, M. S. (2023). Sea water salinity.
- Copernicus-Services (2024). Carbon Storage.
- Cullen, J. (1982). The deep chlorophyll maximum: Comparing vertical profiles of chlorophyll a".
- Diaz, Knowles B, J. C.-L. C. B. K. H. L. N. F.-H. E. K. S. B. M. (2021). Seasonal mixed layer depth shapes phytoplankton physiology, viral production, and accumulation in the north atlantic.
- Doney, S. C., Fabry, V. J., Feely, R. A., and Kleypas, J. A. (2009). Ocean acidification: the other co2 problem. *Annual review of marine science*, 1(1):169–192.
- DYNECO (2011). *Plancton: un universo marino diverso y desconocido*.
- Energía-Atómica, O. I. (2024). *Efectos del dióxido de carbono en los océanos*.
- Fai PBA, Kenko DBN, T. N. M. M. K. K. R. J. (2023). *Use of multivariate analysis to identify phytoplankton bioindicators of stream water quality in the monomodalequatorial agroecological zone of cameroon. Environmental Monitoring and Assessment*.
- García, U. (2019). Introducción a las Redes Neuronales.

- Good, S.; Fiedler, E. M. C. M. M. A. R. R. R.-J. J. S. T. W. J. W. J. W. M. (2020). OSTIA: Global Ocean OSTIA Sea Surface Temperature and Sea Ice Reprocessed.
- Han, C. (2023). Phytoplankton & Ocean Health.
- Hu, C., Lee, Z., and Franz, B. (2012). Chlorophyll algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, 117(C1).
- Jiménez, F. (2019). Nutrientes en el océano.
- Kirk, J. T. (1994). *Light and photosynthesis in aquatic ecosystems*. Cambridge university press.
- Kostadinov, T., Siegel, D., and Maritorena, S. (2010). Global variability of phytoplankton functional types from space: assessment via the particle size distribution. *Biogeosciences*, 7(10):3239–3257.
- Kostadinov, T. S. (2016). Carbon-based phytoplankton size classes retrieved via ocean color estimates of the particle size distribution.
- Kramer, S. J., Siegel, D. A., Maritorena, S., and Catlett, D. (2022). Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales. *Remote Sensing of Environment*, 270:112879.
- los Andes Colombia, U. (2023). *Departamento de geociencia de la Universidad los Andes Colombia*.
- Lueker, T. J., Dickson, A. G., and Keeling, C. D. (2000). Ocean pco₂ calculated from dissolved inorganic carbon, alkalinity, and equations for k₁ and k₂: validation based on laboratory measurements of co₂ in gas and seawater at equilibrium. *Marine chemistry*, 70(1-3):105–119.
- McCarthy, J. (1996). *Defending IA research: A collection of essays and reviews*.
- Minsky, M. (1969.). *Perceptrons*.
- Moore TS, Sathyendranath, S., Brewin, R. J., Brockmann, C., Brotas, V., Calton, B., Chuprin, A., Cipollini, P., Couto, A. B., Dingle, J., Doerffer, R., et al. (2019). An ocean-colour time

- series for use in climate studies: the experience of the ocean-colour climate change initiative (oc-cci). *Sensors*, 19(19):4285.
- NASA's-Earth-Observatory (2017). Temperatura superficial del mar.
- NOAA-NESDIS-National (2007). WORLD OCEAN ATLAS 2001.
- of Calgary, U. (2015). Ocean acidification.
- Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., Alikas, K., Kangro, K., Gurlin, D., Hà, N., et al. (2020). Seamless retrievals of chlorophyll-a from sentinel-2 (msi) and sentinel-3 (olci) in inland and coastal waters: A machine-learning approach. *Remote Sensing of Environment*, 240:111604.
- Peters, K. (2017). La historia de la clorofila y los cloroplastos.
- Philip, Svein Sundby, H.-O. P. (2014). Net Primary Production in the Ocean.
- Philip W. Boyd, Svein Sundby, H.-O. P. (2014). Net Primary Production in the Ocean.
- Polovina, J. J., Dunne, J. P., Woodworth, P. A., and Howell, E. A. (2011). Projected expansion of the subtropical biome and contraction of the temperate and equatorial upwelling biomes in the north pacific under global warming. *ICES Journal of Marine Science*, 68(6):986–995.
- Rizzuto, S., Thrane, J.-E., Baho, D. L., Jones, K. C., Zhang, H., Hessen, D. O., Nizzetto, L., and Leu, E. (2020). Water browning controls adaptation and associated trade-offs in phytoplankton stressed by chemical pollution. *Environmental Science & Technology*, 54(9):5569–5579.
- Rouhiainen, L. (2018). Inteligencia artificial. *Madrid: Alienta Editorial*, pages 20–21.
- Siegel, D., Buesseler, K., Doney, S. C., Sailley, S., Behrenfeld, M. J., and Boyd, P. (2014). Global assessment of ocean carbon export by combining satellite observations and food-web models. *Global Biogeochemical Cycles*, 28(3):181–196.
- Sonic, B. L. (2024). *El Oxígeno disuelto y la calidad del agua*.
- Soranno, P. A., Bacon, L. C., Beauchene, M., Bednar, K. E., Bissell, E. G., Boudreau, C. K., Boyer, M. G., Bremigan, M. T., Carpenter, S. R., Carr, J. W., et al. (2017). Lagos-ne: a

- multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of us lakes. *GigaScience*, 6(12):gix101.
- State-Water, R. C. B. T. (2019). The pH of water.
- Sun, X., Zhang, Y., Shi, K., Zhang, Y., Li, N., Wang, W., Huang, X., and Qin, B. (2022). Monitoring water quality using proximal remote sensing technology. *Science of the Total Environment*, 803:149805.
- Tiwari, S. P., Adhikary, S., and Banerjee, S. (2022). Estimation of chlorophyll-a from oceanographic properties-an indirect approach. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 6872–6875. IEEE.
- University-Calgary (2015). Ocean acidification.
- Van Heukelem, L. and Thomas, C. S. (2001). Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *Journal of Chromatography A*, 910(1):31–49.
- Vandermeulen, R. A., Mannino, A., Craig, S. E., and Werdell, P. J. (2020). 150 shades of green: Using the full spectrum of remote sensing reflectance to elucidate color shifts in the ocean. *Remote Sensing of Environment*, 247:111900.
- Wang, C. and Weisberg, R. H. (2000). The 1997–98 el niño evolution relative to previous el niño events. *Journal of Climate*, 13(2):488–501.
- Webb, P. (2023). Introduction to Oceanography.
- Álvarez, D. O. (2024). Ciclos biogeoquímicos.