UC | Facultad de Ciencias

Técnicas de inteligencia artificial aplicadas a la asimilación de datos de oleaje

(Artificial intelligence techniques applied to the assimilation of wave data)

Trabajo de Fin de Máster

para acceder al

MÁSTER EN DATA SCIENCE

Autor: Alejandro González Valle

Director\es: Ezequiel Cimadevilla, Antonio Tomás Sampedro

Septiembre – 2024

RESUMEN

El presente trabajo de fin de máster aborda la aplicación de técnicas de inteligencia artificial para la asimilación de datos de oleaje, un problema recurrente en la ingeniería portuaria.

Se han implementado y comparado diversos modelos, de regresión (lineal, no lineal y direccional no lineal), redes neuronales y un modelo de agrupamiento K-Means, con el objetivo de calibrar y mejorar la precisión de los datos de oleaje de modelos numéricos a partir de datos de oleaje instrumentales.

Los resultados obtenidos demostraron que los modelos seleccionados ofrecen mejoras significativas en la reducción del error y la dispersión de los datos, destacándose el modelo K-Means como el más efectivo. La investigación también resalta la importancia de incluir el parámetro de dirección del oleaje para capturar la variabilidad específica de éste.

Este trabajo no sólo confirma la utilidad de las técnicas de inteligencia artificial en la asimilación de datos de oleaje, sino que también abre la puerta a futuras investigaciones para continuar optimizando e incorporando modelos e integrar nuevas variables a ellos.

Palabras clave: asimilación de datos, oleaje, modelo de regresión, calibración direccional, K-Means.

ABSTRACT

This master's thesis addresses the application of artificial intelligence techniques for wave data assimilation, a recurrent problem in port engineering. Various models, including regression models (linear, nonlinear, and directional nonlinear), neural networks, and a K-Means clustering model, were implemented and compared to calibrate and improve the accuracy of wave data from numerical models using instrumental wave data.

The results demonstrated that the selected models offer significant improvements in reducing error and data dispersion, with the K-Means model standing out as the most effective. The research also highlights the importance of including the wave direction parameter to capture its specific variability.

This work not only confirms the utility of artificial intelligence techniques in wave data assimilation but also establishes a basis for future research to continue optimizing and incorporating models and integrating new variables into them.

Keywords: data assimilation, wave, regression model, directional calibration, K-Means.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

El oleaje es un fenómeno natural que se manifiesta como una serie de ondas o movimientos ondulatorios en la superficie del agua, principalmente generado por la acción del viento sobre ella. Estas ondas se desplazan horizontalmente, y su tamaño y forma pueden variar dependiendo de la velocidad, duración y distancia sobre la cual ejerce la fuerza el viento. Se caracteriza principalmente por la altura de la ola, el período entre ellas (ver figura 1), y la dirección en la que se mueven.

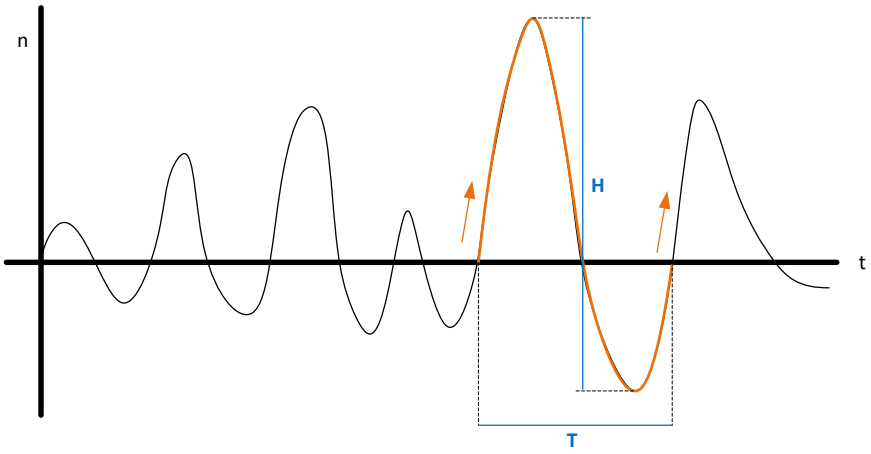

Figura 1: Esquema del estudio del oleaje en los dominios del tiempo

A continuación, se describen dichas características:

- Período (T): Es el tiempo entre el paso de dos crestas sucesivas por un mismo punto.
- Altura (H): Es la distancia entre la cresta y el valle de la ola.
- Altura Significante (Hs): Es el promedio de 1/3 de las olas más altas observadas en una serie en un período de tiempo determinado.

El oleaje está relacionado directamente con las condiciones atmosféricas y oceánicas que conforman el clima marítimo en una región. Para su estudio, es necesario contar con datos históricos de variables oceanográficas que sean de alta calidad, con una resolución temporal y espacial adecuada. Sin embargo, es poco frecuente disponer de una fuente de datos que cumpla con todos estos requisitos en el lugar exacto de interés.

Actualmente, con el avance de los modelos numéricos capaces de reproducir el oleaje en cualquier ubicación a partir de datos atmosféricos, es posible simular

largas series temporales de oleaje con resoluciones espaciales y temporales adecuadas. Sin embargo, estos modelos no siempre reflejan fielmente las condiciones reales. Por esta razón, es habitual combinar la información de los modelos numéricos con datos provenientes de otras fuentes más fiables como datos instrumentales para corregir sus limitaciones; a este proceso se le llama asimilación (ver figura 2).

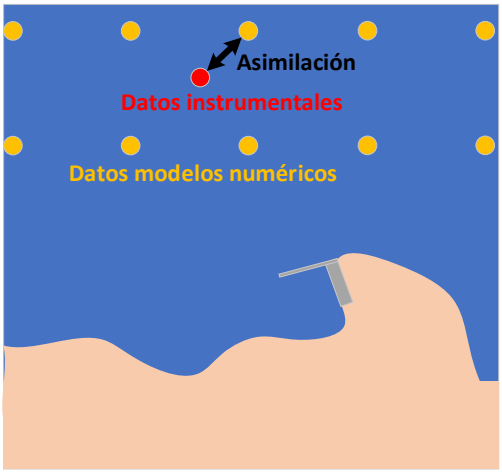

Figura 2: Esquema del proceso de asimilación

El proceso de asimilación se basa en buscar una relación para determinar la variable objetivo a partir de la variable original sin corregir. Para ello (Tomas, 2009) plantea utilizar diferentes técnicas: usar una relación paramétrica a partir de regresiones lineales, utilizar relaciones paramétricas a partir de una relación potencial o aplicar relaciones no paramétricas.

Todas estas técnicas se engloban bajo la rama de la estadística tradicional. Hoy en día, debido al crecimiento de los recursos computacionales y a la masividad de datos con los que se cuenta, aparece un nuevo paradigma llamado inteligencia artificial (ver figura 3) en el que tiene cabida nuevas técnicas estadísticas como las redes neuronales.

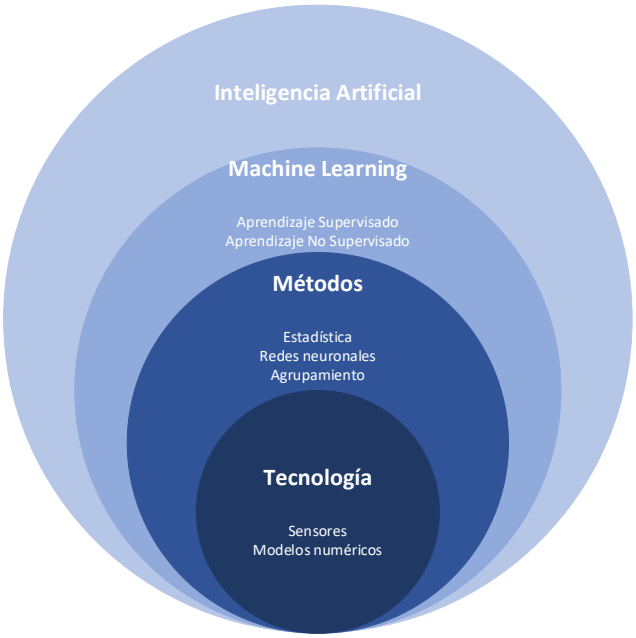

Figura 3: Visión global de la inteligencia artificial

Estas nuevas técnicas han demostrado ser herramientas poderosas en una amplia gama de aplicaciones, desde la clasificación de patrones hasta la predicción, lo que sugiere que se pueda esperar un rendimiento prometedor al aplicarse en este ámbito.

2. OBJETIVOS

El objetivo principal de este trabajo es identificar la mejor técnica para llevar a cabo la asimilación de datos de oleaje, es decir, la corrección de datos provenientes de modelos numéricos o reanálisis en función de datos de observaciones de la naturaleza o instrumentales. A partir de este objetivo surgen otros objetivos secundarios:

- Recopilar fuentes de información numéricas e instrumentales representativos de las distintas características de oleaje del ámbito marítimo español
- Homogenizar las diferentes fuentes de información recopiladas de diferente forma.
- Aplicar los principios FAIR (Findable, Accessible, Interoperable, Reusable) a las fuentes de información utilizadas.
- Identificar un conjunto de técnicas adecuadas para la asimilación de los datos.
- Implementar las técnicas seleccionadas.
- Evaluar las técnicas seleccionadas mediante la aplicación de diferentes métricas.

3. METODOLOGÍA

Para abordar cualquier problema en el ámbito de la inteligencia artificial, es recomendable seguir la metodología descrita en la figura 4.

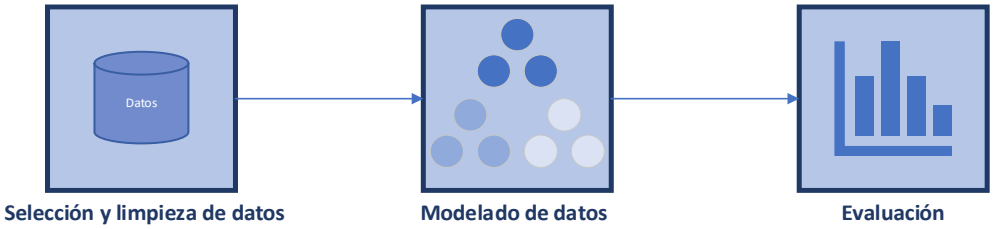

Figura 4: Esquema de la metodología global

En el presente trabajo, se ha adaptado esta metodología general en una versión más específica, con el fin de hacer más eficiente el análisis mediante la aplicación de diversas técnicas a conjuntos de datos provenientes de diferentes fuentes.

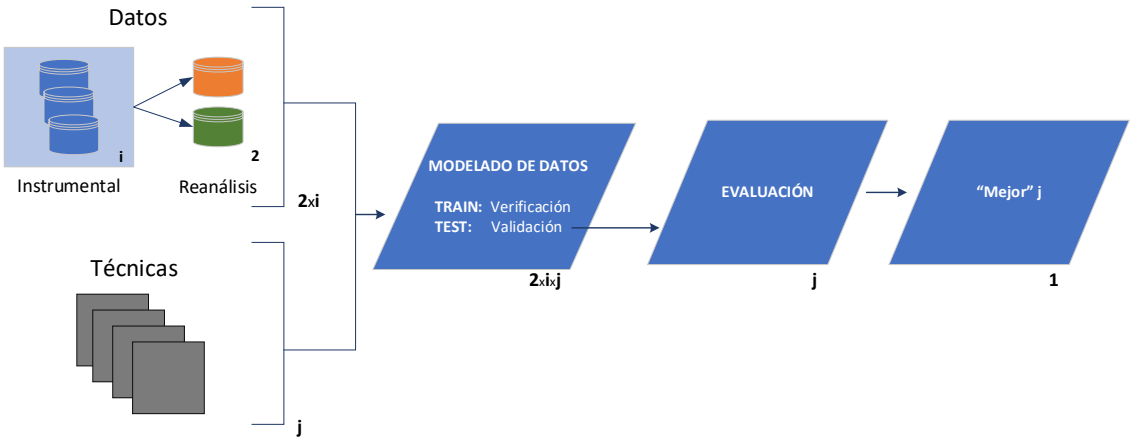

Figura 5: Esquema de la metodología específica

Como se observa en la figura 5, se dispone de datos instrumentales de dimensión *i,* que equivalen a las posiciones donde se ha registrado series temporales de oleaje. También se dispone de datos de reanálisis de dimensión 2, que contienen datos numéricos de oleaje que serán preprocesados para obtener la serie temporal de oleaje en el punto de interés (2*i). Todo ello se presenta detalladamente en el capítulo 3.1. En la siguiente etapa en paralelo, se identifican múltiples técnicas, de dimensión *j*, las cuales se describen en el capítulo 3.2.

El núcleo del trabajo consiste en aplicar estas técnicas a los datos mencionados, lo que se conoce como modelado de datos, obteniendo en una matriz de resultados de dimensión 2*i*j. En esta etapa, los datos son divididos en dos, por un lado, datos de entrenamiento y, por otro lado, datos de validación para conseguir una buena generalización y que el modelo no tienda al sobreajuste. El capítulo 3.3 trata el modelado de datos en detalle.

Finalmente, se procederá a evaluar las técnicas empleadas a partir de unas métricas establecidas que se expone en el capítulo 3.4. El objetivo de identificar la mejor técnica *j,* cuyos resultados se detallan en el capítulo 3.5.

3.1. DATOS DE OLEAJE

Una base de datos de oleaje es un conjunto estructurado de información que recoge y almacena datos relacionados con las características del oleaje en una determinada zona. Se clasifican en función del método por el cual se ha recopilado dicha información.

Las fuentes de información basadas en observaciones de la naturaleza se denominan instrumentales e incluyen, entre otras, boyas oceanográficas. Las boyas oceanográficas son dispositivos flotantes equipados con sensores que miden diversas variables oceanográficas y meteorológicas en la superficie del mar. Estas boyas proporcionan datos en tiempo real con alta resolución temporal, siendo precisas y fiables para medir parámetros como altura de ola, periodo de ola y dirección de ola. Sin embargo, su cobertura espacial está limitada a las áreas donde están desplegadas, y requieren mantenimiento constante, siendo susceptibles a daños por condiciones meteorológicas extremas.

En cuanto a las fuentes de información sintéticas se obtienen a partir de modelos numéricos y son también llamadas fuentes de información de reanálisis. Se destacan los modelos de oleaje globales, como el WAVEWATCH III desarrollado por el Environmental Modeling Center (EMC) y el National Center for Environmental Prediction (NCEP), o el modelo de MFWAM desarrollado por Météo-France (MF), los cuales simulan las condiciones de oleaje a escala global utilizando datos meteorológicos y oceanográficos.

En conclusión, tanto las fuentes de información instrumentales como de reanálisis son fundamentales para la asimilación de datos de oleaje. Integrar ambas fuentes proporciona una visión más precisa de las condiciones del oleaje.

3.1.1. DATOS INSTRUMENTALES

Una fuente de información de datos instrumentales es el conjunto de datos formado por las medidas procedentes de la Red de Boyas de Exteriores (REDEXT). (Puertos del Estado, 2024)

3.1.1.1. DATOS DE BOYAS REDEXT

Las boyas de esta red se caracterizan por estar fondeadas lejos de la línea de costa, en aguas abiertas y profundas (más de 200 metros de profundidad). Por tanto, las medidas de oleaje de estos sensores no están afectadas por efectos locales batimétricos y se puede considerar que cada boya proporciona observaciones representativas de grandes zonas litorales.

El almacenamiento de estos datos se complementa con un control de calidad que garantiza que los valores disponibles se han obtenido en condiciones de correcto funcionamiento. Los datos están en abierto tras una petición a Puertos del Estado.

A continuación, en la tabla 1 se muestra información de cada boya perteneciente a REDEXT.

Nombre	Lon (º)	Lat (º)	Prof. (m)	Inicio	Fin
Bilbao-Vizcaya	-3.04	43.64	580	07-11-1990	10-07-2024
Cabo de Peñas	-6.18	43.75	615	09-06-1998	10-07-2024
Estaca de Bares	-7.68	44.12	1800	19-07-1996	10-07-2024
Villano-Sisargas	-9.21	43.80	386	12-05-1998	10-07-2024
Cabo Silleiro	-9.43	42.12	600	06-07-1998	23-04-2020
Golfo de Cádiz	-6.96	36.49	450	27-08-1996	10-07-2024
Alborán	-5.03	36.27	530	17-02-1997	25-02-2016
Cabo de Gata	-2.34	36.57	536	27-03-1998	10-07-2024
Cabo de Palos	-0.31	37.65	230	18-07-2006	03-04-2020

Valencia	-0.20	39.51	260	15-09-2005	15-06-2020
Tarragona	1.47	40.69	688	20-08-2004	10-07-2024
Cabo Begur	3.65	41.90	1200	27-03-2001	10-07-2024
Dragonera	2.10	39.56	141	29-11-2006	10-07-2024
Mahón	4.42	39.71	300	29-04-1993	10-07-2024
Gran Canaria	-15.80	28.20	780	20-06-1997	25-06-2020
Tenerife	-16.61	28.00	710	01-04-1998	25-06-2020

Tabla 1: Información de boyas REDEXT

Figura 6: Localización de boyas REDEXT

La información de la boya no solo se almacena a bordo, sino que se transmiten vía satélite con periodicidad horaria a Puertos del Estado. En la figura 7 se observa un esquema cronológico de la información de cada boya REDEXT.

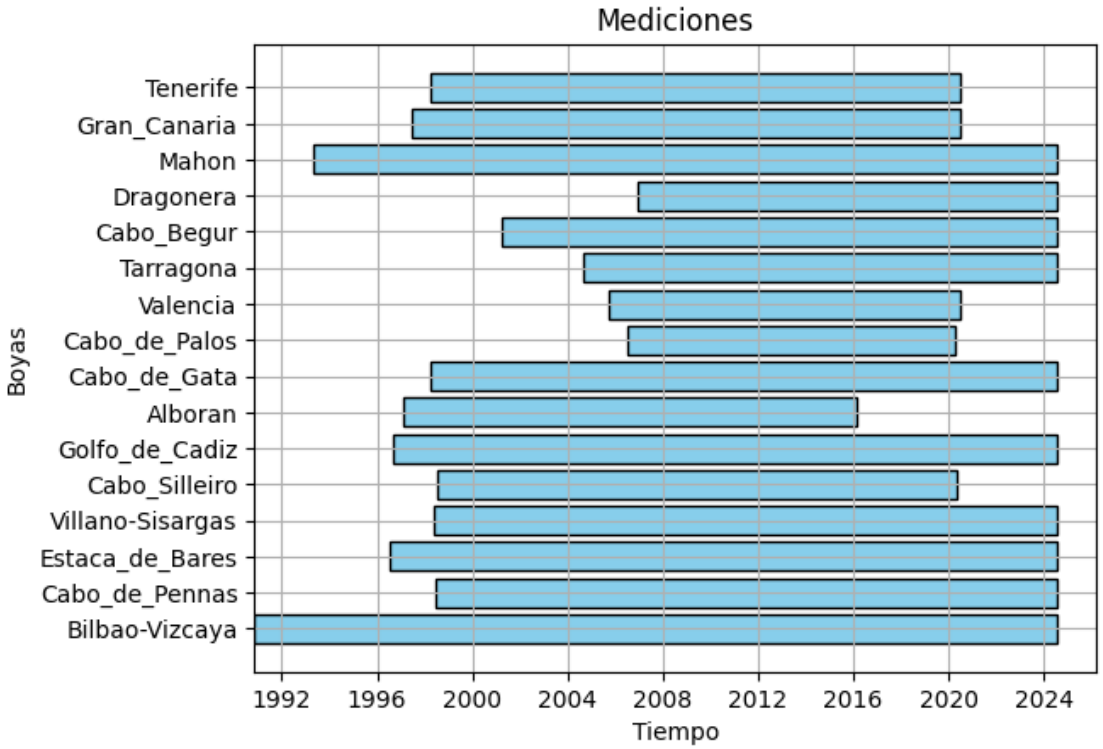

Figura 7: Cronograma de datos de boyas REDEXT

El conjunto de boyas REDEXT proporciona variables de naturaleza oceanográfica, meteorológica y de oleaje. El presente trabajo se centra en las variables relacionadas con el oleaje, las cuales se detallan en la tabla 2.

Variable	**Descripción**	**Unidades**
hs	Altura significante	m
tm02	Periodo medio	s
tp	Periodo de pico	s
hmax	Altura máxima	m
thmax	Periodo asociado a la altura máxima	s
dir	Dirección media	º
dirp	Dirección media en el pico de energía	º
dsp	Dispersión de la dirección en el pico de energía	º

Tabla 2: Variables de los datos de boyas REDEXT

3.1.2. DATOS DE REANÁLISIS

Las fuentes de información de datos de reanálisis utilizados son las siguientes: por un lado, datos de reanálisis de Copernicus (Copernicus, 2024) y, por otro lado, datos de reanálisis de IHCantabria (IHCantabria, 2021).

3.1.2.1. DATOS REANÁLISIS DE COPERNICUS

El IBI-MFC proporciona un producto de reanálisis de oleaje de alta resolución para la zona Iberia-Bizkaia-Irlanda (IBI) llamado IBI_MULTIYEAR_WAV_005_006. El sistema del modelo es gestionado por Nologin con el apoyo del CESGA en términos de recursos de supercomputación.

La configuración del modelo plurianual se basa en el modelo MFWAM que cubre la misma región que el producto de análisis. Se alimentan de los vientos horarios del ECMWF, en concreto, se ve forzado por los datos de viento del reanálisis ERA5. Como condiciones de contorno, está anidado al oleaje espectral del reanálisis GLOBAL.

La información de oleaje generada cubre una extensión espacial en longitud desde -19º a 5º y en latitud desde 26º a 56º con una resolución de 0.027º en ambos espacios.

El producto está disponible en línea y se difunde a través de la Unidad de Difusión del Servicio Marítimo de Copernicus tras controles de calidad automáticos y humanos. https://data.marine.copernicus.eu/product/IBI_MULTIYEAR_WAV_005_006/services

La resolución temporal es de una hora. En la figura 8 se observa un esquema cronológico de los datos de reanálisis IBI.

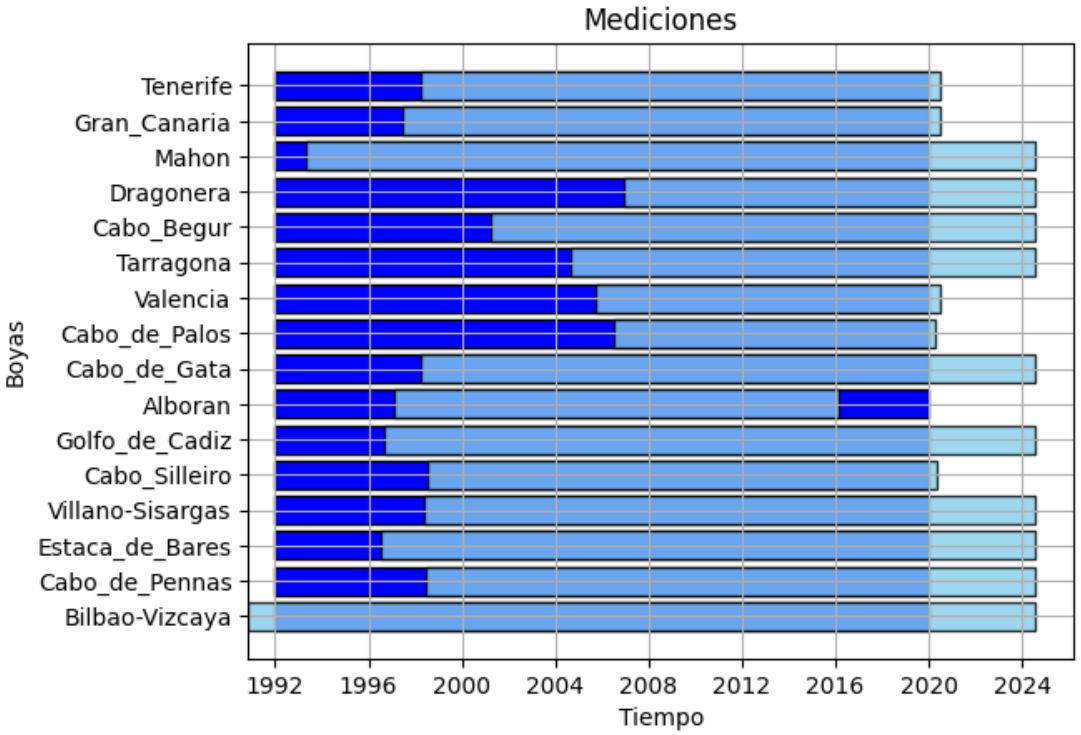

Figura 8: Cronograma de datos IBI_MULTIYEAR_WAV_005_006

Por último, se detallan en la tabla 3 las diferentes variables de oleaje que proporcionan la base de datos.

Variable	Descripción	Unidades
VHM0	Altura significativa de la ola superficial del mar	m
VTM10	Periodo medio de las olas de la superficie del mar a partir del segundo momento de frecuencia de la densidad espectral de varianza	s
VTPK	Periodo de la ola superficial del mar en el máximo de la densidad espectral de la varianza	s
VTM02	Periodo medio de la ola superficial del mar a partir del momento de frecuencia inversa de la densidad espectral de la varianza	s
VPED	Ola de la superficie del mar desde la dirección en el máximo de densidad espectral de varianza	º
VMDR	Ola de la superficie del mar desde la dirección	º

Tabla 3: Variables de los datos IBI_MULTIYEAR_WAV_005_006

3.1.2.2. DATOS REANÁLISIS DE IHCANTABRIA

El producto GOW (Global Ocean Waves) es un conjunto de reconstrucciones históricas de oleaje generadas con el modelo numérico WaveWatch III (Tolman, 1991). El sistema del modelo es gestionado por IHCantabria.

Existen múltiples productos GOW con diferentes características. Esta diversidad se debe a avances en el estado del arte, como, por ejemplo, aparición de nuevos forzamientos y desarrollo de nuevas parametrizaciones de los procesos físicos del oleaje. Como consecuencia, es común que existan múltiples productos GOW para una misma localización. Las necesidades del usuario son las que marcan cuál es el producto más adecuado para cada caso.

En este caso, se ha seleccionado el producto más reciente, GOW CFS. Los forzamientos utilizados proceden del reanálisis global CFSR y de su continuación CFSv2. La versión de GOW CFS (Perez, Menendez, & Losada, 2017), cuenta con una malla global de resolución de 0.5º en ambos espacios, anidada a esta tiene otra malla global de 0.25º en ambos espacios que cubre las zonas cercanas a la costa, y además cuenta con otras dos mallas centradas en los polos con una resolución de 0.º5 en longitud y 0.25º en latitud. La principal ventaja es que la calidad de este GOW es muy superior a la de los anteriores, debido a la calidad de CFSR y a que se utiliza la versión 4.18 de WWIII.

El producto es sometido a controles de calidad y no está disponible a todo el mundo, se difunde a través de peticiones a IHCantabria.

La resolución temporal es de una hora. En la figura 9 se observa un esquema cronológico de los datos.

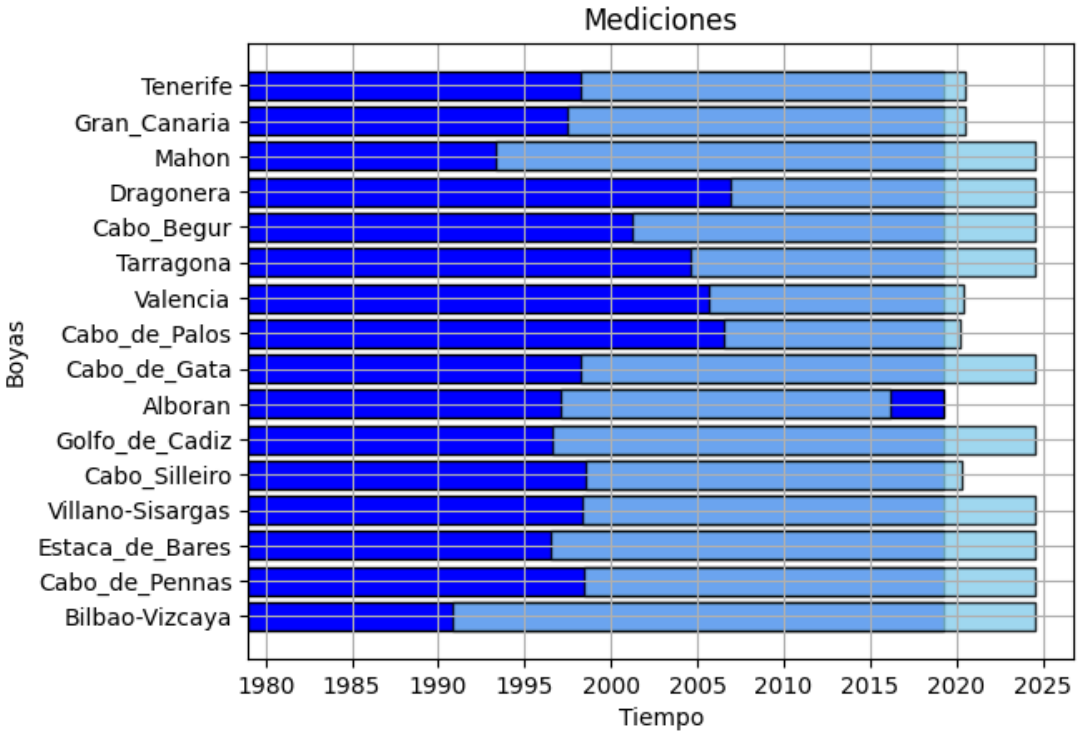

Figura 9: Cronograma de datos GOW

Por último, se detallan en la tabla 4 las diferentes variables de oleaje que se encuentran en la base de datos de reanálisis GOW.

Variable	**Descripción**	**Unidades**
hs	Altura significante	m
tm01	Onda de viento de superficie mar periodo medio de varianza densidad espectral primer momento de frecuencia.	
tm02	Onda de viento de superficie mar periodo medio de varianza densidad espectral segundo momento de frecuencia.	s
fp	Frecuencia de pico	1/s
dir	Dirección media	º
spr	Dispersión de la dirección en el pico de energía	º

Tabla 4: Variables de los datos GOW

3.1.3. PRINCIPIOS FAIR

Los principios FAIR son un conjunto de directrices diseñadas para mejorar la gestión, accesibilidad, interoperabilidad y reutilización de los datos científicos. A continuación, se describen dichos principios:

- **Findable** (Localizable): Los datos deben ser fáciles de encontrar para los usuarios, tanto humanos como máquinas. Esto implica la necesidad de metadatos descriptivos y el uso de identificadores únicos y persistentes como un Digital Object Identifier (DOI).
- **Accessible** (Accesible): Deben ser accesibles para los usuarios de manera clara a través de protocolos de comunicación bien definidos y estandarizados. Los niveles de acceso puedan variar (por ejemplo, acceso abierto, acceso bajo ciertas condiciones, etc.).
- **Interoperable** (Interoperable): Los datos deben ser presentados en un formato que permita su combinación, lo que requiere la adopción de estándares y vocabularios comunes.
- **Reusable** (Reutilizable): Los datos deben estar bien descritos y documentados para que puedan ser reutilizados en diferentes contextos y por distintos usuarios, sin necesidad de contacto adicional con los autores originales.

A continuación, en la tabla 5 se evalúa en qué medida las bases de datos descritas anteriormente cumplen con los principios FAIR.

	Boya REDEXT	IBI_MULTIYEAR_WAV_005_006	GOW
Findable	✗	✓	✗
Accessible	✗	✓	✓
Interoperable	✓	✓	✓
Reusable	✓	✓	✓

Tabla 5: Evaluación de los principios FAIR

Por lo que se puede concluir que la única base de datos que cumple con los principios FAIR es la base de datos perteneciente a Copernicus.

3.1.4. PRE-PROCESADO DE LOS DATOS

Se han identificado los nodos de la fuente de información de reanálisis que más cerca se encuentran de la localización de las boyas de REDEXT. Para ello se ha creado un script que ordena los nodos de reanálisis por la distancia euclídea a los puntos de las boyas REDEXT y se queda con el menor en cada caso. En la tabla 6 se puede ver el resultado.

Nombre	Boya (Lon, Lat) (º)	IBI	GOW
Bilbao-Vizcaya	(-3.04,43.64)	(-3.00,43.60)	(-3.00,43.63)
Cabo de Peñas	(-6.18,43.75)	(-6.20,43.80)	(-6.13,43.75)
Estaca de Bares	(-7.68,44.12)	(-7.70,44.10)	(-7.63,44.13)
Villano-Sisargas	(-9.21,43.80)	(-9.20,43.80)	(-9.25,43.75)
Cabo Silleiro	(-9.43,42.12)	(-9.40,42.10)	(-9.38,42.13)
Golfo de Cádiz	(-6.96,36.49)	(-7.00,36.50)	(-7.00,36.50)
Alborán	(-5.03,36.27)	(-5.00,36.30)	(-5.00,36.23)
Cabo de Gata	(-2.34,36.57)	(-2.30,36.60)	(-2.38,36.63)
Cabo de Palos	(-0.31,37.65)	(-0.30,37.70)	(-0.25,37.625)
Valencia	(-0.20,39.51)	(-0.20,39.50)	(-0.25,39.50)
Tarragona	(1.47,40.69)	(1.50,40.70)	(1.50,40.75)
Cabo Begur	(3.65,41.90)	(3.70,41.90)	(3.63,41.88)
Dragonera	(2.10,39.56)	(2.10,39.60)	(2.13,39.50)
Mahón	(4.42,39.71)	(4.40,39.70)	(4.34,39.75)
Gran Canaria	(-15.80,28.20)	(-15.80,28.20)	(-15.75,28.25)
Tenerife	(-16.61,28.00)	(-16.60,28.00)	(-16.63,28.00)

Tabla 6: Localización de extracción de los datos

Las series de oleaje resultantes de cada fuente de información son válidas, aunque no estén exactamente en la misma localización ya que los nodos se encuentran en aguas abiertas, lo que equivale a nodos representativos de grandes áreas espaciales.

La calidad y consistencia de los datos son fundamentales para asegurar la validez de los resultados y análisis posteriores. Para ello, es necesario realizar un conjunto de operaciones de sincronización de los datos recopilados.

Cada fuente de información facilita los datos con un control de calidad apto para su uso, pero dado que cubren diferentes periodos y registros, uno de los primeros desafíos es alinear temporalmente los datos. Este paso es necesario para garantizar que todas las series de datos compartan un mismo conjunto de fechas coincidentes, y así poder llevar a cabo un análisis comparativo.

El procedimiento seguido se puede desglosar en los siguientes pasos:

1. Identificación de fechas con datos faltantes, se realiza un análisis inicial para identificar todas las fechas registradas en cada una de las bases de datos.
2. Eliminación de fechas no coincidentes, una vez identificadas las fechas con datos faltantes, se procede a eliminarlas de las bases de datos. Esto asegura que las fechas restantes son comunes a todas las fuentes.
3. Validación de la consistencia temporal, después de la limpieza, se valida que las series de datos resultantes fueran temporalmente consistentes, es decir, que todas las bases de datos tuvieran registros completos para las fechas restantes. (ver figura 10)

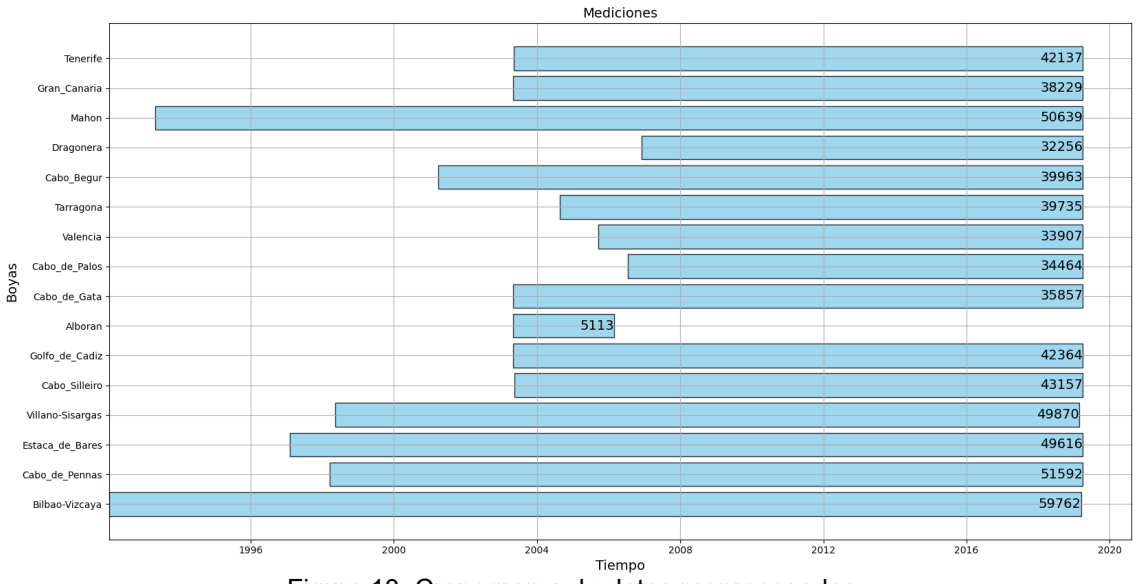

Figura 10: Cronograma de datos preprocesados

Para finalizar se ha realizado un análisis descriptivo de las variables de interés en las diferentes fuentes de información de cada localización para identificar patrones, anomalías y tendencias en los datos, lo que facilita la correcta interpretación de los resultados. Los resultados se encuentran en el repositorio GIT. (https://github.com/alexgonzvalle/TFM/tree/main/plot/data_procesed)

3.1.5. CONCLUSIONES

Se han descrito y preprocesado las fuentes de información llegando a unas primeras conclusiones:

- La localización de Alborán fue excluida debido a su estado fuera de servicio, lo cual limitaba la disponibilidad de datos actualizados y fiables.
- Respecto a la resolución temporal de las fuentes de información, después de realizar el proceso de limpieza, van desde 32.256 elementos para la localización de Dragonera a 59.762 elementos para la localización de Bilbao-Vizcaya por lo que es aceptable para afrontar el problema.
- En cuanto a las variables de oleaje utilizadas, se han identificado aquellas comunes a ambas fuentes de información: altura de ola significante (Hs), periodo medio (Tm02), periodo o frecuencia de pico (Tp/Fp), y dirección media de las olas (Dm). Estas variables son importantes para comprender la dinámica del oleaje.

3.2. TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

El aprendizaje automático abarca una variedad de técnicas diseñados para abordar diferentes tipos de problemas mediante diferentes enfoques.

Respecto al tipo de problema nos podemos encontrar con:

- Predicción, la estimación de valores futuros basados en datos históricos.
- Clasificación, asignar etiquetas a datos en función de características.
- Asociación, descubrir relaciones significativas entre variables en grandes conjuntos de datos.
- Segmentación o clustering, agrupar datos en subconjuntos homogéneos.
- Reducción de la dimensión, simplificar conjuntos de datos al reducir el número de variables, manteniendo la información esencial.

Estos problemas se resuelven mediante enfoques que se agrupan en:

- Aprendizaje supervisado, donde los modelos se entrenan con datos etiquetados para predecir resultados conocidos.
- Aprendizaje no supervisado, que se enfoca en descubrir patrones ocultos en datos no etiquetados.

Técnica	Enfoque	Problema
Regresión	Aprendizaje Supervisado	Predicción Clasificación
Red neuronal	Aprendizaje Supervisado	Clasificación Predicción
	Aprendizaje No Supervisado	Segmentación
KNN	Aprendizaje Supervisado	Clasificación Predicción
Árbol de decisión	Aprendizaje Supervisado	Clasificación Predicción
Red probabilística	Aprendizaje Supervisado	Predicción Clasificación
	Aprendizaje No Supervisado	Asociación
K-Means	Aprendizaje No Supervisado	Segmentación
PCA	Aprendizaje No Supervisado	Reducción de la dimensión

Tabla 7: Técnicas de aprendizaje automático

Para resolver el proceso de la asimilación se van a utilizar los dos enfoques. Por un lado, técnicas de aprendizaje supervisado, regresiones (3.2.1) y redes neuronales (3.2.2). Por otro lado, técnicas de aprendizaje no supervisado, se usará un método de segmentación (3.2.3) para agrupar los datos.

3.2.1. REGRESIÓN

La regresión es una técnica fundamental en el ámbito de la inteligencia artificial y el aprendizaje automático supervisado, ampliamente utilizada para modelar y analizar relaciones entre variables. Al emplear métodos de regresión, es posible establecer relaciones matemáticas entre variables observadas:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}) + \in = \hat{\boldsymbol{y}} + \in \qquad [1]$$

Regresión lineal

La regresión lineal tiene como objetivo principal determinar una ecuación lineal que mejor ajuste los datos observados, permitiendo así predecir el valor de la variable dependiente a partir de los valores de las variables independientes.

En su forma más sencilla, la regresión lineal se denomina regresión lineal simple, donde se considera una sola variable independiente. Además, para este estudio, se obliga a que la recta pase por el punto (0,0). La ecuación que representa esta relación es:

$$\boldsymbol{y = \beta x} \qquad [\,2\,]$$

Donde:

- *y* es la variable dependiente.
- *x* es la variable independiente.
- β es el coeficiente de regresión, que indica el cambio esperado en *y* por cada unidad de cambio en *x*.

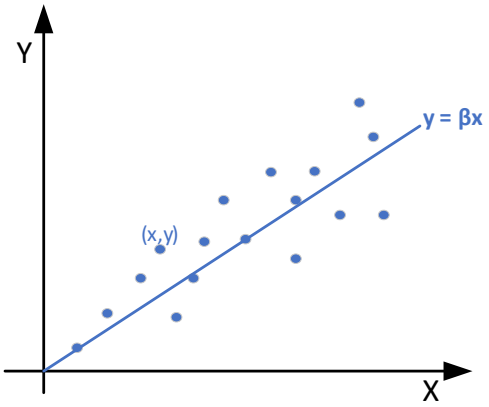

Figura 11: Esquema de la regresión lineal simple

Regresión no lineal

La regresión no lineal se utiliza cuando la relación entre la variable dependiente y una o más variables independientes no puede ser adecuadamente descrita por una línea recta.

A diferencia de la regresión lineal, que se basa en una recta, la regresión no lineal puede tomar muchas formas, dependiendo de la naturaleza de la relación entre las variables, por ejemplo, basado en una relación potencial donde se obliga a pasar por el punto (0,0):

$$y = \beta x^{\gamma} \qquad [3]$$

Donde:

- *y* es la variable dependiente.
- *x* es la variable independiente.
- β es un coeficiente que escala la relación.
- γ es el exponente que describe cómo *y* cambia con respecto a *x*.

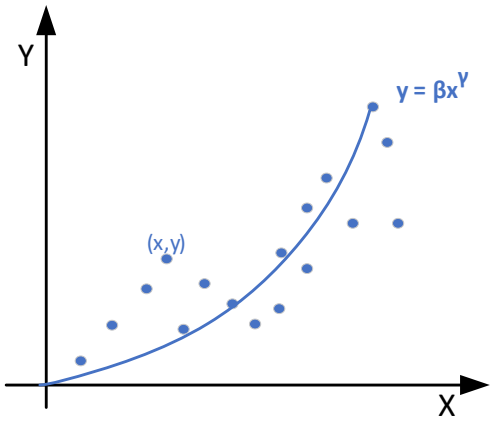

Figura 12: Esquema de la regresión no lineal

3.2.2. RED NEURONAL

Una red neuronal es una técnica de aprendizaje automático supervisado que emula la toma decisiones de forma similar al cerebro humano, replicando la manera en que las neuronas biológicas colaboran para reconocer patrones, evaluar opciones y llegar a conclusiones.

Estas redes neuronales se componen de capas de nodos o neuronas artificiales: una capa de entrada, una o varias capas ocultas y una capa de salida. Cada nodo se conecta a otro nodo y tiene su propio peso y umbral asociados. Cuando la salida de un nodo supera el valor umbral establecido, el nodo se activa y trasmite la información a la siguiente capa de la red; de lo contrario, no se trasmite nada a la siguiente capa de la red. Un esquema general de una red neuronal quedaría de la siguiente manera:

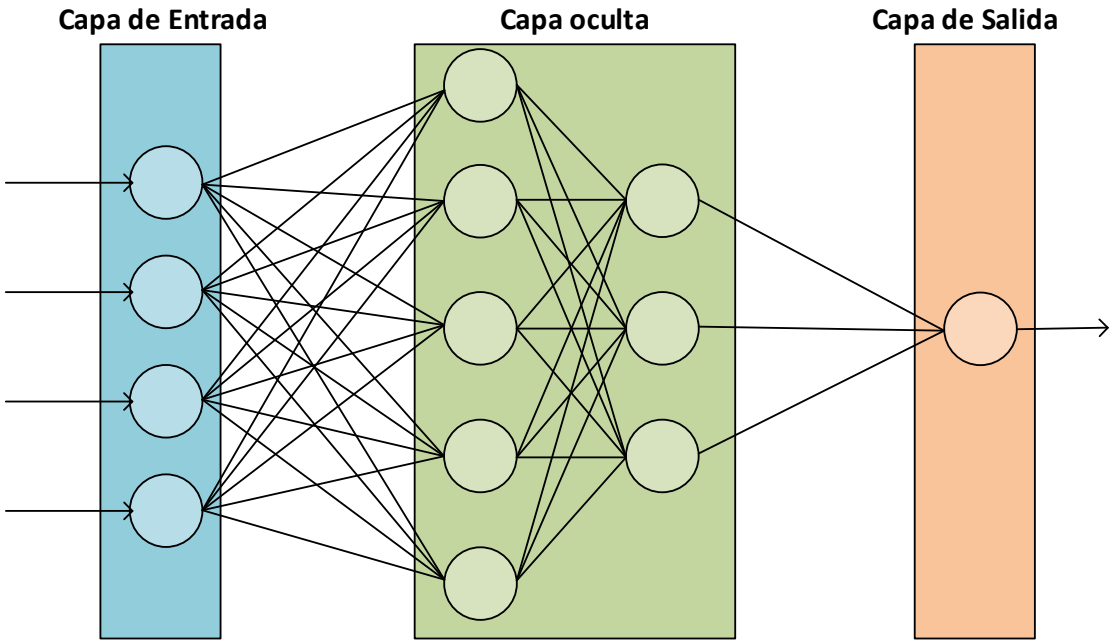

Figura 13: Esquema general de red neuronal

Y cada neurona de cada capa se detalla en el siguiente esquema:

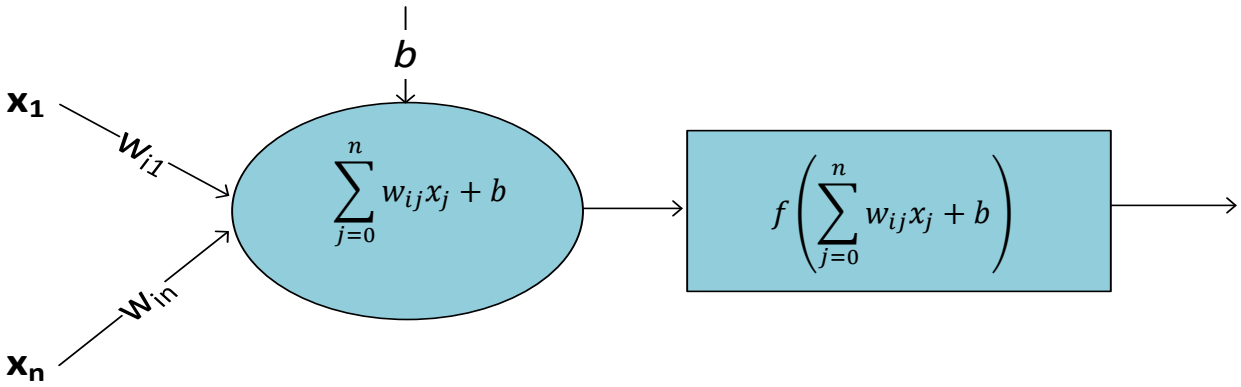

Figura 14: Esquema general de una neurona

Donde:

- x_j son los datos de entrada.
- w_{ij} son los pesos.
- b es el sesgo.
- f es la función de activación.

3.2.3. SEGMENTACIÓN

La segmentación es una técnica de aprendizaje no supervisado que agrupa los datos en subconjuntos (clústers) de tal manera que los datos en el mismo grupo son similares entre sí.

K-Means

Es uno de los algoritmos iterativos más utilizados, el cual, agrupa considerando la distancia euclídea.

$$\boldsymbol{d(x_i, x_{i'}) = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = \|x_i - x_{i'}\|^2}$$

$$\boldsymbol{W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} \|x_i - x_{i'}\|^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \overline{x_k}\|^2} \quad [4]$$

El objetivo es encontrar K centroides solución del siguiente problema de optimización:

$$\boldsymbol{\min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \overline{m_k}\|^2} \quad [5]$$

Para ello una vez definido el parámetro K, número de clúster:

- Se realiza una primera asignación definida aleatoriamente.
- Se repite hasta converger o alcanzar el número máximo de iteraciones:
  - Se estima el centroide de cada conglomerado.
  - Se redefine los clústeres considerando los nuevos centroides.

3.3. MODELADO DE DATOS

Como se ha ido definiendo en este documento, la asimilación de datos es un procedimiento de comparación entre dos fuentes de información, de manera que se modifican para tratar de ajustarse, con la mayor exactitud posible, a la realidad.

En este capítulo se va a presentar la aplicación de las técnicas antes descritas que constituyen los *j* modelos de asimilación, en este caso 5. Cada modelo ha

sido aplicado en las *i* localizaciones anteriormente descritas, en concreto, 15. Y para las dos bases de datos mencionadas, lo que lleva a 30 resultados para cada modelo. Con todos estos resultados, se tiene en total 150 calibraciones diferentes.

Como se ha mencionado anteriormente, para obtener cada resultado, se ha dividido los datos de entrada en datos de entrenamiento y datos de validación.

En este estudio, ya que se usan series temporales, se ha decidido dividir los datos sin perder el componente temporal. Por ello, se ha dividido los datos seleccionando las mismas fechas para la validación y respetando una relación máxima de aproximadamente 70% de entrenamiento y 30% de validación y una mínima de aproximadamente 50% entrenamiento y 50% validación. La fecha seleccionada ha sido 01/01/2013, así que los datos de cada localización anteriores a esta fecha serán usados para entrenamiento, y los posteriores para validación, en la figura 15 se muestra como quedan para cada localización.

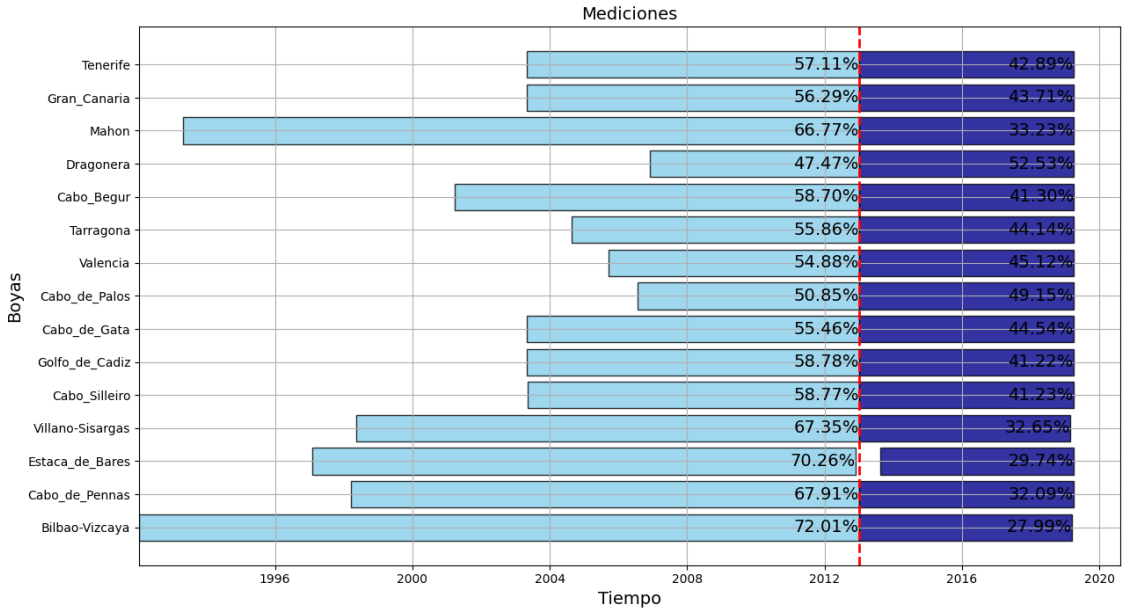

Figura 15: Cronograma de datos divididos en entrenamiento y validación

Todos los códigos generados se pueden encontrar en el repositorio GIT. (https://github.com/alexgonzvalle/TFM.git)

3.3.1. REGRESIÓN LINEAL

Este modelo, obtenido de la tesis (Tomas, 2009), tiene como objetivo buscar la función de ajuste (y_{cal}) entre la altura de la ola observada (y) y la altura de la ola calculada numéricamente (x).

Para obtener la relación de asimilación del modelo de regresión, hay que destacar que no se incluyen el término independiente, para evitar faltas de homogeneidades en la asimilación de los valores próximos a cero:

$$y = \beta_0 + \beta_1 x \approx \beta_1 x \rightarrow x = \frac{1}{\beta_1} y \qquad [6]$$

$$y_{cal} = b_0 + b_1 x \approx b_1 x = b_1 \left(\frac{1}{\beta_1} y\right)$$

Una vez planteadas las ecuaciones de partida, se iguala la variable objetivo de asimilación a las variables predictoras para obtener la relación de asimilación.

$$y_{cal} \equiv Y \rightarrow b_1 \frac{1}{\beta_1} y = y \rightarrow b_1 \frac{1}{\beta_1} = 1 \rightarrow b_1 = \beta_1 \qquad [7]$$

Despejando la ecuación llegamos a la expresión final:

$$y_{cal} = \beta_1 x \qquad [8]$$

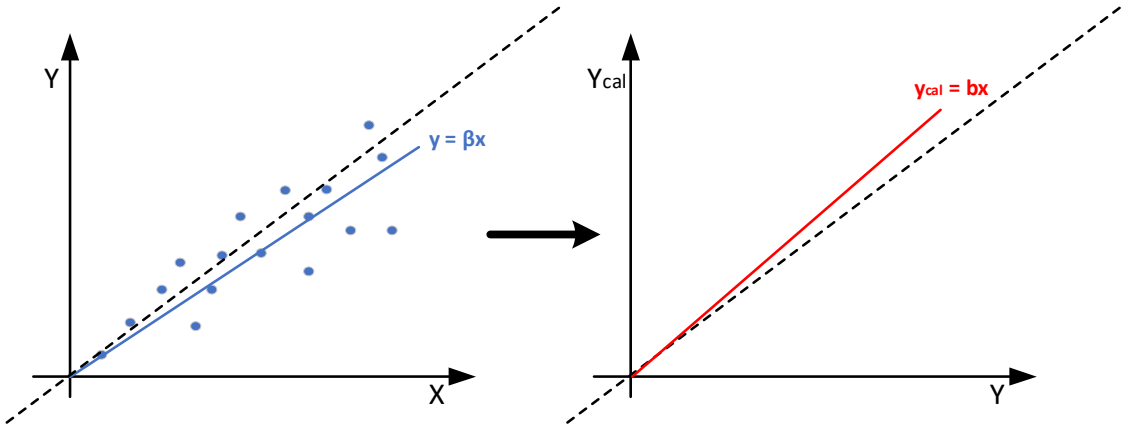

Figura 16: Relación de asimilación a partir de la regresión lineal

3.3.2. REGRESIÓN NO LINEAL

Como en el caso anterior, este modelo, obtenido de la tesis (Tomas, 2009), busca la función de ajuste (y_{cal}) entre la altura de la ola observada (y) y la altura de la ola calculada numéricamente (x).

Primero se plantean las ecuaciones de partida para la regresión no lineal:

$$y = \beta x^{\gamma} \rightarrow \frac{y}{\beta} = x^{\gamma} \rightarrow \frac{y^{1/\gamma}}{\beta^{1/\gamma}} = x \qquad [9]$$

$$y_{cal} = bx^{c} = b\left(\frac{y^{1/\gamma}}{\beta^{1/\gamma}}\right)^{c} = b\frac{y^{c/\gamma}}{\beta^{c/\gamma}}$$

Se iguala la variable objetivo a las variables predictoras para obtener la relación.

$$y_{cal} \equiv Y \rightarrow b\frac{y^{c/\gamma}}{\beta^{c/\gamma}} = y \qquad [10]$$

Para que esta igualdad sea cierta para cualquier valor de *Y*, los exponentes y los coeficientes deben ser iguales. Esto nos lleva a dos ecuaciones:

$$\frac{c}{\gamma} = 1 \rightarrow c = \gamma \qquad [11]$$

$$\frac{b}{\beta^{c/\gamma}} = 1 \rightarrow \frac{b}{\beta^{c/c}} = 1 \rightarrow b = \beta$$

Despejando la ecuación llegamos a la expresión final:

$$y_{cal} = \beta x^{\gamma} \qquad [12]$$

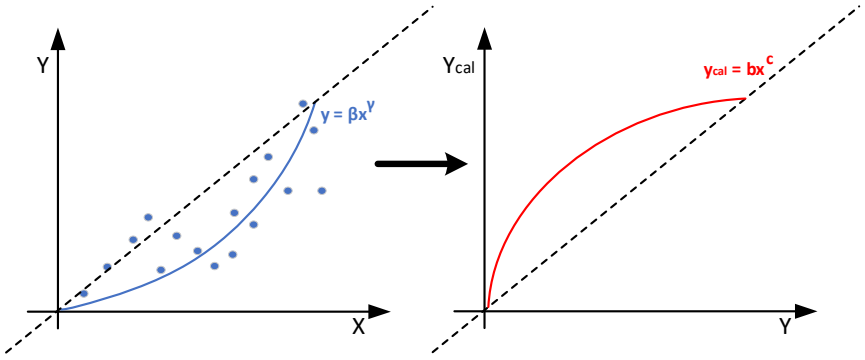

Figura 17: Relación de asimilación a partir de la regresión no lineal

3.3.3. REGRESIÓN NO LINEAL DIRECCIONAL

Hasta el momento, los modelos de calibraciones presentados corrigen la altura del oleaje sin considerar la dirección de procedencia del oleaje. En el IHCantabria se ha ido trabajando en un modelo de asimilación donde se agregan los datos por direcciones específicas y se aplican correcciones a cada una de ellas (Mínguez, Espejo, Tomas, Méndez, & Losada, 2011).

Este modelo de asimilación se basa en la siguiente expresión:

$$\boldsymbol{y_{cal}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{f}(\boldsymbol{a^R}, \boldsymbol{b^R};\ \boldsymbol{H_s^R}, \boldsymbol{\theta}) = \boldsymbol{a^R}(\boldsymbol{\theta})(\boldsymbol{H_s^R})^{\boldsymbol{b^R(\theta)}} \qquad [13]$$

Donde:

- $x = H_s^R$, es la altura de la ola calculada numéricamente.
- $\boldsymbol{a^R}(\boldsymbol{\theta})$y $\boldsymbol{b^R}(\boldsymbol{\theta})$ son los parámetros dependientes de la dirección del oleaje.

El objetivo de este modelo es minimizar el error entre la función calculada y la variable objetiva que es la altura de la ola observada (y).

3.3.4. RED NEURONAL

Este modelo de asimilación se basa en la altura y dirección del oleaje. Dentro de la red neuronal hay diferentes parámetros a calibrar: la función de activación, el optimizador, la tasa de aprendizaje, las capas ocultas y el número de neuronas por cada capa oculta, a continuación, se detalla cómo se han definido cada uno.

La función de activación establecida para cada neurona en cada capa ha sido la función ReLU (Rectified Linear Unit) ya que es eficiente computacionalmente y evita el problema del desvanecimiento del gradiente. Se define matemáticamente de la siguiente manera:

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{ReLU}(\boldsymbol{x}) = \mathbf{max}\,(\mathbf{0}, \boldsymbol{X}) \qquad [14]$$

A la vista del análisis descriptivo de las fuentes de información, respecto a la variable altura de ola, se distribuyen principalmente en sus valores medios teniendo menos registros en los valores extremos. Por ello, se ha incluido una capa de regularización para obtener un modelo mejor generalizado.

Respecto a los parámetros restantes se ha generado un script que realice una combinación de entre diferentes opciones establecidas de cada parámetro con el objetivo de encontrar la configuración que mejor ajuste en términos de error cuadrático medio (MSE).

Se han definido las siguientes opciones:

- Optimizador: adam, sgd, rmsprop
- Numero de capas ocultas: 1, 5, 10, 15
- Numero de neuronas por capa oculta: 10, 50, 100, 200.

Respecto a las opciones de los optimizadores se explica brevemente cada uno a continuación: Stochastic Gradient Descent (SGD) es la versión estocástica del descenso del gradiente. Root Mean Square Propagation (RMSprop) es un optimizador adaptativo. Adaptive Moment Estimation (Adam) combina las ventajas de RMSprop y el método del momentum. La tasa de aprendizaje de cada optimizador se ajusta internamente en el proceso.

Se ha ejecutado dicho script en cada localización y para cada modelo de reanálisis. Los resultados obtenidos se han rankeado por su menor valor en MSE en cada localidad y en cada modelo de reanálisis, a continuación, se ha agrupado por cada configuración y sumado sus rankings.

Optimizador	$\boldsymbol{\eta}$	**Capas**	**Neuronas**	**Ranking**
sgd	0.006	1	100	1 (94)
adam	0.007	1	50	2 (187)
rmsprop	0.007	1	10	3 (216)

Tabla 8: Evaluación de parámetros del modelo de red neuronal

Se ha optado por seleccionar la configuración con mejor ranking, cuya expresión matemática es la siguiente:

$$\boldsymbol{y_{cal} = w_2 \cdot ReLU(w_1 X + b_1) + b_2} \qquad [15]$$

Donde:

- w_2 es un vector de pesos de tamaño 1x100.
- b_2 es un escalar de sesgo de tamaño 1.
- w_1 es una matriz de pesos de tamaño 100x2.

- b_1 es un vector de sesgos de tamaño 100x1.
- X es un vector $[x_1, x_2]^T$, que equivale a H_s y Dir_m calculado numéricamente.

3.3.5. K-MEANS

Este modelo de asimilación se basa en la altura y dirección del oleaje. Se divide en dos pasos, en primer lugar, se realiza una segmentación para después aplicar a cada grupo una corrección.

Para realizar la segmentación se ha usado el algoritmo K-Means, como parámetro a definir se encuentra el número de grupos o clústeres, llamado K. Como en el modelo anterior, se ha realizado un script que realiza un ranking para obtener el mejor K, entre 2 y 500, en base al menor MSE para cada localización y cada fuente de información de reanálisis.

K	Ranking
200	1 (3389)
201	2 (3484)
204	3 (3558)

Tabla 9: Evaluación del parámetro K del modelo de K-Means

Se ha optado por fijar el parámetro K en 200. A modo de ejemplo, en la figura 18 se muestra la evolución de K respecto al MSE para la localización del Golfo de Cádiz y la fuente de información GOW. Como se puede apreciar el punto mínimo establece el número de grupos optimo, para este caso K_{opt}=88, que no difiere en gran medida en términos de MSE del K establecido.

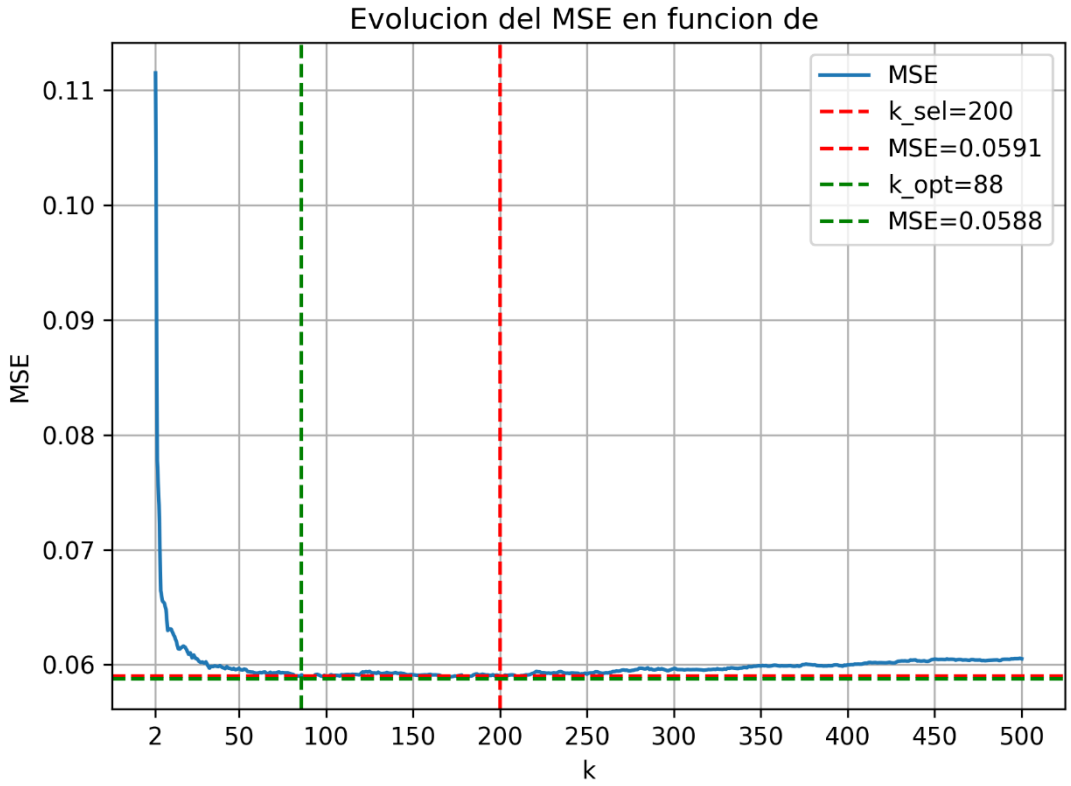

Figura 18: Evolución de MSE en función de k para el golfo de Cádiz

El segundo paso en este modelo es realizar una calibración de cada grupo (k) entre la altura de la ola observada (y) y la altura de la ola calculada numéricamente (x). Para ello se define la siguiente relación:

$$\boldsymbol{y_{cal}(\mathrm{k})_{k=2}^{200} = \frac{\sum_{i=0}^{k} \frac{y(k)}{x(k)}}{k} * x(k)} \qquad [\ 16\ ]$$

3.4. EVALUACIÓN DE LAS TÉCNICAS

La evaluación de los modelos de asimilación utilizados se determina mediante diversas métricas estadísticas que permiten cuantificar el ajuste del modelo a los datos observados. Las métricas seleccionadas son: el sesgo (bias), el error cuadrático medio (RMSE), el coeficiente de correlación de Pearson y el índice de dispersión. A continuación, se describen en detalle cada una de estas métricas.

3.4.1. BIAS

El sesgo es una medida de la tendencia del modelo a sobrestimar o subestimar los valores verdaderos. Se calcula como la diferencia promedio entre los valores predichos por el modelo y los valores observados. Un sesgo cercano a cero indica que el modelo no presenta una tendencia sistemática de error, mientras que un sesgo positivo o negativo sugiere una sobreestimación o subestimación, respectivamente. La fórmula para calcular el sesgo es:

$$\boldsymbol{Bias} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\mathbf{1=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}) \qquad [17]$$

3.4.2. RMSE

El error cuadrático medio (Root Mean Square Error, RMSE) es una métrica que mide la magnitud promedio de los errores entre los valores observados y los valores predichos por el modelo. Es una medida de precisión que penaliza los errores grandes de manera más severa. Un RMSE cercano a 0 indica un mejor ajuste del modelo a los datos observados. La fórmula para calcular el RMSE es:

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\mathbf{1=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}} \qquad [18]$$

3.4.3. COEFICIENTE DE CORRELACIÓN DE PEARSON

El coeficiente de correlación de Pearson es una medida de la fuerza y la dirección de la relación lineal entre dos variables. En el contexto de la regresión, mide la correlación entre los valores observados y los valores predichos. El coeficiente de Pearson varía entre -1 y 1, donde un valor de 1 indica una correlación positiva perfecta, -1 indica una correlación negativa perfecta y 0 indica ninguna correlación lineal. La fórmula para calcular el coeficiente de Pearson es:

$$\boldsymbol{\rho} = \frac{\sum_{\mathbf{1=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{\hat{y}_i} - \bar{\hat{\boldsymbol{y}}})}{\sqrt{\sum_{\mathbf{1=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \bar{\boldsymbol{y}})^{\mathbf{2}}}\sqrt{\sum_{\mathbf{1=1}}^{\boldsymbol{n}}(\boldsymbol{\hat{y}_i} - \bar{\hat{\boldsymbol{y}}})^{\mathbf{2}}}} \qquad [19]$$

Donde $\bar{y}$ e $\bar{\hat{y}}$ son las medias de los valores observados y predichos, respectivamente.

3.4.4. ÍNDICE DE DISPERSIÓN

El índice de dispersión es una métrica que evalúa la variabilidad relativa de los errores del modelo. Se calcula como la razón entre la varianza de los errores y la varianza de los valores observados. Un índice de dispersión cercano a 1 indica que el modelo tiene una variabilidad de errores comparable a la variabilidad de los datos observados. La fórmula para calcular el índice de dispersión es:

$$SI = \frac{\frac{1}{n}\sum_{1=1}^{n}(y_i - \bar{\hat{y}})^2}{\frac{1}{n}\sum_{1=1}^{n}(y_i - \bar{y})^2} = \frac{RMSE}{\bar{X}} \quad [20]$$

3.5. COMPARATIVA DE MODELOS

Se han obtenido fichas resumen del resultado de cada modelo, en cada localización y para cada base de datos de reanálisis que se pueden encontrar en https://github.com/alexgonzvalle/TFM/tree/main/plot/model.

En ellas se muestran en la primera fila una rosa de oleaje con los datos de la base de datos instrumental (Hs_{Boya}), una rosa de oleaje con los datos de reanálisis (Hs_{IBI} o Hs_{GOW}) y una rosa de oleaje con los datos resultantes ($Hs_{Calibrada}$). En la segunda fila se muestra los datos de entrenamiento de Hs_{Boya} frente a la Hs de reanálisis, y de igual forma para los datos de prueba con el resultado de las métricas definidas para cada base de datos, por último, se muestra una rosa de oleaje con el factor de correlación que el modelo aplica a la base de datos de reanálisis. Este último grafico es de gran utilidad ya que da información de cómo aplicar el modelo en casos futuros.

A continuación, se presentan los resultados obtenidos para Dragonera en el modelo de reanálisis IBI.

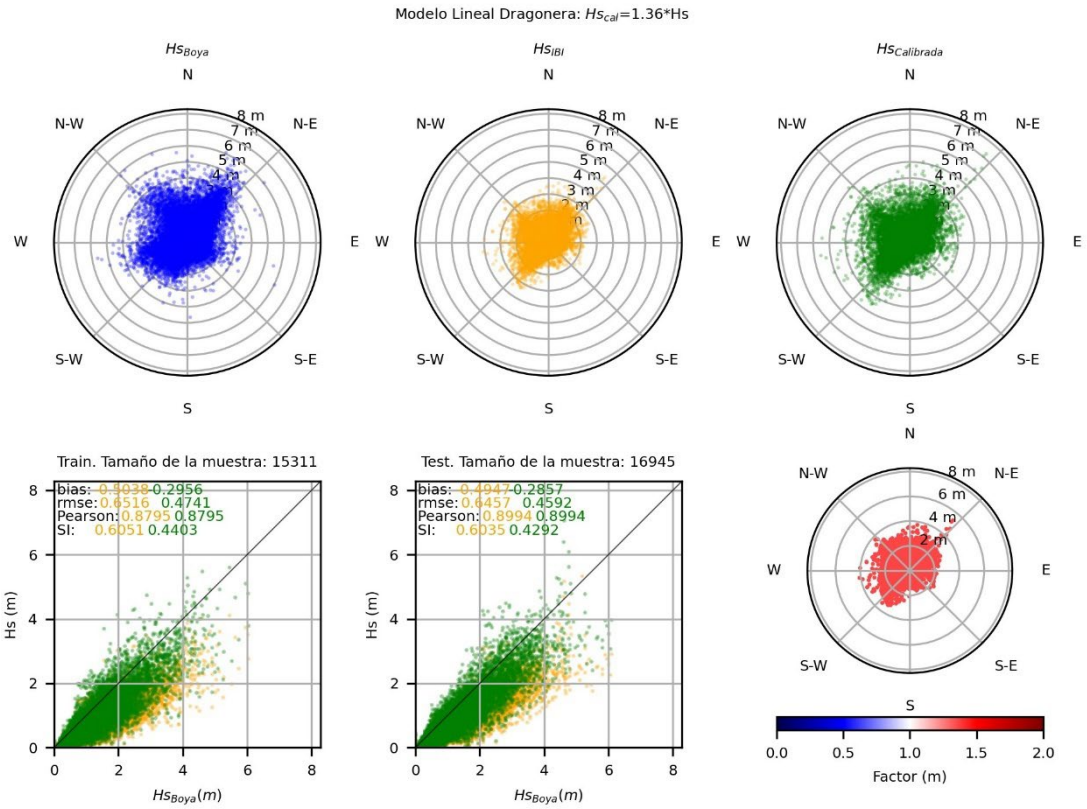

Figura 19: Resultados del modelo lineal para Dragonera con datos IBI

En este caso, en la figura 19 se aprecia como el modelo lineal mejora todas las métricas establecidas respecto a los datos sin transformar. Respecto al factor de calibración se observa cómo es constante para todas las direcciones y alturas de ola y de valor 1.36.

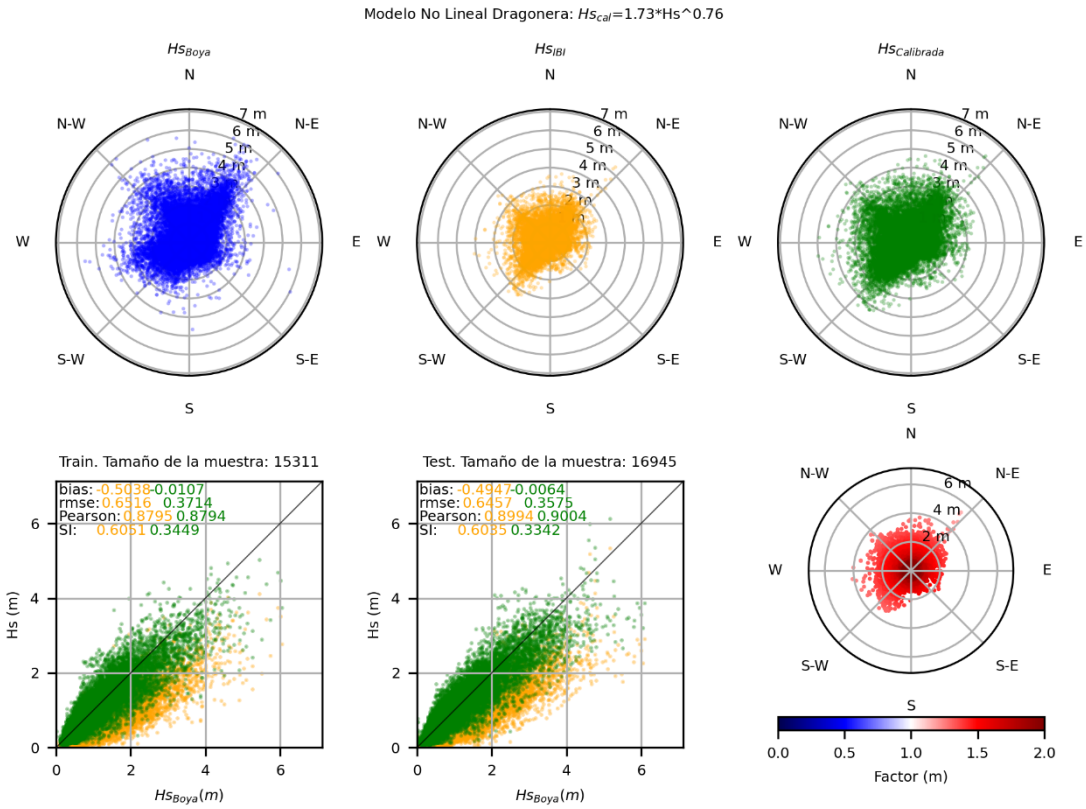

Figura 20: Resultados del modelo no lineal para Dragonera con datos IBI

El modelo no lineal mejora las métricas establecidas respecto a los datos sin transformar como se aprecia en la figura 20, además, obtiene mejores resultados que el modelo lineal. Respecto al factor de calibración se observa como ya no es constante si no que varía en función a la altura de ola oscilando entre 1.73 y 2.

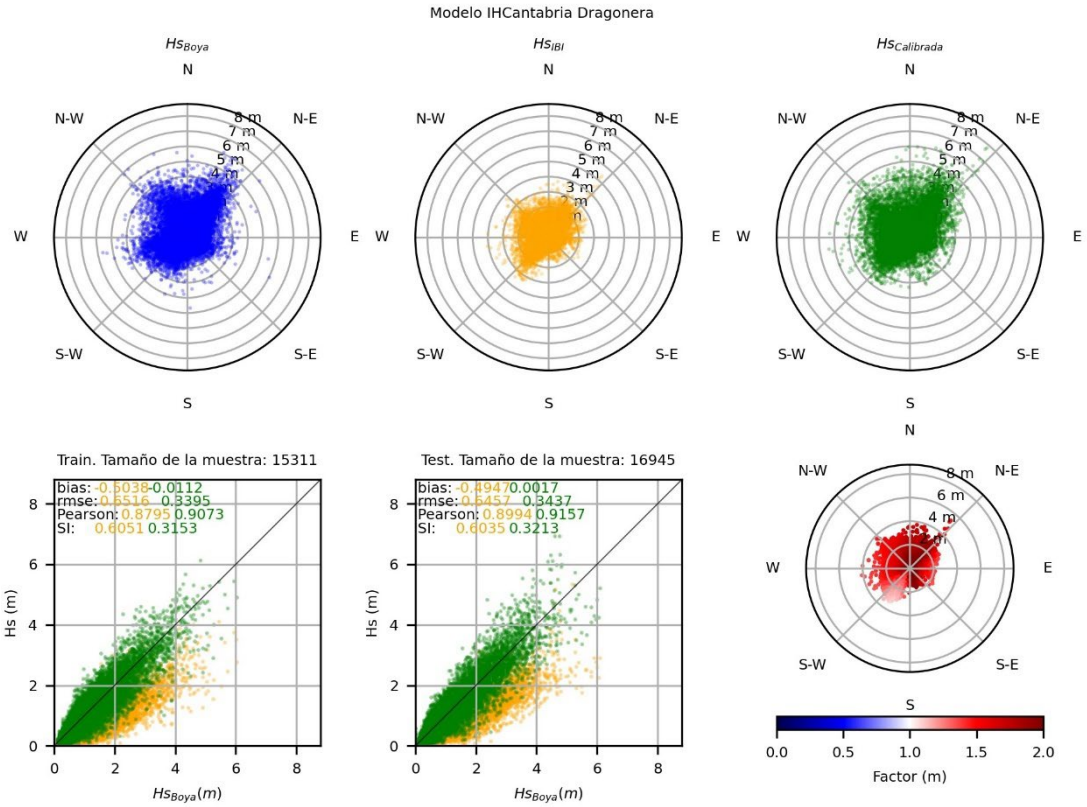

Figura 21: Resultados del modelo no lineal direccional para Dragonera con datos IBI

El modelo direccional no lineal, presentado en la figura 21, mejora las métricas de los anteriores modelos. Respecto al factor de calibración se observa cómo se obtiene una calibración direccional que le da valor y sentido al modelo. Por ejemplo, se observa como los oleajes entre Sur y Suroeste están sobreestimados en la fuente de reanálisis por lo que se le aplica un factor de correlación menor que al resto.

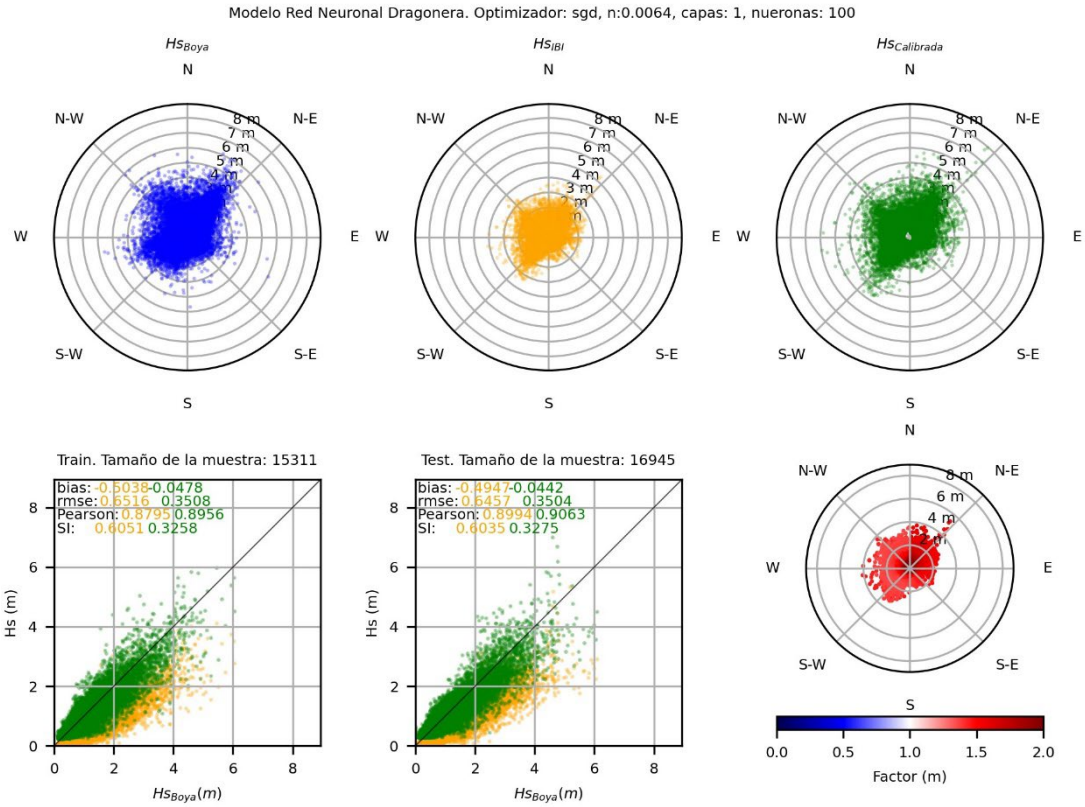

Figura 22: Resultados del modelo red neuronal para Dragonera con datos IBI

El modelo de red neuronal, figura 22, obtiene métricas similares al modelo direccional no lineal. Respecto al factor de calibración se observa en este caso que los oleajes con componentes del Oeste y mayor altura de ola se les aplica un menor factor que a los del Este. Una cosa negativa del modelo es que a los oleajes con altura de ola cercana a 0 se les aplica una factor de calibración alto.

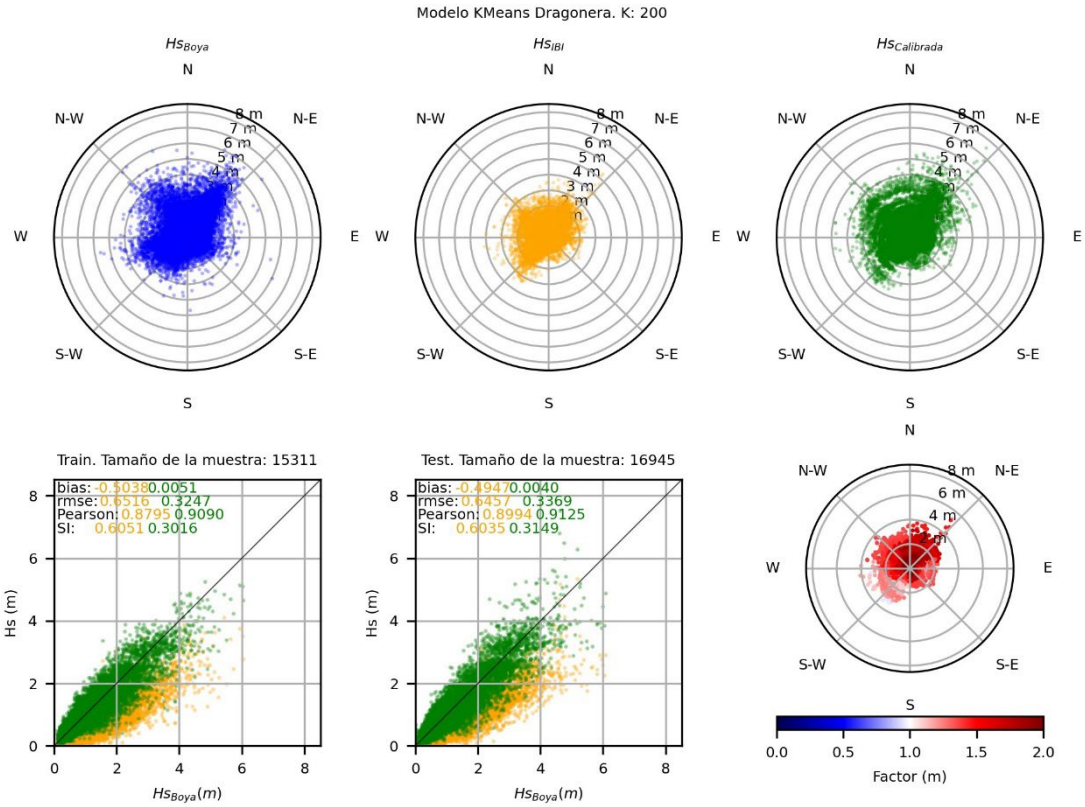

Figura 23: Resultados del modelo K-Means para Dragonera con datos IBI

El modelo de agrupamiento K-Means, figura 23, mejora o iguala todas las métricas de los modelos anteriores. Respecto al factor de calibración se observa que es más rico según las direcciones e intensidad del oleaje y replica en gran medida la corrección direccional no lineal.

Como se ha comentado, se ha aplicado el modelo para las 15 localización y para las dos fuentes de información. La evaluación de cada métrica se recoge en el ANEXO I - Resultados. A modo de resumen comparativo, la tabla 10, presenta la media de cada métrica para cada modelo.

Modelo	Bias	RMSE	Pearson	SI
Lineal	0.121 ± 0.09	0.37 ± 0.07	0.91 ± 0.05	0.27 ± 0.10
No Lineal	0.021 ± 0.05	0.33 ± 0.06	0.91 ± 0.05	0.24 ± 0.07
Dir. No Lineal	0.013 ± 0.07	0.34 ± 0.07	0.93 ± 0.04	0.24 ± 0.07
Red neuronal	-0.026 ± 0.02	0.31 ± 0.06	0.92 ± 0.04	0.23 ± 0.06
K-Means	0.028 ± 0.05	0.32 ± 0.07	0.93 ± 0.04	0.23 ± 0.07

Tabla 10: Evaluación de las métricas para cada modelo

Para seleccionar el mejor respecto al problema de asimilación, se visualiza a través de un gráfico boxplot, figura 24, que muestra la distribución a través de cinco puntos clave: el primer cuartil (Q1), la mediana, el tercer cuartil (Q3), los valores mínimos y máximos (excluyendo atípicos), y los valores atípicos, que se representan como puntos fuera de los bigotes. La caja refleja la dispersión central (IQR), y la posición de la mediana dentro de la caja indica la simetría (si está centrada) o sesgo de la distribución. Los bigotes muestran el rango de los datos normales, mientras que los puntos fuera de ellos destacan valores extremos. Los valores que se refieren a Base indican el valor inicial de cada métrica con el que se compara cada modelo.

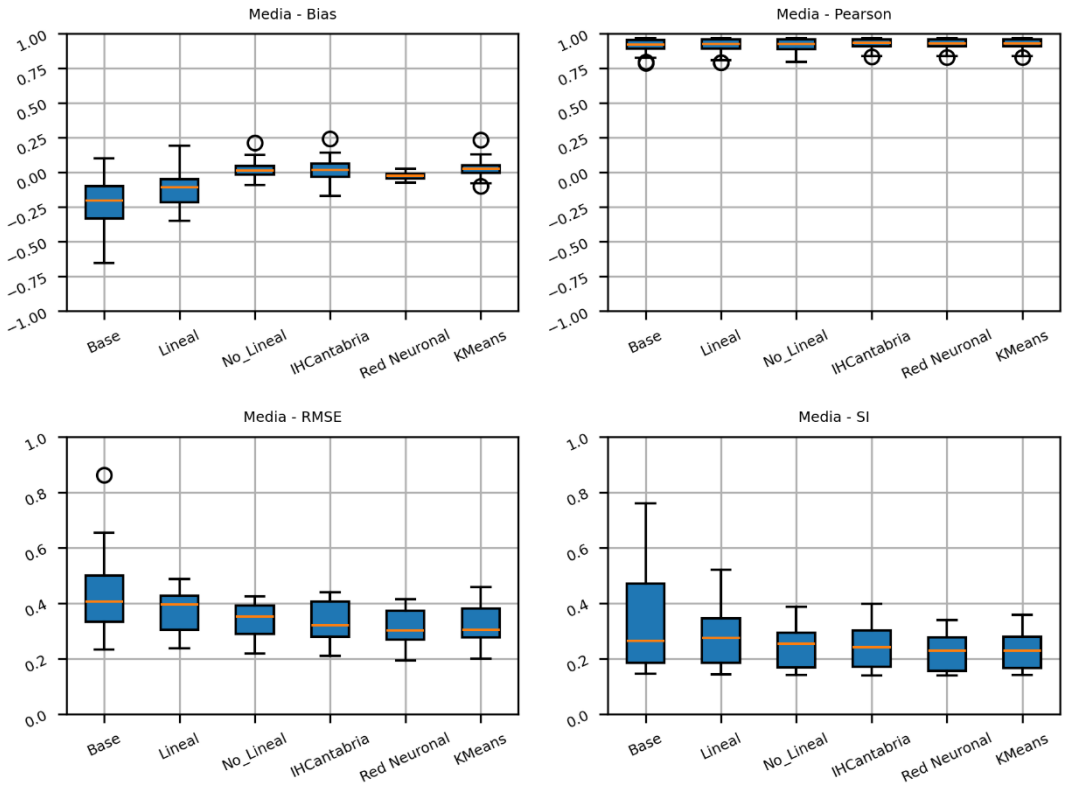

Figura 24: Comparativa de la evaluación de los modelos

Respecto al Bias, todos los modelos tienen sesgos cercanos a cero y mejoran el de los datos sin calibrar, destacar los modelos direccional, red neuronal y K-Means con los valores más cercanos a cero y con menor variabilidad, destacando sobre los demás el modelo de red neuronal.

Respecto al RMSE, los modelos lineal y no lineal tienen un valor más cercano a los datos sin calibrar, aunque con menor variabilidad que el de los datos sin calibrar. El resto de los modelos presentan valores menores y más similares entre sí, indicando una mejor precisión en sus predicciones.

Respecto al coeficiente de correlación de Pearson, todos los modelos muestran coeficientes muy altos, cercanos a 1, lo que sugiere una fuerte correlación lineal, similar al valor de referencia.

Respecto, al índice de dispersión, el modelo lineal tiene un valor más elevado, el modelo no lineal y el modelo direccional tienen un valor similar y algo más elevado que el modelo de red neuronal y el modelo K-Means que presentan valores menores, lo que indica una menor variabilidad y una mayor consistencia.

En resumen, los modelos mejoran en todas las métricas en comparación con la inicial, siendo el modelo de red neuronal y el modelo K-Means los que muestran un rendimiento superior y más consistente.

4. CONCLUSIONES Y LÍNEAS FUTURAS

En este trabajo se ha demostrado el rendimiento de diferentes técnicas de inteligencia artificial aplicadas a la asimilación de datos de oleaje. A continuación, se presentan las conclusiones específicas:

- Se han obtenido y preprocesado diferentes fuentes de información de modelos numéricos obteniendo los datos cercanos a la fuente de información instrumental, con una resolución temporal y espacial apta para enfrentarse a problemas de la ingeniería de costa.
- El modelo lineal y el modelo no lineal son modelos monoparamétricos que no se pueden extrapolar a otras dimensiones. Respecto al modelo direccional no lineal es un modelo multiparamétrico que tampoco se puede extrapolar a otras dimensiones. Por último, el modelo de red neuronal y el modelo de K-Means son modelos multiparamétricos que si se pueden extrapolar a otras dimensiones.
- La inclusión de la variable dirección dentro de los modelos ha permitido mejorar el valor de las métricas contribuyendo a una asimilación de datos más precisa y ajustada a las condiciones reales.
- El modelo de red neuronal, en general, ha encontrado la configuración optima en la red más sencilla planteada.
- El modelo de red neuronal y el modelo K-Means ha demostrado un buen funcionamiento en la generalización de su configuración demostrando no ser dependiente de la localización y fuente de información.
- El modelo de red neuronal y el modelo K-Means al aplicarse a series continuas pueden presentar discontinuidades en la serie de resultados.
- El modelo direccional, el modelo de red neuronal y el modelo K-Means han mostrado un mayor rendimiento al mejorar significativamente todas las métricas en comparación con los datos no calibrados y con el resto de los modelos.

- Seleccionar un modelo ganador generalizado no es evidente, dependiendo de la localización o fuente de información a ajustar varía el modelo ganador.

Líneas Futuras

Para futuros trabajos, se sugiere explorar las siguientes líneas de investigación:

- **Integración de nuevas variables:** Incorporar otras variables meteorológicas y oceanográficas que puedan influir en la dinámica del oleaje, como el periodo o la velocidad y dirección del viento, para mejorar la precisión de los modelos.

- **Mejora de modelos:** seguir explorando los diferentes parámetros de configuración de las técnicas seleccionadas para conseguir unos resultados más consistentes.

- **Incluir diferentes modelos híbridos:** Desarrollar e implementar modelos híbridos que combinen las fortalezas de diferentes técnicas de inteligencia artificial, como la combinación de redes neuronales con técnicas de agrupamiento para captar tanto las relaciones lineales como no lineales de los datos.

- **Ampliación de modelos:** incorporar nuevas técnicas que integren la dependencia temporal de los datos como puede ser las redes neuronales recurrentes, más específicamente, Long Short-Term Memory (LSTM).

Estas líneas futuras permitirán no sólo mejorar los modelos actuales, sino también ampliar el conocimiento sobre la asimilación de datos de oleaje, contribuyendo a un mejor manejo y predicción de los fenómenos oceánicos.

5. BIBLIOGRAFÍA

Copernicus. (2024). *Atlantic -Iberian Biscay Irish- Ocean Wave Reanalysis.* doi:https://doi.org/10.48670/moi-00030

IHCantabria, I. d. (2021). *Global Ocean Wave (GOW).* Obtenido de https://ihdata.ihcantabria.com/wave-data/

Mínguez, R., Espejo, A., Tomas, A., Méndez, F., & Losada, I. (2011). Directional calibration of wave reanalysis databases using instrumental data. *JOURNAL OF ATMOSPHERIC AND OCEANIC TECHNOLOGY*.

Perez, J., Menendez, M., & Losada, I. (2017). GOW2: A global wave hindcast for coastal applications. *Coastal Engineering 124C*. doi:10.1016/j.coastaleng.2017.03.005

Puertos del Estado. (2024). *La red de boyas costeras: descripción de la red y las series temporales disponibles (REDEXT).* Obtenido de https://bancodatos.puertos.es/BD/informes/INT_2.pdf

Tolman, H. L. (1991). A third-generation model for wind waves on slowly varying, unsteady and inhomogeneous depths and currents. *J. Phys. Oceanogr*, 21, 782-797.

Tomas, A. (2009). Metodologías de calibración de bases de datos de reanálisis de clima marítimo (Tesis doctoral). Obtenido de http://www.tesisenred.net/TDR-0628110-134140

ANEXO I – RESULTADOS

BIAS

Nombre	IBI						GOW					
	Base	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*	*Base*	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*
Bilbao-Vizcaya	-0.300	-0.095	-0.017	-0.130	-0.036	-0.003	-0.089	0.021	0.031	0.076	-0.011	0.031
Cabo de Peñas	-0.158	-0.263	-0.068	-0.172	-0.047	-0.079	-0.314	-0.055	0.018	-0.038	-0.055	0.015
Estaca de Bares	-0.194	-0.258	-0.091	-0.157	-0.045	-0.100	0.066	-0.010	0.027	-0.059	-0.012	0.036
Villano-Sisargas	-0.177	-0.160	-0.015	-0.049	-0.031	-0.018	0.101	0.067	0.125	0.104	-0.019	0.129
Cabo Silleiro	-0.214	-0.071	-0.031	-0.046	-0.038	-0.028	-0.017	0.091	0.125	0.142	0.010	0.131
Golfo de Cádiz	-0.296	-0.338	0.008	-0.009	-0.047	0.042	-0.087	-0.166	-0.005	0.013	-0.013	0.003
Cabo de Gata	-0.495	-0.193	0.010	-0.007	-0.024	0.022	-0.139	-0.062	-0.027	0.022	-0.042	-0.029
Cabo de Palos	-0.437	-0.263	0.011	-0.028	-0.037	0.016	-0.088	-0.065	0.046	0.040	-0.020	0.051
Valencia	-0.530	-0.297	0.036	0.034	-0.044	0.061	-0.339	-0.158	0.041	0.065	-0.013	0.068
Tarragona	-0.498	-0.349	-0.024	-0.035	-0.053	-0.031	-0.181	-0.098	0.033	0.030	-0.015	0.039
Cabo Begur	-0.654	-0.206	-0.008	0.013	-0.076	0.024	-0.240	0.194	0.214	0.244	0.026	0.235
Dragonera	-0.499	-0.286	-0.006	0.002	-0.044	0.004	-0.234	-0.047	0.063	0.075	-0.021	0.062
Mahón	-0.483	-0.216	-0.024	0.031	-0.043	0.006	-0.127	0.017	0.048	0.080	-0.022	0.053
Gran Canaria	-0.233	-0.137	-0.037	-0.031	-0.011	-0.039	-0.001	-0.030	0.060	0.074	-0.026	0.064
Tenerife	-0.083	-0.098	0.047	0.054	0.005	0.052	-0.187	-0.113	0.042	0.062	-0.005	0.039

Tabla 11: Evaluación de BIAS

RMSE

Nombre	IBI						GOW					
	Base	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*	*Base*	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*
Bilbao-Vizcaya	0.48	0.38	0.37	0.41	0.37	0.37	0.36	0.38	0.38	0.39	0.37	0.38
Cabo de Peñas	0.38	0.45	0.37	0.38	0.34	0.36	0.51	0.42	0.42	0.42	0.41	0.40
Estaca de Bares	0.44	0.49	0.42	0.43	0.39	0.41	0.38	0.40	0.40	0.44	0.40	0.41
Villano-Sisargas	0.42	0.43	0.40	0.40	0.38	0.38	0.41	0.41	0.42	0.43	0.40	0.43
Cabo Silleiro	0.40	0.36	0.36	0.37	0.36	0.36	0.36	0.42	0.42	0.43	0.37	0.42
Golfo de Cádiz	0.46	0.48	0.34	0.30	0.29	0.28	0.27	0.30	0.26	0.24	0.24	0.24
Cabo de Gata	0.64	0.40	0.36	0.32	0.30	0.30	0.33	0.29	0.29	0.28	0.29	0.29
Cabo de Palos	0.53	0.39	0.30	0.31	0.28	0.28	0.27	0.27	0.27	0.29	0.27	0.28
Valencia	0.63	0.41	0.30	0.32	0.27	0.29	0.45	0.34	0.31	0.28	0.26	0.28
Tarragona	0.60	0.46	0.30	0.30	0.28	0.28	0.30	0.26	0.24	0.24	0.23	0.23
Cabo Begur	0.86	0.47	0.42	0.43	0.41	0.41	0.46	0.40	0.40	0.44	0.32	0.46
Dragonera	0.65	0.46	0.36	0.34	0.35	0.34	0.40	0.31	0.31	0.28	0.28	0.28
Mahón	0.65	0.43	0.37	0.38	0.37	0.36	0.33	0.30	0.31	0.31	0.30	0.31
Gran Canaria	0.33	0.28	0.24	0.24	0.24	0.25	0.24	0.25	0.26	0.27	0.25	0.26
Tenerife	0.23	0.24	0.22	0.21	0.19	0.20	0.29	0.24	0.22	0.21	0.20	0.20

Tabla 12: Evaluación de RMSE

Coeficiente de pearson

Nombre	IBI						GOW					
	Base	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*	*Base*	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*
Bilbao-Vizcaya	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
Cabo de Peñas	0.95	0.96	0.96	0.96	0.96	0.96	0.94	0.94	0.94	0.94	0.95	0.95
Estaca de Bares	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
Villano-Sisargas	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.95	0.96	0.96
Cabo Silleiro	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.96
Golfo de Cádiz	0.85	0.86	0.86	0.91	0.90	0.91	0.92	0.92	0.92	0.93	0.93	0.93
Cabo de Gata	0.88	0.88	0.88	0.91	0.92	0.92	0.91	0.93	0.93	0.93	0.93	0.93
Cabo de Palos	0.90	0.89	0.89	0.89	0.90	0.90	0.92	0.91	0.91	0.90	0.91	0.91
Valencia	0.83	0.84	0.82	0.84	0.85	0.84	0.83	0.81	0.80	0.88	0.86	0.86
Tarragona	0.86	0.87	0.87	0.89	0.89	0.89	0.91	0.92	0.92	0.93	0.93	0.93
Cabo Begur	0.91	0.91	0.91	0.92	0.92	0.92	0.94	0.96	0.96	0.96	0.95	0.95
Dragonera	0.89	0.90	0.90	0.92	0.91	0.91	0.92	0.93	0.93	0.95	0.94	0.94
Mahón	0.92	0.94	0.94	0.94	0.94	0.94	0.95	0.96	0.96	0.96	0.96	0.96
Gran Canaria	0.93	0.93	0.93	0.93	0.93	0.93	0.92	0.92	0.93	0.92	0.93	0.93
Tenerife	0.79	0.79	0.80	0.83	0.84	0.84	0.79	0.79	0.80	0.84	0.83	0.83

Tabla 13: Evaluación del coeficiente de pearson

Índice de dispersión

Nombre	IBI						GOW					
	Base	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*	*Base*	*Lineal*	*No Lineal*	*IHC*	*ANN*	*K-Means*
Bilbao-Vizcaya	0.25	0.19	0.18	0.20	0.18	0.18	0.18	0.19	0.19	0.19	0.18	0.18
Cabo de Peñas	0.19	0.21	0.17	0.18	0.16	0.17	0.25	0.20	0.20	0.20	0.19	0.19
Estaca de Bares	0.18	0.19	0.16	0.16	0.15	0.16	0.15	0.15	0.15	0.17	0.15	0.16
Villano-Sisargas	0.17	0.16	0.15	0.15	0.14	0.14	0.16	0.15	0.16	0.16	0.15	0.16
Cabo Silleiro	0.17	0.14	0.14	0.15	0.14	0.14	0.15	0.16	0.17	0.17	0.15	0.17
Golfo de Cádiz	0.38	0.38	0.27	0.24	0.23	0.23	0.22	0.24	0.20	0.19	0.19	0.19
Cabo de Gata	0.62	0.39	0.35	0.31	0.29	0.29	0.32	0.29	0.28	0.27	0.28	0.28
Cabo de Palos	0.52	0.39	0.29	0.31	0.28	0.28	0.26	0.27	0.27	0.29	0.26	0.27
Valencia	0.76	0.50	0.37	0.40	0.34	0.36	0.55	0.43	0.39	0.35	0.32	0.34
Tarragona	0.67	0.52	0.34	0.34	0.32	0.32	0.34	0.30	0.28	0.28	0.26	0.26
Cabo Begur	0.66	0.35	0.32	0.33	0.31	0.32	0.35	0.30	0.31	0.33	0.24	0.35
Dragonera	0.60	0.43	0.33	0.32	0.33	0.31	0.38	0.29	0.29	0.26	0.26	0.26
Mahón	0.50	0.31	0.27	0.28	0.27	0.26	0.26	0.22	0.23	0.23	0.22	0.22
Gran Canaria	0.20	0.16	0.14	0.14	0.14	0.14	0.15	0.15	0.15	0.16	0.14	0.15
Tenerife	0.27	0.28	0.25	0.25	0.23	0.23	0.33	0.28	0.25	0.25	0.23	0.24

Tabla 14: Evaluación del Índice de dispersión