LA IMPORTANCIA DEL DATO EN LA SIMULACIÓN FLUIDODINÁMICA DE PLATAFORMAS FLOTANTES PARA ENERGÍAS RENOVABLES MARINAS

JESÚS MARÍA BLANCO

Universidad del País Vasco

ANGELA BERNARDINI

NAITEC

LANDER GALERA-CALERO

Universidad del País Vasco

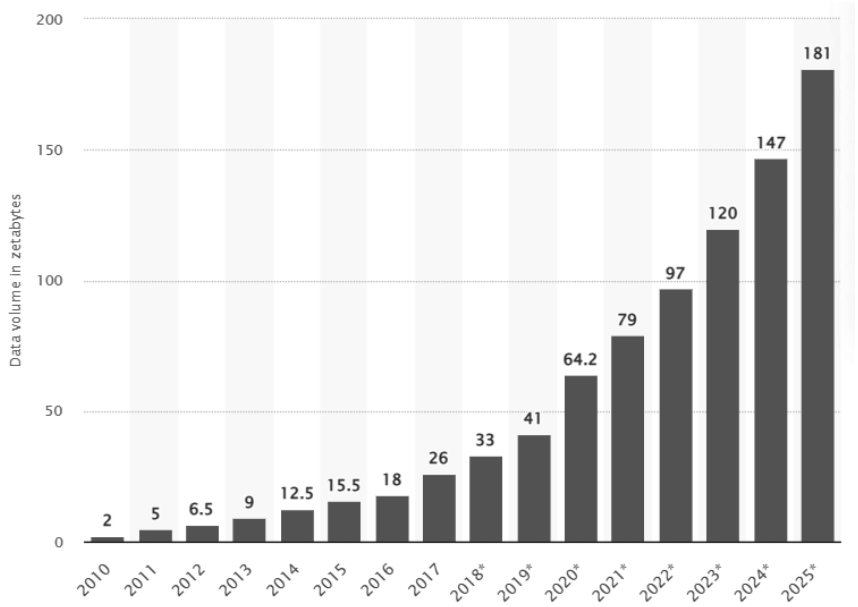
El dato puede ser muchas cosas: números, medidas, observaciones, declaraciones, valores, cualidades, opiniones y cualquier cosa que se pueda recopilar y comparar. Es una colección de hechos introducida de tal forma que un ordenador puede procesarlos. El dato no es otra cosa que información. La información es poder y su explotación es fuente de ventajas competitivas.

En 2017, The Economist publicó un artículo cuyo título decía «El recurso más valioso del mundo ya no es el petróleo, sino los datos». Desde su publicación, el tema ha generado mucha discusión, y «Los datos son el nuevo petróleo» se ha convertido en un slogan. El problema es que la discusión generalmente se enfoca en por qué esto es algo malo. Existen preocupaciones legítimas sobre cómo los gigantes tecnológicos están explotando lo que saben sobre nosotros, pero al mismo tiempo, existen innumerables formas en las que todos estos datos pueden mejorar (y de hecho lo hacen) el mundo. Merece la pena detallar algunos ejemplos:

Un equipo de Google publicó en la revista «Nature» una investigación que compara la precisión en la detección del cáncer de mama realizada por sistemas de aprendizaje automático con la llevada a cabo por radiólogos expertos (Scott *et al*, 2020). Los autores del estudio aseguran que su sistema de inteligencia artificial (IA) superó tanto las decisiones históricas tomadas por los radiólogos que habían evaluado los mamogramas con anterioridad, como las decisiones que tomaron otros seis radiólogos que debían interpretar 500 casos elegidos de forma aleatoria, y registró una reducción del 3,5% en los falsos positivos y del 8,1% en los falsos negativos. Ese sistema de IA prepara el camino para que los ensayos clínicos mejoren la precisión y la eficiencia del cribado del cáncer de mama. Para esto, los investigadores tuvieron que acumular grandes cantidades de datos a partir de los cuales entrenaron sus modelos de aprendizaje automático.

Según el Smart City Index (2020), Singapur, seguida por Helsinki y Zúrich, es la ciudad más inteligente y eficiente del planeta en términos de movilidad, salud, seguridad y productividad. La isla lidera el ranking por haber implantado medidas tan punteras como: soluciones inteligentes de control de tráfico, taxis autónomos, video vigilancia inteligente para detectar actividad delictiva o el Smart Health TeleRehab, un programa gracias al cual todos los habitantes de la tercera edad cuentan con dispositivos especiales

FIGURA 1
EXPLOSIÓN DE LA GENERACIÓN DE DATOS



Fuente: Statista

para realizar consultas médicas en cualquier momento. La idea es aprovecharse de la enorme cantidad de datos generados por la población, las redes wifi y los dispositivos conectados para implementar mejoras en los servicios de una ciudad y tratar de convertirla en más sostenible, conectada y segura.

El mundo ha cambiado completamente, y también han cambiado los datos que hacen que el mundo gire. A medida que el COVID-19 recorría el mundo haciendo estragos casi todos los aspectos de la vida, desde el trabajo hasta el ejercicio, se trasformaron de forma abrupta, haciendo que la gente dependiera cada vez más de aplicaciones móviles e internet para socializar, educar y entretenerse. Los datos se generan constantemente en clics en anuncios, reacciones en las redes sociales, acciones, viajes, transacciones, contenido de transmisión y mucho más. Examinar estos datos puede ayudar a comprender un mundo que se mueve a velocidades cada vez mayores.

Los dispositivos inteligentes conectados que interactúan entre sí y con nosotros mientras recopilan todo tipo de datos, han experimentado un claro auge desde la última década (en 2021 hay más de 10 mil millones de dispositivos IoT (Internet of Things) activos; se estiman que superarán los 25,4 mil millones en 2030), siendo uno de los principales impulsores de la explosión de generación de datos (Figura 1).

Está claro que los datos realmente son el nuevo petróleo y que el principal impacto en la humanidad es cómo los datos pueden mejorar nuestras vidas. Pero tampoco hay que darles todo el mérito a los datos, porque a veces engañan o puede que nos dejemos engañar.

Los datos pueden ser irrelevantes: Los datos irrelevantes son aquellos que en realidad no se necesitan y no encajan en el contexto del problema que se está tratando de resolver. Los datos pueden ser incorrectos. Los datos casi siempre son incorrectos y se necesita averiguar en qué porcentaje y grado. Los datos pueden interpretarse incorrectamente. Debido a que los datos deben interpretarse y analizarse, existe una elevada probabilidad de que se interpreten incorrectamente, es decir los datos se pueden usar incorrectamente. Esto es consecuencia del punto anterior, pero se trata de acciones que son respuestas de algoritmos. Un ejemplo claro es el algoritmo COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), por su acrónimo en inglés, que en español puede traducirse como Administración de Perfiles de Criminales para Sanciones Alternativas, correspondiente al sistema de prisiones de Estados Unidos. Este programa evalúa la probabilidad de que un criminal se convierta en reincidente. Hay docenas de estos algoritmos de evaluación de riesgos en uso y existen herramientas líderes a nivel internacional ofrecidas por proveedores comerciales.

Sin embargo, existen estudios como (Christoph *et al*, 2021) sobre estas herramientas con el objetivo de descubrir la precisión subyacente de su algoritmo de reincidencia y probar si el algoritmo está sesgado contra ciertos grupos. En general todos ellos llegan a la conclusión de que los acusado de raza negra son mucho mas propenso a ser juzgado incorrectamente, presentan mas probabilidad de ser eveluados incorrectamente de bajo riesgo. Cuando una mala interpretación de los datos lleva a conclusiones erró-

neas, se dice que los datos son engañosos. A menudo, esto es el resultado de datos incompletos o falta de contexto.

De hecho, lo importante no es el dato, sino el problema que se quiere resolver con ello. ¡El contexto importa!. Los datos ni siquiera son el punto de partida. El punto de partida es siempre la pregunta y la hipótesis. Cuanto más claro se tenga lo que se quiere lograr, más claro serán los datos que se necesitan. No será cualquier dato, sino los datos que son específicos de esa hipótesis. Luego se analiza, se construye, se predice y se sacan conclusiones, llegando entonces al siguiente problema. Normalmente en este punto, en caso de no tener datos disponibles, empieza un proceso de reflexión para determinar qué datos considerar, es decir qué variables medir. Un científico de datos puede participar en esta reflexión aportando intuición matemático-estadística, pero normalmente el dueño de la información, siendo éste el que más conoce su aplicación, proceso y/o producto, encuentra fácilmente y tiene claras cuales son estas variables. Otra cosa es que quiera y/o pueda considerarlas todas… Es decir, hay variables que no se consideran porque es caro medirlas, otras porque resultan poco influyentes, otras porque no es posible actuar sobre ellas, etc. Una vez seleccionadas las variables la recopilación de los datos utiliza todas las fuentes relevantes a través de casi cualquier método, desde la entrada manual y el web scraping hasta la captura de datos de sistemas y dispositivos en tiempo real. Por otro lado, la ingeniería virtual es ya una herramienta estándar (McCorkle *et al*, 2006) y (Huang *et al*, 2004). Las simulaciones permiten el desarrollo, la validación y la prueba de productos y procesos ahorrando tiempo, reduciendo costes y aumentando la calidad. Por esta razón es común utilizar modelos virtuales para crear datos que alimentarán algoritmos predictivos (Bernardini *et al*, 2012) y (Parra *et al*, 2010).

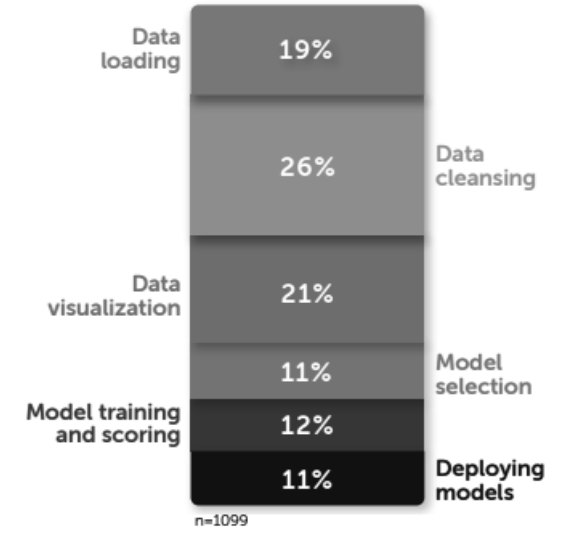
El problema a resolver está claro, así como los datos a medir, sin embargo, la cuestión ahora es el volumen de datos necesarios, la pregunta del millón. La mayoría de las empresas no hacen Big Data, no están preparadas para la captura de zeta bytes de datos. Recompilar datos cuesta tiempo y dinero. Sin embargo, Banco y Brill, (2001), en un estudio sobre procesamiento del lenguaje natural ya pusieron de manifiesto que no resulta ganador quien tiene el mejor algoritmo, sino quien tiene más datos. Es de conocimiento común que muy pocos datos de entrenamiento dan como resultado una aproximación pobre. ¿Pero cuanto entonces? En realidad, el número de datos depende del problema. Pero hay alguna regla general:

- Depende del espacio de inferencia. Si se necesita predecir el consumo energético de una pequeña empresa, solo se necesitará la información de esa pequeña empresa, sin embargo, si se necesita predecir la calidad del producto de una multinacional, presente en todo el mundo, se necesitará la información de cada una de las plantas de la multinacional.
- Modelos complejos requieren más datos. Por ejemplo, la regresión es un modelo más simple que un árbol de decisión. La regresión logística asume cierta relación entre las características y el valor a predecir. Por otro lado, un RandomForest es más complejo en el sentido de que utiliza un conjunto de árboles de decisión para realizar las predicciones.
- El número de datos tiene que superar el número de atributos. Existen métodos heurísticos estadísticos, principalmente para problemas de clasificación, que permiten calcular un tamaño de muestra adecuado, el número de ejemplos tiene que ser mayor que el cuadrado del número de características, o al menos un orden de magnitud más. Sin embargo, todos parecen factores de escala ad-hoc.

Diríamos entonces que cuantos más atributos más datos y, en general, que se utilicen tantos datos como se pueda. Cuanto más grande sea el tamaño de la muestra, mejores serán los resultados, mientras esto no afecte el rendimiento.

Los datos darán respuestas fiables si son fiables y de calidad. La cantidad importa, pero ¡importa aún más la calidad!. En este campo, es común escuchar la expresión «basura adentro, basura afuera», o «GIGO» como abreviatura de Garbage In-Garbage Out. Intuitivamente, esto lleva a pensar que se trata de una cuestión de precisión, es decir, si los datos son correctos o no. Sin embargo, la calidad de los datos va mucho más allá. Incluyen factores (dimensiones) adicionales como: la integridad, la consistencia y la puntualidad según Catarci y Scannapieco, (2002). Desde la perspectiva de la investigación, muchos académicos han identificado diversas metodologías y marcos para evaluar y mejorar la calidad de los datos a través de diferentes técnicas y estrategias (Batini *et al*, 2009) y (Pipino *et al*, 2002). Cabe notar que hubo ciertas discrepancias en la definición de las dimensiones de la calidad de los datos, tanto que desde 1985 hasta 2009 se propusieron unas cuarenta diferentes. Pero centrémonos en las relevantes. El término exactitud se refiere al grado en que los datos reflejan con precisión un evento u objeto descrito. La integridad de los datos es el porcentaje de todos los datos requeridos disponibles en un conjunto de datos. Pocos conjuntos de datos están completos al 100%. La integridad está entonces relacionada con los valores faltantes, valores que existen en el mundo real pero que no están disponibles en una recopilación de datos. Con el fin de caracterizar la integridad, es importante comprender por qué falta el valor. El valor puede faltar porque existe, pero no se conoce, o porque no existe, o porque no se sabe si existe. Coherencia de los datos significa que existe lógica en la medición de las variables en to-

FIGURA 2
CÓMO EMPLEAN SU TIEMPO LOS CIENTÍFICOS DE DATOS



Fuente: Anaconda, 2020

dos los conjuntos de datos. Esto se convierte en una preocupación especialmente cuando los datos se recopilan de múltiples fuentes. Las discrepancias en el significado de los datos entre las fuentes de datos pueden crear conjuntos de datos inexactos y poco fiables. Otro aspecto importante de los datos es su actualización a lo largo del tiempo. Si la información está disponible justo cuando se necesita, los datos son puntuales. La puntualidad se puede medir como el tiempo entre el momento en que se espera la información y el momento en que está disponible para su uso lo cual puede resultar un aspecto crítico en algunas centrales termoeléctricas (Blanco, *et al*, 2013) y (Vázquez, *et al*, 2015).

El paradigma GIGO toma una importancia especial en un contexto predictivo. Es decir, si modelos de aprendizaje automático, o más en general de Inteligencia Artificial, se entrenan en conjuntos de datos que carecen de calidad, difícilmente serán exitosos e incluso podrán devolver predicciones incorrectas que a su vez llevarán a conclusiones y toma de decisiones erróneas.

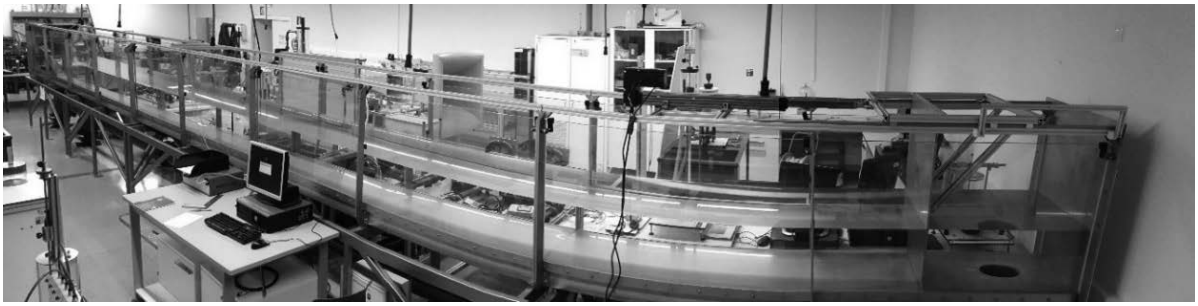
En la calidad de los datos juega un papel importante su limpieza, el proceso de preparar los datos para el análisis a fin de proporcionar resultados precisos. Como dijo Jeffrey Heer, profesor de informática en la Universidad de Washington y cofundador de Trifacta, empresa dedicada al desarrollo de aplicaciones para la exploración y preparación de datos, entrenar un algoritmo con datos sin procesar y pensar obtener información valiosa es un mito. Esta tarea puede llegar a ser tediosa, ya que no se trata simplemente de borrar información para dejar espacio a nuevos datos, sino de encontrar la forma de maximizar la precisión del conjunto sin eliminar necesariamente la información. Además, si bien se sigue una metodología general para limpiar los datos, los pasos pueden variar según el tipo de datos. Esto explica el resultado de la encuesta realizada por Anaconda (2020), según la cual las tareas de preparación hoy en día siguen ocupando la mayor parte del tiempo de los científicos de datos (Figura 2).

PREPARACIÓN DE LOS DATOS

El primer paso en el tratamiento de los datos debería ser la eliminación de los datos irrelevantes. Los datos irrelevantes son aquellos que pueden no encajar en el contexto del problema. Eliminar columnas no necesarias puede reducir tiempos de procesamiento y devolver una salida de forma más rápida. A veces, cuando se combinan conjuntos de datos puede producirse una duplicación. Los valores duplicados son similares a los irrelevantes. Aumenta el volumen de datos, lo que a su vez incrementa los tiempos de procesamiento. Los errores estructurales incluyen errores ortográficos, nomenclaturas incongruentes, uso incorrecto de palabras, etc. Estos pueden afectar el análisis porque, si bien pueden ser obvios para los humanos, la mayoría de los algoritmos de aprendizaje automático no reconocerían los errores. De manera similar, es necesario estandarizar fechas, direcciones, números de teléfono, etc. A menudo aparecen valores atípicos, es decir valores que no parecen encajar dentro de los datos que se están analizando. Sin embargo, el hecho de que exista un valor atípico no significa que sea incorrecto. Si hay una razón legítima para eliminarlo, por ejemplo si se trata de una entrada incorrecta, hacerlo ayudará.. En un conjunto de datos siempre faltan valores, por lo que es necesario encontrar la forma de manejarlo. Existen dos formas de hacerlo, o bien se pueden eliminar las observaciones que tienen valores faltantes, pero al hacerlo se eliminará o perderá información o se pueden determinar los valores faltantes en base a otras observaciones, pero existe la posibilidad de perder la integridad de los datos ya que se estaría operando desde suposiciones y no desde observaciones reales (Loshin, 2013). Después de la limpieza, los datos se transforman en un formato adecuado. Dicha transformación puede ser simple o muy compleja según los cambios necesarios. La estandarización, la conversión de caracteres, la codificación, la combinación de campos, la conversión de unidades de medidas a un formato estándar o la agregación, son algunas de las tareas involucradas en la transformación de datos.

Hay que decir que existe una gran variedad de herramientas disponibles en el mercado para respaldar la limpieza y transformación de datos. Por lo general estas herramientas se concentran en dominios específicos, como limpiar los campos nombre y dirección, o una fase de limpieza específica, como la eliminación de duplicados. Debido a su dominio restringido, las herramientas especializadas suelen funcionar muy bien, pero aún requieren un esfuerzo manual de programación para abordar un amplio espectro de pro-

FIGURA 3
TANQUE DE OLAS UBICADO EN EL LABORATORIO DE MECÁNICA DE FLUIDOS DEL DEPARTAMENTO DE INGENIERÍA ENERGÉTICA DE LA ESCUELA DE INGENIERÍA DE BILBAO



Fuente: Elaboración propia

blemas de transformación y limpieza (Rahm, 2020). Los nuevos enfoques para la limpieza de datos se basan en una mayor interactividad con el usuario que va construyendo gradualmente el conjunto de datos a través de una interfaz gráfica de usuario (Herramientas ETL) que chequean paso a paso si los cambios conducen a discrepancias, lo que lleva a una limpieza de datos más eficaz (Han *et al*, 2012). Herramientas que parecen muy prometedoras. Cabe notar que hasta ahora sólo existe una pequeña investigación, concentrada entre el 2000 y 2010, sobre la limpieza de datos, aunque la gran cantidad de herramientas existentes es un claro indicativo tanto de la importancia como de la dificultad del problema.

APLICACIÓN A LAS ENERGÍAS RENOVABLES MARINAS

Estudios previos a nivel de laboratorio

La importancia de los datos, y su correcto uso, transciende a todas las áreas imaginables de nuestra sociedad, como bien ha sido explicado anteriormente. Por ello, los datos pueden ser utilizados para luchar contra una de las mayores crisis que estamos afrontando como sociedad, el cambio climático. Como un claro ejemplo de la importancia de los datos frente a este problema fué recogido en el estudio de Mark Z. Jacobson (2017) en el que se mostraba la posibilidad de que 139 países basaran el 100% de su sistema energético en energías renovables. Lejos de presentar algo parecido, en este artículo los autores se centran en una tecnología concreta, y todo lo que le rodea, la cual consideran que va a tener un papel muy importante en este contexto, como es la eólica flotante marina.

Esta tecnología, que apenas está llegando a oídos del público en general, se encuentra en un momento crítico de desarrollo y madurez en el que el correcto uso se antoja vital a la hora de hacerla económicamente competitiva con respecto a otro tipo de energías, tanto renovables cómo no renovables. Así como los datos están a nuestro alrededor en nuestro día a día, también se pueden obtener infinidad de ellos en el estudio de este tipo de estructuras. Sin embargo, saber qué datos son los importantes para cada caso que se desee estudiar, es fundamental.

En este artículo, expondremos varios casos en los que el uso de datos experimentales y computacionales ha sido útil para el desarrollo y testeo de infraestructuras así como los diferentes retos que se tienen que afrontar para seguir evolucionando en su estudio.

Un claro ejemplo de esto es el trabajo desarrollado por Izquierdo *et al*. (2021). En este artículo el uso de los datos utilizados ha partido de intensas campañas experimentales-computacionales, siendo a su vez comparados con valores teóricos a escala de laboratorio, usando el canal de olas ó wave flume de 12,5 x 0,6 x 0,7 m (largo x ancho x alto) de la Escuela de Ingeniería de Bilbao que se muestra en la Figura 3. Los millones de datos obtenidos en diferentes puntos del canal sobre la variación de la superficie libre han de ser analizados y verificados antes de poder ser utilizados para su estudio. Una vez estos datos fueron verificados, el estudio del sistema de reflexión se pudo realizar tanto de manera computacional como experimental, permitiendo así, un estudio comparativo, y el funcionamiento del sistema de extinción a final del tanque, pudiendo así, obtener la mejor posición del sistema de extinción para cada tipo de condiciones del canal.

Además, el profundo conocimiento de los datos y variables analizadas permitieron a los investigadores encontrar tendencias que, a priori, no resultaban tan claras, pudiendo obtener más datos aprovechables para investigaciones futuras. Además, el conocimiento teórico de diferentes fórmulas permitió analizar el problema desde diferentes puntos de vista, pudiendo verificar el correcto funcionamiento tanto del tanque experimental como computacional.

Aproximaciones a la escala real

A pesar de que los canales de olas de dimensiones como el que acabamos de mencionar de la

Universidad del País Vasco están diseñados para trabajar en TRL (Technology Readiness Level) bajos, este tipo de estudios se hacen para llevarlas a última instancia a infraestructuras de mayores dimensiones. Un claro ejemplo es el estudio realizado por Wind *et al* (2020) en el que se resalta la importancia de la verificación y la validación previas al análisis computacional debido a la posible obtención de malos datos. Este trabajo presenta un concienzudo análisis experimental previo en el que se han tomado los datos de movimiento, fuerza del conversor de energía, presiones en la superficie de la boya así como el desplazamiento de la superficie libre del agua en varios puntos. Todo ello con el objetivo de realizar una comparativa numérica con los datos obtenidos y su posterior verificación y validación a escala 1:1.

Ejemplos como estos dejan clara la gran importancia que tiene la generación de datos, su correcta toma y análisis en el diseño de estructuras que, en un futuro próximo alcanzarán cientos de metros en varias dimensiones. Es importante resaltar que los estudios experimentales, aunque necesarios, son muy costosos y que, por ello, junto al incremento de la potencia computacional que se está observando en este siglo, la idea de los *digital twins, ó* gemelos digitales, está ganando mucha fuerza. Estos gemelos permiten, una vez validados, el análisis de diferentes tipos de condiciones que no se han podido hacer en los tanques experimentales, además de aportar un mayor conocimiento del comportamiento de las estructuras. Es importante resaltar que esto no es algo que solo se esté desarrollando en este área de la industria si no que está ganando fuerza a escalas de ciudades como Singapur, donde han conseguido realizar un gemelo digital de la ciudad que les permita estudiar cómo futuros cambios, como por ejemplo la inclusión de un nuevo bloque de oficinas o residencias, puede a afectar a diferentes áreas de la propia ciudad, como el tráfico, la luz, o la propia circulación del viento en las proximidades del nuevo edificio lo cual puede ser importante para la extinción de incendios, control de la temperatura, etc.

Esto es uno de los ejemplos que deja clara la gran importancia de los datos en el diseño de estructuras que, en un futuro, alcanzarán cientos de metros en varios sentidos. A pesar de que la infraestructura está diseñada para trabajar en TRLs bajos, el estudio presentado es algo que se debe hacer cada vez que prototipos a escala van a ser estudiados en canales de olas (paso previo a tener prototipos a escalas más grandes que son testeados en mar abierto).

La importancia de los datos experimentales para la verificación de las simulaciones en procesos de desarrollo e investigación queda fuera de toda duda. Hoy en día es imposible entender cualquier investigación con un enfoque computacional sin tener datos experimentales para verificarla y validarla. Este paso es muy importante a la hora de aportar credibilidad a estudios basados en las simulaciones de temas que no se han podido tratar durante las campañas experimentales, así como de profundizar en el conocimiento del comportamiento de la propia estructura.

Es importante resaltar que gran parte del diseño de este tipo de estructuras, así como de su optimización mediante diversas iteraciones se basa exclusivamente en modelos computacionales, resaltando así la grandísima importancia que tienen los datos y su calidad a la hora de verificarlos y ajustarlos para tener valores más realistas. Un claro ejemplo es la utilización de los datos experimentales como patrón a la hora de comparar diferentes enfoques teóricos en las simulaciones. Por ejemplo, los autores utilizaron los datos de desplazamiento de la superficie libre en diferentes puntos del canal de olas, así como el coeficiente de reflexión, coeficiente que indica cuanta energía de las olas incidentes es reflejada por el sistema de extinción, para hacer una comparativas de los diferentes modelos de turbulencia que se podían aplicar al sistema de ecuaciones RANS (Reynolds-Averaged Navier-Stokes), uno de los enfoques más populares de la dinámica de fluidos computacional, CFD (Computational Fluid Dynamics) por sus siglas en inglés, según muestra Galera-Calero *et al*, (2020).

En este escenario, una de las tecnologías que más rápidamente están ganando en popularidad es la inteligencia artificial o «machine learning». Es cierto que la importancia de esta tecnología no ha parado de crecer en los últimos años, así como la creencia errónea de que pueda abordar cualquier tipo de problema.

Sin embargo, enfocándonos en la CFD, es importante destacar que las redes neuronales, una de las muchísimas herramientas de la inteligencia artificial, no son herramientas mágicas y que su correcta implementación, así como la obtención de los datos para las mismas es un proceso que puede llegar a ser muy costoso. Sin desmerecer su gran importancia y potencial en el área de CFD es importante destacar que, en palabras de Vinuesa y Brunton, (2021) «*Los métodos de ML (Machine Learning), como el aprendizaje profundo, suelen ser caros de entrenar y requieren grandes cantidades de datos. Por tanto, es importante identificar las áreas en las que el ML supera a los métodos clásicos, establecidos desde hace décadas, y pueden ser más precisos y eficientes*». Por ello, aparte de la importancia de la calidad y cantidad de los datos, es necesario entender que problemas queremos solucionar y si su implementación es, a fin de cuentas, una mejora de los métodos clásicos.

Finalmente, los datos obtenidos en los diversos experimentos que se deben realizar para entender las estructuras flotantes deben ser correctamente analizados y saber que, sin su correcta aplicación pueden llevar a un entendimiento parcial y equivocado del comportamiento que tendrán estas estructuras en condiciones reales, algo que, a día de hoy, todavía es muy difícil de verificar.

Investigación en torno a eólica flotante

La energía eólica flotante es una de las tecnologías renovables más prometedoras. Sin embargo, presenta varios problemas que deben abordarse urgentemente como son las interacciones naturales entre la plataforma, la turbina y los amarres. Además, el movimiento de la plataforma bajo la acción de las olas afecta en gran medida al rendimiento de la turbina, lo que hace que funcione en ángulos sub-óptimos con respecto al viento y, en última instancia, reduce la potencia de salida del sistema y aumenta su costo nivelado de energía (LCOE). En nuestra investigación, proponemos un nuevo modelo acoplado para investigar el comportamiento de la turbina eólica marina, la plataforma flotante y sus líneas de amarre, en el dominio del tiempo.

Sin embargo, el uso del CFD en este tipo de estructuras es extremadamente costoso, habiendo muy pocas publicaciones al respecto, como puede ser la de Zhou *et al* (2020) en el que, a pesar de ser uno de los trabajos más completos de los últimos años, en donde se enfoca tanto en la dinámica del cuerpo flotante, la tensión de las líneas de amarre, el empuje de la turbina eólica así como la producción de energía de la misma entre otros; no se simulan por completo las líneas de amarre si no que son definidas como fuerzas que restringen el movimiento de la propia plataforma. Esto es debido a que un estudio de estas características requiere de un grandísimo coste computacional.

Aun así, en los últimos años se ha podido ver un incremento en servicio clústeres, donde marcas tan conocidas como Amazon, mediante Amazon web services (AWS), Microsoft, mediante AZURE, o IBM, mediante IBM cloud, ofrecen acceso a sus servidores para incrementar la capacidad computacional para todo aquel que lo necesite.

En esta línea, parte de los autores consiguieron una beca de 15,000 dólares, a través del programa AI for Earth de Microsoft, gracias al proyecto de acrónimo (MARIA, 2020) del título en inglés Monitoring floAting platfoRms in offshore wind through artificIAl intelligence. El proyecto, comenzó con el proceso de simulación de una plataforma flotante real para eólica offshore, en el que se tuvieron en cuenta mediante diversas fases todas las fuerzas y movimientos que afectan al comportamiento de la misma. Con estas simulaciones se pretende entender cómo se comportan plataformas de este tipo mediante diferentes condiciones medioambientales y cómo este comportamiento puede afectar al rendimiento de la turbina y, por lo tanto, al coste al que logren generar la electricidad que es en definitiva el objetivo último de estos sistemas. Es importante resaltar que, a pesar de que estas plataformas vayan a ser instaladas en zonas muy alejadas de la costa, donde el viento es mucho más constante, rápido y de mejor calidad, el propio movimiento de estos sistemas afectará negativamente a su rendimiento en comparación con sus equivalentes, los sistemas de eólica offshore fijados en el fondo marino, ya que las turbinas estarán trabajando mucho tiempo en ángulos de incidencia sub-óptimos. El acceso a esta cloud AZURE permitió acelerar el proceso de las simulaciones de CFD en las primeras instancias de la investigación. Esto es debido gracias al acceso al uso de nodos de hasta 48 núcleos y 300 GB de memoria RAM que permitían poner varias simulaciones de forma paralela.

El proyecto, dirigido por el profesor de la Escuela Dr. Jesús Mari Blanco junto al Dr. Gregorio Iglesias, profesor del University College Cork de Irlanda ha seguido su evolución después de la etapa de AZURE gracias a los servicios de servidor de la propia Universidad del País Vasco, el servido ARIÑA, en el que se han tenido acceso a capacidades de hasta 320 cores y 640 Gb de memoria RAM. Este clúster tiene un total de 3728 cores, Intel Xeon con memorias entre 16 y 512 GB por nodo, lo que lo hizo una herramienta ideal para el proceso de simulación de esta investigación.

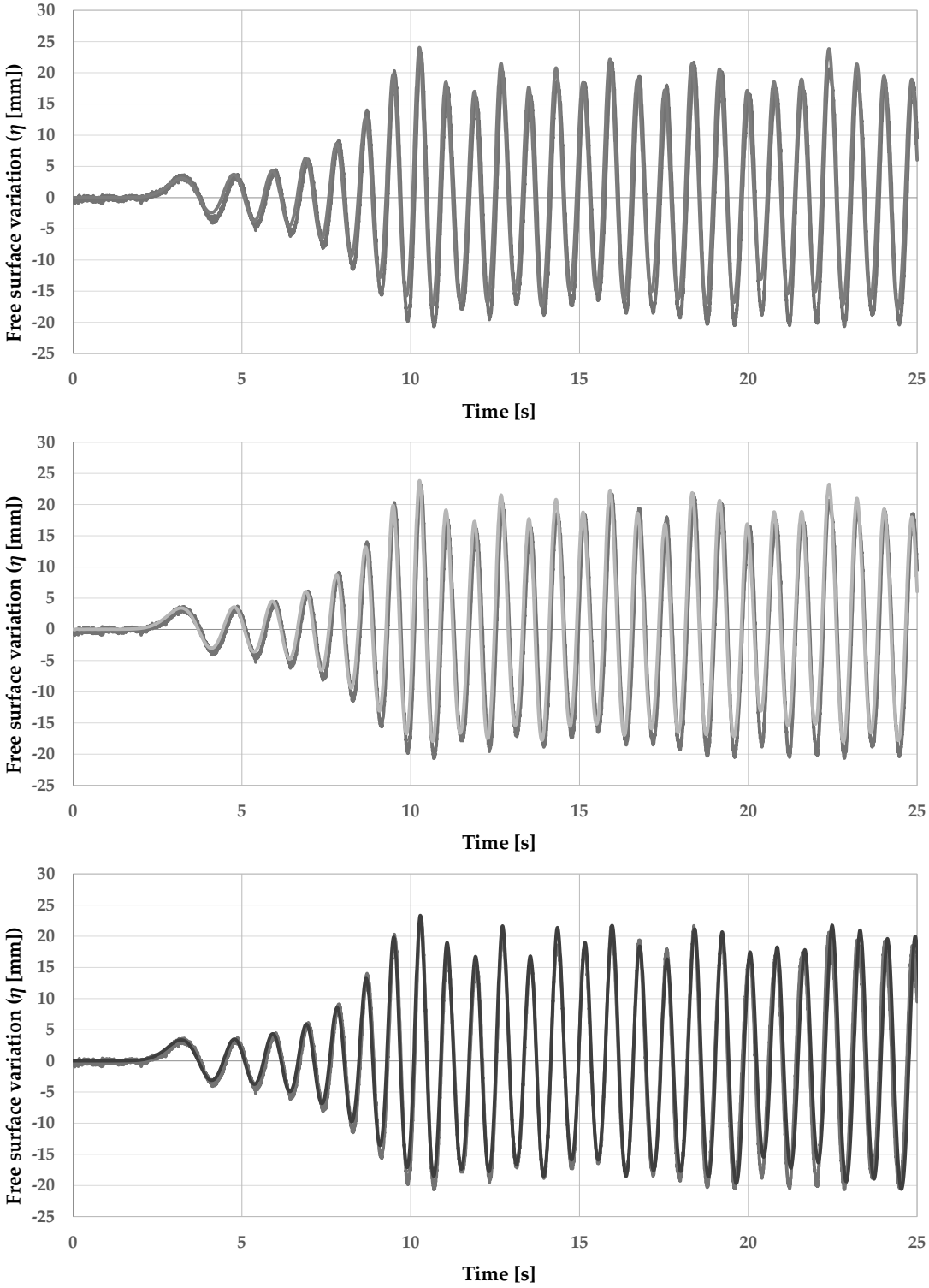
El objetivo final de esta investigación es poder definir el ML que permita acelerar las simulaciones fluidodinámicas manteniendo la precisión de la CFD. Sin embargo, como bien se ha comentado anteriormente es importante la obtención de datos de calidad que no requieran un coste computacional más alto que las aproximaciones que se realizan actualmente.

Caso de estudio

El objetivo de este estudio es, por tanto, la definición de un acople de estas características que permita tener una definición del comportamiento del sistema lo más realista posible. El primer paso de esta investigación será por tanto la validación del acople comparando los resultados del movimiento de la plataforma mediante simulaciones con los datos obtenidos a escala de laboratorio. Esta parte del estudio, permitirá la optimización tanto de una red neuronal en el uso del intercambio de datos así como de la malla computacional del CFD. Una vez la primera fase se haya dado por finalizada, se comenzará con la creación de un código que consiga acoplar los softwares que estudian tanto la turbina, la plataforma y las líneas de fondeo. Mientras tanto, la primera forma de acople será testeada para diferentes plataformas y estados de mar, así como la combinación de diferentes velocidades y direcciones de viento, para cada estado concreto.

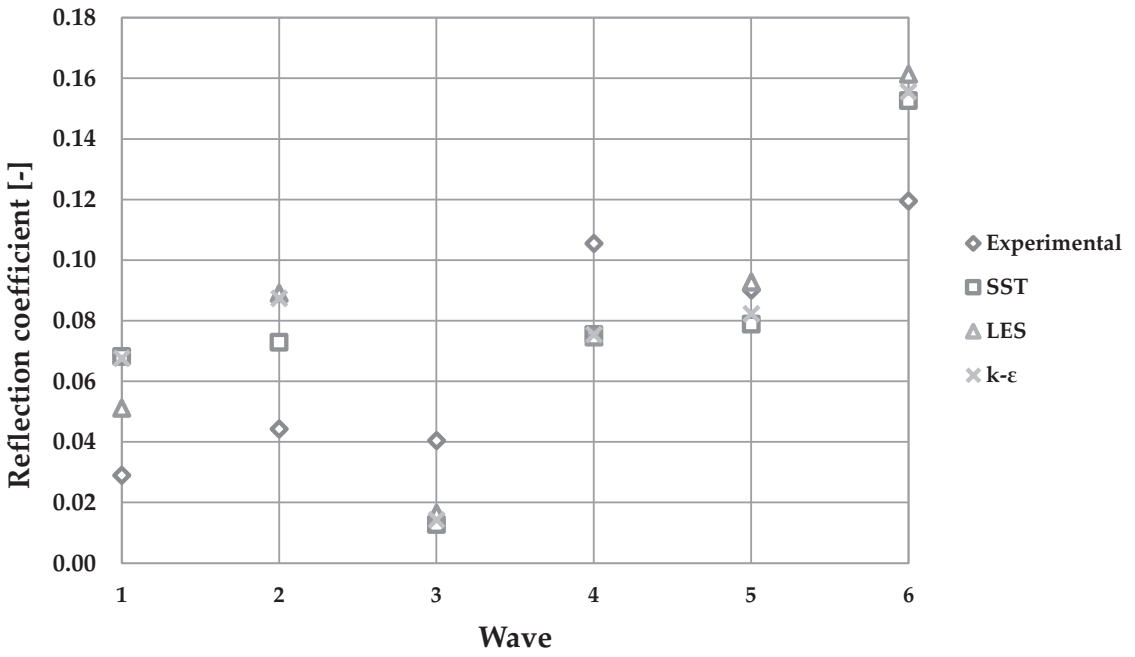
Con todo ello, se pretende comprender el comportamiento preciso de diferentes plataformas y la correspondiente potencia generada por la turbina eólica para cada estado de mar, lo que llevará al cálculo de la potencia generada total en un intervalo de tiempo determinado, lo cual conlleva el estudio de situaciones extremas, tan importantes a la hora de calcular la vida útil tanto de una única

FIGURA 4
COMPARACIÓN DEL DESPLAZAMIENTO DE LA SUPERFICIE LIBRE DEL CANAL DE OLAS CON LOS MODELOS DE TURBULENCIA K-E (ARRIBA), SST (CENTRO) Y LES (ABAJO) CON LAS MEDIDAS EXPERIMENTALES (LÍNEA AZUL)



Fuente: Galera-Calero *et al* (2020)

FIGURA 5
COMPARACIÓN DE COEFICIENTES DE REFLEXIÓN CON LOS DIFERENTES MODELOS DE TURBULENCIA ENSAYADOS EN COMPARACIÓN CON LOS VALORES EXPERIMENTALES



Fuente: Galera-Calero *et al* (2020)

turbina eólica flotante, como de un conjunto (farm), lo que se traduce en una mejor definición de la producción de la turbina. En la Figura 4 se muestran los resultados de las simulaciones de la oscilación que sufre la plataforma mediante la aplicación de tres modelos de turbulencia diferentes (K-ε, SST (Shear Stress Transport) y LES (Large Eddy Simulations) respectivamente), para un mismo período de tiempo, comparándolos con las medidas tomadas experimentalmente.

Durante este proyecto se manejaron ingentes cantidades de datos. Comenzando por el principio, varios análisis computacionales fueron realizados para una plataforma flotante comparándolos con datos experimentales. El primer paso fue hacer un estudio de los diferentes modelos de turbulencia que se pueden aplicar en el software de fluidodinámica computacional, para ver cómo los diferentes modelos de turbulencia pueden afectar al resultado final.

La comparación de la superficie libre con los diferentes modelos de turbulencia fue utilizada para calcular el coeficiente de reflexión en cada una de ellas y verificar que los métodos para analizar la reflexión eran útiles. Ver Figura 5.

Se puede ver que los diferentes modelos de turbulencia se comportaron de manera bastante similar durante todos los ensayos, siempre estando ligeramente desplazados con respecto a los valores experimentales.

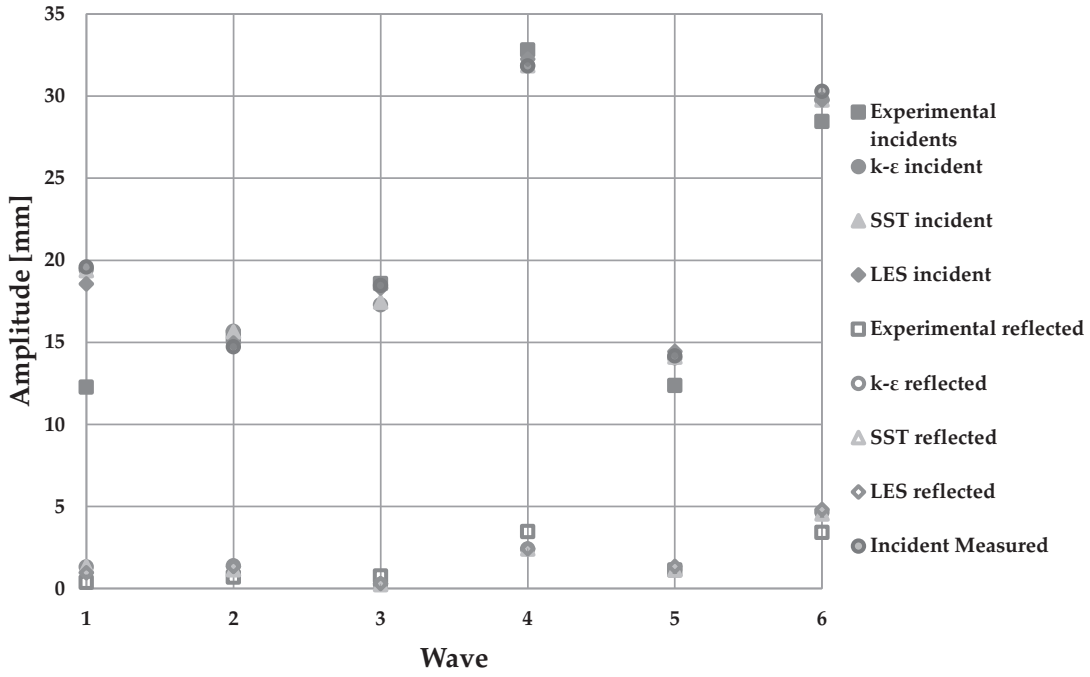
Para verificar que los métodos utilizados eran los correctos a la hora de hacer las mediciones para calcular los coeficientes, éstos se usaron para hacer las mediciones tanto de la ola incidente como de la ola reflejada. Además, la propia medición de la ola incidente experimental fue colocada para comprobar si los métodos eran adecuados.

Después de ese paso previo, un estudio mucho más profundo se realizo para encontrar las posiciones idóneas del sistema de extinción del canal de olas así como para poder entender de una manera más fiable su comportamiento y la relación que mantiene respecto a las olas incidentes y la profundidad de los ensayos realizados.

Se encontraron tendencias muy interesantes que tenían en relación la profundidad del ensayo, la altura a la que está puesta la playa, la longitud de las olas estudiadas y la forma de cómo rompen las olas en el sistema de extinción, medidas mediante el número de Iribarren (Ir) que tiene en cuenta la inclinación del sistema de extinción así como la relación entre la altura y la longitud de ola. El número de Iribarren es un parámetro adimensional que se utiliza para modelar los efectos de olas de gravedad superficiales en playas y estructuras costeras.

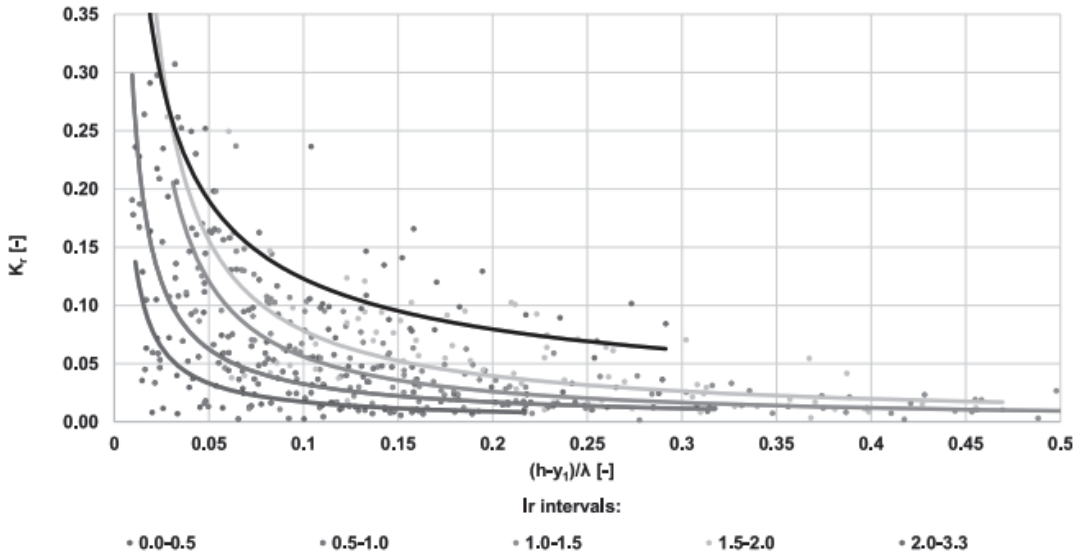
En la Figura 6 se muestran los resultados en amplitud de diferentes olas tanto incidentes como reflejadas según las mediciones computacionales y experimentales, frente a la medición de la ola incidente realizada en laboratorio.

FIGURA 6
MEDICIÓN DE OLA INCIDENTE Y REFLEJADA MEDIANTE LOS MÉTODOS DE ESTUDIO DE REFLEXIÓN FRENTE A LA MEDICIÓN DE LA OLA INCIDENTE HECHA EN EL LABORATORIO



Fuente: Galera-Calero *et al* (2020)

FIGURA 7
RESULTADOS DE LOS COEFICIENTES DE REFLEXIÓN RESPECTO A LA ALTURA DE LA PLAYA EN RELACIÓN A LA PROFUNDIDAD DEL ENSAYO Y EL NÚMERO DE IRIBARREN

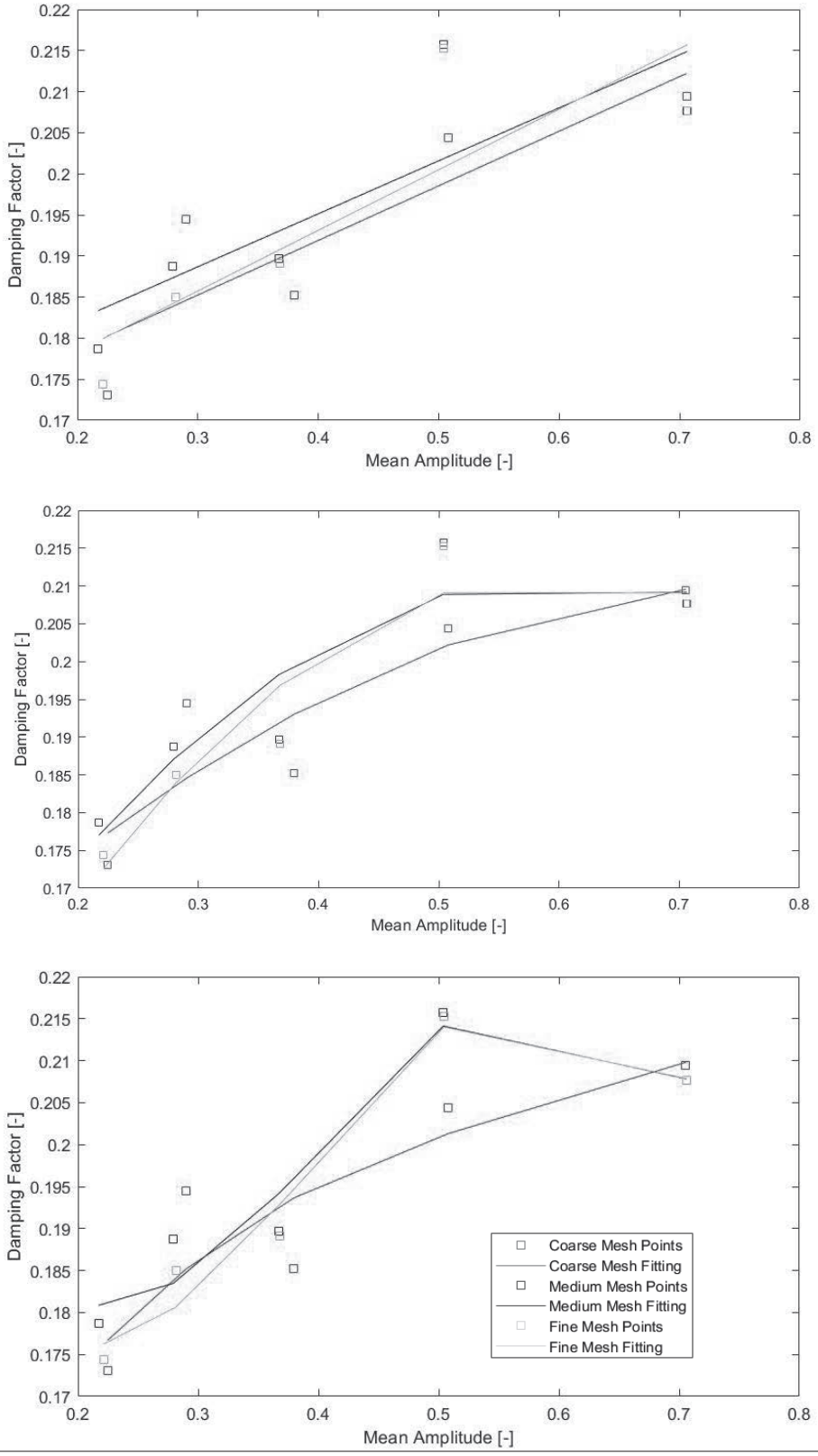


Fuente: Izquierdo *et al* (2021)

Como se puede observar en la Figura 7, se hizo un estudio experimental intensivo sobre el comportamiento del tanque a la hora de hacer estudios de plataformas flotantes con respecto a la altura de la zona de extinción (playa) y profundidad del ensayo para distintos números de Iribarren.

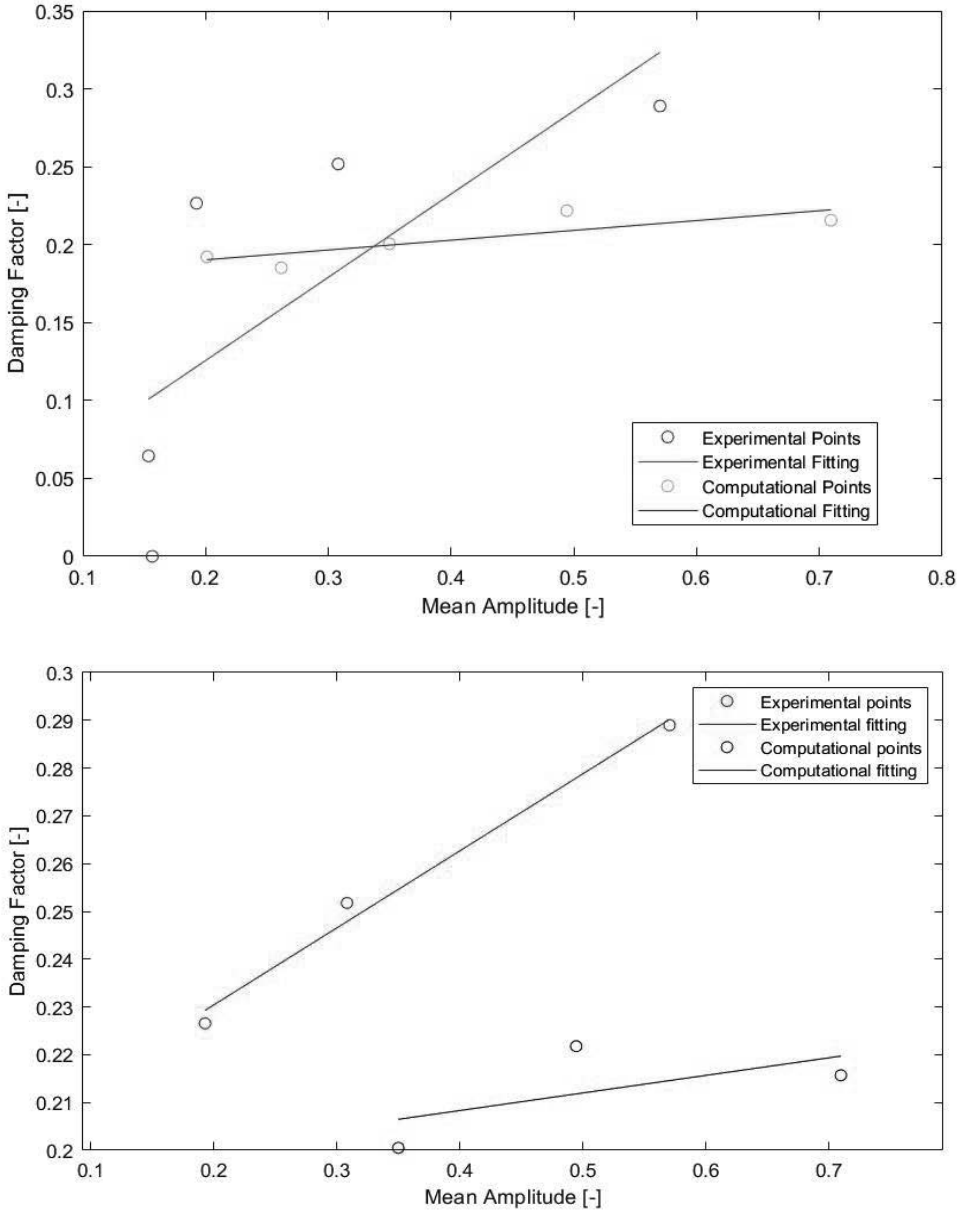
En la investigación, se tuvo en cuenta un estudio inicial del comportamiento de una plataforma flotante en la que se manejan los datos de los primeros ensayos que se realizan en este tipo de estructuras flotantes como son los test de caídas y la respuesta frente a oleaje regular.

FIGURA 8
COMPARACIÓN DE TENDENCIAS LINEALES (ARRIBA), CUADRÁTICAS (CENTRO) Y CÚBICAS (ABAJO) CON LOS RATIOS DE AMORTIGUAMIENTO PARA DIFERENTES TAMAÑOS DE MALLA DE UNA PLATAFORMA FLOTANTE



Fuente:Galera-Calero *et al* (2021)

FIGURA 9
COMPARACIÓN NUMÉRICO-EXPERIMENTAL DEL TEST DE CAÍDA DE UNA PLATAFORMA FLOTANTE TENIENDO EN CUENTA SI LOS DATOS SON O NO ADECUADOS



Fuente:Galera-Calero *et al* (2021)

Antes de poder hacer las simulaciones finales, los estudios de convergencia de malla y tiempo han de hacerse para asegurarse de que los datos obtenidos en las simulaciones son fiables y no son datos defectuosos debido a una mala definición. Este estudio se puede hacer mediante los diferentes ratios de amortiguamiento de un test de caída, donde se compara cómo se va amortiguando el movimiento en cada oscilación. En la figura 8 se pueden observar los datos y las diferentes concepciones que se pueden tomar utilizando tendencias (ajustes de interpolación) que pueden ser o no adecuadas (lineales, cuadráticas y cúbicas respectivamente) en base a tres diferentes calidades de malla computacional (fina, media y grosera).

Por ejemplo con la tendencia lineal puede parecer que la malla gruesa y la media tienen un comportamiento similar, mientras que se demuestra que la media es mucho más próxima a la fina cuando se aumenta el grado del ajuste.

Finalmente, en la figura 9 se presenta la comparación numérico-experimental en la que se puede ver la importancia de ser conocedor de los datos que se están analizando y decidir cuáles de ellos deben ser utilizados. En la parte superior, algunos de los datos experimentales marcan cero para el factor de amortiguamiento, algo totalmente irreal, lo cual es debido a que sólo se está analizando un grado de libertad de

los 6 posibles. Por ello, en la última oscilación, la plataforma parece que se mueve más en vertical debido al «ruido» proveniente de otros movimientos.

Esto, sin embargo, no puede ser reproducido en las simulaciones computacionales ya que, en aras de mantener lo más bajo posible el coste computacional de las simulaciones, se impuso un plano de simetría que delimitaba el movimiento de la plataforma en 3 grados de libertad en vez de en 6, tal y como se muestra en la parte inferior de la figura, por lo que ese «ruido» no podía ser incluido en los datos. Por ello, solo se hizo la comparación con las mediciones de las primeras oscilaciones, viendo así una mejora sustancial en la comparativa numérico-experimental.

CONCLUSIONES

Los datos generados incesantemente por humanos y máquinas han experimentado un crecimiento exponencial a lo largo de la última década. En 2020, la cantidad de datos creados alcanzó un nuevo récord. El crecimiento fue mayor de lo esperado anteriormente debido a la pandemia de COVID-19, ya que muchas más personas trabajaron, aprendieron y se entretuvieron en el hogar. El análisis llevado a cabo por Statista, pronostica que la cantidad de datos creados, capturados, copiados y consumidos a nivel mundial crezca a más de 180 zettabytes en el 2025. Este nuevo panorama ofrece la posibilidad de acceder a nuevas y únicas oportunidades de negocios, potenciadas por la misma abundancia de datos y las perspectivas que éstos proporcionan. Por otro lado, requiere decisiones estratégicas sobre la recopilación, utilización y ubicación de toda la información generada.

La creciente disponibilidad de datos ha impulsado los avances en técnicas analíticas y en tecnologías, al frente del aprendizaje automático. El requisito clave para el aprendizaje automático es el correcto manejo de grandes cantidades de datos, que son necesarios para entrenar algoritmos. La calidad de las predicciones por otro lado depende directamente de la calidad de los datos, de forma que por esta razón datos precisos y de alta calidad, son condición necesaria para generar un valor añadido que sea a su vez fiable.

En este artículo se han resaltado las ventajas ofrecidas por un correcto tratamiento de los datos, se ha dado respuesta a aquellas preguntas que normalmente surgen a la hora de empezar un proyecto de análisis basado en datos y se ha orientado al lector en los pasos a seguir para obtener datos de calidad.

Se ha presentado la importancia que tienen los datos en el mundo actual y, más concretamente mediante una aplicación en el área de las simulaciones fluidodinámicas de plataformas marinas flotantes, para obtención de energía eólica mediante turbinas, un área tecnológica emergente que necesita del correcto tratamiento de una ingente cantidad de datos como son los meteorológicos, velocidad y dirección del viento, temperatura, datos asociados al estado de la mar como altura y amplitud de las olas, así como aquellos asociados con el movimiento de la plataforma en los tres ejes coordenados tanto en desplazamiento como en giro, así como del movimiento de las palas de la propia turbina marina (velocidad de giro de las palas y ángulo de orientación de la nacelle), todo ello en períodos regulares de tiempo y por último las fuerzas de todos los amarres en las tres direcciones coordenadas.

De todo ello, se puede ver lo importante que es llevar a cabo un buen análisis de todos los datos a la hora de mejorar las simulaciones del comportamiento de dichas estructuras flotantes, así como de la correcta evaluación de la cantidad de los mismos. De esta forma, se ha demostrado que una mayor profundización a la hora de hacer los análisis de estas estructuras, así como el hecho de aumentar el número de ensayos resulta absolutamente necesaria para mejorar la calidad de los datos que permitan predecir su comportamiento con un alto grado de precisión ya que de ello depende directamente la correcta evaluación de la producción de energía y con ello el LCOE de las propias turbinas marinas.

En relación a la investigación en sí, los estudios realizados han servido para marcar donde hay grandes áreas de mejora en las simulaciones, como es principalmente en la consideración de todos los grados de libertad y de los datos necesarios debido al gran número de comparativas realizadas ya en las primeras fases de la investigación.

Esto permitirá el futuro uso de redes neuronales para acelerar este tipo de simulaciones y poder profundizar en algunos sistemas mucho más complejos en los que se tenga en cuenta el efecto de muchas más fuerzas externas que afecten a su movimiento, como podrían ser las ocasionadas por roturas de algunos de los anclajes debidas a fenómenos extremos, que ayude en la correcta optimización de los mismos, minimizando así las restricciones de movimientos y por ende las limitaciones a la producción de energía, así como minimizar los costes de mantenimiento asociados a este tipo de estructuras.

AGRADECIMIENTOS

Los autores agradecen la ayuda de MICROSOFT concedida al proyecto MARIA dentro del programa AI for Earth. Asimismo agradecen al grupo de investigación de Gobierno Vasco IT1514-22 (ITSAS REM) por su apoyo en la elaboración del presente trabajo.

REFERENCIAS

AI for EARTH, MICROSOFT, 2020: https://www.microsoft.com/en-us/ai/ai-for-earth (last accessed Nov. 2021)

Anaconda-SODS-Report-2020: https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-SODS-Report-2020-Final.pdf (last accessed Nov. 2021)

Banko, M., Brill, E. Mitigating the paucity-of-data problem: exploring the effect of training corpus size on classifier performance for natural language processing. *Proc. 1st Internal Conference on human language technology research*, 1-5, San Diego, CA, (2001).

Batini, C. Cappiello, C. Francalanci, C., Maurino, A., Methodologies for data quality assessment and improvement, *ACM Computing Surveys (CSUR)*, 41, pp. 16-23, (2009).

Bernardini, A., Asensio, J., Olazagoitia, J.L., Biera, J., Evolutionary Neural Networks for Product Design Tasks, *Hybrid Artificial Intelligent Systems Lecture Notes in Computer Science*,7208, 2012, pp 421-428.

Blanco, J.M.; Vazquez, L.; Peña, F.; Díaz, D., New investigation on diagnosing steam production systems from multivariate time series applied to thermal power plants, Applied Energy, 101, 2013: pp. 589-599. doi: 10.1016/j.apenergy.2012.06.060.

Catarci, T., Scannapieco, M. Data quality under the computer science perspective. Archivi Computer, 2, 2002, pp. 1-15.

Christoph E., Grgić-Hlača, N. Machine Advice with a Warning about Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court, *Journal of Legal Analysis*, 13, 1, 2021, pp 284-340.

Galera-Calero, L.; Blanco, J.M.; Izquierdo, U.; Esteban, G.A. Performance Assessment of Three Turbulence Models Validated through an Experimental Wave Flume under Different Scenarios of Wave Generation. J. Mar. Sci. Eng. 2020, 8, 881, doi: 10.3390/jmse8110881.

Han, J., Kamber, M., Pei, J., Data Preprocessing*, Data Mining (Third Edition)*, pp. 83-124, 2012.

Huang, G., Bryden, KM, McCorkle, DS, Interactive Design using CFD and Virtual Engineering, *Actas de la 10ª Conferencia de optimización y análisis multidisciplinario de AIAA / ISSMO*, AIAA-2004-4364, Albany, (2004).

Izquierdo, U.; Galera-Calero, L.; Albaina, I.; Vázquez, A.; Esteban, G.A.; Blanco, J.M. Experimental and Numerical Determination of the Optimum Configuration of a Parabolic Wave Extinction System for Flumes. Ocean Eng. 2021, 238, 109748, doi: 10.1016/j.oceaneng.2021.109748.

Jacobson, M.Z.; Delucchi, M.A.; Bauer, Z.A.F.; Goodman, S.C.; Chapman, W.E.; Cameron, M.A.; Bozonnat, C.; Chobadi, L.; Clonts, H.A.; Enevoldsen, P.; *et al*. 100% Clean and Renewable Wind, Water, and Sunlight All-Sector Energy Roadmaps for 139 Countries of the World. Joule, 2017, 1, pp. 108-121, doi: 10.1016/j.joule.2017.07.005.

Loshin, D., Data Quality, *Business Intelligence (Second Edition)*, pp. 165-187, (2013).

MARIA project, 2020: https://www.energias-renovables.com/eolica/microsoft-presta-su-inteligencia-artificial-a-un-20200518 (last accessed Nov. 2021)

McCorkle, DS, Bryden, KM, Using the Semantic Web to Enable Integration with Virtual Engineering Tools, *Actas del 1er Taller Internacional de Fabricación Virtual* (27), Washington, DC, (2006).

Parra, C., Olazagoitia, J.L., Biera, J., Development of intelligent tools to eliminate squeal noise in brake systems*, 6th European Conference on Braking JEF 2010*, Lille, France, (2010).

Pipino, L.L., Lee, Y.W., Wang, R.Y., Data quality assessment, *Communications of the ACM*, 45, pp. 211-218, (2002).

Rahm, E., Do, H., Data Cleaning: Problems and Current Approaches, *Computer Science IEEE Data Eng. Bull.*, 2000.

Scott Mayer M., Marcin Sienieky Shravya S., International evaluation of an AI system for breast cancer screening, Nature, 577, 2020, pp 89-94.

Smart City Index 2020 by IMD Business School: https://www.imd.org/smart-city-observatory/Home/ (last accessed Nov. 2021)

Vázquez, L.; Blanco, J.M.; Ramis, R.; Peña, F.; Díaz, D., Robust methodology for steady state measurements estimation based framework for a reliable long term thermal power plant operation performance monitoring, Energy, 93, 1, 2015: pp. 923-944. doi: 10.1016/j.energy.2015.09.044

Vinuesa, R.; Brunton, S.L. The Potential of Machine Learning to Enhance Computational Fluid Dynamics. ArXiv abs/211002085 Phys. 2021.

Windt, C.; Faedo, N.; García-Violini, D.; Peña-Sanchez, Y.; Davidson, J.; Ferri, F.; Ringwood, J.V. Validation of a CFD-Based Numerical Wave Tank Model of the 1/20th Scale Wavestar Wave Energy Converter. Fluids, 2020, 5, 3, 112, doi: 10.3390/fluids5030112.

ANEXO: NOMENCLATURA

CFD	Computational Fluid Dynamics
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
ETL	Extracción Transformación Carga (Load)
GIGO	Garbage In - Garbage Out (basura adentro – basura afuera)
IA	Inteligencia Artificial
IoT	Internet of Things
Ir	Número de Iribarren
LCOE	Levelized Cost Of Energy
LES	Large Eddy Simulations
MARIA	Monitoring floating platfoRms in offshore wind through artificIAl intelligence
ML	Machine Learning
RANS	Reynolds-Averaged Navier-Stokes
SST	Shear Stress Transport
TRL	Technology Readiness Level