INFORME FINAL

SECRETARÍA EJECUTIVA DE LA COMISIÓN COLOMBIANA DEL OCÉANO

PLAN NACIONAL DE EXPEDICIONES CIENTÍFICAS MARINAS

1 Datos generales del proyecto

Título del proyecto	Presencia de Vibrio spp en la subregión Sanquianga-Gorgona y su relación con las condiciones hidrográficas
Expedición Científica	ECP2021-I Bocas de Sanquianga
Investigadores / Filiación	Jenny Parada, Christian Bermúdez-Rivas, Yadi Moreno, Fredy Castrillón, S2 Brainer Ángel Centro de Investigaciones Oceanográficas e Hidrográficas del Pacífico
Institución responsable	Centro de Investigaciones Oceanográficas e hidrográficas del Pacífico
Instituciones aliadas	
Correos electrónicos	cbermudezr@dimar.mil.co
Fecha de entrega	21/04/2023

1.1 Resumen

Vibrio spp. es un género de bacterias gramnegativas, curvas o rectas, móviles por flagelos, que se encuentran comúnmente en ambientes acuáticos. Algunas especies de *Vibrio* pueden causar infecciones en humanos y animales, incluyendo la enfermedad del cólera, gastroenteritis y sepsis. La identificación precisa de estas bacterias es importante para la prevención y el control de enfermedades relacionadas, así como para la evaluación de la calidad del agua y los alimentos de origen acuático. Sin embargo, otro factor importante para estos tipos de agentes patógenos es la posibilidad de identificar cuáles son las variables que condicionan su presencia y proliferación. Muchos estudios se han enfocado en tratar de relacionar variables ambientales con la incidencia (presencia-ausencia) de esta bacteria, encontrando que muchas relaciones no son del todo claras. Estas variables que pueden ser potencialmente determinantes pueden ser variables fisicoquímicas o biológicas como la temperatura, salinidad, pH, presencia de zooplancton y productividad primaria.

En este trabajo se relacionaron variables hidrográficas y biológicas con la incidencia del género *Vibrio* spp. en la subregión Sanquianga-Gorgona. Para evaluar estas relaciones y desarrollar modelos de predicción espacial, se utilizaron cuatro algoritmos de aprendizaje automatizado supervisado (*Machine Learning*), los cuales se basan en la detección automática de patrones en los datos y predecir su comportamiento en diferentes escenarios. Los algoritmos de regresión logística y *Random forest* presentaron los mejores desempeños para detectar la presencia y la ausencia de *Vibrio* spp. basados en las variables predictivas de temperatura y salinidad. En los escenarios de predicción de los períodos mareales, ambos algoritmos respondieron de una

manera similar donde en marea alta, toda el área tiene presencia de *Vibrio* spp. y en marea baja solo hay ausencia en una de las bocanas del río Sanquianga.

2 Sinopsis técnica (Máximo 1500 palabras)

Las especies del género Vibrio son bacilos gramnegativos que tienen una amplia distribución en la naturaleza. En ambientes marinos y estuarinos, los *Vibrios* se aíslan comúnmente del sedimento, la columna de agua, el plancton y los mariscos. Los mariscos que a menudo albergan especies de Vibrio incluyen mariscos bivalvos (ostras, almejas y mejillones), cangrejos, camarones y langostinos. Los *Vibrios* son bacilos aerobios curvos y móviles que poseen un flagelo polar. La mayor parte de las especies del género *Vibrio* son halotolerantes y el cloruro de sodio (NaCl) a menudo estimula su multiplicación. El género está constituido por más de 115 especies. Todas las especies son principalmente acuáticas y su distribución depende de la temperatura, la concentración de sodio Na+, el contenido de nutrientes del agua y de las plantas y animales presentes. Se ha encontrado que sólo once especies causan infecciones en humanos, provocando diarrea o infecciones extra-intestinales, pero algunas como V. cholerae pueden causar ambas. La mayoría de las infecciones están relacionadas con la exposición al agua o a través del consumo de peces y mariscos. Cualquier cepa de V. cholerae puede causar diarrea, pero solo los serogrupos O1 y O139 han causado pandemias de cólera.

La especie *V. cholerae*, produce colonias convexas, lisas y redondas que son opacas y granulosas bajo luz transmitida. *V. cholerae* y la mayor parte de los demás *Vibrios*, se multiplican bien a una temperatura de 37°C en muchas clases de medios que contienen sales minerales y asparagina como fuentes de carbono y nitrógeno. *V. cholerae* se multiplica bien en agar de tiosulfato-citrato-bilis-sacarosa (TCBS, thiosulfate citrate bile sucrose), produce colonias amarillas que son fácilmente visibles sobre el fondo verde oscuro del agar. Los *Vibrio* son oxidasa positivos, lo que los distingue de las bacterias gram negativas entéricas. Es característico que los *Vibrio* se multipliquen a un pH muy alto (8,5 a 9,5) y que rápidamente sean destruidos por ácido. Por tanto, los cultivos que contienen hidratos de carbono fermentables se vuelven estériles con rapidez.

En este trabajo se evaluó la relación de la incidencia de *Vibrio* spp. con 17 variables hidrográficas en la subregión Sanquianga-Gorgona usando modelos lineales generalizados y algoritmos de aprendizaje automático como Vectores de Soporte (*Support Vector Machine* – SVM), el Bosque Aleatorio (*Random Forest* - RF) y el Vecino más cercano K (*K-Nearest Neighbors* KNN). Una vez entrenados y ajustados los modelos, los mejores ajuste se utilizaron para predecir espacialemente la incidencia *Vibrio* spp tanto en la marea alta como en la marea baja. Para determinar cuales fueron las mejores variables de respuesta se realizó un procedimiento de reducción dimensional por medio de un análisis de componentes principales; la evaluación de la incidencia y su relación con los componentes principales no arrojó ningún resultado satisfactorio. Todos estos análisis se desarrollaron con el lenguaje de programación R (R Core Team, 2022) y para efectos de reproducibilidad se creó un repositorio digital de los scripts en el sitio web de github (https://www.github.com).

3 Cumplimiento de objetivos

3.1 Objetivo general

Objetivo general	Evaluar la presencia del género bacteriano Vibrio spp (Vibrionales: Vibrionaceae), en la subregión Sanquianga-Gorgona y determinar cuáles son las variables hidrográficas que se relacionan con su presencia.	Porcentaje de avance	% 100

Resultado obtenido	Dificultades	Observaciones
Se logró evaluar la relación entre la incidencia de *Vibrio* spp. y las variables hidrográficas. La temperatura y la salinidad fueron las principales variables predictivas en los modelos.	El procesamiento de las muestras en un espacio con poca esterilidad como fue el laboratorio del buque causó que alguno de los cultivos de las muestras se perdieran debido a la colonización por hongos.	A pesar de no lograrse la obtención de todas las muestras, se pudo obtener resultados adecuados.

Comisión Colombiana del Océano · Construyendo País Marítimo
www.cco.go
Avenida Ciudad de Cali No. 51-66
Bogotá D.C., 4 de mayo

3.2 Objetivos específicos

Objetivo específico	Colectar muestras de agua para aplicar las pruebas confirmatorias para el diagnóstico de *Vibrio* spp.	Porcentaje de avance	67%
Resultado obtenido	Dificultades	Observaciones	
Se colectaron las 36 muestras planeadas.	Debido a las condiciones de esterilización del laboratorio del buque, algunas muestras se perdieron debido a la colonización por hongos.	Los análisis se realizaron con solo 24 datos obtenidos de las muestras que no se perdieron por el procesamiento.	
Objetivo específico	Espacializar la presencia y ausencia de las especies de Vibrio en el área de las desembocaduras de los ríos Amarales, Guascama y Sanquianga en la subregión Sanquianga-Gorgona.	Porcentaje de avance	%
Resultado obtenido	Dificultades	Observaciones	
Se obtuvo el mapa de incidencia de *Vibrio* spp. Tanto para marea alta como para marea baja en el delta del río Sanquianga	Ninguna		
Objetivo específico	Relacionar las condiciones hidrografías entre la marea baja y la marea alta con la presencia de *Vibrio* spp.	Porcentaje de avance	%
Resultado obtenido	Dificultades	Observaciones	
Se logró evaluar la relación entre las condiciones hidrográficas y biológicas del área con la incidencia de *Vibrio* spp. usando los algoritmos de aprendizaje automático. La temperatura y la salinidad fueron las variables más importantes a la hora de ajustar los modelos. Solo dos algoritmos respondieron a los datos obtenidos de manera satisfactoria.	El número de datos no fue suficiente para dar un resultado con el nivel de confianza que se hubiese querido obtener, sin embargo, es una muy buena aproximación a lograr hacer evaluaciones con un número demuestras y datos más grandes.	El entrenamiento de estos algoritmos en este trabajo se puede seguir usando para alimentar futuras evaluaciones considerando más datos de los que se obtuvieron en este trabajo.	

4 Introducción

La familia *Vibrionaceae* está compuesta por una amplia variedad de bacterias heterótrofas autóctonas que, a diferencia de otros géneros bacterianos, la mayoría de sus especies se encuentran en ambientes marinos y estuarinos debido a su necesidad de sodio para crecer (Oliver & Oliver, K, 2007). Dentro del género *Vibrio*, existen alrededor de 115 especies confirmadas de bacilos Gram-negativos, de las cuales aproximadamente una docena pueden ser potencialmente infecciosas para los seres humanos (Wong et al., 2019).

Las especies de *Vibrio* muestran diferencias en su distribución y abundancia a nivel global (Thompson et al., 2006). En general, las especies de *Vibrio* son más abundantes y diversas en ambientes tropicales y se puede encontrar en el medio todo el año. En contraste, en los ambientes boreales, como los del Ártico, las especies de *Vibrio* son menos abundantes y menos diversas debido a las bajas temperaturas, aunque durante el invierno se puede haber presencias de organismos viables, pero no cultivables. Además, los ambientes boreales suelen estar dominados por especies de *Vibrio* psicrófilas, que pueden crecer a bajas temperaturas, mientras que en los ambientes tropicales predominan las especies de *Vibrio* mesófilas, que pueden crecer a temperaturas más cálidas (Thompson et al., 2006).

Las especies de *Vibrio* están asociadas a la causa de enfermedades en peces, camarones y corales y también en humanos (Ceccarelli & Colwell, 2014; Rosenberg & Falkovitz, 2004; Thompson et al., 2006). La especie patogénica más común del género *Vibrio* es *V. cholerae*, causante del Cólera, esta enfermedad se puede producir por la ingesta de alimentos o agua contaminada con este patógeno (Thompson et al., 2006). Otras de las especies dentro de este género que puede causar enfermedades graves es la *V. parahaemolyticus* y *V. vulnificus* causantes de la "Vibriosis". Las infecciones por *Vibrio* en humanos se asocian a menudo con el consumo reciente de marisco, ya que los *Vibrios* se encuentran comúnmente en las aguas de estuario y en una variedad de mariscos. Algunos estudios han demostrado que las muestras de comida con productos marinos como los camarones, pescado crudo y los moluscos pueden contener altas Unidades Formadoras de Colonias (UFC) de *Vibrio* spp., con varias especies como *V. alginolyticus, V. parahaemolyticus, V. vulnificus, V. fluvialis, V. mimicus* y *V. cholera*. Vibrio *parahaemolyticus* está reconocido como la principal causa de gastroenteritis bacteriana asociada al consumo de pescado y marisco en muchas partes del mundo (Oliver & Oliver, K, 2007).

Para comprender la epidemiología de estas enfermedades y el riesgo de contaminación que pueden presentar estos patógenos para los asentamientos humanos, es esencial entender su ecología y cómo las condiciones ambientales están relacionadas con su presencia para poder prevenir la toma de alimentos de dichos lugares y emitir alertas tempranas (Córdoba Meza et al., 2021). Algunos estudios han identificado la temperatura, salinidad y concentración de nutrientes (nitritos, nitratos, fosfatos y silicatos) en el medio como los principales factores que influyen en la distribución y abundancia de estas especies (Wong et al., 2019), sin embargo, estas relaciones no son del todo consistentes con otras variables y dependen mucho de la resolución taxonómica (Takemura et al., 2014). Además, se ha observado que las especies de Vibrio establecen asociaciones estrechas con organismos planctónicos, especialmente crustáceos como los copépodos (Turner et al., 2009). En estas interacciones, las bacterias aprovechan la quitina exoesquelética de los organismos planctónicos como fuente de nutrientes para obtener carbono y nitrógeno, lo que les brinda una ventaja competitiva sobre las especies que no utilizan esta estrategia ecológica.

La mayoría de los estudios que evalúan las relaciones entre *Vibrio* spp y las condiciones ambientales se han centrado en análisis basados en regresiones lineales entre la magnitud de las variables y la abundancia (Heidelberg et al., 2002). Sin embargo, son escasos los estudios que relacionan las variables ambientales con la presencia de este género (Takemura et al., 2014). Existen algunos trabajos que han evaluado estas relaciones con algoritmos no lineales (Baker-Austin et al., 2013; Escobar et al., 2015) e incluso se han desarrollado sistemas de alerta temprana a partir del uso de estos algoritmos (Brumfield et al., 2021).

Este estudio busca evaluar la incidencia del género *Vibrio* spp. y determinar, mediante modelos lineales generalizados y algoritmos de aprendizaje automático, las variables hidrográficas que pueden explicar su presencia.

5 Metodología

5.1 Metodología efectiva de muestreo

Entre los días 28 de abril y el 07 de mayo de 2021, a bordo del buque oceanográfico ARC "Providencia", se visitaron 18 estaciones de muestreo en marea alta y marea baja (Figura 1) repartidas en tres transectos en las bocas del delta del río Sanquianga, Boca Amarales, Boca Guascama y Boca Sanquianga (6 estaciones c/u) (Figura 2). Durante cada una de las visitas se colectaron las muestras para sembrar los cultivos de *Vibrio* spp. a partir de un litro de agua que se tomó directamente en una botella esterilizada. Para las muestras de agua para determinar las variables fisicoquímicas se utilizó una botella Niskin de 10 litros (Figura 3A) atada a una cuerda con un mensajero para el cierre automático. En esta botella se tomaron 4 litros de agua para determinar la clorofila *a*, el pH, los nutrientes (nitritos, nitratos, fosfatos, silicatos) en el agua. La transparencia se midió con un disco Secchi al momento de la toma de las muestras y los perfiles de temperatura, salinidad, densidad, y oxígeno disuelto se midieron con una sonda CTDO.

Para la determinación de la riqueza y diversidad del fitoplancton, se filtraron alrededor de 20 litros de agua a través de una red de 50um y se tomó una muestra de 500 ml para ser analizadas en el laboratorio del Centro de Investigaciones Oceanográficas e Hidrográficas del Pacífico en Tumaco. Las muestras de zooplancton fueron colectadas con una red tipo bongo de 300 y 500um de ojo de malla con arrastres horizontales a 2 nudos durante 5 minutos; estas muestras se preservaron en formol.

Comisión Colombiana del Océano · Construyendo País Marítimo
www.cco.go
Bogotá D.C., 4 de mayo
Avenida Ciudad de Cali No. 51-66

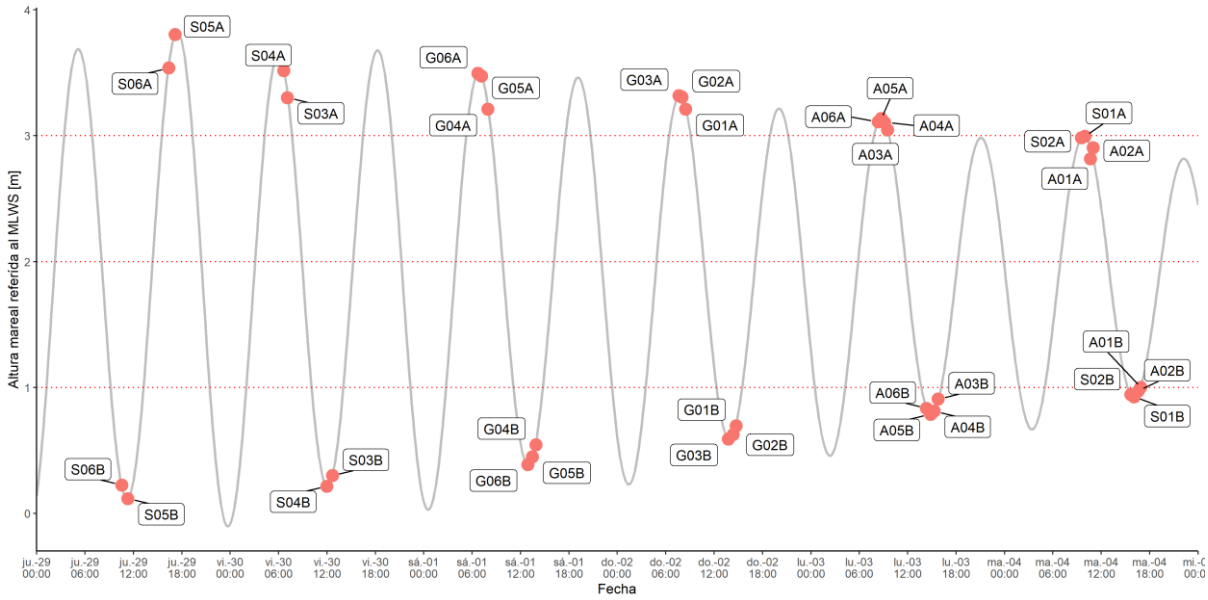

Figura 1. Comportamiento de la onda mareal entre el 28 de abril y el 07 de mayo de 2021, señalando el momento de colecta de las muestras en cada una de las estaciones.

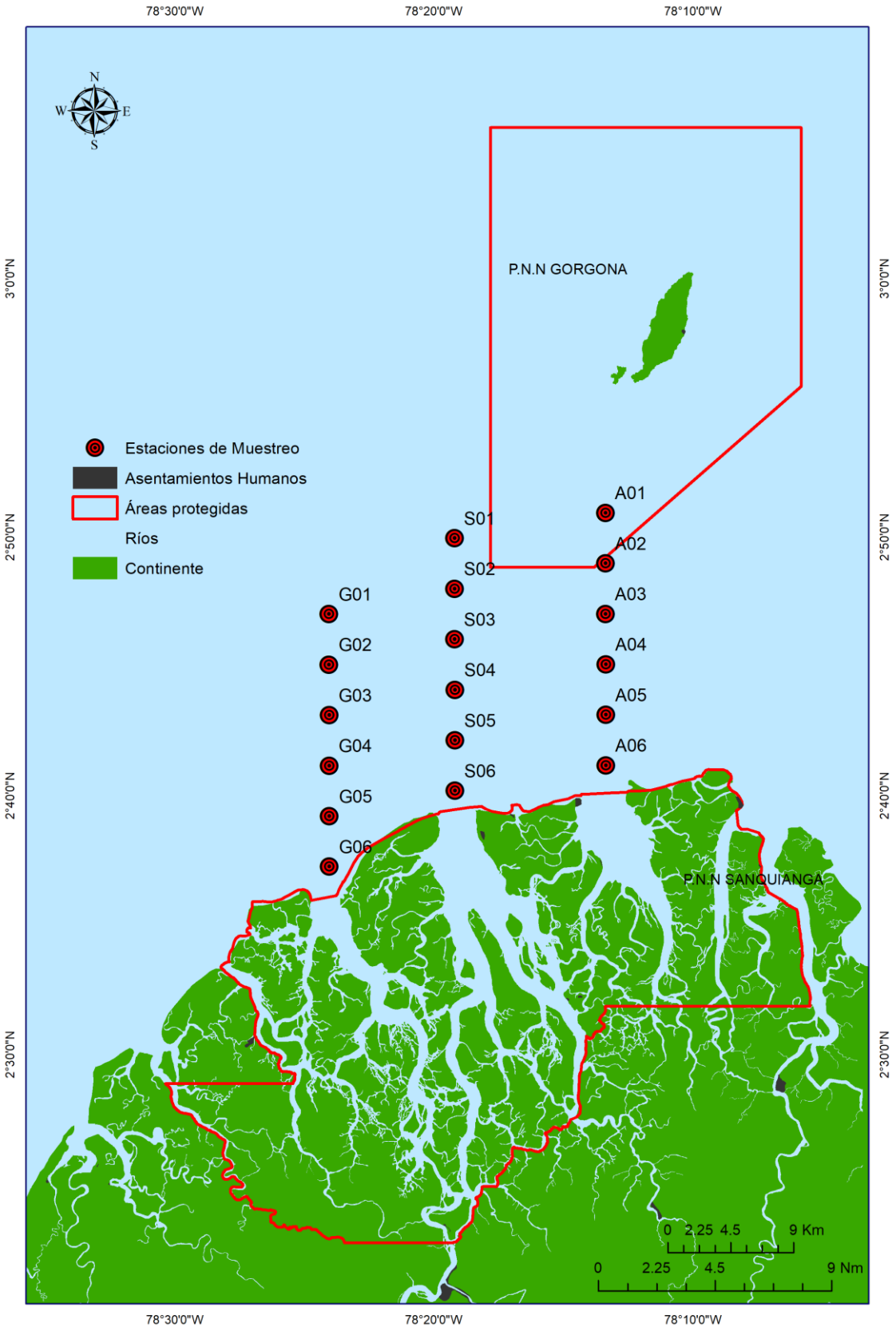

Figura 2. Área de estudio. Región Sanquianga-Gorgona.

5.2 Metodología de procesamiento y análisis de resultados

Análisis de muestras

Muestras microbiologías

En este estudio, se utilizó una unidad de filtración previamente esterilizada con una bomba de vacío y expuesta a la luz UV durante 10 minutos para garantizar su esterilidad antes de su uso en el procedimiento de ensayo. Se colocó un filtro de membrana estéril en la unidad de filtración con la cuadrícula hacia arriba y se homogenizó vigorosamente la muestra antes de agregarla al embudo de filtración. Se aplicó vacío para hacer pasar la muestra a través del filtro de membrana de 0,45um, realizando diluciones si era necesario. Se observó que todo el líquido pasara a través del filtro y se enjuagó la superficie interior del embudo con agua peptonada estéril 0,1% o agua tamponada de dilución estéril. Después de la filtración, se retiró el filtro de membrana utilizando pinzas estériles y se incubó en Agua Peptona Alcalina al 1% a 35 ± 2 ºC durante 6 a 8 horas. Se repitió este procedimiento para todas las muestras restantes, realizando enjuagues con agua destilada estéril y exponiendo los embudos a luz UV por 10 minutos para evitar la contaminación cruzada por arrastre.

Después de haber transcurrido el tiempo de incubación, se tomó una asa de aro de inoculación (Figura 3B) y se sembró por agotamiento en la superficie del Agar TCBS. Las placas de Petri con el Agar TCBS se incubaron invertidas durante 18 a 24 horas a 35 ± 2 °C. Luego, se observaron las colonias presuntivas de *Vibrio* spp en el Agar TCBS, las cuales se caracterizan por tener un tamaño de 2-3 mm de diámetro, textura lisa, color amarillo y ser ligeramente aplanadas, con el centro opaco y la periferia translúcida.

Muestras químicas

Durante la expedición a bordo del buque, se llevaron a cabo las mediciones de Oxígeno Disuelto (OD), salinidad y pH. Para determinar la OD, se utilizó un dosificador Metrhom modelo Multidosimat y para medir la salinidad y el pH se utilizó un multiparametro Schott modelo Handylab multi 12. Además, se realizó un pretratamiento de las muestras destinadas a los ensayos de nutrientes y clorofila a, y se continuó con el tratamiento analítico de las mismas en el laboratorio del Centro de Investigaciones Oceanográficas e Hidrográficas del Pacífico.

En el Laboratorio del CCCP se analizaron las muestras de nutrientes, pH, salinidad, Solidos Suspendidos Totales (SST), Oxígeno Disuelto (OD) y clorofila a, siguiendo los métodos verificados en el laboratorio de química en conformidad con la norma NTC ISO/IEC 17025:2017, la cual establece los requisitos generales para la competencia de los laboratorios de ensayo y calibración. Las determinaciones analíticas de nitritos, nitratos y silicatos se realizaron mediante métodos colorimétricos descritos por Bendschneider & Robinson (1952) y reducción con cadmio-cobre y metol-sulfito, respectivamente, tal como se describen en Strickland & Parsons (1972). Para determinar los fosfatos, se empleó el método del ácido ascórbico según lo publicado Murphy & Riley (1958). La medición de pH y salinidad se realizó mediante los métodos 4500-H+ B y 2510 B, respectivamente. Para determinar los Sólidos Suspendidos Totales se utilizó el método 2540 D, mientras que para el Oxígeno Disuelto se usó el método yodométrico 4500-O B. Finalmente, la determinación de clorofila a se realizó aplicando el método tricromático 10200 H, todos estos procedimientos se describen en el Standard Methods for the Examination of Water and Wasterwater (Lipps et al., 2023).

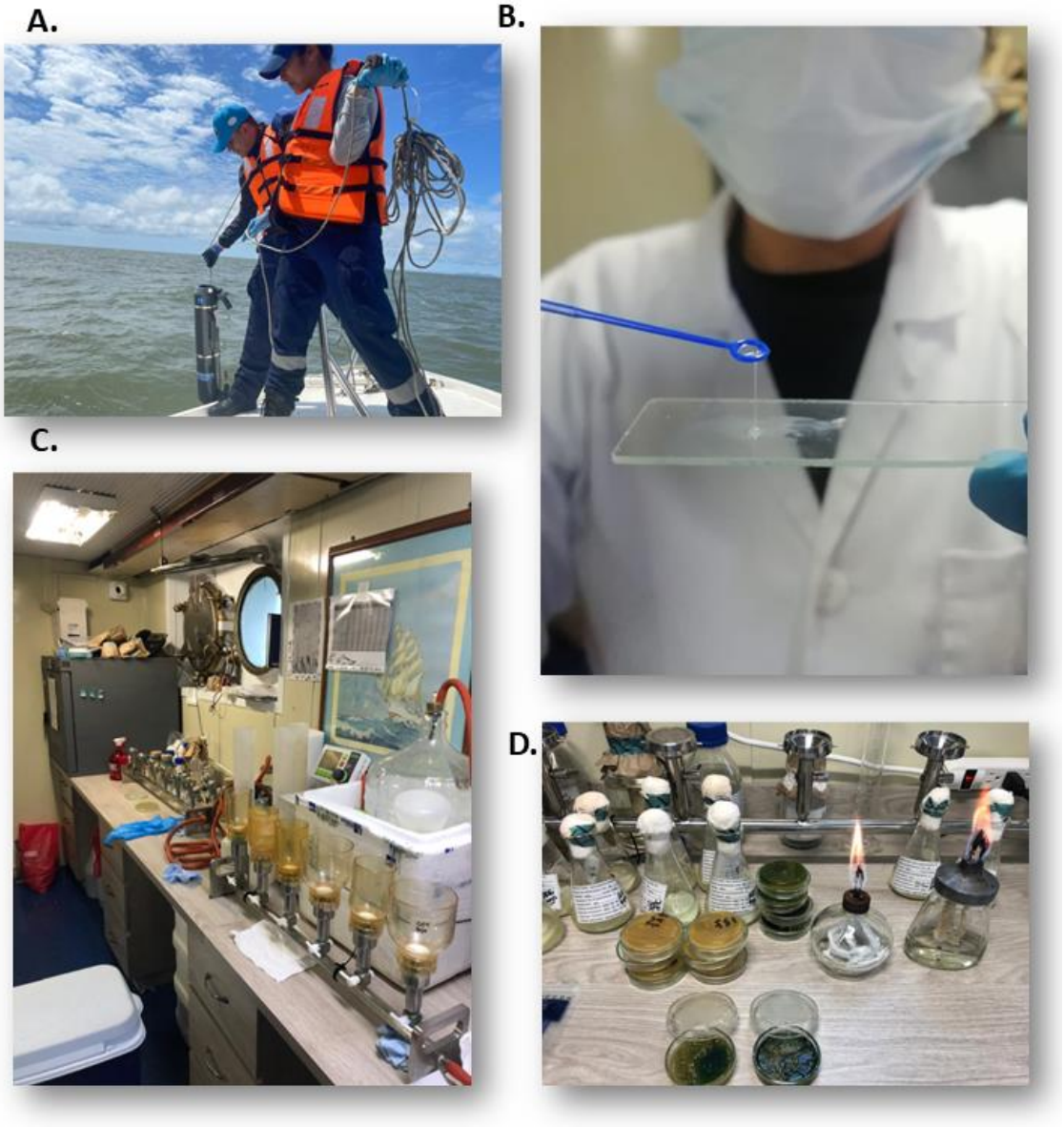

Figura 3. Toma y procesamiento de muestras. A. Toma de muestras de agua en las estaciones de muestreo. B. Proceso de inoculación con asa de aro en el medio de cultivo. C. Laboratorio de procesamiento de las muestras de microbiología. D. Siembra de las muestras de *Vibrio* spp.

Muestras biológicas

En el laboratorio del Área de Protección del Medio Marino (APROMM) del CIOH Pacífico, se cuantificaron las células de fitoplancton en una cámara Sedgwick-Rafter de 1 ml, utilizando un microscopio invertido Leica DMi1 con objetivos de 10X y 40X para observar las células con mayor detalle cuando era necesario. Se realizaron tres réplicas por muestra y se identificaron las especies utilizando claves taxonómicas y descripciones (Balech, 1988; Cupp, 1943; Morales-Pulido & Aké-Castillo, 2019; Tomas et al., 2010). Además, se verificó el estado taxonómico de las especies,

autores y sinonimia mediante la base de datos Algaebase (Guiry & Guiry, 2023). Del ensamblaje del fitoplancton presente en el área se extrajo la densidad celular en el agua dada en células por litro y se calcularon dos medidas de diversidad a partir de los números de Hill, la diversidad de orden cero (^0D) y la diversidad de orden uno (^1D) (Chao et al., 2014).

$$^qD = (\sum_{i=1}^{S} P_i^q)^{1/(1-q)}$$

Donde S es el número de especies en el ensamblaje y la especie *i* tiene la abundancia relativa p_i, *i* = 1, 2, ..S. El parámetro q determina la sensibilidad de la medida a las frecuencias relativas. Para q = 0 las abundancias no intervienen en la sumatoria de la anterior (Chao et al., 2014). Para el cálculo de esto número se usó el paquete de R "*vegan*" (Oksanen et al., 2022) y se obtuvieron de cada estación de muestreo a partir de una matríz de densidad de células.

Análisis de datos

En total se obtuvieron datos para 17 variables ambientales hidrográficas, seis variable químicas: Nitratos (NO2 -) [µM], Nitratos (NO3 -) [µM], Fosfatos (PO4-3) [µM], Silicatos (SiO2) [µM], pH, Oxígeno disuelto [mg O2.L-1]; cinco variables físicas Transparencia del agua (m), Solidos Suspendidos totales [mg.L-1], Temperatura superficial del mar (°C), Salinidad superficial del mar (PSU) y Densidad del agua superficial [kg.m-3]; cinco variables biológicas Clorofila a superficial [µg.L-1], Densidad del fitoplancton superficial [Cel.L-1], Diversidad de primer y segundo orden (0D y 1D números de Hill q=0 – Riqueza de especies; q = 1 exp (H')).

Se utilizó el paquete de lenguaje de programación R "*tidyverse*" " (Wickham et al., 2019) para el manejo y limpieza de datos. Para evaluar las diferencias significativas entre las mareas, se aplicó el procedimiento estadístico no paramétrico conocido como *Multiresponse Permutation Procedure* (MRPP). Esta prueba compara la disimilitud media entre los grupos observados con la disimilitud media esperada si las observaciones estuvieran distribuidas al azar. Si la disimilitud media observada es significativamente mayor que la esperada al azar, se concluye que los grupos son significativamente diferentes. El paquete "*vegan*" de R (Oksanen et al., 2022) se utilizó para realizar este análisis, el cual es adecuado para datos no paramétricos y resistentes a valores atípicos debido a que se basa en la permutación de los datos. Luego de esta prueba, se espacializaron todas las variables medidas a partir de las estaciones de muestreo, tanto en marea baja como en marea alta, utilizando el algoritmo de Barnes (Koch et al., 1983) para obtener mapas de distribución de las magnitudes de las variables. Este análisis se llevó a cabo utilizando el paquete "*oce*" de R (Kelley & Richards, 2022). A partir de esta comparación general, se realizaron comparaciones particulares para detectar diferencias particulares para cada variable utilizando la prueba no paramétrica de Wilcoxon usando el paquete "stats" de R (R Core Team, 2022).

Con el propósito de reducir la cantidad de variables estudiadas y descartar aquellas que pudieran presentar correlaciones fuertes, se utilizó una correlación múltiple de Spearman mediante el paquete "stats" de R (R Core Team, 2022). Luego, se llevó a cabo un análisis multivariado de componentes principales para identificar las variables más importantes y transformar los datos a un espacio de menor dimensionalidad. Para este análisis, se utilizaron los paquetes "stats" de R (R Core Team, 2022) y el paquete "factoextra" para el manejo de los resultados y gráficos (Kassambara & Mundt, 2020).

Para evaluar y crear modelos de relación entre la incidencia de Vibrio y las variables obtenidas, se utilizó un modelo General Linealizado (GLM) de tipo binomial con el paquete de R "pROC" (Robin et al., 2011), utilizando la función *logit* para modelar la relación entre la probabilidad de éxito de la variable dependiente y las variables predictoras (Herrera et al., 2023). Este modelo se basa en la distribución binomial y utiliza la función *logit* para modelar la relación entre la probabilidad de éxito de la variable dependiente y las variables predictoras.

$$p(X) = \beta_0 + \beta_1 X$$

donde pi es la probabilidad de éxito para la observación *i*, β0 es la intersección, β1 son los coeficientes de regresión para las variables predictoras X , respectivamente.

La función logit se define como:

$$\text{logit}(p) = \ln(p/1 - p)$$

donde *p* es la probabilidad de éxito.

Como alternativa a la aproximación logística se usaron algoritmos de aprendizaje automático para evaluar la misma relación entre la incidencia de *Vibrio* spp. y las variables ambientales. Estos incluyen los Vectores de Soporte (*Support Vector Machine* – SVM), el Bosque Aleatorio (*Random Forest* - RF) y el Vecino más cercano K (*K-Nearest Neighbors* KNN). El *SVM* es un algoritmo que busca encontrar la mejor forma de separar dos clases de datos mediante la identificación de un hiperplano que maximice el margen de separación entre las clases. Los puntos de datos que se encuentran en el margen máximo se denominan vectores de soporte y son fundamentales para la construcción del modelo (Zhou, 2021). El *RF* es un algoritmo de aprendizaje supervisado que utiliza una colección de árboles de decisión para clasificar objetos. Es un método de conjunto que utiliza la votación para determinar la clase final de un objeto y utiliza la muestra aleatoria y el subconjunto de características para reducir la varianza y el sobreajuste (James et al., 2014; Zhou, 2021). El algoritmo *KNN* clasifica nuevos puntos de datos en función de su proximidad a los puntos de datos existentes en un conjunto de entrenamiento. *KNN* es un método basado en instancias, lo que significa que el algoritmo no construye un modelo explícito, sino que almacena todos los puntos de datos de entrenamiento y los utiliza para clasificar nuevos puntos de datos.

Una vez obtenidos los ajustes de todos los modelos, se evaluó su capacidad predictiva en relación a los resultados de la incidencia de *Vibrio* spp. Inicialmente, se realizó una partición aleatoria de los datos en un conjunto de entrenamiento (75%) y un conjunto de evaluación (25%). Para ajustar los modelos, se utilizó el conjunto de entrenamiento y se aplicó una validación cruzada con 10 particiones. Para comparar los modelos, se emplearon el área bajo la curva (*AUC*) y una matriz de confusión como medidas de precisión. El *AUC* es una medida común para evaluar la capacidad discriminativa de un modelo, mientras que la matriz de confusión permite evaluar el desempeño del modelo en términos de los errores de clasificación (James et al., 2014).

Para efectos de reproducibilidad, todos los análisis están disponibles en el repositorio público de *Github: https://github.com/ChrisBermudezR/Vibrio_ExpPacifico2021*

6 Resultados

Incidencia de *Vibrio* spp.

De las 36 muestras tomadas solo se logró tener resultados y cultivar exitosamente el 66% (24 muestras - Figura 4) debido a que 12 muestras fueron atacadas por hongos y se debieron desechar.

Se registró una presencia de *Vibrio* spp. del 83.3%, mientras que la ausencia fue del 16.67%. Durante la marea alta, se detectaron 13 estaciones con presencia de *Vibrio* spp. y una con ausencia, mientras que durante la marea baja se identificaron seis estaciones con presencia y tres con ausencia. En general, la marea alta representó el 65% de todas las presencias en todo el muestreo, mientras que la marea baja representó el 35% del total de presencias. En cuanto a las ausencias, el 25% se registró durante la marea alta y el 75% durante la marea baja (Figura 5).

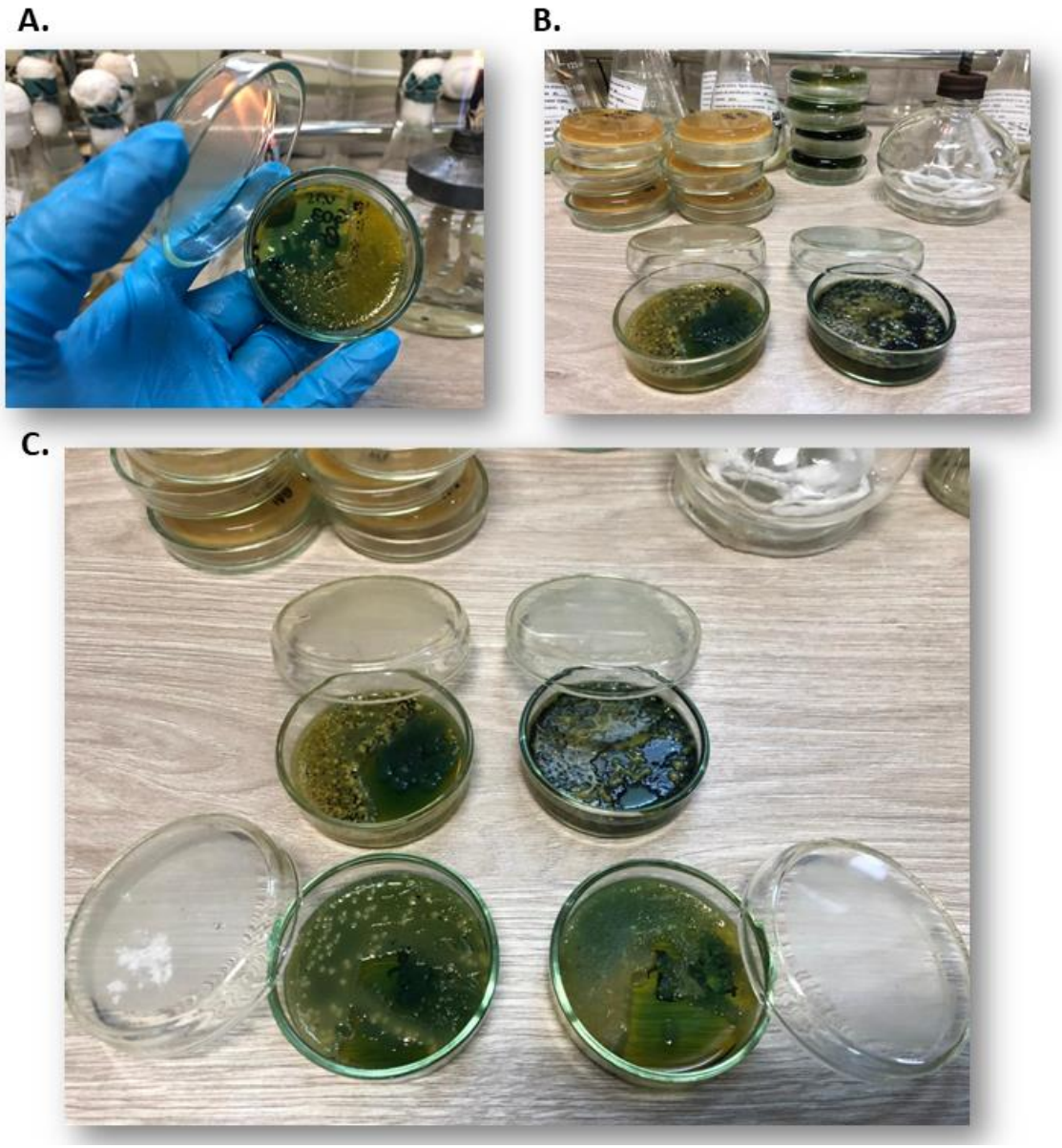

Figura 4. Resultados de los cultivos con hagar TCBS mostrando la confimación visual de la presencia de *Vibrio* spp. (A, B y C – color verde típico de este tipo de cultivo).

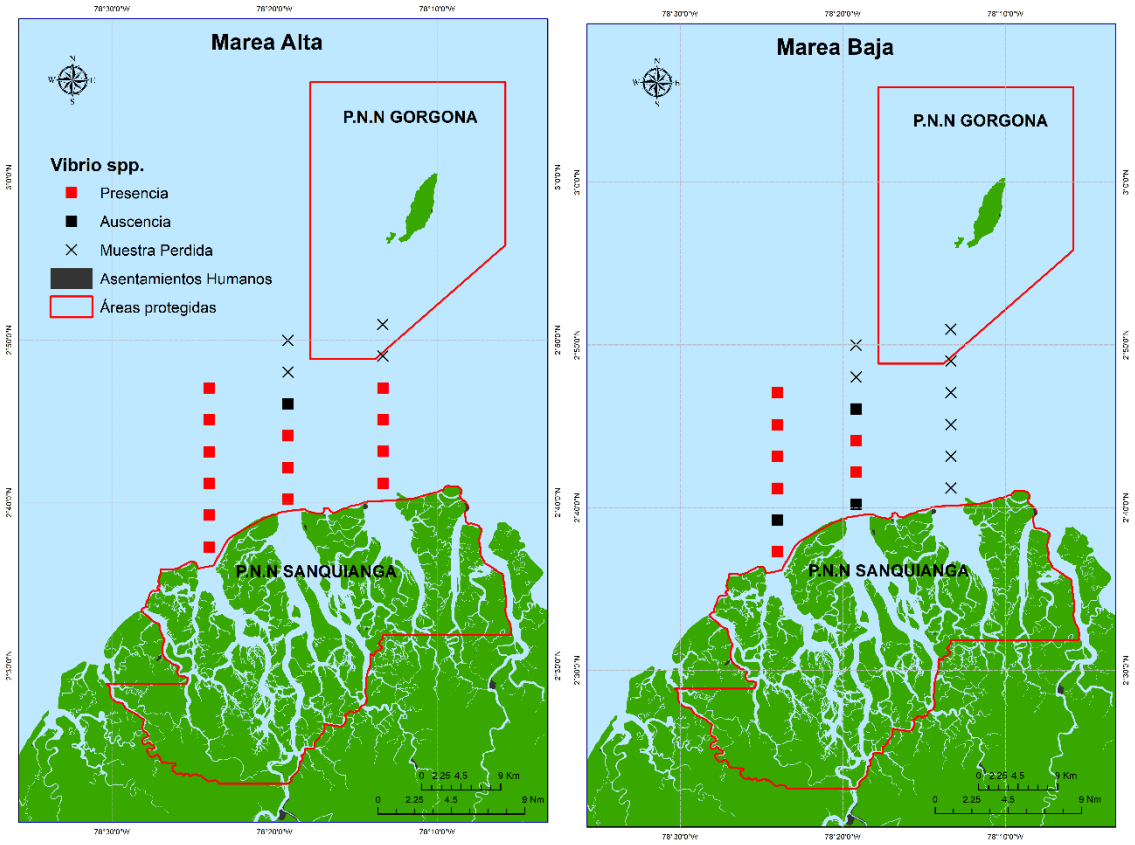

Figura 5. Incidencia de *Vibrio* spp. en la subregión Sanquianga-Gorgona durante mareas alta y baja.

Análisis de datos

La prueba de MRPP no detectó diferencias significativas entre la marea alta y la marea baja en el área durante el período de tiempo de la toma de muestras (A = 0.0163, p = 0.154). Sin embargo, la técnica de espacialización con el algoritmo de Barnes para todas las variables mostró una distribución espacial diferente para cada variable en cada período mareal.

En la marea alta, los valores de nitritos oscilaron entre 0.05 y 0.77 µM, con una media de 0.1856 µM, mientras que en la marea baja los valores estuvieron en el rango de 0.06 a 0.92 µM, con una media de 0.4206 µM. Se observa que los valores de nitritos en la marea baja son más altos en general que en la marea alta. La prueba de Wilcoxon, que se realizó para evaluar si había una diferencia significativa en los valores de nitritos entre ambas mareas (W = 67, p< 0.01) lo que indica que hay una diferencia significativa en los valores de nitritos entre las dos mareas. En la distribución espacial, los nitritos mostraron mayores concentraciones en la bocana Amarales en ambas mareas (Figura 6), pero para los nitritos la distribución tuvo una tendencia espacial contraria, presentándose mayor concentración en la marea baja en la bocana de Guascama (Figura 6). En marea alta, los niveles de nitratos oscilaron entre 0.27 y 2.28 µM, con una media de 0.8717 µM. Por otro lado, en marea baja, los niveles de nitratos variaron entre 0.19 y 4.31 µM, con una media de 1.716 µM. La prueba de Wilcoxon reveló una diferencia significativa entre los niveles de nitratos en marea alta y marea baja (W = 96, p< 0.05), indicando que la concentración de nitratos es estadísticamente diferente entre ambos períodos (Tabla 1).

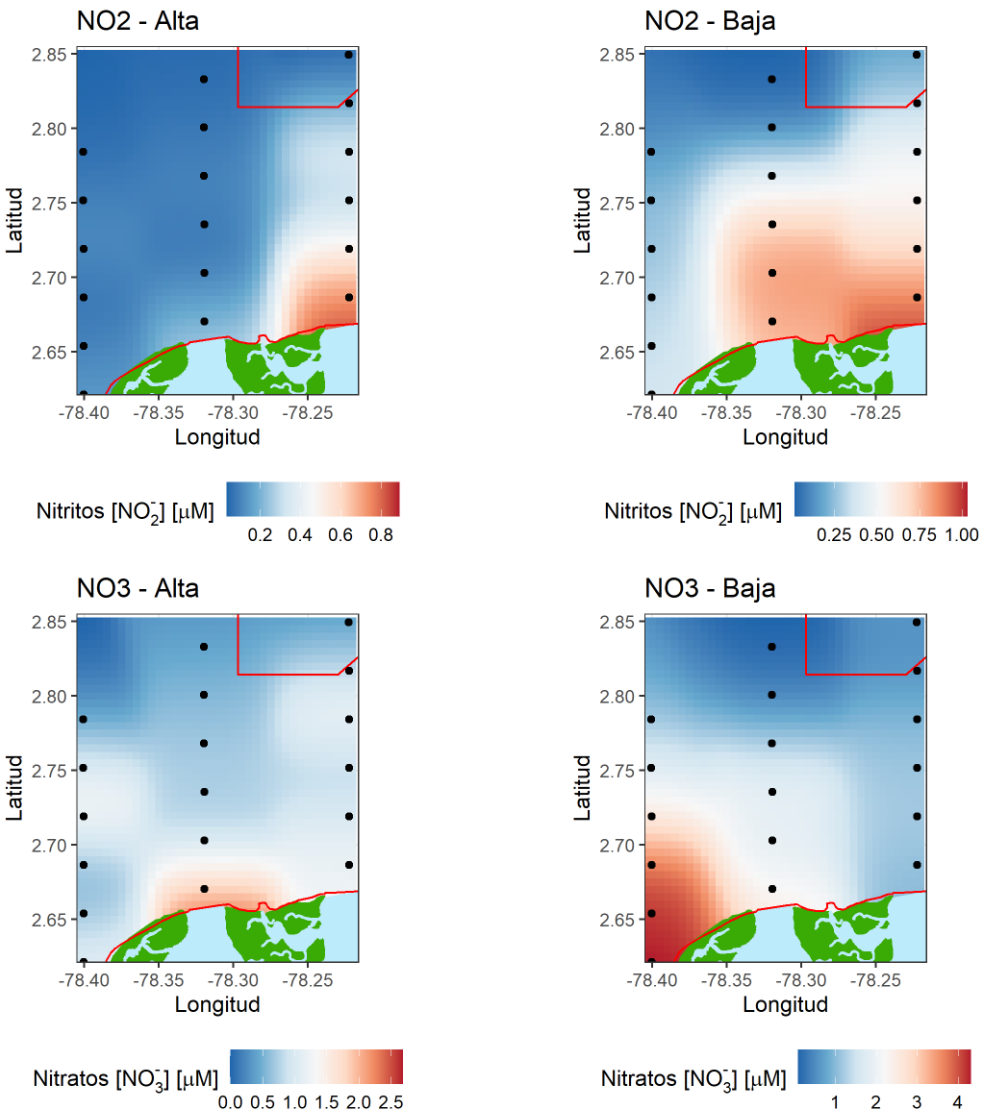

Figura 6. Mapas de la concentración de nitritos y nitratos en el área de estudio durante la marea alta y la marea baja.

En la marea alta, la concentración de fosfatos osciló entre 0.06 µM y 0.42 µM, con una mediana de 0.19 µM y una media de 0.1972 µM. Por otro lado, en la marea baja, la concentración de fosfatos osciló entre 0.06 µM y 0.91 µM, con una mediana de 0.325 µM y una media de 0.3083 µM. Se observó una variación en la dispersión de los fosfatos entre la marea alta y baja, siendo más notables las concentraciones en la bocana Sanquianga durante la marea baja, sin embargo no se encontraron diferencias significativas entre los períodos mareales ($W = 111$, $p = 0.109$)

Por otro lado, se encontraron las mayores concentraciones de silicatos en la bocana Guamales (Figura 7) con diferencias significativas entre los períodos mareales ($W = 55$, $p < 0,01$). Durante la marea alta, se encontró que los valores de silicatos oscilaron entre un mínimo de 3.51 µM y un máximo de 44.44 µM, con una mediana de 11.49 µM y una media de 14.55 µM. Mientras tanto, durante la marea baja, se encontró que los valores de silicatos oscilaron entre un mínimo de 6.86 µM y un máximo de 78.47 µM, con una mediana de 27.64 µM y una media de 34.50 µM (Tabla 1).

Comisión Colombiana del Océano · Construyendo País Marítimo
www.cco.go
Bogotá D.C., 4 de mayo
Avenida Ciudad de Cali No. 51-66

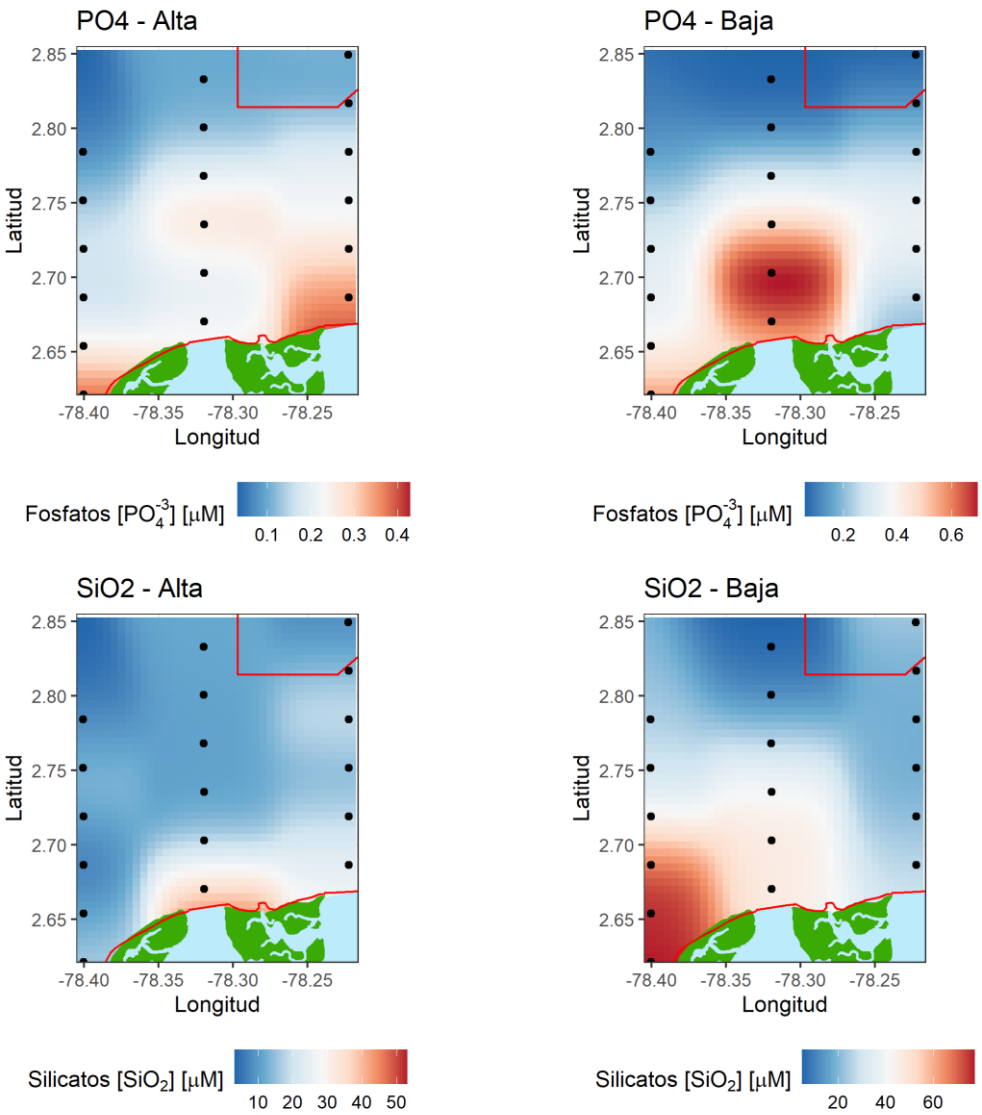

Figura 7. Mapas de la concentración de fosfatos y silicatos en el área de estudio durante la marea alta y la marea baja.

Durante la marea alta, el nivel promedio de clorofila fue de 1.021 µg.L^{-1}, mientras que durante la marea baja, el nivel promedio fue de 2.031 µg.L^{-1}. Se realizó una prueba de Wilcoxon y se encontró que hay una diferencia significativa en los niveles de clorofila entre la marea alta y la marea baja ($W = 39$, $p < 0.01$). En cuanto al pH, durante la marea alta, el nivel promedio fue de 8.191 y durante la marea baja fue de 8.128. Aunque hay una pequeña diferencia, la prueba de Wilcoxon no mostró una diferencia significativa entre los niveles de pH de la marea alta y la marea baja ($W = 207.5$, $p = 0.1537$) (Tabla 1).

Se observó que la concentración de clorofila a mostró una tendencia a ser mayor en las zonas costeras, con una mayor presencia en la marea baja en las bocanas de Sanquianga y Amarales, aunque solo se registró en la boca de Amarales durante la marea alta. Por otro lado, se encontraron valores más bajos de pH en la marea baja de la bocana Guamales (Figura 8).

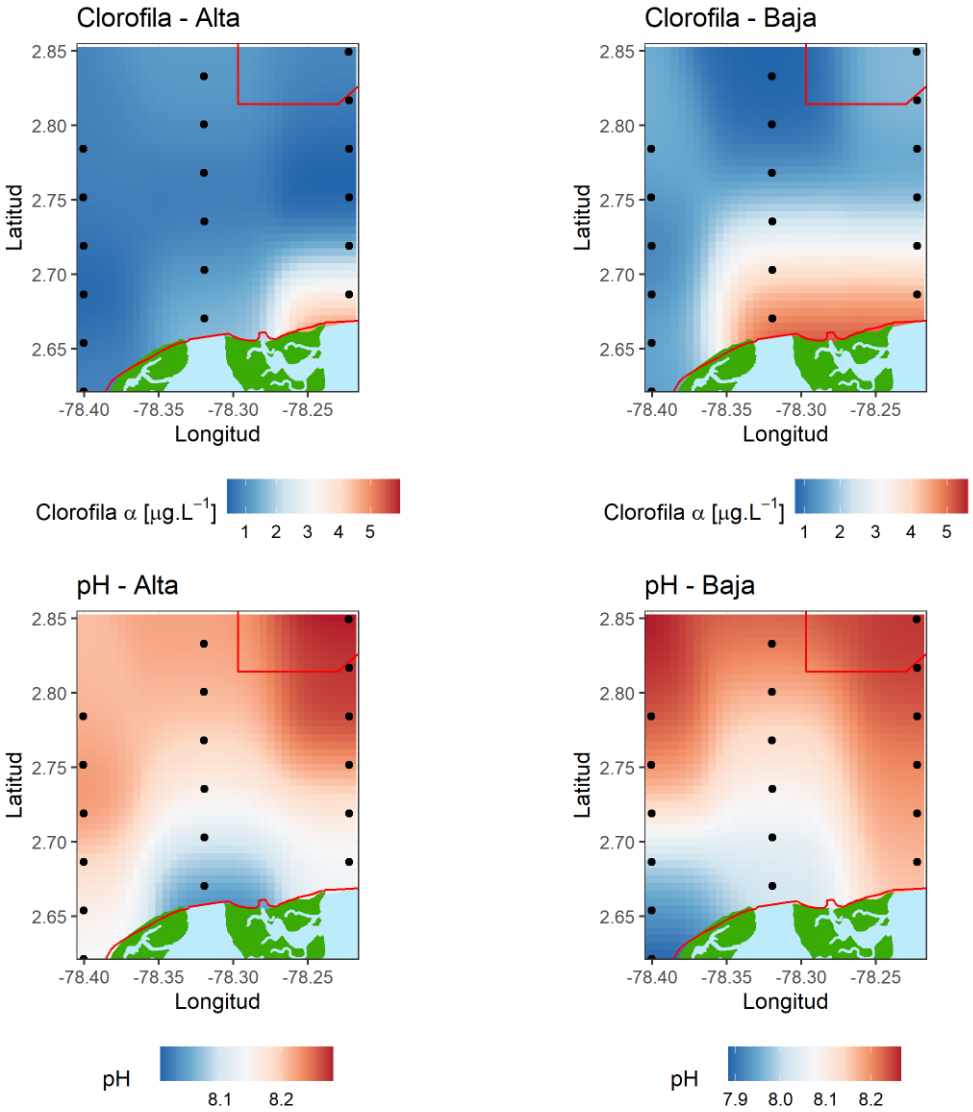

Figura 8. Mapas de la concentración de clorofila a y el pH en el área de estudio durante la marea alta y la marea baja.

La concentración de oxígeno disuelto superficial tuvo una tendencia a ser mayor hacia el frente oceánico y con mayor concentración en la marea alta. En la boca Sanquianga se encontró la menor concentración de oxígeno superficial durante la marea baja, presentado diferencias signifivativas para ambas mareas ($W = 252$, $p < 0{,}01$). La transparencia tendió a ser mayor hacia el frente oceánico y mucho menor hacia la costa (Figura 9) sobre todo en marea alta. Durante la marea alta, la transparencia mínima fue de 0.6 metros, con un primer cuartil de 1.975 metros, una mediana de 2.5 metros, una media de 4.367 metros y un tercer cuartil de 6 metros. Por otro lado, durante la marea baja, se observó una transparencia mínima de 0.6 metros, un primer cuartil de 0.925 metros, una mediana de 1.35 metros, una media de 2.336 metros y un tercer cuartil de 2.375 metros. Además, se registró un valor máximo de 8.5 metros en la marea baja y de 13 metros en la marea alta. Estos resultados indican una transparencia significativamente mayor durante la marea alta en comparación con la marea baja ($W = 233$, $p < 0{,}05$) (Tabla 1).

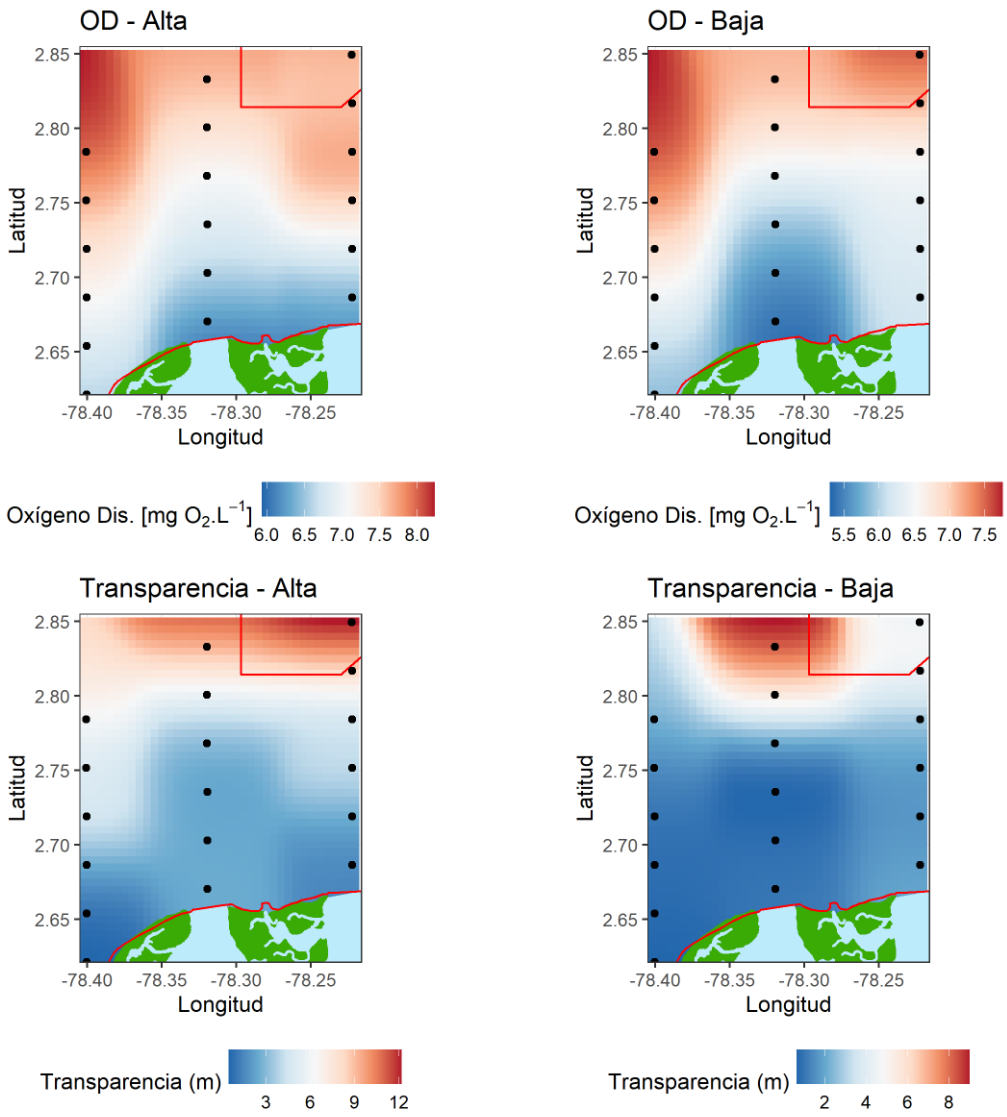

Figura 9. Mapas de la concentración de oxígeno disuelto superficial y la transparencia del agua en el área de estudio durante la marea alta y la marea baja.

Se encontró la mayor concentración de sólidos suspendidos en la bocana de Sanquianga en ambas mareas (Figura 10). y no se encontraron diferencias significativas en las mediciones (W = 110, p = 0.103). En la marea alta, la concentración media de sólidos suspendidos fue de 38.04 mg.L^{-1}, con una mediana de 24.98 mg.L^{-1} y un máximo de 160.97 mg.L^{-1}. En la marea baja, la concentración media fue de 48.73 mg.L^{-1} con una mediana de 35.88 mg.L^{-1}y un máximo de 188.05 mg.L^{-1}.

En cuanto a la temperatura superficial, en marea alta la temperatura mínima fue de 24.67°C, mientras que la temperatura máxima registrada fue de 27.91°C y la media fue de 27.02°C, con una mediana de 27.16°C. Para la marea baja se registraron temperaturas en el rango de 25.47 °C a 30.02 °C, con una temperatura media de 27.36 °C. El 25% de las mediciones se encontraban por debajo de 26.90 °C, mientras que el 75% de las mediciones estaban por debajo de 27.85 °C. La temperatura más baja se registró en 25.47 °C y la más alta en 30.02 °C. Se observó un patrón distinto entre la marea alta y la marea baja. Durante la marea alta, se registró una lengua fría que se extendía desde el oeste del área de estudio y atravesaba la porción media, mientras que, durante la marea baja, se registró la temperatura más baja en la bocana de Sanquianga (Figura

10) sin embargo no se encontraron diferencias significativas para la temperatura (W = 139, p =0.48) (Tabla 1).

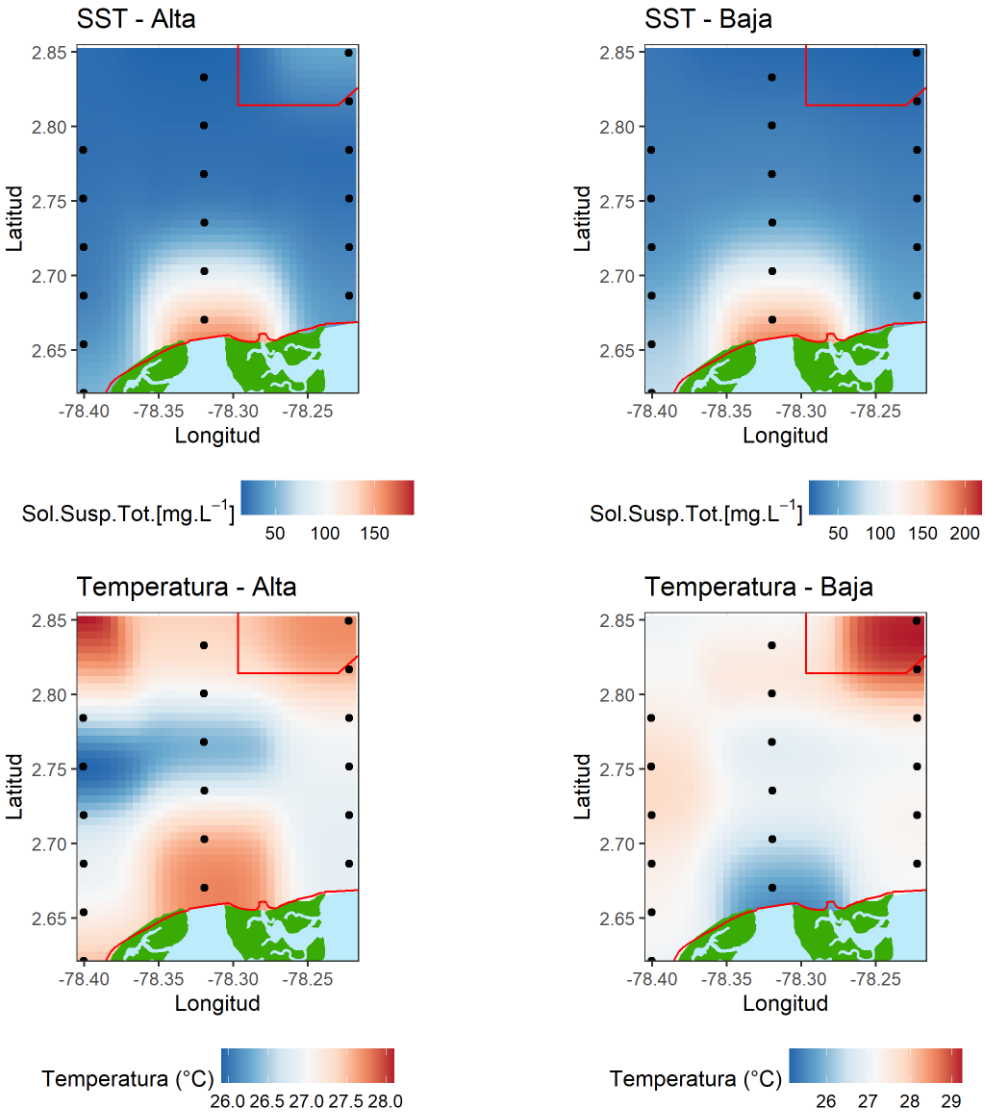

Figura 10. Mapas de la concentración de solidos suspendidos totales y la temperatura superficial en el área de estudio durante la marea alta y la marea baja.

En marea alta, la salinidad de agua de mar varió entre un mínimo de 22.725 PSU y un máximo de 29.474 PSU, con una media de 23.456 PSU. Por otro lado, durante marea baja, la salinidad varió entre un mínimo de 0.3501 PSU y un máximo de 30.1194 PSU, con una media de 17.0694 PSU. La prueba de Wilcoxon mostró una diferencia significativa entre las muestras de marea alta y marea baja para la salinidad (W=225, $p < 0.05$). La menor salinidad se registró en la bocana de Sanquianga para ambas mareas y en marea baja el flujo de agua en la superficie registró una muy baja salinidad (Figura 11).

Para la densidad, durante marea alta la media fue de 1014 kg.m^{-3}, con valores que oscilaron entre un mínimo de 1003 kg.m^{-3}y un máximo de 1018 kg.m^{-3}. Durante marea baja, la media de densidad fue de 1009 kg.m^{-3}, con valores que oscilaron entre un mínimo de 997 kg.m^{-3}y un máximo de 1019

kg.m^{-3}. La prueba de Wilcoxon también mostró una diferencia significativa entre las muestras de marea alta y marea baja para la densidad (W=228, p<0.05) (Tabla 1).

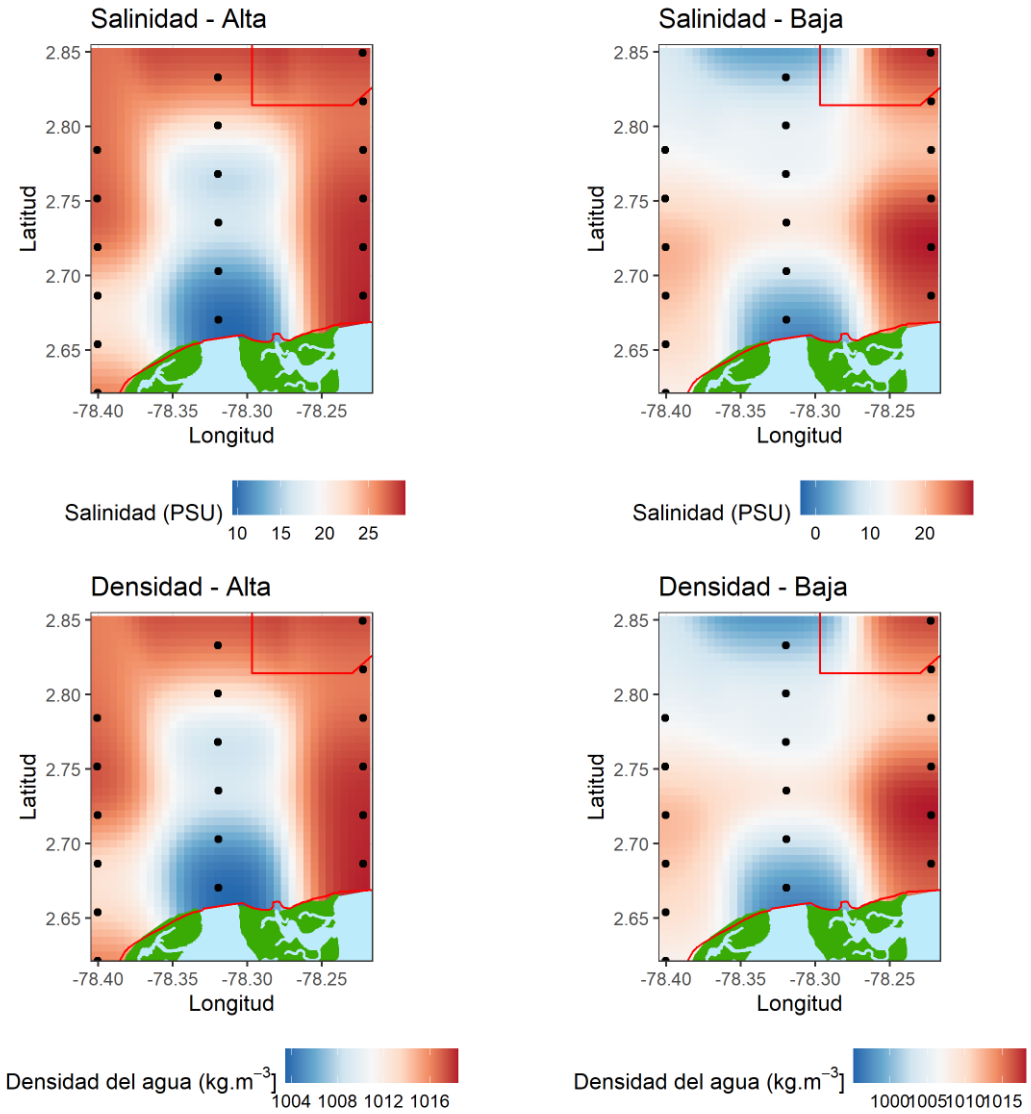

Figura 11. Mapas de la salinidad superficial y la densidad del agua en el área de estudio durante la marea alta y la marea baja.

En marea alta, la densidad del fitoplancton presentó un mínimo de 360 Cel.L^{-1} y un máximo de 29298 Cel.L^{-1}, con una media de 5292 Cel.L^{-1}. En comparación, en marea baja la densidad del fitoplancton tenía un mínimo de 760 Cel.L^{-1}y un máximo de 14414 Cel.L^{-1}, con una media de 4879 Cel.L^{-1}. Sin embargo, la prueba de Wilcoxon no encontró una diferencia significativa en la densidad del fitoplancton entre las dos mareas (W = 119, p = 0.181). La mayor distribución de la densidad se presentó en la boca del Sanquianga en ambas mareas.

Por otro lado, en marea alta el peso húmero del zooplancton colectado con la red de 500 µm presentó un mínimo de 17.56 g.m^{-3} y un máximo de 1913.18 g.m^{-3}, con una media de 281.21 g.m^{-3}. En comparación, en marea baja el peso húmero tenía un mínimo de 0.5078 g.m^{-3}y un máximo de 580.9613 g.m^{-3}, con una media de 63.9584 g.m^{-3}. La prueba de Wilcoxon encontró una diferencia significativa en el peso húmero del zooplancton colectado con la red de 500 µm entre las dos

mareas (W = 268, p< 0.01). El peso húmedo del zooplancton colectado con la red de 500 µm fue mayor sobre todo en el transecto de la bocana Sanquianga en marea alta cercana a la costa y en marea baja lejana a la costa en las condiciones más oceánicas (Figura 12) (Tabla 1).

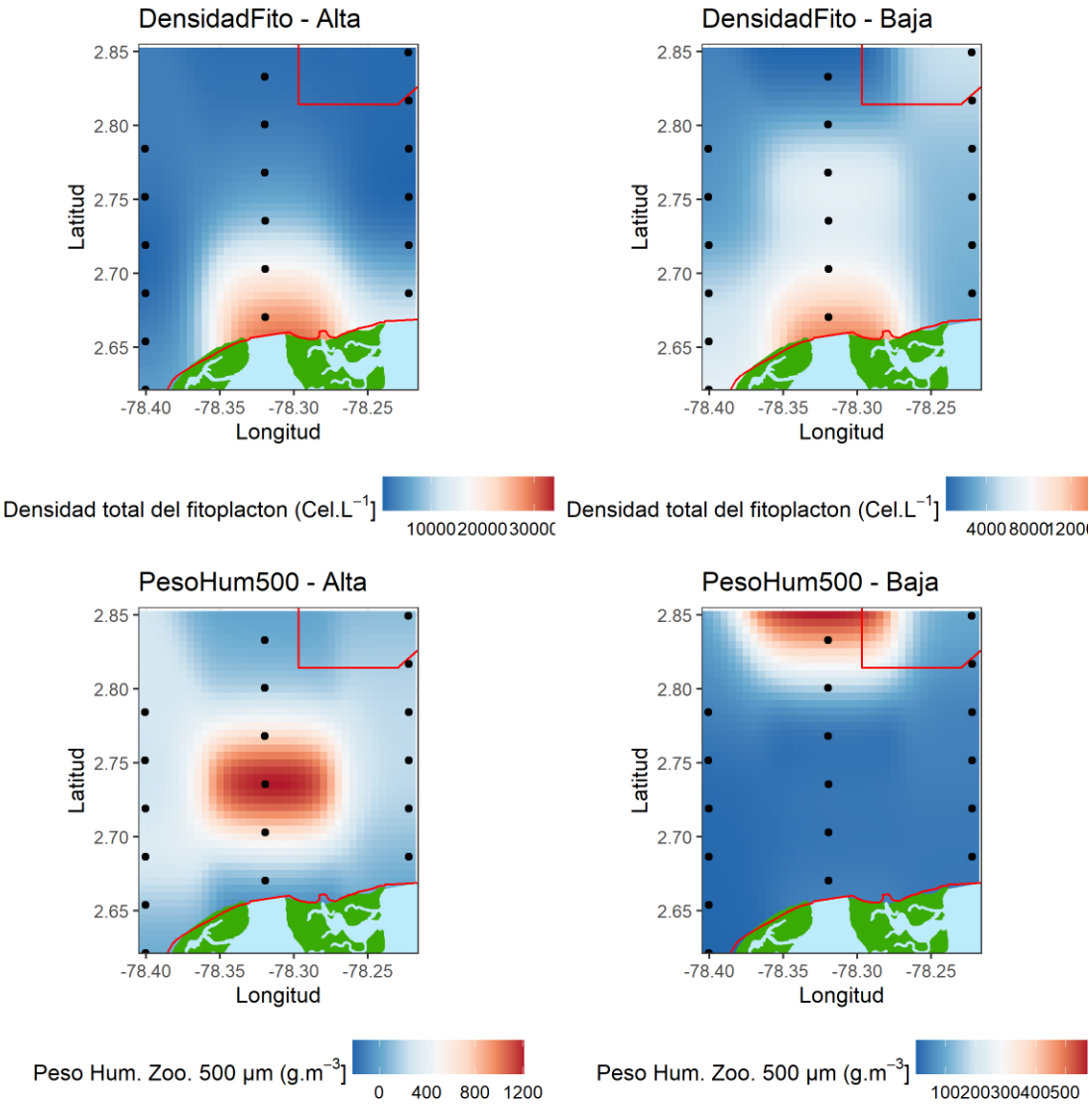

Figura 12. Mapas de la densidad de fitoplancton y el peso húmedo del zooplancton capturado con red de 500um en el área de estudio durante la marea alta y la marea baja.

Durante la marea alta, se registraron valores de Peso Húmero del zooplancton medido con la red de 300 µm que van desde 75.95 g.m^{-3} hasta 3070.97 g.m^{-3}, con una mediana de 518.93 g/m3 y una media de 680.68 g.m^{-3}. Por otro lado, la diversidad de orden cero q0 durante la marea alta fue de un mínimo de 8 especies y un máximo de 42 especies, con una mediana de 24.50 especies y una media de 24.00 especies. En contraste, durante la marea baja, los valores de Peso Húmero van desde 18.94 g.m^{-3}hasta 1772.80 g.m^{-3}, con una mediana de 689.79 g.m^{-3}y una media de 633.80 g.m^{-3}. La diversidad de orden cero q0 durante la marea baja fue de un mínimo de 12 especies y un máximo de 49 especies, con una mediana de 25.50 especies y una media de 27.50 especies. No se encontraron diferencias significativas entre ambos períodos mareales para el Peso Húmero del

zooplancton entre marea alta y marea baja (W=150, p =0.7193), ni tampoco en la diversidad de orden cero q0 (W=133, p =0.3666) (Figura 13) (Tabla 1).

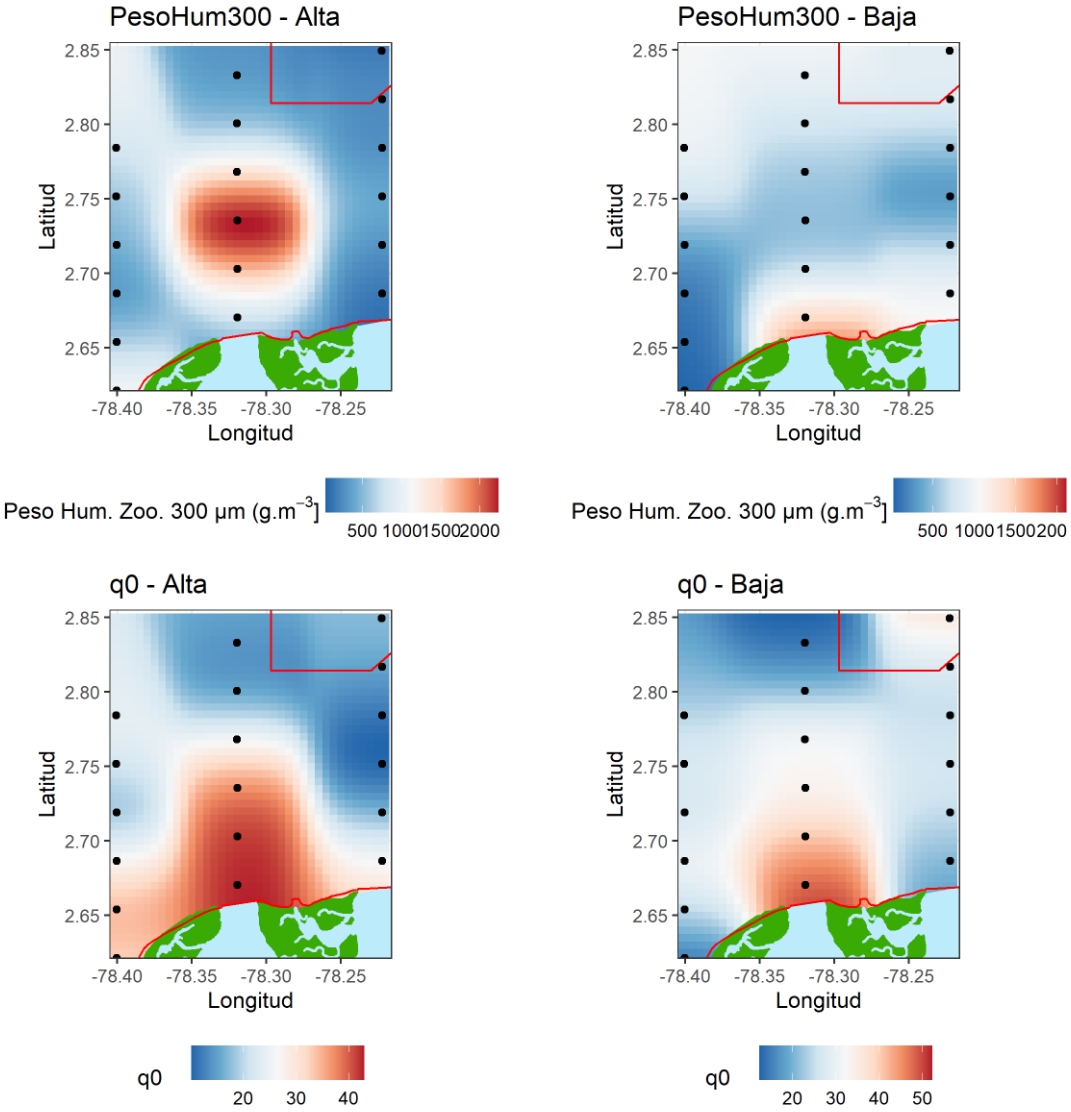

Figura 13. Mapas del peso húmedo del zooplancton capturado con red de 300um y la diversidad de fitoplancton de orden q = 0 medida con los números de Hill, en el área de estudio durante la marea alta y la marea baja.

Durante marea alta, se observó que la diversidad de orden uno (^{0}D q = 1) varió entre un mínimo de 2.495 y un máximo de 21.415, con una media de 9.674. Por otro lado, durante marea baja, los valores diversidad de orden uno osciló entre un mínimo de 2.301 y un máximo de 16.556, con una media de 9.694.

Se llevó a cabo una prueba de Wilcoxon para comparar los valores de q1 entre marea alta y marea baja, y se obtuvo un valor de W=150 y un p-valor de 0.7193, lo que indica que no hay una diferencia significativa entre los valores de q1 en ambas condiciones de marea (Figura 14) (Tabla 1).

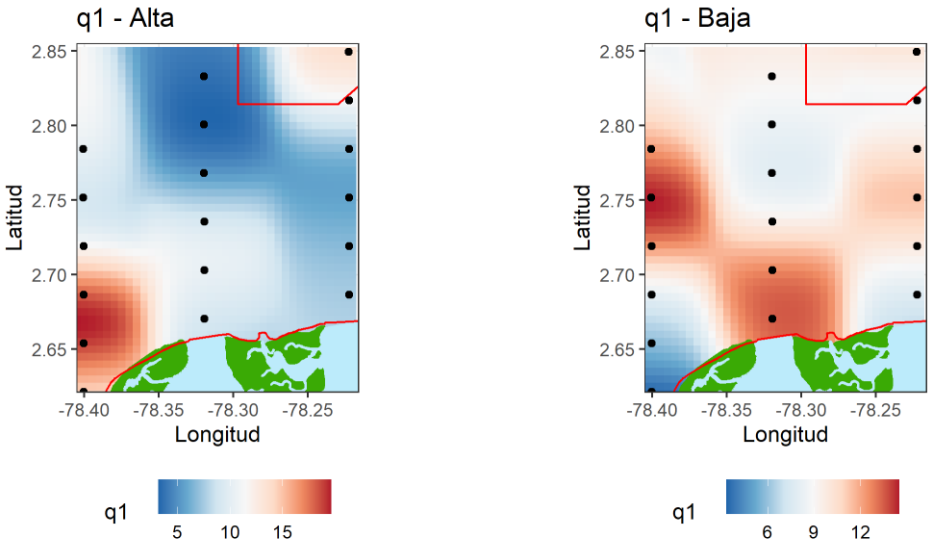

Figura 14. Mapa la diversidad de fitoplancton de orden q = 1 medida con los números de Hill, en el área de estudio durante la marea alta y la marea baja.

El análisis de correlación múltiple mostró que existen 12 correlaciones fuertes (-0.75 < r > 0.75) entre las variables estudiadas (Figura 15). En general, se observan correlaciones positivas significativas (correlaciones fuertes) entre los nutrientes nitritos, nitratos, fosfatos, silicatos y clorofila a. En particular, las correlaciones más fuertes se observaron entre la densidad y la salinidad (r=1.0), nitratos y silicatos (r = 0.93), el pH y el oxígeno disuelto (r=0.93), los nitritos y los fosfatos (r = 0.79), y los sólidos suspendidos totales y la clorofila (r = 0.76). Además, se observaron correlaciones negativas significativas (correlaciones débiles a fuerte) entre las variables ambientales, como los sólidos suspendidos y el oxígeno disuelto (r=-0.88) y los sólidos suspendidos y el pH (r=-0.85), el oxígeno disuelto y los fosfatos (r=-0.83) y los silicatos con el peso húmedo de zooplancton colectado con la red de 500 µm (r= -0.81) y la densidad de fitoplancton y el pH (r=-0.77).

La transparencia, la temperatura la salinidad, la densidad, el peso húmedo del zooplancton colectado con la red de 300 µm y la diversidad de orden uno fueron las únicas variables que no presentaron correlaciones fuertes (r = ± 0.75). Cabe destacar que se encontraron correlaciones moderadas positivas entre los nutrientes y la clorofila, y correlaciones negativas débiles a moderadas entre las variables físicas y los nutrientes y la clorofila (Tabla 1).

Tabla 1. Valores de la prueba del Wilcoxon para evaluar las diferencias de los valores de cada variable entre los períodos de marea alta y baja entre el 28 de abril y el 07 de mayo de 2021, en la subregión de Sanquianga -Gorgona.

Variable	W	Probabilidad
Nitritos [µM]	*67*	< 0.01
Nitratos [µM]	*96*	< 0.05
Fosfatos [µM]	*111*	0.109
Silicatos [µM]	*55*	< 0.01
Clorofila a [$\mu g.L^{-1}$]	*39*	< 0.01
pH	*207*	0.1537
Oxígeno disuelto [mg $O_2.L^{-1}$]	*252*	< 0.01
Transparencia (m)	*233*	< 0.05
Solidos Suspendidos Totales [$mg.L^{-1}$]	*110*	0.103
Temperatura (°C)	*139*	0.48
Salinidad (PSU)	*225*	< 0.05
Densidad del agua ($kg.m^{-3}$)	*228*	< 0.05
Densidad celular de Fitoplancton ($Cel.L^{-1}$)	*119*	0.18
Peso Húmedo del zooplancton 500 µm ($g.m^{-3}$)	*268*	< 0.01
Peso Húmedo del zooplancton 300 µm ($g.m^{-3}$)	*150*	0.7193
^0D	*133*	0.3666
^1D	*150*	0.7193

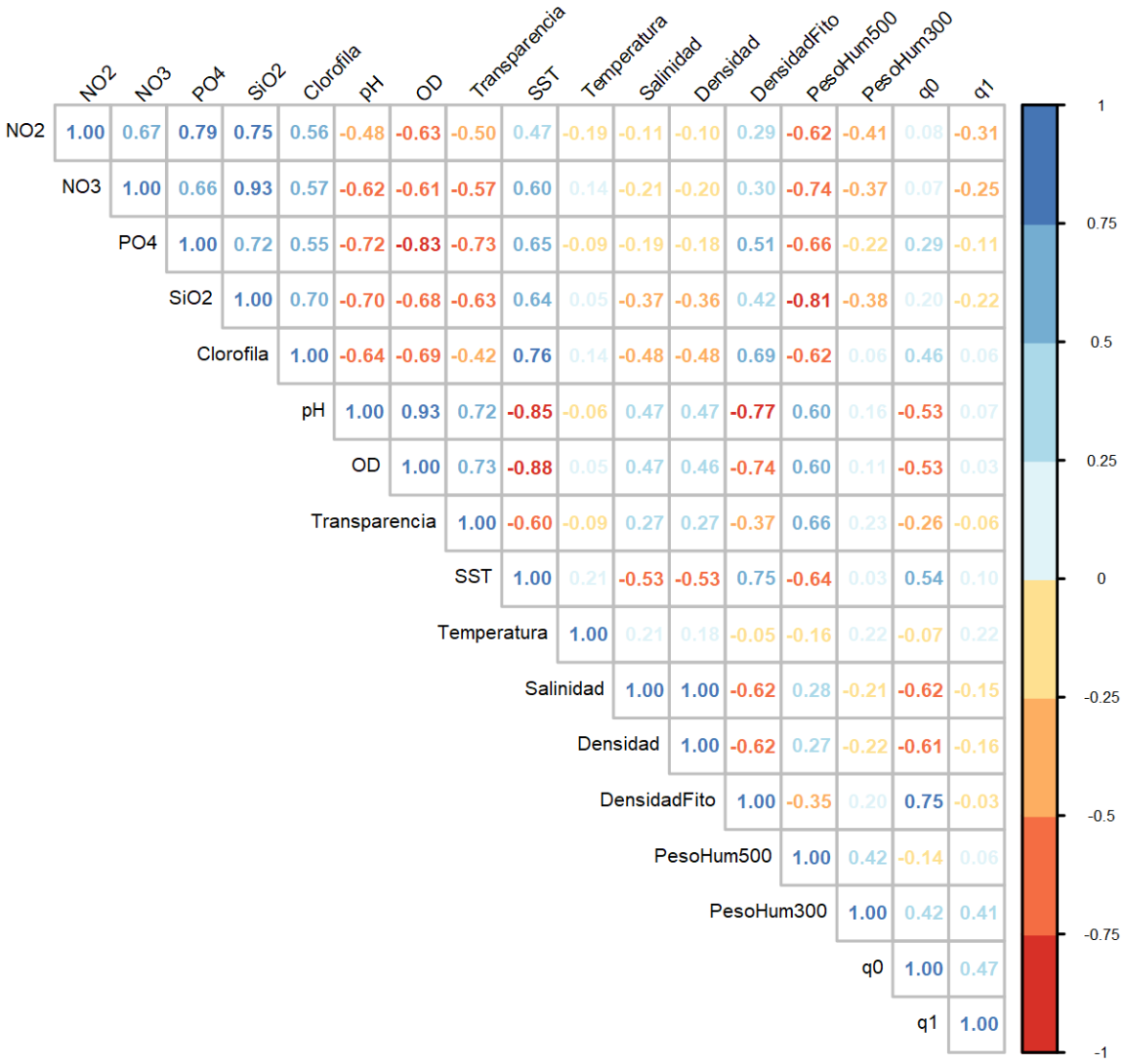

Figura 15. Matriz de correlación múltiple entre las variables medidas en la subregión Sanquianga-Gorgona.

En el análisis de componentes principales, se detectaron cuatro correlaciones fuertes fuertes (-0.75 < r > 0.75) dentro del primer componente, dos correlaciones positivas y dos correlaciones negativas (Figura 16). Las correlaciones positivas fuertes se presentaron con el oxígeno disuelto y el pH y las negativas se presentaron con los silicatos y los sólidos suspendidos totales. Para el lsegundo componente las mayores correlaciones se presentaron para las variables del peso húmedo del zooplancton medido con la red de 300 um y la diversidad de orden uno del fitoplancton, sin embargo, estas correlaciones fueron moderadas (r=0.75 y r=0.71 respectivamente).

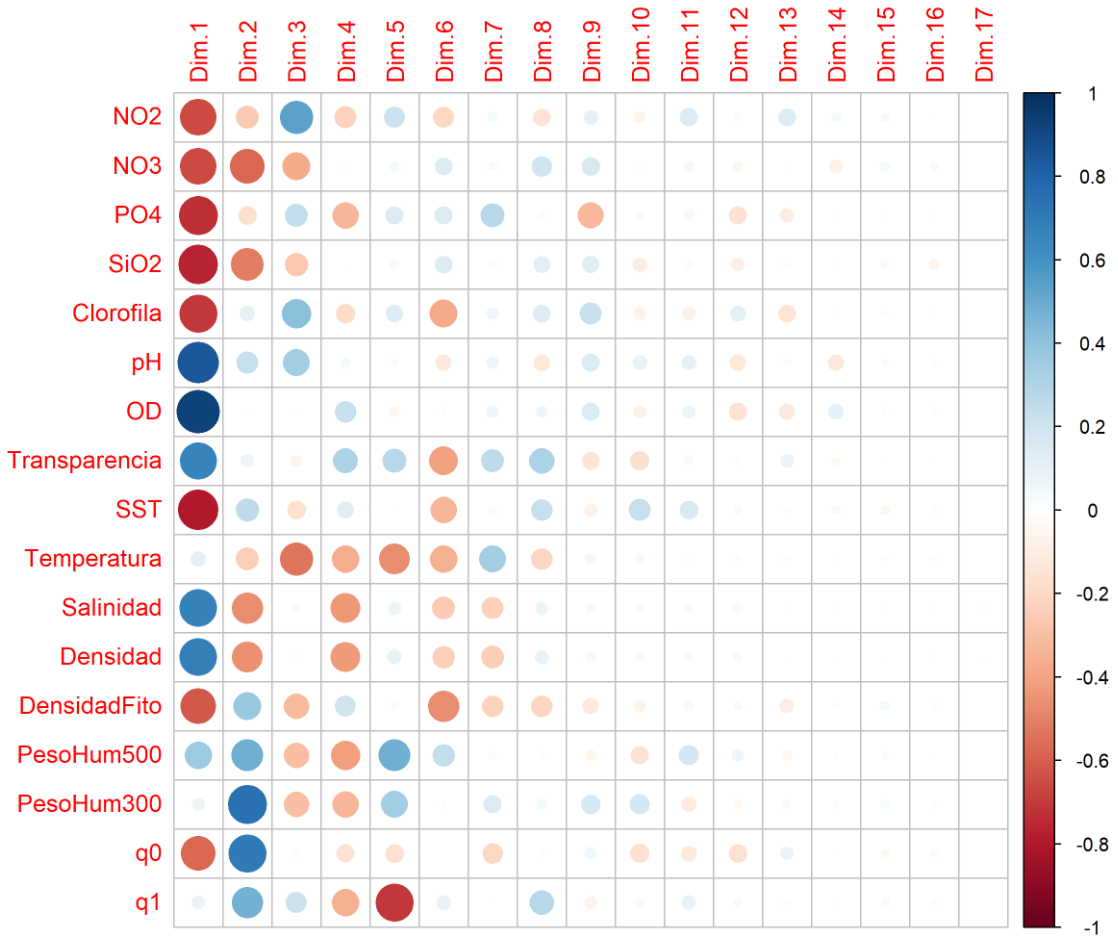

Figura 16. Correlaciones de las variables en el análisis de componentes principales por cada dimensión.

Los tres primeros componentes contribuyeron con un 67 % de la variabilidad total de información contenida en todas las variables de estudio (Figura 17), del primer componente se concluye que los mayores aportantes son las variables del oxígeno disuelto (12%), el pH (10%) y los sólidos suspendidos totales (9%) (Figura 17). Para el segundo componente, los mayores aportantes fueron: el peso húmedo del zooplancton (300 µm), la diversidad de orden cero del fitoplancton (16%) y los nitratos (11%). Para el componente tres, la temperatura y los nitritos tuvieron un aporte similar (19%) y la clorofila y los nitratos tuvieron un aporte de cerca del 10% del total del componente (Figura 17).

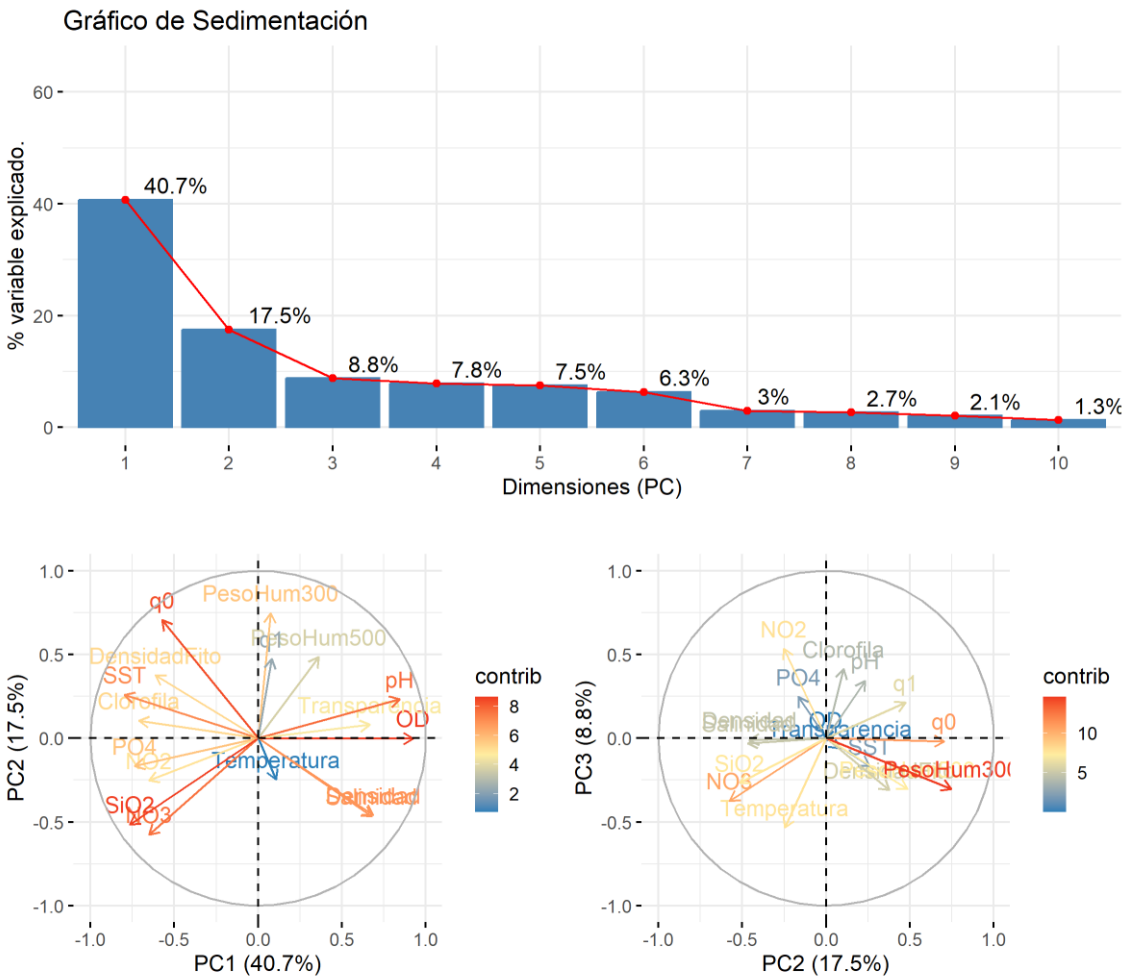

Figura 17. Gráfico de sedimentación (Scree plot) y de dispersión del análisis de componentes principales identificando las principales variables del estudio.

Construcción de los modelos de clasificación.

Al realizar los análisis de clasificación con los cuatro algoritmos seleccionados, se excluyeron algunas de las variables que tuvieron un alto índice de correlación para ser incluidos en el entrenamiento de los modelos a través de los cuatro algoritmos usados.

Las principales variables que respondieron con gran importancia para explicar los patrones resaltados por los modelos fueron la salinidad, la temperatura y los nitratos. Sin embargo, no hay consistencia entre todos los algoritmos frente a la respuesta de estas variables y la importancia varía entre los algoritmos.

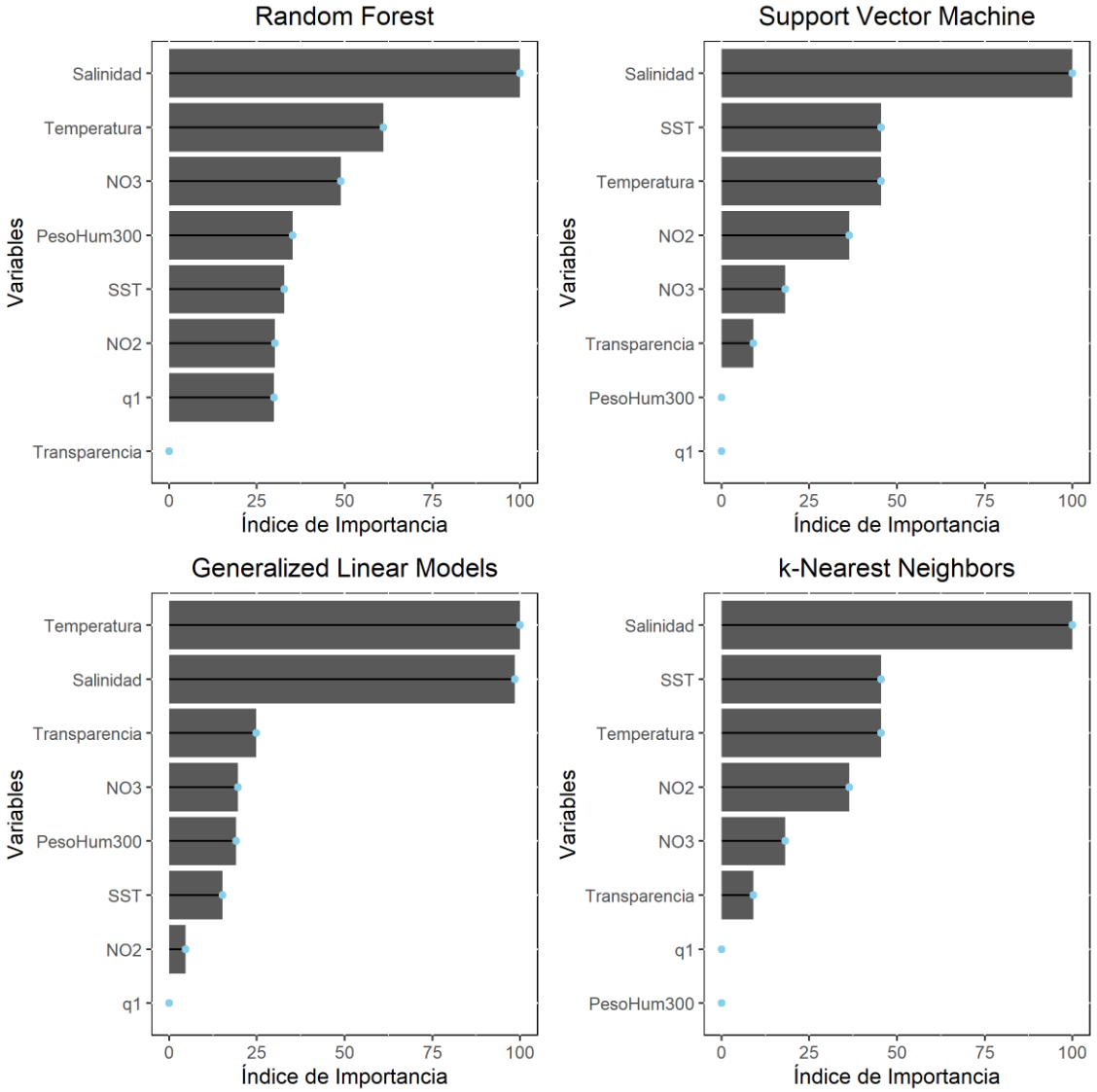

Figura 18. Índice de importancia de las variables usadas para el ajuste de los modelos clasificatorios usando los algoritmos de Random Forest, Support Vector Machine, Generalized Linear Models (regresión logística) y el k-Nearest Neighbors.

De los tres algoritmos usados, los de *Support Vector Machine* y el k-*Nearest Neighbors* no demostraron una respuesta de clasificación diferente al de un clasificador aleatorio ya que el área bajo la curva (AUC) no fue superior al 50% (Figura 19). El modelo general linealizado con la regresión logística presentó una precisión del 83% evaluado con la matriz de confusión y un AUC del 90%. Para el algoritmo de regresión logística (GLM) el área bajo fue de un 100% y la evaluación a través de la matriz de confusión dio como resultado una precisión de 100% (Figura 19).

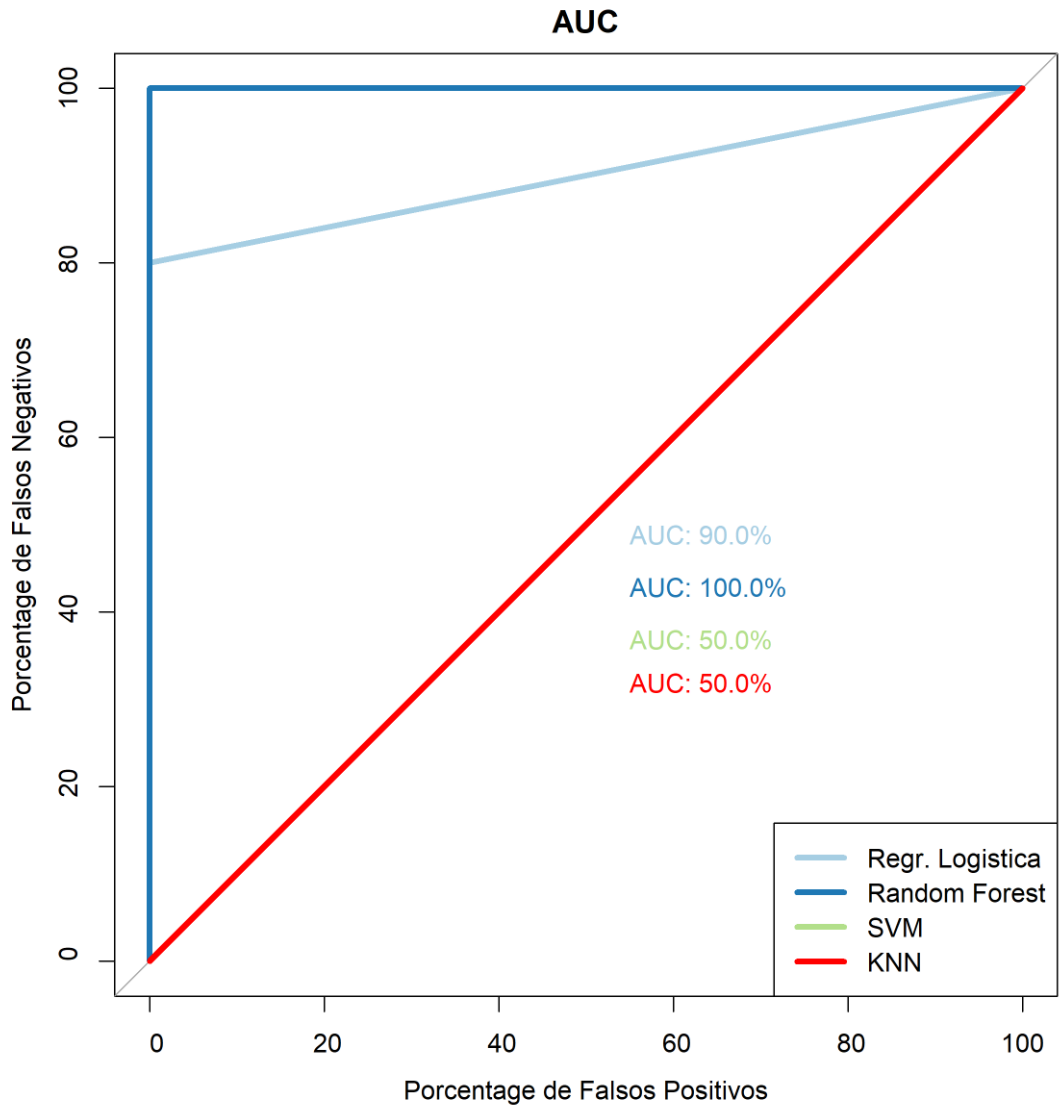

Figura 19. Gráficos ROC (*Receiver Operating Characteristic*), mostrando los valores del área bajo la curva para los cuatro algoritmos usados para los ajustes de los modelos.

Después de evaluar diferentes algoritmos, se encontró que tanto el Random Forest como la regresión logística presentaron el mejor ajuste y capacidad predictiva. Por lo tanto, se seleccionaron las variables más importantes para ajustar ambos algoritmos y evaluar su relación de manera individual, con el objetivo de determinar la significancia del ajuste.

Al realizar la evaluación con los tres primeros componentes de análisis del PCA, no se encontró ningún ajuste significativo utilizando la regresión logística (PC01: AIC: 22.04, p <0.1, PC02: AIC: 24.67, p >0.1, PC03: AIC: 25.37, p >0.1), para el algoritmo de *Random Forest* la tasa de error fuera

de la bolsa (OOB) estimada fue mayor del 20%. El error de clasificación se calculó utilizando una matriz de confusión, que muestra que se clasificaron correctamente 16 de las 18 observaciones. El error de clasificación fue del 10.33% para la clase 1 (Factor = 1) y del 75% para la clase 0 (Factor = 0) para el componente uno que tuvo las mejores métricas. El AUC para la regresión logística y para el *Random forest* estuvieron por debajo de 80%.

Al evaluar la temperatura superficial individualmente, no se encontró un ajuste significativo utilizando la regresión logística (AIC: 17.96, p >0.1), para el algoritmo de *Random Forest* la tasa de error fuera de la bolsa (OOB) estimada fue del 16.67%. El error de clasificación se calculó utilizando una matriz de confusión, que muestra que se clasificaron correctamente 15 de las 18 observaciones. El error de clasificación fue del 13.33% para la clase 1 (Factor = 1) y del 33.33% para la clase 0 (Factor = 0). El AUC para la regresión logística y para el *Random forest* estuvieron por debajo de 80% (Figura 20).

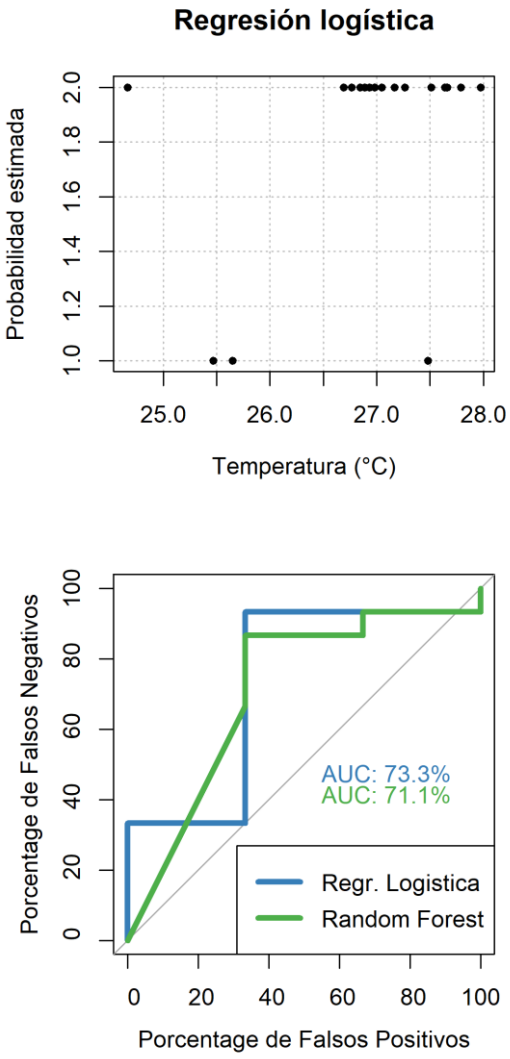

Figura 20. Ajuste de la temperatura superficial del mar con el algoritmo de regresión logística y de *Random forest*, mostrando los valores calculados de AUC.

Comisión Colombiana del Océano · Construyendo País Marítimo
www.cco.go
Bogotá D.C., 4 de mayo
Avenida Ciudad de Cali No. 51-66

Con la evaluación individual de la salinidad, el algoritmo de regresión logística tuvo un ajuste con una significancia mayor a p > 0.05 pero menor a p< 0.1, lo que indica una respuesta de la incidencia de *Vibrio* spp. con la salinidad. Para el algoritmo de *Random Forest*, La tasa de error OOB estimada fue del 22.22%, lo que indica que el modelo clasificó incorrectamente alrededor de una quinta parte de los datos. La matriz de confusión muestra que se clasificaron correctamente 14 de las 18 observaciones. El error de clasificación fue del 13.33% para la presencia de *Vibrio spp.* (Factor = 1) y del 66.67% para la ausencia (Factor = 0), lo que indica que el modelo tuvo más dificultades para clasificar correctamente la ausencia.

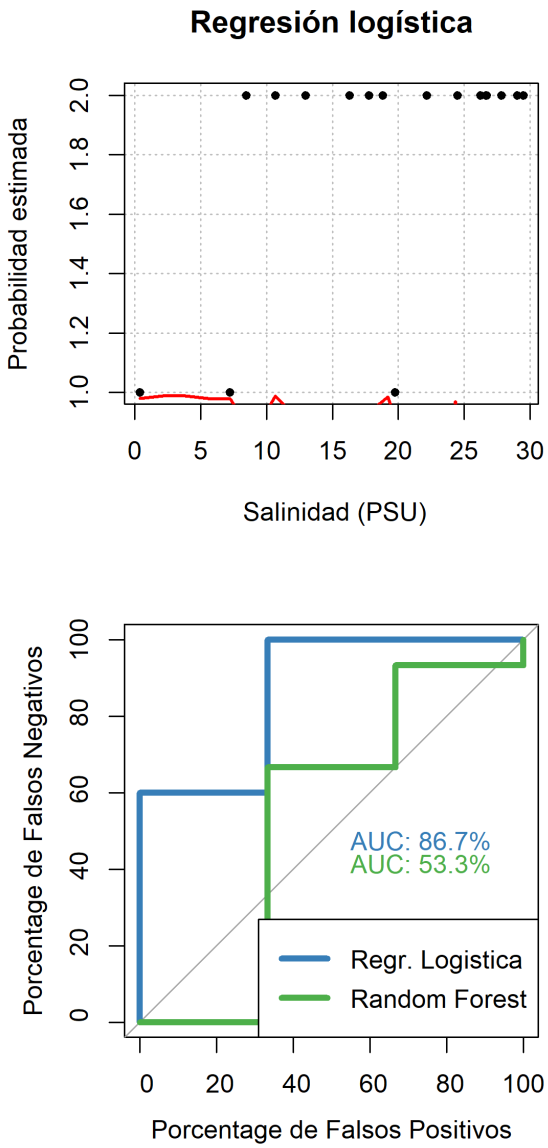

Figura 21. Ajuste de la salinidad superficial del mar con el algoritmo de regresión logística y de *Random forest*, mostrando los valores calculados de AUC.

A partir de la predicción de encuentro con los ajustes de los modelos tanto de *Random Forest* como la regresión logística (Figura 22), se realizó una predicción espacial con el modelo entrenado con las 8 principales variables. En marea alta, la probabilidad de detección de presencia de *Vibrio* spp. estuvo por encima del 80 % para la regresión logística y del 60% para el *Random Forest*. En la marea baja se presentaron las menores probabilidades de detección cerca a la bocana del Sanquianga (Figura 22).

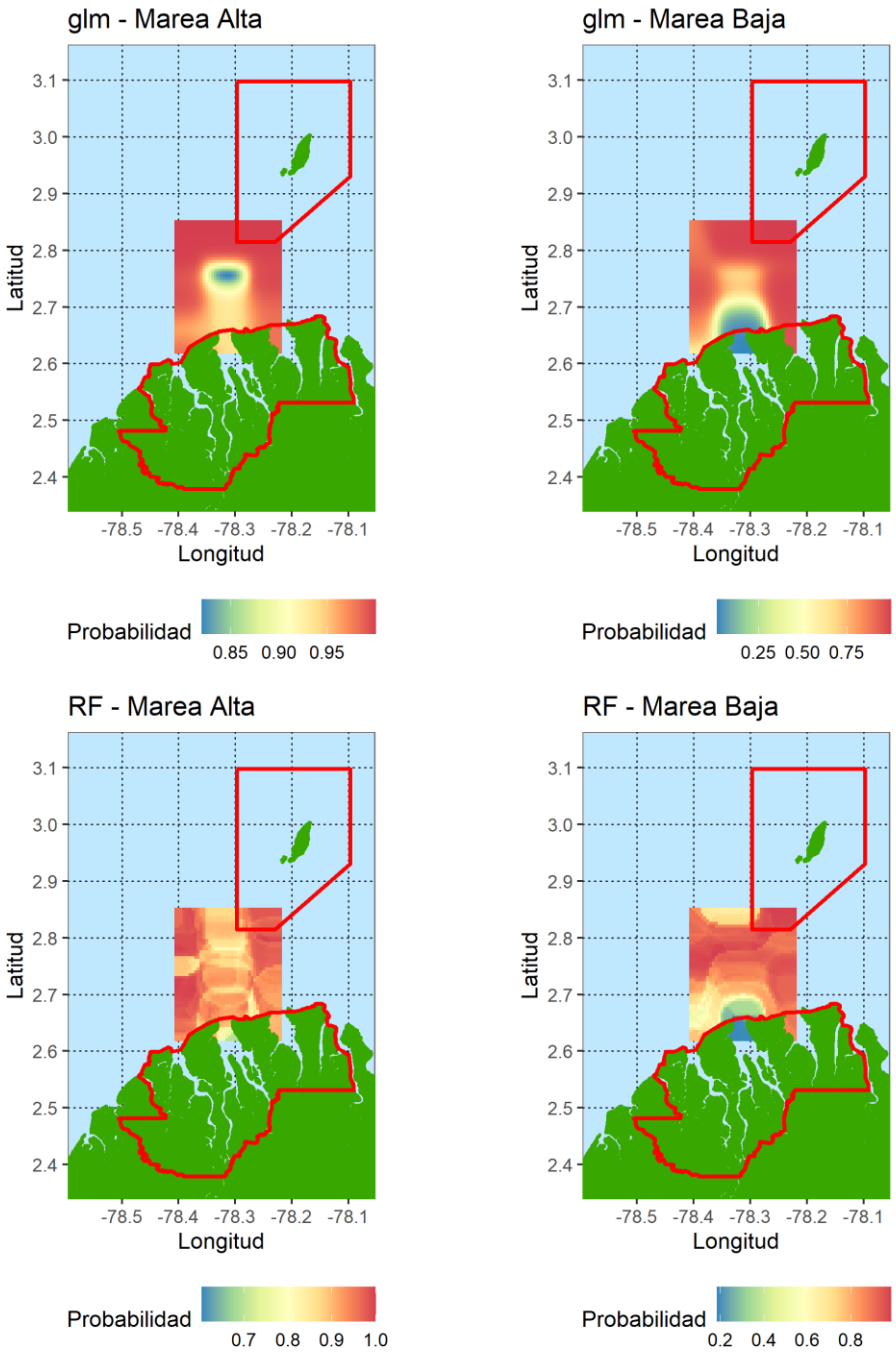

Figura 22. Predicción basada en la probabilidad de encuentro con el ajuste del modelo basados en la regresión logística y *Random Forest*.

La predicción basada en la incidencia para ambos algoritmos, arrojó un escenario en marea alta donde la presencia de *Vibrio* spp. en la zona cubre todo el espacio. Para la marea baja, el escenario predicho por ambos escenarios muestra una presencia en toda la zona exceptuando la bocana del Sanquianga, donde la predicción marca ausencia de *Vibrio* spp (Figura 23).

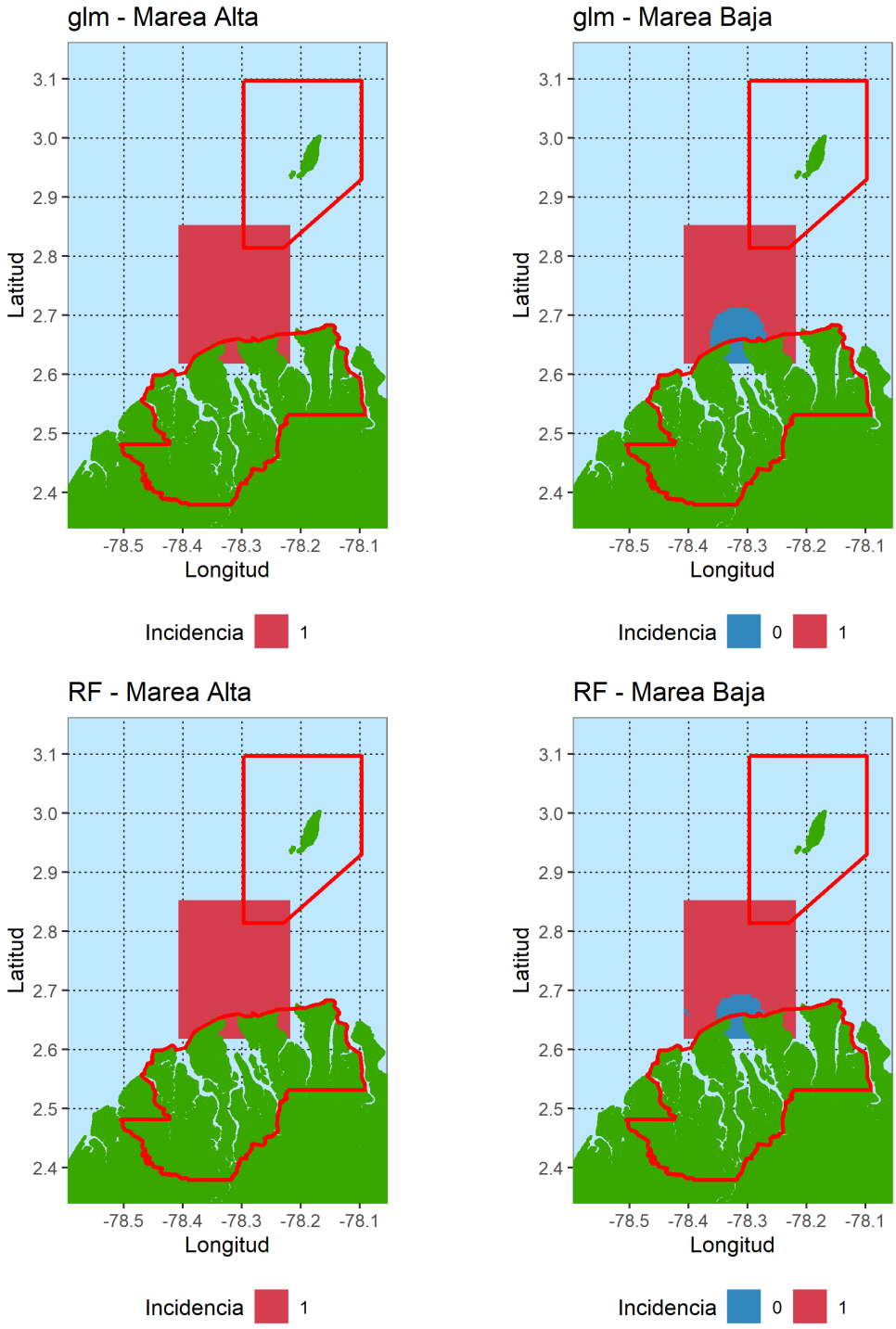

Figura 23. Predicción basada en la incidencia de encuentro con el ajuste del modelo basados en la regresión logística y *Random Forest*.

7 Discusión

Aunque se planeó inicialmente tomar 36 muestras para capturar el comportamiento de la incidencia durante la marea alta y baja, la falta de asepsia en el laboratorio del buque "ARC PROVIDENCIA" impidió obtener la lectura final de todas las muestras, lo que resultó en una disminución en el número de casos disponibles para el modelamiento. No obstante, se encontró una alta presencia de Vibrio spp. en todas las muestras revisadas, cercana al 80%.

Se evaluaron un total de 17 variables hidrográficas, de las cuales la mitad no mostraron diferencias significativas entre los períodos mareales. Sin embargo, se observaron diferencias significativas en la concentración de nutrientes nitrogenados y silicatos, mientras que los fosfatos no presentaron diferencias. En cuanto a las variables relacionadas con el Fitoplancton, ninguna presentó diferencias significativas, excepto la concentración de clorofila, que presentó correlaciones positivas y moderadas con los nutrientes en general. Además, la concentración de oxígeno disuelto estuvo fuertemente relacionada con la concentración de fosfatos y presentó diferencias significativas en los cambios mareales. La transparencia, densidad del agua y salinidad también mostraron diferencias significativas en los cambios mareales, y las dos últimas presentaron una correlación muy fuerte, ya que la medida de densidad depende totalmente de la medida de conductividad que se obtiene con la sonda CTDO utilizada para calcular la salinidad.

A partir del análisis de componentes principales, se puede concluir que las variables que más contribuyen a la variabilidad del sistema son el oxígeno disuelto, el pH y los sólidos suspendidos totales. En general, el oxígeno disuelto presentó correlaciones negativas moderadas y fuertes con todos los nutrientes y las condiciones físicas, así como con algunas variables biológicas como la clorofila, el peso húmedo del zooplancton medido con la red de 500 um y la densidad celular del fitoplancton. Cabe destacar que, al analizar la relación entre la incidencia de Vibrio y los tres componentes del análisis, no se encontró ninguna relación significativa.

Son pocos los estudios que han abordado la relación entre las condiciones de marea y la presencia y distribución de *Vibrio* spp. Intentos por relacionar las condiciones ambientales con la abundancia y distribución de especies de *Vibrio* han sido numerosos (Blackwell & Oliver, 2008; Colwell, 1996; Grimes et al., 2009; Huq et al., 2013; Johnson et al., 2010; Kelly, 1982; Lobitz et al., 2000; Stauder et al., 2010; Zimmerman et al., 2007) con algún éxito en cuanto a la demostración de estas relaciones con especies específicas como *V.cholerae* y la abundancia de los copépodos (Colwell, 1996). En un estudio previo, Heidelberg et al. (2002) encontraron una correlación positiva significativa entre la altura de la marea y la abundancia de *Vibrio vulnificus* y *Vibrio cholerae-mimicus*. Específicamente, mediante un modelo de regresión lineal, los autores observaron un aumento en el coeficiente de determinación ajustado (R^2) para la altura de la marea en relación con la abundancia de estas especies bacterianas. En nuestro trabajo, debido al desbalance de datos, no se pudo evaluar esta relación entre la presencia de *Vibrio* spp. y los datos de marea mediante modelos soportados de relaciones categóricas bidireccionales. No obstante, se recomienda considerar esta aproximación en futuros estudios que exploren la relación entre estas variables.

Los resultados del ajuste de los modelos por medio de los algoritmos de clasificación supervisada utilizados en este trabajo, arrojaron resultados que concuerdan con otros estudios que han explorado estas relaciones. Las variables que mas respondieron al ajuste de los modelos fueron la salinidad y la temperatura coincidiendo con lo que Takemura et al. (2014) encontraron en su trabajo de meta-análisis. Estos autores se enfocaron en estudiar las asociaciones físicas de *Vibrio*

a diferentes escalas ambientales y taxonómicas, encontrando que, para la abundancia, la temperatura y la salinidad tienen el mayor poder explicativo. Si bien en este trabajo no se consideró la abundancia, estas variables respondieron como buenos predictores para examinar la incidencia de este género de bacteria.

Con respecto a la aproximación para lo construcción de los modelos utilizando algoritmos de clasificación supervisada, hasta el momento son muy poco los estudios que han usado este tipo de algoritmos (Escobar et al., 2015), los cuales pueden ser una muy buena alternativa para estudiar las complejas relaciones entre los diferentes escalas de estudio. Por ejemplo, lo incipiente respuesta que se obtuvo en este trabajo con respeto a la explicación de la incidencia de *Vibrio* spp. puede deberse al hecho que solo se está considerando el género. Esto podría estar enmascarando una relación mas estrecha entre las necesidades de cada especie y las condiciones ambientales (Johnson et al., 2012). Escobar et al. (2015) encontraron una relación muy estrecha entre la clorofila a y la presencia de *V. cholerae* a nivel global, utilizando un algoritmo basado en la Maximización de la entropía (Maxent) en el espacio geográfico (Elith et al., 2011), sin embargo, hay que tener en cuenta que en este trabajo, dicho algoritmo fue entrenado en una escala de variación muy amplia para los predictores o variables ambientales. Una de las dificultades de usar este tipo de algoritmos en trabajos tan localizados como el presente, es que los algoritmos no tienen la oportunidad de ser expuestos a la variabilidad que pueden tomar las variables en un rango de escala regional y por esta razón se pueden enmascarar las respuestas. Una forma de evitar esto, es tener muchos más datos para entrenar los algoritmos con un mayor rango de variabilidad y así obtener mejores posibilidades de predicción.

El entrenamiento de los algoritmos permitió la predicción espacial de la incidencia en el área de estudio, lo que se evidenció con los mejores ajustes de los algoritmos de regresión lineal y de *Random Forest* fue que para la marea alta, las condiciones de temperatura y salinidad son las más optimas para que haya presencia en toda el área de *Vibrio*, pero en la marea baja, sobre todo en las bocas de los ríos donde la salinidad es baja, se predice que puede a ver ausencia.

Por el momento, no todos los algoritmos han tenido la capacidad de encontrar una relación entre las variables consideradas y la incidencia de *Vibrio* spp. esto solo se logrará continuando con la adición de datos en otros escenarios dentro de la costa colombiana donde se evalúen estas mismas relaciones. Las enfermedades producidas por las especies del género *Vibrio*, como la gastroenteritis (*V. parahaemolyticus*), la infección de las heridas (*V. vulnificus*) y el cólera (*V. cholerae*), han sido prevalentes en la costa tropical del pacífico suramericano, sobre todo de esta última, desde su reaparición en la costa de Perú en la década de los 90's asociada con condiciones ambientales cálidas del ENOS ("El Niño" Oscilación Sur) (Colwell, 1996; Gabastou et al., 2002). Esta es una de las razones por las cuáles es importante estudiar las relaciones entre las condiciones ambientales del agua de mar y la incidencia de este género, para desarrollar un entendimiento de su ecología y poder establecer sistemas de alerta temprana basados en este conocimiento (Brumfield et al., 2021).

8 Conclusiones

Si bien no se lograron obtener todos los datos planeados, el uso de estos permitió lograr ajustar algunos modelos de manera aceptable con los cuales se pudo hacer algunas predicciones para la incidencia de *Vibrio* spp. en marea alta y baja. Se encontró que no todos los algoritmos de clasificación usados respondieron a los datos obtenidos y que los algoritmos de *Support vector machine* y K-*Nearest neighbour* no son adecuado cuando no se cuenta con muchos datos. LA regresión logística y el *Random forest* fueron mas sensibles a la clasificación con pocos datos como casos de entrenamiento. Sin embargo, se debe ser precavido con las predicciones debido a que estas se realizaron bajo un escenario bastante restringido con unas condiciones de variabilidad para todas las variables condicionadas muy localmente. Este tipo de trabajo puede ser usado como base para iniciar el levantamiento de un conjunto de datos mucho grande, que abarque una variabilidad mayor para la temperatura y la salinidad y con una resolución taxonómica mayor, para relacionar mejor cada especie dentro del género *Vibrio* a sus condiciones restrictivas.

9 Productos generados

Repositorio de análisis de los datos de este estudio:

https://github.com/ChrisBermudezR/Vibrio_ExpPacifico2021

10 Literatura citada

Baker-Austin, C., Trinanes, J. A., Taylor, N. G. H., Hartnell, R., Siitonen, A., & Martinez-Urtaza, J. (2013). Emerging Vibrio risk at high latitudes in response to ocean warming. *Nature Climate Change*, *3*(1), 73-77. https://doi.org/10.1038/nclimate1628

Balech, E. (Enrique). (1988). *Los dinoflagelados del Atlántico sudoccidental*. Ministerio de Agricultura, Pesca y Alimentación; Instituto Español de Oceanografía. http://hdl.handle.net/10508/993

Bendschneider, K., & Robinson, R. J. (1952). *A new spectrophotometric method for the determination of nitrite in sea water* (Technical Report No. 8; Project NR 083 012, p. 18). Office of Naval Research.

Blackwell, K. D., & Oliver, J. D. (2008). The ecology of Vibrio vulnificus, Vibrio cholerae, and Vibrio parahaemolyticus in North Carolina Estuaries. *The Journal of Microbiology*, *46*(2), 146-153. https://doi.org/10.1007/s12275-007-0216-2

Brumfield, K. D., Usmani, M., Chen, K. M., Gangwar, M., Jutla, A. S., Huq, A., & Colwell, R. R. (2021). Environmental parameters associated with incidence and transmission of pathogenic *Vibrio spp* . *Environmental Microbiology*, *23*(12), 7314-7340. https://doi.org/10.1111/1462-2920.15716

Ceccarelli, D., & Colwell, R. R. (2014). Vibrio ecology, pathogenesis, and evolution. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00256

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45-67. https://doi.org/10.1890/13-0133.1

Colwell, R. R. (1996). Global Climate and Infectious Disease: The Cholera Paradigm. *Science*, *274*(5295), 2025-2031. https://doi.org/10.1126/science.274.5295.2025

Córdoba Meza, T., Espinosa Díaz, L. F., & Vivas Aguas, L. J. (2021). Ocurrencia Y Distribución De Vibrio cholerae Cultivable En La Ciénaga Grande De Santa Marta, Caribe Colombiano. *Acta Biológica Colombiana*, *27*(2). https://doi.org/10.15446/abc.v27n2.92057

Cupp, E. E. (1943). *Marine Planktonic Diatoms of the West Coast of North America Bulletin of the Scripps Institution of Oceanography* (Vol. 5). University of California Press.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists: Statistical explanation of MaxEnt. *Diversity and Distributions*, *17*(1), 43-57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015). A global map of suitability for coastal Vibrio cholerae under current and future climate conditions. *Acta Tropica*, *149*, 202-211. https://doi.org/10.1016/j.actatropica.2015.05.028

Gabastou, J.-M., Pesantes, C., Escalante, S., Narváez, Y., Vela, E., García, L., Zabala, D., & Yadon, Z. E. (2002). Características de la epidemia de cólera de 1998 en Ecuador, durante el fenómeno de «El Niño». *Revista Panamericana de Salud Pública*, *12*(3), 157-164. https://doi.org/10.1590/S1020-49892002000900003

Grimes, D. J., Johnson, C. N., Dillon, K. S., Flowers, A. R., Noriea, N. F., & Berutti, T. (2009). What Genomic Sequence Information Has Revealed About Vibrio Ecology in the Ocean—A Review. *Microbial Ecology*, *58*(3), 447-460. https://doi.org/10.1007/s00248-009-9578-9

Guiry, M. D., & Guiry, G. M. (2023). *AlgaeBase. World-wide electronic publication*. National University of Ireland, Galway. https://www.algaebase.org

Heidelberg, J. F., Heidelberg, K. B., & Colwell, R. R. (2002). Seasonality of Chesapeake Bay Bacterioplankton Species. *Applied and Environmental Microbiology*, *68*(11), 5488-5497. https://doi.org/10.1128/AEM.68.11.5488-5497.2002

Herrera, S. E. L., Díaz, N. O. F., & Lozano, J. D. T. (2023). *Métodos numéricos con aplicación a la ingeniería-2da edición*. Ecoe Ediciones.

Huq, A., Hasan, N., Akanda, A., Whitcombe, E., Colwell, R., Haley, B., Alam, M., Jutla, A., & Sack, R. B. (2013). Environmental Factors Influencing Epidemic Cholera. *The American Journal of Tropical Medicine and Hygiene*, *89*(3), 597-607. https://doi.org/10.4269/ajtmh.12-0721

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer New York. https://books.google.com.co/books?id=at1bmAEACAAJ

Johnson, C. N., Bowers, J. C., Griffitt, K. J., Molina, V., Clostio, R. W., Pei, S., Laws, E., Paranjpye, R. N., Strom, M. S., Chen, A., Hasan, N. A., Huq, A., Noriea, N. F., Grimes, D. J., & Colwell, R. R. (2012). Ecology of Vibrio parahaemolyticus and Vibrio vulnificus in the Coastal and Estuarine Waters of Louisiana, Maryland, Mississippi, and Washington (United States). *Applied and Environmental Microbiology*, *78*(20), 7249-7257. https://doi.org/10.1128/AEM.01296-12

Johnson, C. N., Flowers, A. R., Noriea, N. F., Zimmerman, A. M., Bowers, J. C., DePaola, A., & Grimes, D. J. (2010). Relationships between Environmental Factors and Pathogenic Vibrios in the Northern Gulf of Mexico. *Applied and Environmental Microbiology*, *76*(21), 7076-7084. https://doi.org/10.1128/AEM.00697-10

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses* (version 1.0.7) [R package]. https://CRAN.R-project.org/package=factoextra

Kelley, D., & Richards, C. (2022). *oce: Analysis of Oceanographic Data* (version 1.7-10) [R package]. https://CRAN.R-project.org/package=oce

Kelly, M. T. (1982). Effect of temperature and salinity on Vibrio (Beneckea) vulnificus occurrence in a Gulf Coast environment. *Applied and Environmental Microbiology*, *44*(4), 820-824. https://doi.org/10.1128/aem.44.4.820-824.1982

Koch, S. E., desJardins, M., & Kocin, P. J. (1983). An interactive Barnes objective map analysis scheme for use with satellite and conventional data. *J. Clim. Appl. Meteorol.*, *22*(9), 1487-1503.

Lipps, W. C., Braun-Howland, E. B., & Baxter, T. E. (2023). *Standard Methods for the Examination of Water and Wastewater*. APHA Press.

Lobitz, B., Beck, L., Huq, A., Wood, B., Fuchs, G., Faruque, A. S. G., & Colwell, R. (2000). Climate and infectious disease: Use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Proceedings of the National Academy of Sciences*, *97*(4), 1438-1443. https://doi.org/10.1073/pnas.97.4.1438

Morales-Pulido, J. M., & Aké-Castillo, J. A. (2019). Coscinodiscus y Coscinodiscopsis (Bacillariophyceae) del Parque Nacional Sistema Arrecifal Veracruzano, golfo de México. *Revista mexicana de biodiversidad*, *90*.

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M. D., Durand, S., … Weedon, J. (2022). *vegan: Community Ecology Package* (version 2.6-2) [R package]. https://CRAN.R-project.org/package=vegan

Oliver, J., & Oliver, K. (2007). Vibrio Species. En *Food Microbiology: Fundamentals and Frontiers* (3rd ed). ASM Press.

R Core Team. (2022). *stats: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). *PROC: an open-source package for R and S+ to analyze and compare ROC curves*.

Rosenberg, E., & Falkovitz, L. (2004). The Vibrio shiloi/Oculina patagonica model system of coral bleaching. *Annu. Rev. Microbiol.*, *58*, 143-159.

Stauder, M., Vezzulli, L., Pezzati, E., Repetto, B., & Pruzzo, C. (2010). Temperature affects Vibrio cholerae O1 El Tor persistence in the aquatic environment via an enhanced expression of GbpA and MSHA adhesins: Temperature affects V. cholerae in the aquatic environment. *Environmental Microbiology Reports*, *2*(1), 140-144. https://doi.org/10.1111/j.1758-2229.2009.00121.x

Strickland, J. D. H., & Parsons, T. R. (1972). *A practical handbook of seawater analysis*. Fisheries research board of Canada.

Takemura, A. F., Chien, D. M., & Polz, M. F. (2014). Associations and dynamics of Vibrionaceae in the environment, from the genus to the population level. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00038

Thompson, F. L., Austin, B., Swings, J. G., & American Society for Microbiology (Eds.). (2006). *The biology of vibrios*. ASM Press.

Tomas, C. R., Hasle, G. R., Syvertsen, E. E., Steidinger, K. A., & Tangen, K. (2010). *Identifying Marine Diatoms and Dinoflagellates* (C. R. Tomas, Ed.). Academic Press.

Turner, J. W., Good, B., Cole, D., & Lipp, E. K. (2009). Plankton composition and environmental factors contribute to Vibrio seasonality. *The ISME Journal*, *3*(9), 1082-1092. https://doi.org/10.1038/ismej.2009.50

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the

tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wong, Y. Y., Lee, C. W., Bong, C. W., Lim, J. H., Narayanan, K., & Sim, E. U. H. (2019). Environmental control of Vibrio spp. Abundance and community structure in tropical waters. *FEMS Microbiology Ecology*, *95*(11), fiz176. https://doi.org/10.1093/femsec/fiz176

Zhou, Z.-H. (2021). *Machine learning*. Springer Nature.

Zimmerman, A. M., DePaola, A., Bowers, J. C., Krantz, J. A., Nordstrom, J. L., Johnson, C. N., & Grimes, D. J. (2007). Variability of Total and Pathogenic *Vibrio parahaemolyticus* Densities in Northern Gulf of Mexico Water and Oysters. *Applied and Environmental Microbiology*, *73*(23), 7589-7596. https://doi.org/10.1128/AEM.01700-07