DESDE 2013
https://repository.uaeh.edu.mx/revistas/index.php/icbi/issue/archive
Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI
ISSN: 2007-6363

Publicación Semestral Pädi Vol. 13 No. 26 (2026) 111–118

Modelo de regresión basado en redes neuronales artificiales para la estimación de la concentración de clorofila-a en el Pacífico Tropical Oriental frente a Perú. Artificial Neural Network-Based Regression Model for Estimating Chlorophyll-a Concentration in the Eastern Tropical Pacific off the Coast of Peru.

J. M. Talamantes-Murillo [a], J. E. Luna-Taylor [a], E. D Sanchéz-Pérez [b,*], M. A. Castro-Liera [a], I. M Santillán-Méndez [a]

[a] *División de Estudios de Posgrado e Investigación, Tecnológico Nacional de México/I. T. de La Paz, 23080, La Paz, B. C. S., México.*
[b] *Secihti-Instituto Politécnico Nacional-Centro Interdisciplinario de Ciencias Marinas (IPN-CICIMAR). Avenida IPN s/n, Col. Playa Palo de Santa Rita, La Paz, Baja California Sur, 23096, México.*

Resumen

El fitoplancton es un microorganismo marino que aporta más del 50 % del oxígeno del planeta y constituye el eslabón primario en las cadenas tróficas marinas. Sin embargo, la actividad antropogénica ha generado el vertido de grandes volúmenes de contaminantes en los océanos, alterando la abundancia y distribución de estos organismos, en un contexto donde aún se desconoce la magnitud de su relación con el cambio climático. Con el objetivo de aportar herramientas para abordar esta problemática, en este trabajo se desarrolló un modelo computacional basado en redes neuronales artificiales para predecir la concentración de biomasa fitoplanctónica (clorofila-a) a partir de variables fisicoquímicas y físicas. El modelo fue entrenado con datos de la plataforma *Copernicus Marine Service*, correspondientes al Pacífico Tropical Oriental frente a Perú. Los resultados muestran un excelente desempeño, con una precisión del 98.5 %, un coeficiente de determinación (R^2) de 0.9994, un error cuadrático medio de 0.0005 mol/m^2 y un error absoluto medio de 0.0100 mol/m^2. Estos resultados confirman la efectividad del modelo propuesto, cumpliendo con el objetivo planteado y ofreciendo una herramienta útil para el monitoreo ambiental marino.

Palabras Clave: clorofila, fitoplancton, modelos predictivos, redes neuronales, inteligencia artificial.

Abstract

Phytoplankton is a marine microorganism that contributes more than 50 % of the planet?s oxygen and represents the primary link in marine trophic chains. However, anthropogenic activity has led to the discharge of large volumes of pollutants into the oceans, altering the abundance and distribution of these organisms, within a context where the extent of their relationship with climate change remains uncertain. To provide tools to address this issue, this study developed a computational model based on artificial neural networks to predict phytoplankton biomass concentration (chlorophyll-a) from physicochemical and physical variables. The model was trained with data from the *Copernicus Marine Service*, corresponding to the Eastern Tropical Pacific off the coast of Peru. The results show excellent performance, with an accuracy of 98.5 %, a coefficient of determination (R^2) of 0.9994, a mean squared error of 0.0005 mol/m^2, and a mean absolute error of 0.0100 mol/m^2. These findings confirm the effectiveness of the proposed model, fulfilling the objective of the study and providing a useful tool for marine environmental monitoring.

Keywords: chlorophyll, phytoplankton, predictive models, neural networks, artificial intelligence.

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-a, la cual actúa como catalizador en la conversión de energía solar en energía química (Behrenfeld y Falkowski, 1997). La clorofila-a ha sido históricamente usada co-

*Autor para correspondencia: esanchezp@secihti.mx
Correo electrónico:esanchezp@secihti.mx (Elvia Denisse Sanchéz Pérez), jorge.lt@lapaz.tecnm.mx (Jorge Enrique Luna Taylor), m23310001@lapaz.tecnm.mx (Juan Manuel Talamantes Murillo)

Historial del manuscrito: recibido el 13/05/2025, última versión-revisada recibida el 19/09/2025, aceptado el 24/09/2025, en línea (postprint) desde el 26/09/2025, publicado el 05/01/2026. **DOI:** https://doi.org/10.29057/icbi.v13i26.15172

mo 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 (Rrs (λ)) en múltiples longitudes de onda. Esta capacidad permite derivar no solo la concentración de clorofila-a, sino también indicadores de materia orgánica disuelta cromófora (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 Rrs(λ), 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 hiperespectrales de reflectancia remota (Rrs(λ)) 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.*, 2024a).

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.

2. Materiales y Métodos

2.1. Colección de datos

El área de estudio abarca la región del Pacífico Tropical Oriental frente a la costa peruana (0.8-12 °S; 76°-87 °O; Figura 1). El análisis comprende el período 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 y Weisberg, 2000).

Para el desarrollo del modelo (entrenamiento y validación), se emplearon dos conjuntos de datos de acceso abierto en formato NetCDF, obtenidos de la plataforma *Copernicus Marine Service*. Variables biogeoquímicas provenientes del producto *Global Ocean Biogeochemistry Hindcast*: (`https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_BGC_001_029/description`, Tabla 1). Variables físicas obtenidas del *Global Ocean Ensemble Physics Reanalysis* (`https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_ENS_001_031/description`, Tabla 2). Ambos conjuntos presentaron una resolución espacial de 1/4° (4 km/píxel). Adicionalmente, se incorporaron datos de irradiancia superficial del proyecto *NASA Prediction of Worldwide Energy Resources* (`https://power.larc.nasa.gov/`), esenciales para el análisis de productividad primaria. La variable objetivo seleccionada para el entrenamiento del modelo fue la concentración de clorofila-a (g/L), un proxy clave de la biomasa fitoplanctónica en estudios biogeoquímicos marinos.

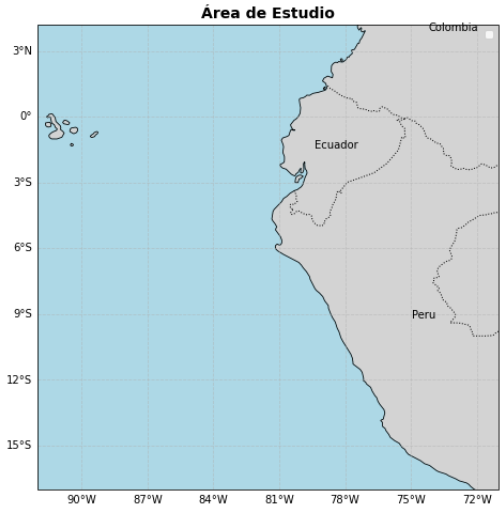

Figura 1: Pacífico Tropical-Subtropical frente a Perú

Tabla 1: Variables biogeoquímicas obtenidas desde la plataforma de Copernicus Marine Service

Parámetros	Descripción	Unidades
Fe	Hierro disuelto	$mmol\ m^{-3}$
NO_3	Nitratos	$mmol\ m^{-3}$
$nppv$	Producción neta de biomasa expresada como carbono	$mg\ m^{-3}$
O_2	Oxigeno disuelto	$mmol\ m^{-3}$
pH	Potencial de hidrogeno	–
$phyc$	Concentración de fitoplancton expresada como carbono	$mmol\ m^{-3}$
PO_4	Fosfatos	$mmol\ m^{-3}$
Si	Silicatos	$mmol\ m^{-3}$
CO_2	Presión parcial de dioxido de carbono	Pa

Tabla 2: Variables físicas obtenidas desde la plataforma de Copernicus Marine Service

Parámetros	Descripción	Unidades
$mlotst$	Espesor de la capa mixta oceánica definido por sigma theta	m
so	Salinidad del agua de mar	psu
$thetao$	Temperatura potencial del agua de mar	C
uo	Velocidad del agua del mar hacia el este	m/s
vo	Velocidad del agua del mar hacia el norte	m/s
zos	altura de la superficie del mar sobre el geoide	m

2.2. 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 proxis 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 (IPCC, 2021). Por ejemplo, nutrientes como el nitrato (NO_3), el hierro disuelto (Fe), el fosfato (PO_4) y el silicato (Si) son macronutrientes esenciales para la síntesis de biomoléculas en el fitoplancton (Sathyendranath S. *et al.*, 2019), la biomasa de carbono (nppv) y carbono fitoplanctónico (phyc) como indicadores clave de la fijación de CO_2 durante la fotosíntesis (Behrenfeld y Falkowski, 1997), el oxígeno disuelto (O_2) 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 (pCO_2) 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 (thetao) regulan la tasa de crecimiento del fitoplancton (Van Heukelem y Thomas, 2001), la salinidad 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 (uo, vo) 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 (mlotst) determina la disponibilidad de nutrientes en la zona eufótica (Soranno P.A. *et al.*, 2017). La altura superficial del mar (zos) 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).

2.3. 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 clorofila-a como variable de salida. A continuación, se muestran los algoritmos utilizados:

2.3.1. Algoritmo 1: Conversión de datos NetCDF a CSV

Se desarrolló el Algoritmo 1 para extraer, validar y transformar los datos almacenados en formato NetCDF a un archivo estructurado en formato CSV.

Algoritmo 1: Conversión de documentos en formato .nc a formato .csv

Entrada: Archivo en formato NetCDF con datos de las variables
Rango 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

data_in ← apertura del archivo NetCDF a través de la librearía H5PY
data_out ←creación y apertura del archivo de salida en formato .CSV

```
for dia in range(DIAS) do
    for lat in range(LATS) do
        for lon in range(LONS) do
            // se extraen datos del archivo NetCDF:
            «variables»← data_in.get("«variable»")[dia, lat, lon]
            // se validan datos:
            if all([es_valido(«variables»)]) then
                // se almacena un registro:
                data_out.write(date, latitude, longitude,
                  «variables»)
        end
    end
end
return data_out
```

2.3.2. Algoritmo 2: Combinación de variables biogeoquímicas y físicas

A partir de los archivos independientes en formato CSV generados previamente, se implementó el Algoritmo 2 para realizar la fusión espacio-temporal de registros mediante una operación 'innerjoin' basada en claves primarias compuestas (fecha, latitud, longitud).

Algoritmo 2: 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

data_out ←creación y apertura del archivo de salida
data_out.write("Fecha, Latitud, Longitud, «Variables mezcladas»")
for *chunk_bio* **in** *pd.read_csv(bio_file, chunksize=chunk_size, usecols=['Fecha', 'Latitud', 'Longitud', '«Variables biogeoquímicas»'])* **do**
  chunk_bio['*Fecha*']←convierte a formato yyyy-mm-dd
  for *chunk_phy* **in** *pd.read_csv(phy_file, chunksize=chunk_size, usecols=['Fecha', 'Latitud', 'Longitud', '«Variables físicas»'])* **do**
    chunk_phy['*Fecha*']←convierte a formato yyyy-mm-dd
    merged_chunk ← *pd.merge*(*chunk_bio*, *chunk_phy*, *on*=['*Fecha*', '*Latitud*', '*Longitud*'], *how*='*inner*')
    merged_chunk ← *merged_chunk*[['*Fecha*', '*Latitud*', '*Longitud*', '«*Variables reordenadas*»']]
    merged_chunk.to_csv(*data_out*, *header* = *False*, *index* = *False*)
    delete merged_chunk
    gc.collect()
  end
end
return *data_out*

Algoritmo 3: Incorporación de la variable de irradiancia al conjunto de datos de variables físicas y biogeoquímicas

Input : Archivo de variables biogeoquímicas y físicas (*input_file*)
Lista de archivos con datos de irradiancia (*irra_file_N*)
Cantidad de archivos de irradiancia (*contIrraFiles*)
Output: Archivo con variables biogeoquímicas, físicas e irradiancia

output_file ←creación y apertura del archivo de salida
output_file.write('Fecha', 'Latitud', 'Longitud', '«Variables con irradiancia»')
data_frame ← *pd.read_csv*(*input_file*, usecols=['Fecha', 'Latitud', 'Longitud', '«Variables»'])
for *registro* **in** *data_frame* **do**
  numIrraFiles = 1
  contIrradiance = 0
  sumIrradiance = 0
  while *numIrraFiles* <= *contIrraFiles* **do**
    for *reg_irra* **in** *data_irra* **do**
      lat_irra ← *reg_irra*['*latitude*']
      lon_irra ← *reg_irra*['*longitude*']
      fecha ← *reg_fecha*['*fecha*']
      irradiance ← *reg_irra*['*irradiance*']
      coordRange = 0,25
    end
    numIrraFiles += 1
  end
  if *contIrradiance* > *0* **then**
    output_file.write(«*variables*»+*str*(*averageIrradiance*))
end
return *output_file*

2.3.3. Algoritmo 3: Incorporación de la variable irradiancia

Una vez integradas las variables biogeoquímicas y físicas, se implementó el Algoritmo 3 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. 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 arbitrario de ± 0.25° en ambas direcciones, permitiendo asignar un valor de irradiancia a cada registro mediante interpolación espacial.

Adicionalmente, la plataforma *NASA POWER* genera un archivo independiente por cada día de lecturas, lo que resultó en una extensa colección de archivos para cubrir el periodo de estudio (1999-2005). Para abordar este desafío, 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 unificada. Este enfoque garantiza la integración precisa de la irradiancia como variable adicional, respetando las limitaciones impuestas por la heterogeneidad en la resolución espacial de las fuentes de datos.

3. Desarrollo 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, reconocida por su flexibilidad y eficiencia en el entrenamiento de redes neuronales. La arquitectura propuesta para la red neuronal se detalla en la Tabla 3. 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 optimiza mediante experimentación iterativa, buscando un equilibrio entre precisión predictiva y eficiencia computacional.

Tabla 3: Arquitectura de la red neuronal utilizada

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

3.1. Entrenamiento del modelo

El proceso de entrenamiento se implementó en el Algoritmo 4, 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) procesado 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 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.

Algoritmo 4: Entrenamiento del modelo

```
Input  : Conjunto de datos de entrenamiento
Output: Mejor modelo generado

learning_rate ← 0,001
loss_function ← MSE_Loss()
optimizer ← Adam(lr = learning_rate)
num_epochs ← 1000
best_test_loss ← float('inf')
patience ← 0
for epoch in range(1, num_epochs) do
    model.train()
    for (data_train,chla_train) in train_database do
        chla_pred ← model(data_train)
        loss_train ← loss_function(chla_pred, chla_train)
        loss_train.backward()
        optimizer.step()
    end
    model.eval()
    total_loss_test ← 0
    for (data_test,chla_test) in test_database do
        chal_pred ← model(data_test)
        loss_test ← loss_function(chla_pred, chla_test)
        total_loss_test ← total_loss_test + loss_test
    end
    current_test_loss ← total_loss_test/len(test_database)
    if current_test_loss < best_test_loss then
        patience ← 0
        best_test_loss ← current_test_loss
        save(model,"model.pth")
    else
        patience ← patience + 1
        if patience > 100 then
            break
        end
    end
end
```

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 4.

Tabla 4: Parámetros del entrenamiento del modelo

Parámetro	Valor
Número de muestras	12,000
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

4. Resultados

4.1. *Evaluación del modelo*

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

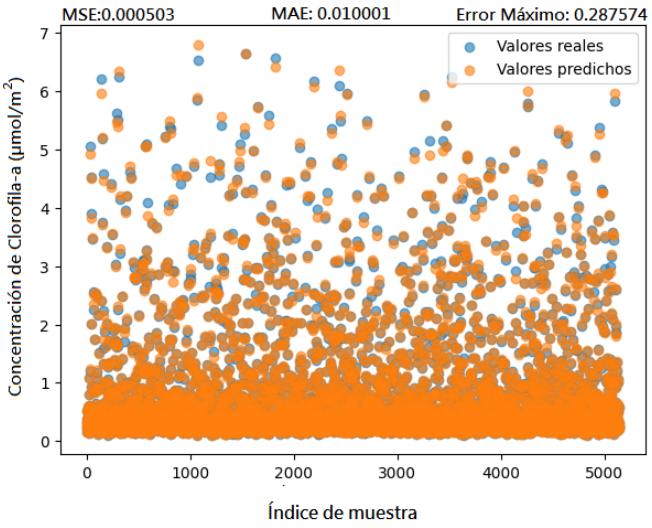

Figura 2: Concentración de clorofila-a (mol/m^2). Valores observados (puntos azules) y valores predichos (puntos naranja).

La Figura 3 muestra la gráfica de correlación entre los valores reales de clorofila-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 clorofila-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. En la Figura se muestra también 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 índica que un modelo no tiene capacidad predictiva, mientras que un valor cercano a uno índica un alto nivel predictivo del mismo.

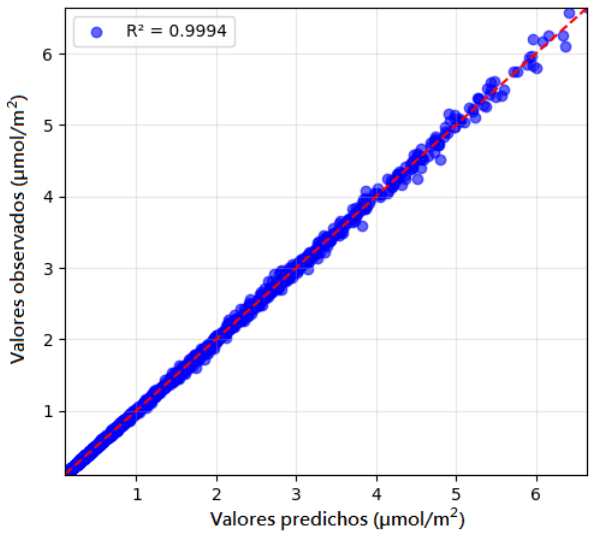

Figura 3: Correlación de los valores observados vs predichos de la concentración de clorofila-a (mol/m^2). Coeficiente de correlación (R2 = 0.9994).

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 clorofila-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 clorofila-a de 0 a 1 mol/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 mol/m^2. En contraste, el segundo grupo abarca muestras con valores > $1mol/m^2$, correspondiente al 17 % restante del conjunto de prueba, y se agrupan en intervalos de 0.5 mol/m^2.

El mapa de calor revela un desempeño detallado del modelo al comparar los valores predichos con los observados. En la Figura 4 se representa el primer grupo, el eje X corresponde a los valores discretizados, usando intervalos de 0.1 mol/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 mol/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.

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

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

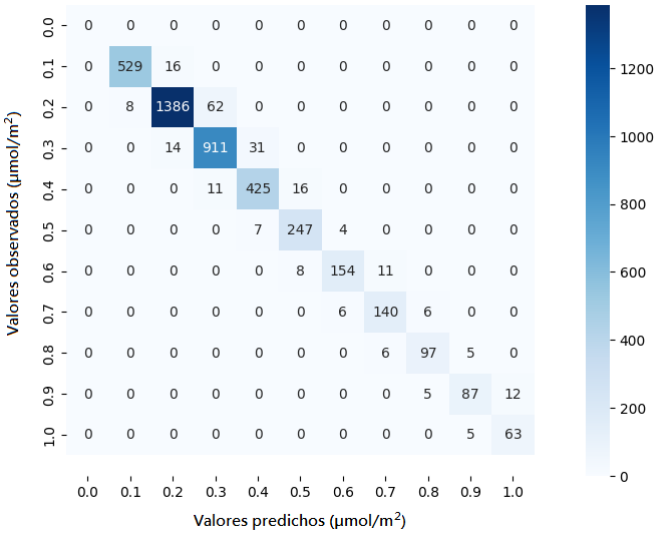

Figura 4: Grupo de muestras con valores de clorofila-a en el rango de 0 a 1 mol/m^2 discretizados en intervalos de 0.1. Entre más concentrado esté el color en la diagonal, mejor es el rendimiento del modelo.

La Figura 5 presenta el mapa de calor correspondiente al segundo grupo (valores > $1mol/m^2$), 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.

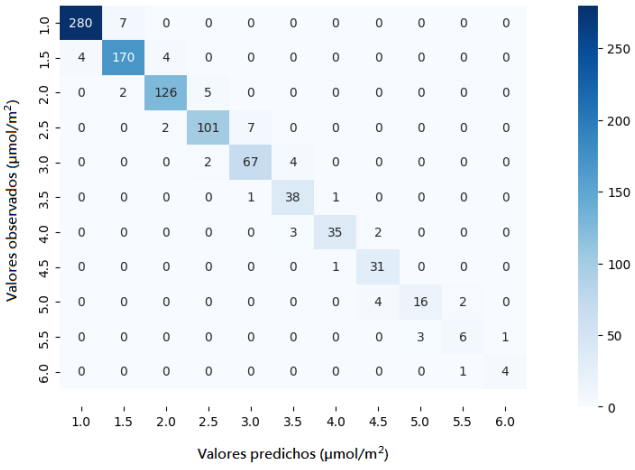

Figura 5: Grupo de muestras con valores de clorofila-a mayor a 1 mol/m^2 discretizados en intervalos de 0.5. Entre más concentrado esté el color en la diagonal, mejor es el rendimiento del modelo.

En la Tabla 5 se muestra la estadística de las pruebas de predicción del modelo y de otros modelos del estado del arte

presentados en (Adhikary *et al.*, 2024b). 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.

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

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

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

Tabla 5: Resultados estadísticos de las pruebas de predicción de la concentración de clorofila-a del modelo propuesto y de otros modelos del estado del arte

Métrica	Modelo Propuesto	Random Forest	Bagging	Extra Trees	HGBR
Error Absoluto Medio (MAE)	0.010	0.0261	0.0279	0.0255	0.0395
Error Cuadrático Medio (MSE)	0.0005	0.0035	0.0039	0.0033	0.0058
Coeficiente 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. Conclusión

El modelo obtuvo un desempeño relevante con un coeficiente de determinación de 0.9994, una media del error cuadrado de 0.0005 y un porcentaje de precisión predictivo del 98.5 %.

Estos resultados sugieren que el modelo es factible de ser empleado como una herramienta de monitoreo de la productividad biológica en ecosistemas marinos particularmente en escenarios de cambio climático. No obstante, su aplicación práctica requiere considerar la disponibilidad y calidad de los datos de entrada. Bajo estas condiciones, el modelo puede contribuir a un monitoreo eficiente y de bajo costo de la biomasa fitoplanctónica y servir como apoyo en la toma de decisiones para la gestión y conservación de los recursos marinos.

Futuras líneas de trabajo contemplan la evaluación del modelo en distintas regiones y periodos, y el análisis con conjuntos de datos más heterogéneos. Estas acciones fortalecerán su capacidad de generalización y consolidarán su utilidad como herramienta de monitoreo ambiental marino.

Agradecimientos

Agradecemos completamente al Consejo Nacional de Ciencia y Tecnología (CONACYT), al Tecnológico Nacional de México (TecNM) y al Centro Interdisciplinario de Ciencias Marinas (CICIMAR), por todo su apoyo en el desarrollo de este proyecto de investigación.

Referencias

Adhikary, S., Tiwari, S. P., y Banerjee, S. (2022). Realtime oil spill detection by image processing of synthetic aperture radar data. En *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–5. IEEE.

Adhikary, S., Tiwari, S. P., Banerjee, S., Dwivedi, A. D., y Rahman, S. M. (2024a). Global marine phytoplankton dynamics analysis with machine learning and reanalyzed remote sensing. *PeerJ*, 12:e17361.

Adhikary, S., Tiwari, S. P., Banerjee, S., Dwivedi, A. D., y Rahman, S. M. (2024b). Global marine phytoplankton dynamics analysis with machine learning and reanalyzed remote sensing. *PeerJ*, 12:e17361.

Bates, N. R., Astor, Y. M., Church, M. J., Currie, K., Dore, J. E., González-Dávila, M., Lorenzoni, L., Muller-Karger, F., Olafsson, J., y Santana-Casiano, J. M. (2014). A time-series view of changing surface ocean chemistry due to ocean uptake of anthropogenic co? and ocean acidification. *Oceanography*, 27(1):126–141.

Behrenfeld, M. J. y Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and oceanography*, 42(1):1–20.

Bracher, A., Vountas, M., Dinter, T., Burrows, J., Röttgers, R., y Peeken, I. (2009). Quantitative observation of cyanobacteria and diatoms from space using phytodoas on sciamachy data. *Biogeosciences*, 6(5):751–764.

Chai, F., Wang, Y., Xing, X., Yan, Y., Xue, H., Wells, M., y Boss, E. (2020). A limited effect of sub-tropical typhoons on phytoplankton dynamics. *Biogeosciences Discussions*, 2020:1–16.

Doney, S. C., Fabry, V. J., Feely, R. A., y Kleypas, J. A. (2009). Ocean acidification: the other co2 problem. *Annual review of marine science*, 1(1):169–192.

Hu, C., Lee, Z., y Franz, B. (2012). Chlorophyll aalgorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, 117(C1).

Kirk, J. T. (1994). *Light and photosynthesis in aquatic ecosystems*. Cambridge university press.

Kostadinov, T., Siegel, D., y Maritorena, S. (2010). Global variability of phytoplankton functional types from space: assessment via the particle size distribution. *Biogeosciences*, 7(10):3239–3257.

Kramer, S. J., Siegel, D. A., Maritorena, S., y Catlett, D. (2022). Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales. *Remote Sensing of Environment*, 270:112879.

Lueker, T. J., Dickson, A. G., y 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.

Pahlevan, Smith B., Schalles J., BindingC., Cao Z., Ma R., Alikas K., Kangro K., Gurlin, Daniela and Ha, Nguyen and others (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.

Rizzuto, S., Thrane, J.-E., Baho, D. L., Jones, K. C., Zhang, H., Hessen, D. O., Nizzetto, L., y Leu, E. (2020). Water browning controls adaptation and associated trade-offs in phytoplankton stressed by chemical pollution. *Environmental Science & Technology*, 54(9):5569–5579.

Sathyendranath S., Brewin J.W., Brockmann C., Brotas V., Calton B., Chuprin A., Cipollini P., Couto A.B., Dingle, J. and Doerffer, R. and others (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.

Siegel, D., Buesseler, K., Doney, S. C., Sailley, S., Behrenfeld, M. J., y Boyd, P. (2014). Global assessment of ocean carbon export by combining satellite observations and food-web models. *Global Biogeochemical Cycles*, 28(3):181–196.

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. and Carr, J.W. and others (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.

Sun, X., Zhang, Y., Shi, K., Zhang, Y., Li, N., Wang, W., Huang, X., y Qin, B. (2022). Monitoring water quality using proximal remote sensing technology. *Science of the Total Environment*, 803:149805.

Tiwari, S. P., Adhikary, S., y Banerjee, S. (2022). Estimation of chlorophyll-a from oceanographic properties-an indirect approach. En *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6872–6875. IEEE.

Van Heukelem, L. y 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., y 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. y Weisberg, R. H. (2000). The 1997–98 el niño evolution relative to previous el niño events. *Journal of Climate*, 13(2):488–501.