DM

Web-based Interface for Environmental Niche Modelling

MASTER DISSERTATION

José Rúben Silva Freitas
MASTER IN INFORMATICS ENGINEERING

UNIVERSIDADE da MADEIRA
A Nossa Universidade
www.uma.pt

October | 2021

Web-based Interface for Environmental Niche Modelling

MASTER DISSERTATION

José Rúben Silva Freitas

MASTER IN INFORMATICS ENGINEERING

ORIENTATION
Marko Radeta

CO-ORIENTATION
Filipe Magno Gouveia Quintal

UNIVERSIDADE da MADEIRA

A Nossa Universidade

www.uma.pt

FCEE

Mestrado em Engenharia Informática

Web-based Interface for Environmental Niche Modelling

José Freitas

supervisionado por

Prof. Dr. Marko Radeta and Prof. Dr. Filipe Quintal

3 de dezembro de 2021

Abstract

Marine species are subject to anthropogenic impacts, such as noise pollution, marine litter, and direct impact collisions. While there are efforts in the marine community and crowd-sourcing to report the occurrence of marine species, not enough projects explore the prediction of where such animals may be.

This dissertation analyzes the state of the art in species distribution modeling (SDM) systems, capable of reporting and predicting marine biodiversity. The proposal implements the algorithms for predicting species through publicly available repositories of data, provides means to ease the upload and management of occurrence points as well as methods for prediction analysis. A web-based user interface is proposed using Ecological Niche Modelling (ENM) as an automated alerting mechanism towards ecological awareness.

Performed user studies evaluate marine biodiversity concerns from fisherman and whale-watching sea-vessels, assessing attitudes, threats, values, and motivation of both samples. Further, biologists and experts on ENMs will evaluate the workflow and interface, reviewing the proposal's potential to enable ecologists to create solutions for their custom problems using simple protocols without the need for any third-party entities and extensive knowledge in programming.

Keywords: Ecological Niche Modelling · Wildlife Monitoring · Web Application Development · Biodiversity Assessments · Species Distribution Modelling · Ecological Awareness.

Resumo

Espécies marinhas estão sujeitas a impactos antropogênicos, tais como poluição sonora, lixo marinho, e colisões com tráfego marinho. Apesar de existirem alguns esforços da comunidade marinha e *crowdsourcing* relativamente ao registo de ocorrências de biodiversidade marinha, não existem projeto suficientes que exploram as previsões de onde estas espécies poderão estar.

Esta dissertação analisa o estado da arte em sistemas de modelação de distribuição de espécies, capazes de relatar e prever biodiversidade marinha. A proposta implementa os algoritmos para prever espécies por meio de repositórios consolidados de dados disponíveis online, fornece meios para facilitar o carregamento e gestão de pontos de ocorrência, bem como métodos para análise das previsões. Uma interface web de utilizador é proposta utilizando *Ecological Niche Modeling* como um mecanismo de alerta automatizado para incrementar a consciência ecológica.

Os estudos do sistema irão avaliar as preocupações relativas a biodiversidade marinha de embarcações de pesca e navios de observação de baleias. Desta forma é possível determinar atitudes, ameaças, valores e motivação de ambas as amostras para com a biodiversidade marinha. Além disso, biólogos e especialistas nesta tipologia de sistemas, avaliarão o fluxo de trabalho e a interface desenvolvida, avaliando o potencial do sistema, permitindo aos ecologistas criar soluções personalizados através de protocolos simples, sem a necessidade de quaisquer entidades terceirizadas e conhecimento em programação.

Keywords: Modelação preditiva de distribuição de espécies · Monitorização de vida selvagem · Desenvolvimento de aplicações *web* · Avaliações de biodiversidade · Consciência Ecológica.

Acknowledgements

When I first started my academic journey I did not have the motivation to go further, however important people in my life supported me and enabled me to achieve and keep grinding towards a better future. Today I can thank them for all that they have done for me. As it all started, I'd like to appreciate my parents who gave me all the tools to succeed, not only in my academic journey but in every aspect of my life. My parents are my biggest supporters, as is Tatiana Abreu who managed to put up with me through all these years. Since we first met, she incentivized me to go further and be the best version of myself. I'd also like to mention all my remaining family and friends who made everything easier and enjoyable. Lastly, I would like to thank all the professors that were part of this journey but most importantly my supervisors, Prof. Marko Radeta (Wave Labs Coordinator) and Prof. Filipe Quintal for all the dedication and support that made this dissertation possible.

This dissertation is part of following projects: **INTERWHALE** - Advancing Interactive Technology for Responsible Whale-Watching, with grant no. PTDC/CCI-COM/0450/2020 by FCT; **LARGESCALE** - Location-based Augmented Reality Gadgets and Environment-friendly Sightseeing of Cultural Attractions for Locals and Excursionists, with grant no. PTDC/CCI-CIF/32474/2017 by FCT and Portuguese National Funds (PIDDAC); **INTERTAGUA** - Interfaces Aquáticas Interativas para Deteção e Visualização da Megafauna Marinha Atlântica e Embarcacoes na Macaronésia usando Marcadores Rádio-transmissores, with grant no. MAC2/1.1.a /385 by Programa de Cooperación INTERREG V-A España-Portugal MAC (Madeira-Azores-Canarias) 2014-2020; **LARSyS** - Laboratório de Robótica e Sistemas de Engenharia, with grant no. UIDB/ 50009/2020 by FCT/MCTES (PIDDAC); **LARSyS** - Laboratório de Robótica e Sistemas de Engenharia, with grant no. UID/EEA/ 50009/2019 by FCT/MCTES (PIDDAC); and **MITIExcell** - Excelencia Internacional de IDT&I nas TIC, with grant no. M1420-01-01450FEDER0000002, by Regional Government of Madeira. Project has been conducted with the support of the Wave Lab[1] - interactive technologies for depicting the anthropogenic impact on marine biosphere.

[1] `http://wave-labs.org`

Table of Contents

List of Figures

List of Tables

Glossary

1. **ENM**. Environmental Niche Modeling also known as Species Distribution Modeling (SDM) or Ecological Niche Modeling, is the process of correlation between environmental variables and species occurrence points, that provide the basis for predictions of biodiversity.
2. **GUI**. Graphical User Interface
3. **CLI**. Command line interface
4. **SatEO**. Satellite Earth Observation
5. **RS**. Remote Sensing
6. **WSN**. Wireless Sensor Networks
7. **GIS**. Geographic Information Systems
8. **VHR**. Very-High-Resolution
9. **VME**. Vulnerable Marine Ecosystems
10. **API**. Application programming interface
11. **HTTP**. Hypertext Transfer Protocol
12. **JSON**. JavaScript Object Notation
13. **NR**. Network Ratio
14. **IV**. Inverse Variation
15. **MRC**. Model-Route-Controller
16. **CSV**. Comma Separated Values
17. **ORM**. Object-relational Mapper
18. **EEA**. European Environment Agency

1 Introduction

Marine biodiversity is the variety of life in the sea, and the ocean contains a great variety of it, providing valuable resources [1]. All these organisms interact in a physical environment towards an ecosystem. These have been affected in many ways throughout history, as for example overexploitation and habitat loss, however, due to recent developments and a human population explosion, these pressures have grown exponentially [2]. Currently, habitat loss, degradation, extraction, pollution, eutrophication [2], the introduction of alien species, and climate change affect the greatest number of marine taxa [3].

Accordingly, the thesis proposes means for which technology could assist the aforementioned concerns, focusing on taxa distribution predictions as a measure towards biodiversity conservation. The proposal adopts Ecological Niche Modelling methods, providing alerting mechanisms of high-density spots of biodiversity, therefore attempting to reduce threats associated with human activities.

Focusing on such aspects, this thesis will produce a web-based system that supports biologists to assess species distributions, as maps of suitability scores. This system requires no third-party entities or extensive knowledge in programming contexts, allowing for straightforward means to manage data towards predictions and promote the abundance of species. Moreover, such a system is designed to assess regular sea sailors' behaviors related to threats, attitudes, and values of marine biodiversity, in the form of ecological awareness inflator. Two case studies will be performed, evaluating the usability of the proposed system and assessing the impact of such data on local vessels on the coast of Madeira island, Portugal.

1.1 Motivation

Concerns include direct impact collision, a consequence of increasing marine traffic, as the most common for North Atlantic Right Whales (hereinafter, right whales), where their population remains endangered. Such species are protected by the U.S. Endangered Species Act and by the International Union for Conservation of Nature [4]. Right whales occupy the western Atlantic Ocean from southern Greenland and the Gulf of St. Lawrence south to Florida. However,

[2] excessive plant and algal growth due to the increased availability of one or more limiting growth factors needed for photosynthesis

the occurrence and movements of the species are not well known and it has declined in recent years [5]. Limited knowledge about the location and movements of right whales hinders efforts to manage human activities such as slowing down of sea vessels, to minimize activities that aggravate the mortality of such species. Motivated to predict the abundance of such species, this dissertation will provide an online platform to use scientific repositories as a source of data for Ecological Niche Modelling (ENM), providing estimates and predictions of marine biodiversity, serving to impact the marine traffic.

1.2 Context

Understanding, predicting, and studying ecological processes is a long-term area of interest [6–8], dating back to 1931. Early studies managed to correlate relationship variables that provide the basis for prediction present on current modelling protocols. Alongside the development of technology, the ability to establish the inter-relation between variables expanded, originating substantial research efforts into ecological predictions [9].

These efforts originated Environmental Niche Modelling (ENM), also known as Ecological Niche Models or Species Distribution Modelling (SDM), consisting in the development of analytic procedures, used to understand and analyze complex ecological processes, usually employed in predictions of ever-evolving ecosystems [10].

Innovation in this area of research is producing an abundance of methods and techniques, overwhelming beginners [11]. A quick search in Google Scholar using "ecological niche modelling" as the only query parameter highlights over 329 000 results, of which, around 17 000 are from 2021 alone (up to 04/08/2021). Following recent tendencies of research, this situation tends to worsen, creating a barrier of initiation that discourages future initiatives, therefore worsening such concerns for entry-level users.

1.3 Protocol

The necessary steps to establish a prediction using ENMs will be referenced as the protocol from now on.

The protocol of establishing an ENM is segmented into multiple tasks, in which all decisions affect the output generated by the algorithms, meaning that the selection of the initial occurrence dataset will impact the environmental variables elected, spatial resolution, the algorithm choice, and undoubtedly the prediction output.

The first of these steps is the acquisition of datasets that combine species occurrences (presence-absence, presence-only, or abundance levels) with environmental covariates (e.g. temperature, salinity, wind, oxygen, etc...) at a sufficiently high spatial resolution [12]. The quality of this data directly impacts the quality of the prediction, considering that reliable data may generate an accurate prediction, whoever flawed data will never generate accurate predictions. To comply with this assertion, respectable data sources must be included (e.g. Copernicus[3] or Global Biodiversity Inventory Facility - GBIF[4]).

Although data credibility is assured, that is not the only concern regarding data gathering. Likewise, the sample size directly impacts the decision-making regarding the data preparation, as a consequence of all algorithms benefiting from a larger number of records [13]. In the case of a smaller dataset, as is the case with a variable that occurs scarcely on the ecosystem, the focus should be redirected to improving the quality of existing occurrence data [14]. Part of the data preparation techniques includes mechanisms of data curation and spatial-temporal filtering [12], such as: (i) data filtering; (ii) duplication removal; (iii) cross-checking; and (iv) georeferencing.

Subsequently, the given data needs to be formatted to fit as variables of the chosen model. We may consider these models as "formulas" in mathematical contexts or "functions" when referring to computational backgrounds. These models receive inputs as variables or arguments, depending on the context, and return an output known as the prediction. A simple formula could look like formula 1, in which y is a function of all other variables and will return the pretended output, while x may diverge between data frames, presence points, ASCII grids, or other variable types.

$$y = \sum_{i=1}^{\infty} xi \tag{1}$$

[3]`https://www.copernicus.eu/en`
[4]`http://www.gbif.org/`

Different modelling methods generate different model objects, namely maps of presence probability, plots, abundance predictions, maps of suitability, species distribution, environmental envelope (Set of environments within which it is believed that the species can persist), and/or others.

Finally, a model evaluation is required, to access the effectiveness and overall prediction credibility of the protocol, however, it is much easier to run a model than it is to assess how good the model is and whether it can be used for a specific purpose. Considering this complexity, generally, multiple algorithms are employed, guaranteeing comparison metrics, for a specific model, promoting discussion of predictions, and help to assess whether there is anything substantially wrong with the protocol implementation [15].

1.4 Limitations

In most studies, the validation and discussion are supported by cross-validation, in which outputs of different models are corroborated with a confirmed prediction of known occurrences. Nonetheless, the process of running multiple models may become extremely burdensome, as a consequence of different pre-processing methods for each protocol, therefore discouraging cross-validation with further models [15].

As previously explained, the protocol may become tiresome, mostly around data formatting to correspond with the algorithm's arguments. Furthermore, each model's output may differ from the other, generating another cumbersome task for the user to compile and compare. Both these tasks are necessary, even tho inconvenient because different algorithms admit distinct objectives and prediction purposes. Currently, there is no available option that grants a reliable normalization of these steps in the protocol.

Likewise, the complexity of employing said protocols develops a border towards the generalization of ENMs in the scientific community, ascending the dependency for specialized software and personal resources.

1.5 Objectives

The primary purpose of this thesis is to develop a web-based interface that is suitable for ENM research, overcoming existing limitations. This research expects to contribute to existing knowledge and promote the implementation of already developed algorithms, providing compelling contributions to the proposed research area.

Based on the already described characteristics, protocols, and limitations with current processes, the following objectives were established:

- **[O1]** - Develop an accessible platform for biologists without programming backgrounds and feature-rich towards marine ecology experts;
- **[O2]** - Provide a robust pipeline with public available scientific data repositories;
- **[O3]** - Supply to the end-user possibilities of multiple model predictions simultaneously;
- **[O4]** - Creation of the data structures which grant the opportunity to study and reuse old predictions, towards comparison metrics.
- **[O5]** - Modular structure that supports extensive expansion for new models;
- **[O6]** - Eliminate system setup dependencies for user protocol employment;
- **[O7]** - Provide easy to use Graphical User Interface (GUI) for model analysis.

Listed requirements proved beneficial in terms of setting up the structure and purpose of the thesis, arguing that more tools are necessary to facilitate ENM protocols for environmental niche modelling. Moreover, it is necessary to guarantee the accuracy and flexibility required by expert researchers, allowing them to perform long-term studies.

1.6 Research Questions

The listed objectives will be engaged alongside research questions, pinpointing the thesis' focus, opportunities, and purpose.

- **[RQ1] - How could the proposed system facilitate the process of employing ENMs, while retaining the flexibility existent with current solutions?**
  In this question, the thesis conducts satisfaction surveys from distinct study subjects, collecting feedback and displaying the effectiveness and usability of the proposed system.

– **[RQ2] - Can the ENMs be used as an alerting mechanism towards ecological problems?**
  The concluding question inquiries the feasibility of using the proposed system as a systematic task, generating efficient prediction outputs, evaluation methods, and increase ecological awareness.

1.7 Structure of the Document

Introduction outlined the context and motivation behind the dissertation, defined the protocol, and highlighted limitations with existing systems. Then supported by previous regards, the objectives were established and research questions were defined.

Following the contextual analysis, a literature review will be developed, in which subsections for predicting biodiversity, remote sensing, ENM, and interface design originated. Here each subsection was compared and correlated with the proposed solution while drafting a state-of-the-art workflow.

The system was then described according to its workflow, technology, system architecture, and the study setup was defined. These include: (i) back-end development software; (ii) modelling scripts from occurrence treatment to prediction reproduction; (iii) front-end interactive components, feeding the models.

The fourth chapter yields the results of usability and environmental awareness tests, as well as the overall statistical measurements computed from said results.

These results and feedback will be discussed and analyzed, supplying findings, interpreting constraints, providing guidelines for future research efforts, and outlining contributions regarding the research question generated in the introduction.

Finally, the final section will clarify the importance of the proposed system, provide future development efforts for the web-based application of ENMs.

2 Related Work

In below, the dissertation provides an overview of used techniques for predicting marine biodiversity occurrences, depicting studies from computational marine ecology. Moreover, it provides an overview of the type of data which can be collected from passive remote sensing, using Satellite Earth Observation (SatEO), as well it specifies the importance of applied predictions in the marine ecology balance. The aforementined, provide means for which the system may accomplish taxa predictions using near-real time remote sensed data to create automated predictions, therefore retaining temporal relevance towards awareness campaigns.

2.1 Predicting Biodiversity

The prediction of marine biology has been the key to the provision of many ecosystem services: marine resources were recently estimated to contribute 16.9% of the animal protein for nutrition worldwide [16]. Quantifying the abundance distribution of marine fauna across spatial areas is important towards the preservation of biodiversity [17, 18], and can provide a baseline against ecological threats such as fisheries, pollution, aquatic industrialization, and others [19].

Aside from biodiversity-related sectors, various industries benefit from protecting biodiversity, including economic, cultural, social, and public-health arguments.

A study regarding the economical reason for conservation of wild nature [20] concluded that the protection of natural habitats generates marked economic benefits, that exceed those obtained from private benefits. Another study concluded an economic evaluation for the conservation of marine biodiversity in the UK [21] and found that a decline in UK marine biodiversity could result in a varying change in the provision of goods and services, including reduced resilience and resistance to change, declining marine environmental health, reduced fisheries potential, and loss of recreational opportunities. Additionally, natural habitats are also proven to be essential towards coastal protection when affected by natural disasters, such as extreme flooding [22, 23], positively impacting the aforementioned sectors.

2.2 Remote Sensing

Remote Sensing (RS) is defined as the science and art of obtaining information about an object, area, or phenomena through the analysis of data acquired by a device that is not in contact with the object, area, or phenomena under investigation. RS has been an important tool to monitor, map, and label various situations without introducing additional intrusive protocols in natural environments. The natural environment is the victim of constant change, which may prove beneficial or not, caused by natural phenomena and/or human activities [24].

There are two types of RS, namely: (i) active; (ii) and passive. Type (i) consists of a detector that emits energy towards the object/area in study, when in contact causes an interaction and consequently a reaction which is captured back by the detector, e.g. RADAR, echolocation, etc [25, 26]. The second type (ii), captures the energy released from the object/area through sensors, as seen in satellite imagery, optical and thermal sensors, etc [27].

With the help of RS, it is possible to collect data relative to inaccessible areas, accordingly, a substantial amount of information becomes available for study. RS has found its applications in hazard assessment, coastal applications, agricultural applications, natural resource management, and many others.

Remotely sensed environmental covariates expand the knowledge regarding marine environment hot-spots and therefore establish areas of marine conservation priority[5]. These measurements have proved an effective tool to identify potential biological and environmental factors that are associated with ocean productivity and the presence and movements of cetaceans [28]. Some example variables inherited from this topic include sea temperature, sea level, salinity, currents, climate variability, water mass, etc.

Aquatic environment research is a growing concern of the scientific community. In recent studies, various aquatic-based monitoring systems have been developed [29–31], in contrast with traditional monitoring systems using oceanographic research vessels, which are expensive, time-consuming, and with low-resolution spectrum. Wireless Sensor Networks (WSN) consists of several dedicated sensor nodes with sensing and computing capabilities and have recently been considered as potentially promising alternatives for monitoring marine environments [32]. WSNs can sense and monitor the physical parameters and trans-

[5] http://www.eomammals.com/

mit the collected data to a central location using wireless communication technologies. When an unusual event is noticed in the networks, an event is detected through the sensor devices placed at distributed locations. This event detection information is passed to the base station and an informed decision is taken.

Apart from marine monitoring, WSN have been widely employed in a variety of application fields related to forest threats [33–35], industrial processes [36,37], agriculture impact [38,39], battlefield surveillance [40,41], intelligent transportation [42,43], smart homes [44,45], animal behavior [46,47], and disaster prevention [48,49]. The majority of use-cases are focused on the collection of environmental covariates, underlining their relevance in the monitorization of marine environments. Some advantages of WSNs consist of unmanned operations, easy deployment, real-time monitoring, and comparatively low cost. Recently, there is an upward tendency towards satellite-based sensing. This technology brings many advantages and promotes inventorying, monitorization, time-frames, and simplifies the integration with other digital services and geographic information systems (GIS) [50]. Satellites have also proven beneficial when there is a need to acquire, analyze and use data documenting the condition of the Earth's resources and environment on a long-term (permanent) basis [51].

All of these characteristics make this technology accessible to earth monitoring studies and systems. An example was a study by Guirado et al. [52], in which aerial Very-High-Resolution (VHR) imagery was gathered from Google Earth, classified with deep learning methods, and later applied in a whale detection model, towards a whale counting system. A similar approach was later employed by C. C. G. Bamford et al. [53] that compared whale density estimations derived from overlapping satellite VHR imagery and a shipborne survey, concluding that satellite imagery can provide useful data on whale occurrence and density estimates.

Satellite surveys represent an exciting development for high-resolution image-based cetacean observation at sea, particularly in inaccessible regions, such as in the Antarctic, where a study [54] detected, differentiated, and quantified the abundance of penguin species by high-resolution satellite imagery, demonstrating that multiple subspecies of penguins colonies can be detected by high-resolution (2-m multi-spectral, 40–50-cm panchromatic) satellite imagery and that under ideal conditions, such imagery is capable of distinguishing among groups of species where they breed contiguously.

There is a vast library of successful marine fauna identification supported by satellite imagery [55–64] presenting opportunities for ongoing and future research.

Land reports also exist, as seen in a study by Duporges et al. [65] that used commercially available VHR satellite imagery and deep learning to detect and count African elephants in heterogeneous landscapes between 2014-2019. Assessing satellite imagery as a tool to track arctic wildlife, such as polar bears [66,67], allows for mapping of vast areas in a maintainable manner, providing increased spatial resolution and increased the ability to detect polar bears, with implications for use with other arctic wildlife.

Several studies show that the performance of satellite imagery can be equal or even better than humans when the quality of data is good, however large-scale applications may require the development of automated detection processes to expedite review and analysis.

2.3 Ecological Niche Modelling

Ecological niche modelling aims to construct a mathematical or computational representation of the environmental tolerances of a species and provide the species potential spatial distribution based on those tolerances. ENMs are an established tool for quantifying biodiversity presence in the environment, while the availability of publicly available databases has augmented the interest in their use, cementing themselves as the default technology for predicting marine biodiversity presence and absence locations.

Over the past years, ENM models have become commonplace in studies of biogeography, conservation biology, ecology, paleontology, and wildlife management [68].

Various protocols are used to predict the spatial distribution of biodiversity based on both presence and absence data [69], predicting occurrences of zone-specific taxa with overwhelmingly positive results.

Comparative studies also exist, such as the study provided by Lozier et al. [70] correlating the distribution of the North American Sasquatch, commonly referred to as Bigfoot, with the black bear, *Ursus americanus*, suggesting that most sightings of this mythical creature were indeed cases of mistaken identity of this local specimen. Another study [71] compared models of distribution for *Lophelia* *pertusa* and the reef habitat it forms. Recent modelling efforts have focused on predicting the distribution of this species, however, the species is widely distributed whereas reef habitat is not, benefiting from a comparative study. Nonetheless, assessment of habitat extent based on species distribution will be overestimated. An ecological approach includes predicting habitat suitability from the occurrence of vulnerable marine ecosystems [72], concluding that each model predicted similar areas of suitable habitat both in the vicinity of known vulnerable marine ecosystems indicator taxa presence locations as well as across broad regions of un-sampled seafloor.

As referenced ENMs are widely used for habitat modelling, as presented in a study from Bombosch et al. [73] that predicted the habitat suitability patterns of humpback and Antarctic minke whales in the Southern Ocean. The spatial prediction maps acquired provided not only a valuable planning tool in the context of seismic activities but were also used to direct more detailed cetacean-habitat studies. A similar study [74] was employed to cold-water corals along the Irish continental slope, resulting in an accuracy of $25km^2$ at a local scale and $4000km^2$ at a regional scale.

Previous studies have demonstrated a need for increased rigor in building and evaluating ENMs [75]. These tools, however, are subject to several methodological issues including the challenge of balancing the model complexity and the performance of model selection criteria [10], and the need to evaluate model accuracy with cross-validation [76, 77].

Although the aforementioned work provides an important approach in predicting the abundance of marine biodiversity, they are still an important tool for alerting the proper authorities. Nevertheless, such an approach remains cumbersome to use as the greater audience does not have the knowledge required to interpret ENMs. Moreover, existing interfaces do not provide a scalable approach in selecting the desired model, which this dissertation proposes.

There is no doubt that prediction protocols are successful when employed in ecological research, however, could such technology be used in other areas, such as marine litter and natural disasters predictions?

2.3.1 Assistance packages

There is an abundance of specialized software that supports the application of ENM protocols, through deep learning and statistical analysis [75, 78–84].

Although there are so many options, none of them managed to both establish scientific credibility and beginner-friendly interactions, absent of convoluted setup [85].

As a result of the continuous growth of species distribution systems, software innovations are staling, with groups of researchers collaborating primarily with others who use the same software, presenting a barrier to ENM's further innovation [84]. The most commonly used software specifically for ENM research are R (81%) and Maxent (64%), when judged on (i) usefulness; (ii) learning; (iii) system capabilities; (iv) user satisfaction. Concluding that the two most popular software for ENM lies at opposite ends of the use complexity continuum [86].

Kuenm [80] was developed towards detailed development of ecological niche models using Maxent, to enable detailed model calibration and selection, final model creation and evaluation, and extrapolation risk analysis. Unlike Kuenm, Scikit-learn [83] enabled an easy comparison of methods between applications integrating a wide range of machine learning algorithms for statistical data analysis. Since it relies on the Python ecosystem, it can easily be integrated into applications outside the traditional range of statistical data analysis.

The duality between geographic and environmental spaces proposed by G. Evelyn Hutchinson provides a powerful way to conceptualize and analyze biogeographical distributions concerning spatial environmental patterns. Hutchinson's duality has been used to classify and map environments; model potential species distributions under past, present, and future climates; study the distributions of invasive species; discover new species; and simulate increasingly more realistic worlds, leading to spatially explicit, stochastic models that encompass speciation, extinction, range expansion, and evolutionary adaptation to changing environments [87].

NicheA [78] is free, open-source software, available under the GNU Lesser General Public License, that reinforces Hutchinson's duality, creating virtual species and ecological niches in multivariate environmental scenarios. This package projects to the geographic space in the form of continuous or binary species distribution models. NicheA allows users to compare data through visualizations of linked environmental and geographic spaces. Highlighting the fact that it incorporates a command-line interface, however, includes a GUI output visualization.

Similarly, BIOMOD [81] offers the possibility to run 10 state-of-the-art modelling algorithms to describe and model the relationships between a given species and its environment, as well as test models with a wide range of approaches, project species distributions into different environmental conditions (e.g. climate or land-use change scenarios) and dispersal functions. It allows assessing species temporal turnover, plot species response curves, and test the strength of species interactions with predictor variables.

Another concern of these assistance packages is evaluating ENMs performance. ENMeval [75] is an R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ENMs, enabling cross-validation workflows with multiple evaluation metrics to aid in selecting optimal model parameters. Efficient implementation of procedures to accomplish these goals, however, requires automation. NichePy [79] is another project focused on delivering modular tools for estimating the similarity of the ecological niche and species distribution models, to evaluate the robustness of models.

There is a big concern related to ENM cross-validation and model analysis. In this scope, projects such as mMWeb [88] and SDMtoolbox [89] provided well-organized user interfaces towards landscape genetic, biogeographic, and species distribution model analyses. The formed combines 18 existing ENMs and their corresponding algorithms to provide a uniform procedure for modelling the potential habitat niche of a species via a web browser, while latter consists of 59 Python script-based GIS tools developed and compiled into a single interface designed to facilitate complicated pre-and post-processing steps commonly required for ENMs and other geospatial analyses. Such systems are critical for robust study design and accurate interpretation of outputs [90].

Ecological and geographical studies are convoluted and possess overwhelming concerns, resulting in complex analyses. Recently, the scientific community has made an effort to extend the accessibility of such studies. As a result, unified structures have been developed, as seen in sdm [91], an object-oriented, reproducible and extensible, platform for environmental niche modeling. Likewise, ssdm [92] is an r package to predict the distribution of species richness and composition based on stacked species distribution algorithms. The object-oriented design from both packages is such that: users can modify existing methods, extend the framework by implementing new methods, and share them to be reproduced by others, propelling modular status.

2.3.2 Interface

Most approaches are accessed as scripts in a CLI (e.g. Python and R) or through graphical user interfaces (GUI). The former one provides to the user the most control operations with the best performance and flexibility, while the latter focus on usability and reduces the learning curve of the system [93, 94]. However, studies should not be limited by only one of the interface's methods, on the contrary, the simplicity of a CLI + GUI interface allows for maximal flexibility and extensibility, without sacrificing usability and credibility [95]. Hence, tools that combine the positive aspects of CLI and GUI methods can help advance ecological research.

A recent study from Sebastian H. and Robin O. [96] evaluated the usage of CLI and GUI interactions with Git[6] that confirmed that beginners generally benefit from GUI clients, however, most change as they get more experience and use either exclusively CLI or both GUIs and CLI together, maximizing processes' efficiency.

Although previous systems extend the accessibility of analyses to a more generalized user-base, few of them have the characteristic of being a GUI and at the same time provide the functions to employ protocols via command-line (i.e. sdm, ssdm).

Wallace [97] is a flexible platform for reproducible modelling of species niches and distributions built for community expansion that addresses these issues specifically for user communities in ecology and the environmental sciences. Wallace provides an example of an innovative platform that guides users from acquiring and processing data to building models and examining predictions. NicheToolBox, or ntbox [85], is a comparable system to Wallace, in that it allows a fast and straightforward means by which to retrieve and manipulate occurrence and environmental data.

Although there are various solution available on the market, none of them managed to establish both scientific credibility and beginner-friendly interactions, absent of convoluted setup. As such it's possible to identify a scarcity of solutions for ecologists without technological backgrounds.

[6] `http://git-scm.com`

Author	Interface	Product
Osorio-Olvera et al. [85]	GUI+CLI	R package
Naimi et al. [91]	GUI+CLI	R platform
Schmitt et al. [92]	GUI+CLI	R package
H. Qiao et al. [88]	GUI	JNI[7] and JRI[8] web platform
Wallace et al. [8]	GUI	R web platform
Brown et al. [89]	GUI	Python toolbox
Qiao et al. [85]	CLI	Java toolkit
Bentlage et al. [79]	CLI	Python package
Cobos et al. [80]	CLI	R package
Muscarella et al. [75]	CLI	R package
Thuiller et al. [81]	CLI	R package
Hijmans et al. [82]	CLI	R package
Pedregosa et al. [83]	CLI	Python module

Table 1: Assistance packages overview, based on their interface category, development, and employment technology.

2.4 Web-based Interfaces

An interface is the communication layer between the user and the system, consequently, it may promote said system or even introduce unexpected limitations, therefore establishing a "bottleneck" [98]. As a result, the development of user interfaces is time-consuming, complex, and difficult to implement and analyze. An average of 48% of the code and 50% of the projects' timeline is directed towards user interface development [99]. The most common problems developers report when developing a user interface are: (i) users' requirements; (ii) consistency; (iii) performance; (iv) and system integrity [99].

In the past decade, various marine-based interfaces have been developed, as in Whale Alert[9], a mobile application formed by a network of institutions focused on detecting and reducing lethal whale ship-strikes worldwide. Some other examples include Wave Labs' mobile apps[10], Litter Reporter, Dive Reporter,

[9] http://www.whalealert.org/

[10] http://wave-labs.org/kits/apps

and Whale Reporter performing underwater and surface aquatic assessments adopting citizen scientists. These apps are report-based systems, with simplistic interfaces and bold actions, as visible in figures 1, 2A, and 2B.

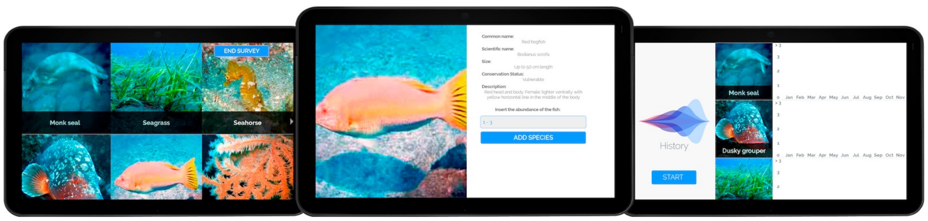

Fig. 1: Dive Reporter GUI

Mobile-Based Marine Tracking projects are being globally promoted by stakeholders, due to the widespread utilization of mobile devices. Successful studies have emerged from this investment, highlighting Andriolo et al. [100] and Jambeck et al. [101], which managed to generate 400,000 reports over an area of 16,000m^2, regarding marine litter mapping and data collection.

Mobile-based marine debris tracking has been positively accepted by end-users, as reported on a study [102] that resulted in an average rating of 4.22/5 regarding Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Attitude Towards Usage (ATU), and Behavior Towards Usage (BTU). This high acceptance also highlights the prospect of using such technologies towards alternative marine systems. Amongst the variables, PU received the highest mean of 4.42 showing that end-users perceived the usefulness of such a system.

Web-applications have also benefited from the mobile application market growth, as seen in platforms such as Cornell Bioacoustics NRW Monitoring[11], WhaleMap[12], Wave Labs[13], Map of life[14], and even the European Environment Agency's (EEA)[15].

Comparatively, mobile applications are more adopted in terms of accessibility, while websites focus on means of search-ability [103], confirming the current

[11] https://portal.nrwbuoys.org/
[12] https://whalemap.ocean.dal.ca/
[13] http://wave-labs.org/
[14] https://mol.org/
[15] https://www.eea.europa.eu/themes/water

scientific tendencies which tend for mobile applications to be focused on reporting and collecting information, while websites focus on data query, visualization, and statistical analysis, mostly due to computing power, interaction methods (e.g. display, duration and time of interactions, workflows, search approaches, layout, and visual clues) and social factors (e.g. environment, user-base, consumer perceptions, and expectations) [104].

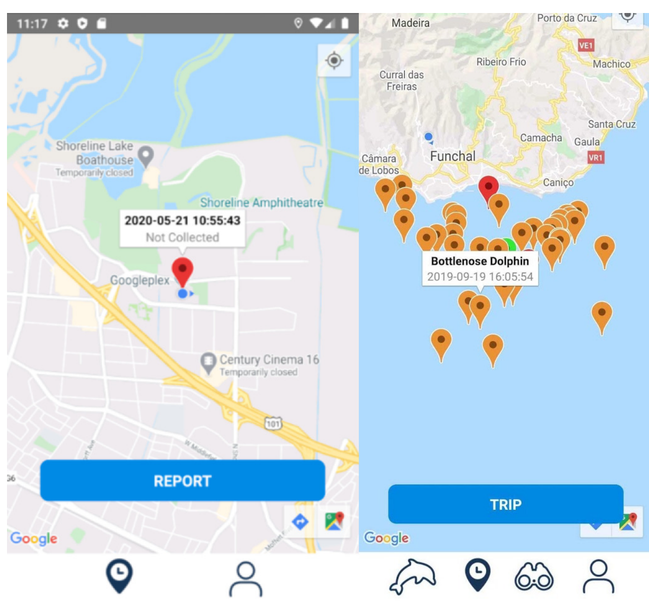

Fig. 2: Mobile application GUI (from left to right): A) Litter Reporter; B) Whale Reporter.

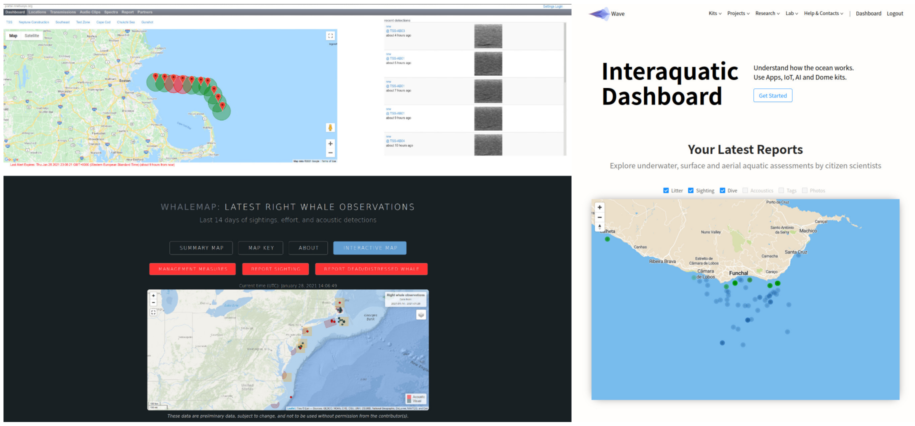

Fig. 3: Web-based interfaces (from top-left to bottom-right): A) Cornell Bioacoustics NRW Monitoring; B) WhaleMap; C) Wave.

2.5 Summary

Notwithstanding the benefits that previous systems have brought into the ecologic scientific community, at the time of writing, no system provides a beginner-friendly "plug-and-play" workflow. All projects, even if some more than others, require previous knowledge of species distribution systems and manual imports of the packages using some sort of programming language (python, java, or R). There is an opportunity for an ecological modelling platform that supports both beginners and experts equally. Moreover, although existing solutions are overwhelming, the proposed apparatus will provide simplistic features that may be filtered by more knowledgeable users.

Additionally, from the literature review, it was not possible to notice any system that includes pipelines of data with online scientific repositories. Lastly, the component responsible to establish the protocol mentioned in section 1.3 will be developed on R programming language, a free software environment for statistical computing, as a result of the tendency towards it, visible in table 1.

This dissertation contributes by providing a state-of-the-art way of employing species distribution models while promoting and expanding ecological awareness.

3 Methodology

This chapter includes detailed descriptions of technologies, techniques, and design choices regarding the proposed system. Initially, in section 3.1 the workflow of the system is presented as is the technology employed, easing the description of future sections by cross-referencing to this section. Then in the following section 3.2, it's describes the required methodology to generate correlative species distribution models, also known as climate envelope models or bio-climatic models, therefore establishing an ENM. Afterward, in section 3.3, the system's apparatus is presented by analyzing the system's architecture, database structure, communication protocol, and logic. Subsequently, data pipelines are described in chapter 3.4, namely Copernicus environmental data and scientific repositories occurrence points. Finally, in section 3.5, the study setup is described, presenting obtained data inquiry, sample size, and the conducted experiments.

3.1 Apparatus

Proposed apparatus includes the interaction workflow and technology regarding the development of the system. The preceding details teh arrangement os tasks necessary to perform predictions, while the latter introduce technological requirements for the development of the dissertation.

3.1.1 Workflow

As observed on previous studies methodologies [75, 78–85, 97], presented in section 2, the dissertation encodes the workflow in four key steps: (i) selecting the occurrence data; (ii) obtain environmental covariates; (iii) fitting one or more algorithms and generating any outputs; (iv) visualize the prediction(s).

This segmentation of the process recognizes an opportunity for parallel workflow between the front-end and back-end, considering that while the client interacts with the system, at each step there are several HTTPS requests, outputting progress outputs, resulting in a significant decrease in user wait-time.

We consider wait-time the number of seconds that the client has to wait until it obtains output results. Without parallelization, the user would complete all ordered steps and then the server would compute as many outputs as required in the request. This approach results in a total wait-time represented by equation 2.

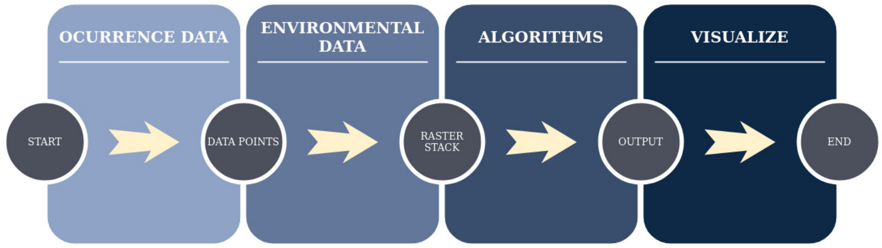

Fig. 4: System workflow diagram

$$waitTime = \sum_{n=1}^{steps} serverTime + userTime \quad (2)$$

On the other hand, with parallelization, the server would compute intermediary steps at the same time that the user executes steps, and therefore reducing the total wait-time. This approach requires common time between the server load and user interactions, meaning that we can subtract this common time to equation 2 and conclude equation 3.

$$waitTime = (\sum_{n=1}^{steps} serverTime + userTime) - commonTime \quad (3)$$

Additional factors may exist in a real use-case, such as latency and client-side rendering, however, common time will always exist meaning that the result from equation 3 will always be lesser than the output from equation 2.

This design choice highly benefits the system's overall performance and consequently user experience. Previous behavior is presented in figure 5, in which the chart on the left represents parallelization and the other the sequential approach.

3.1.2 Technology

As aforementioned, the system will be developed and supported through Wave Labs[16], a web-based application that directly contributes to understanding the impact of human presence on marine environments and provides an insight into

[16] http://wave-labs.org/

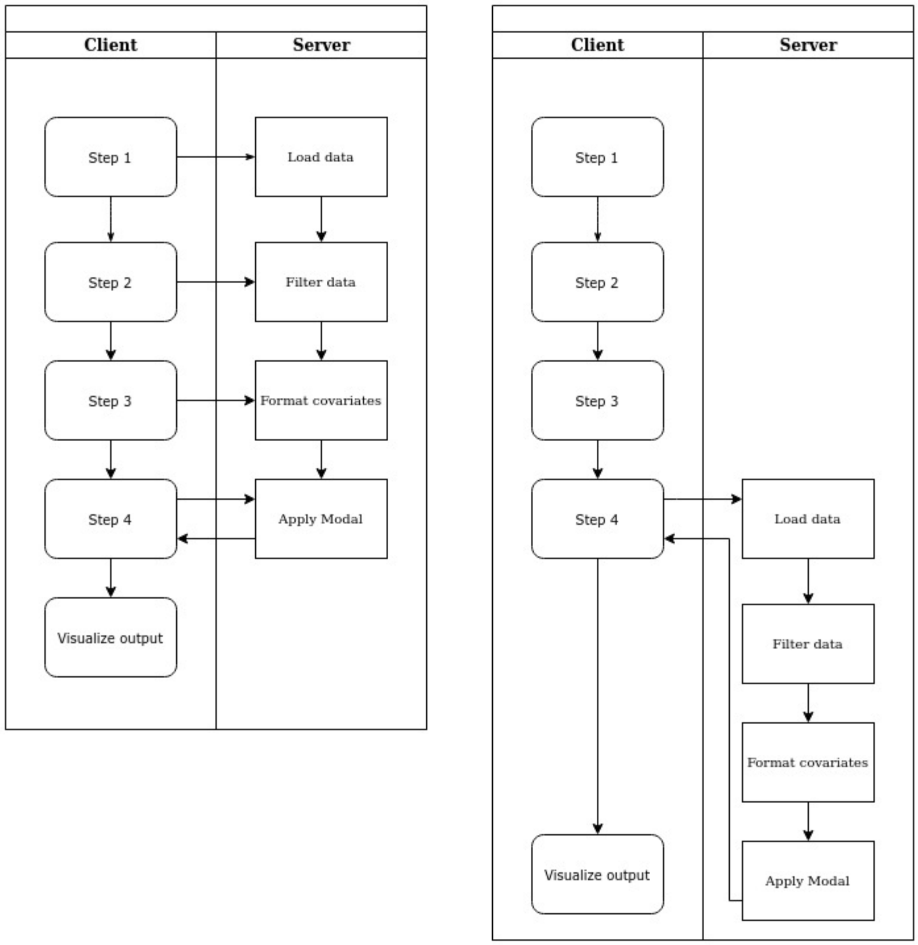

Fig. 5: Wait-time diagram (from left to right): A) With parallelization; B) Without parallelization.

recent cetacean sightings, underwater taxa, and marine litter near Madeira island.

Wave Labs are developed as a Laravel[17] RESTful API structured in a Model-Route-Controller (MRC) structure, using a relational SQL database for storing data, that communicates through HTTPS with a React.js[18] class-based application. The latter controls the workflow on the client-side, managing states with

[17] `https://laravel.com/`
[18] `https://reactjs.org/`

Redux[19] and presenting information with MapBox[20], Ant Design[21], and Styled Components[22].

Accordingly, the dissertation will be based on the same technologies as Wave Labs, while featuring the programming language R to execute the necessary scripts that administer species distribution modelling.

- **Laravel**[23] is a php web application framework for the development of the API logic and structure;
- **React.js**[24] is a JavaScript library for building and managing user interfaces;
- **R**[25] is a statistical computing language and environment for prediction outputting and graphics.

3.2 Environmental Niche Models

This section provides an overview of the protocol established to execute species distribution modelling. As mentioned, ENMs were implemented with the programming language R and executed with a Rscript file containing a set of commands to execute under the R global environment.

The structure of this component is a modular architecture, meaning that each task, as small as it may be, is stored in a capsulized and individual script (e.g. remove duplicates, call a specific algorithm, etc). This option is required to guarantee the maintainability of the project, as a consequence of the ever-recurring upgradability towards new algorithms, changes in protocols, and/or new features.

This decision requires the system to comply with the open-closed principle[26], which defends that software entities should be open for extension but closed for modification. Totalling 13 different scripts that are dependent between themselves, presented in figure 6.

[19] https://react-redux.js.org/
[20] https://www.mapbox.com/
[21] https://ant.design/
[22] https://styled-components.com/
[23] https://laravel.com/
[24] https://reactjs.org/
[25] https://www.r-project.org/
[26] https://en.wikipedia.org/wiki/Open%E2%80%93closed_principle

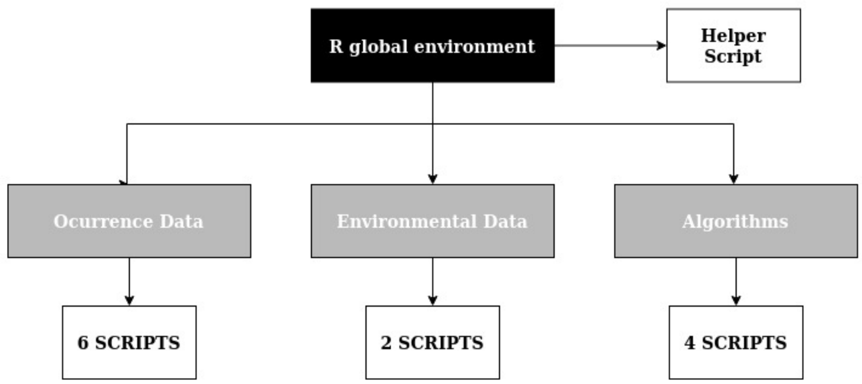

Fig. 6: Modular Architecture

3.2.1 Data preparation

The dissertation currently supports three data input sources, namely: (i) user uploaded comma-separated values (CSV) files; (ii) Wave Labs platform; (iii) public available scientific repositories.

The data collected from source (ii) refers to Wave Labs' application section, which gathers data from citizen scientists. From these applications, two are implemented as a data source, Whale Reporter[27] and Dive Reporter[28];

The first app studies the value of whales and the economic impact of whale watching as a form of ecotourism, for on- and off-shore whale watching, offering an unobtrusive way to spot cetaceans to greater audiences. The latter uses scuba divers as citizen scientists for post-dive biodiversity abundance estimations, used for longitudinal studies necessary to marine biologists.

Regarding data input (iii), these require georeferenced data points for biodiversity occurrences. Following repositories were included:

- Global Biodiversity Information Facility (GBIF)[29]
- Biodiversity Information Serving Our Nation (BISON)[30]

[27] https://play.google.com/store/apps/details?id=com.tigerwhale.sighting
[28] https://play.google.com/store/apps/details?id=com.tigerwhale.dive
[29] https://www.gbif.org/
[30] https://bison.usgs.gov/

- Ocean Biogeographic Information System (OBIS)[31]

As mentioned before, different algorithms sometimes use the different data formats, therefore, data requires normalization. All occurrence data points were stored in comma-separated values (CSV), in a two-column matrix, indexed by their coordinates. At this moment, the system also removes unwanted entries, such as incomplete geo positions and invalid coordinates. Finally, if necessary, rows are converted from strings (eg. xx,x) to numerical values (eg. xx.x) and columns are reordered for (longitude, latitude) to match the algorithm's parameters. Hereinafter, the system can manipulate whatever data it receives from any of the previously mentioned sources without the necessity of preceding knowledge.

Additionally, the system supports mechanisms of data pre-processing, sometimes referred to as data cleaning, for removing duplicate records, sampling bias[32], crop-based on geographic extents, and cross-checking.

3.2.2 Environmental Data

Similar to data preparation, environmental data also involve input data sources, but in this case, it should refer to georeferenced environmental measures (eg. temperature, salinity, etc). Incorporated sources include: (i) user uploaded directory, containing TIF (Tagged Image Format), ASC (armored ASCII file), or NC (ASCII text file) files; (ii) Wordclim's bioclimatic variables derived from the monthly temperature and rainfall values[33]; (iii) Copernicus Marine Environment Monitoring Service (CMEMS)[34].

Wordclim's bioclimatic variables represent annual trends and are presented as follows[35]:

- BIO1 = Annual Mean Temperature
- BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
- BIO3 = Isothermality (BIO2/BIO7) ($\times$100)
- BIO4 = Temperature Seasonality (standard deviation $\times$100)

[31] `https://obis.org/`

[32] Sampling bias is frequently present in occurrence records, but it's possible to remove some of the bias by sub-sampling records, limiting the number of records.

[33] `https://www.worldclim.org/`

[34] `https://marine.copernicus.eu/`

[35] `https://www.worldclim.org/data/bioclim.html`

- BIO5 = Max Temperature of Warmest Month
- BIO6 = Min Temperature of Coldest Month
- BIO7 = Temperature Annual Range (BIO5-BIO6)
- BIO8 = Mean Temperature of Wettest Quarter
- BIO9 = Mean Temperature of Driest Quarter
- BIO10 = Mean Temperature of Warmest Quarter
- BIO11 = Mean Temperature of Coldest Quarter
- BIO12 = Annual Precipitation
- BIO13 = Precipitation of Wettest Month
- BIO14 = Precipitation of Driest Month
- BIO15 = Precipitation Seasonality (Coefficient of Variation)
- BIO16 = Precipitation of Wettest Quarter
- BIO17 = Precipitation of Driest Quarter
- BIO18 = Precipitation of Warmest Quarter
- BIO19 = Precipitation of Coldest Quarter

The usage of all these variables for one specific prediction output may lead to over-fitting of the model, which in turn may result in misrepresentations of species' climate-envelope, for this reason, variables are chosen by the user and provided as arguments [105].

Occurrence data points had to be normalized, likewise, environmental co-variates require the same protocol. However, instead of storing all files in the same format, environmental points also required data to be imported into R's global environment as a RasterStack collection, consisting of RasterLayer objects. These objects are commonly treated as gridded data, or vectors, meaning that the file will be stored as a collection of variables (e.g. temperature, salinity). Now each predictor should be a RasterLayer representing a variable of interest.

Additionally, it's necessary to drop categorical variables from the predictors' stack, because some models cannot use these variables and it would be time-cost inefficient to dynamically apply this feature.

Environmental data is always cropped to fit the geographic extent of the original occurrence points due to performance motives. For any particular prediction, all variables need to have the same geographic extent and resolution. Functions *crop*, *extend* and *aggregate* are used to accommodate such requirements. We now have a set of predictor variables (rasters) and occurrence points.

The next step is to extract the values of the predictors at the locations of the points and combine them into a single data frame.

3.2.3 Modelling algorithms

Once again, for any of the steps associated with species distribution modelling, the process requires normalization and data treatment. As mentioned previously in chapter 3.2.1, the scripts' modular architecture proposes that each algorithm will be encapsulated on its own script, however, this would generate unnecessary code duplication for each new algorithm. There is a helper script that normalizes the protocol for employing algorithms that receive arguments indicating which algorithm(s) it should execute. This behavior is visible in figure 7.

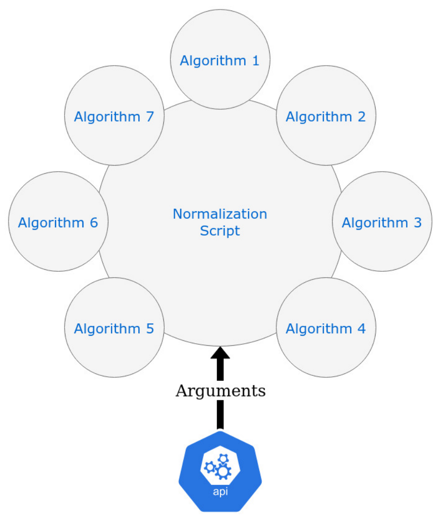

Fig. 7: Helper script that normalizes algorithms step

At the time of deliverable, the system includes two profile algorithms and one machine learning algorithm for modelling methods, respectively BIOCLIM, Domain, and MaxEnt. The developed proposal exclusively accepts presence-only points in the modelling process.

BIOCLIM was the first widely used ENM algorithm for quantifying the environmental niche of species and assessing the likely impacts of climate change on species distributions. Bioclimatic variables developed for BIOCLIM were used to create the Worldclim database, and therefore feeding many other algorithms, due to about 76% of recent published MaxEnt analyses of terrestrial ecosystems using this database [106]. Despite this method's low accuracy when compared to some modern alternatives, it is still used, because of its simplicity and performance and thus usability in teaching species distribution modelling [107]. The BIOCLIM algorithm computes the similarity of a location by comparing the values of environmental variables at any location to a percentile distribution of the values at known locations of occurrence.

The **Domain** algorithm it's another profile algorithm that has been extensively used for species distribution modelling in the past. Similarly to BIOCLIM does not generate extremely precise assessments, however, it's variable sensitive, performs well with limited site data, and is easily implemented. The algorithm computes potential distributions based on a range of values, point-to-point similarity metrics and provides a simple, robust method for modelling distributions of biodiversity [108].

MaxEnt (short for "Maximum Entropy") is the most widely used ENM algorithm. MaxEnt it's a machine learning method and is available as a stand-alone program for modelling species distributions from presence-only species records. Dismo[36] has a method "maxent" that communicates with this program, therefore it is possible to fit it like the profile methods (e.g. BIOCLIM, Domain). That is, you can provide presence points and a RasterStack.

3.2.4 Building the model

Furthermore, it is necessary to create background data, or pseudo-absence points to provide a comparative data set and enable the conditions under which a species occurs to be contrasted against those where it is absent. This data is obtained through a mask of random points limited by the original data's rectangular geographic extent, sample size, and resolution.

Let's create a vector of group memberships with k-fold partitioning and separate occurrence into training and testing groups.

The procedure of k-fold consists in shuffling the dataset randomly and then split it into k groups. Based on algorithm 1, if we define an index that will for sure exist, such as 1, we can establish that the testing group will be identified

[36] Package of methods for species distribution modelling

Algorithm 1 k-fold partitioning

Require: $k \geq 1$, $k \leq 10$, $index = 1$, $data$

$groups \leftarrow$ kfold(data.occurrence, k)
$data.train \leftarrow data.occurrence[groups \neq index]$
$data.test \leftarrow data.occurrence[groups == index]$

by that index, while the remaining will create the training group. This strategy means that the percentage of data for training/testing groups is defined by the value k (e.g. with k = 5, exists 4 groups of training and 1 for testing, or 75% training and 25% for testing). Now repeat the process for pseudo-absence points.

Finally, let's build the species distribution model and predict presence from the model by using the RasterStack with predictor variables to predict a RasterLayer.

Algorithm 2 Build model and predict

Require: $data$, $geographic.extent$

$model$ $\leftarrow$ model(environmental = $data.environmental$, occurrence = $data.train$)
$prediction \leftarrow$ predict(object = $model$, ext = $geographic.extent$)

Afterward, it is possible to plot the raster object generated from the output of the prediction function and store it on the server's file system. This plot can be viewed as a map of suitability scores, as visible in image 8.

Most modelling methods have different measures that can help to assess how well the model fits the data. The process of evaluating the prediction is granted with Area Under The Curve (AUC), Receiver Operating Characteristics (ROC) curve, True Positive Rate (TPR), and environmental variable response plots. The former represents the degree or measure of separability, while the latter curve it's the probability itself. The ROC curve is plotted from the measurements of the true positive rate and false-positive rate.

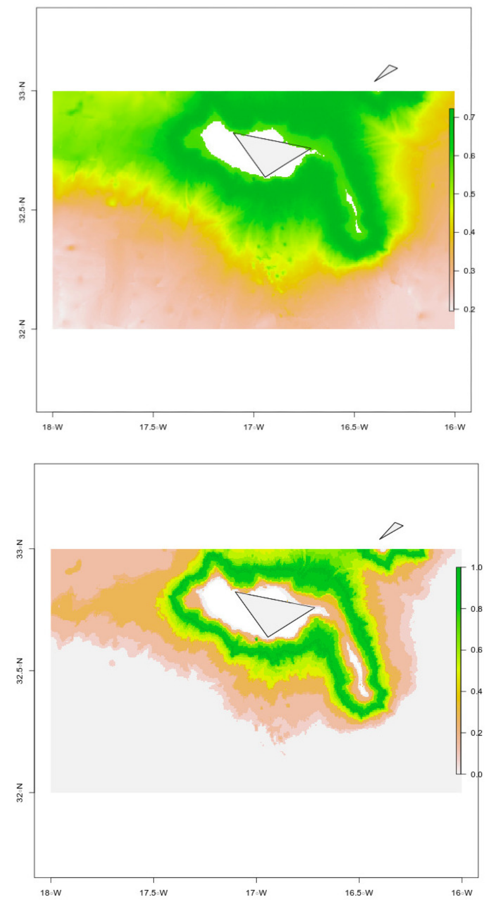

Fig. 8: Prediction plot of the raster object based on temperature collected from Copernicus' "Global Ocean 1/4° Physics Analysis and Forecast updated Daily" and common dolphins occurrences around Madeira's coast line.

3.3 System Architecture

The architecture at figure 9 consists of three (3) components, that define the formal description and representation of the system, namely: (i) client; (ii) server; (iii) and scripts. The client communicates with the server by dispatching actions to the server's routing. Then the indicated can execute scripts through processes tasks in the model, receiving status messages as responses. The server can then return information to the client through resources that are received back in the actions. A communication swim-lane (figure 10) demonstrates the interpolation between the three components of the architecture with an additional lane as process' status.

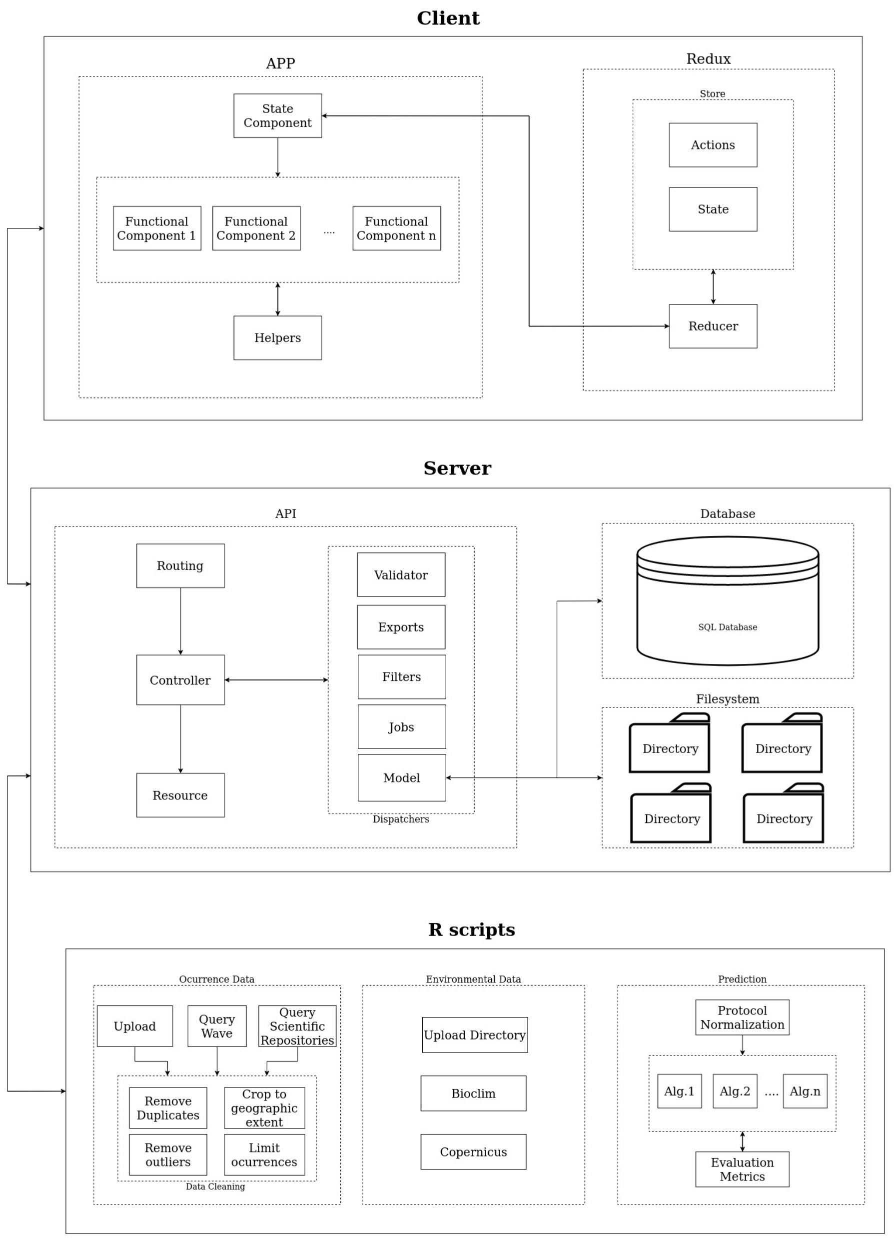

Fig. 9: System Architecture

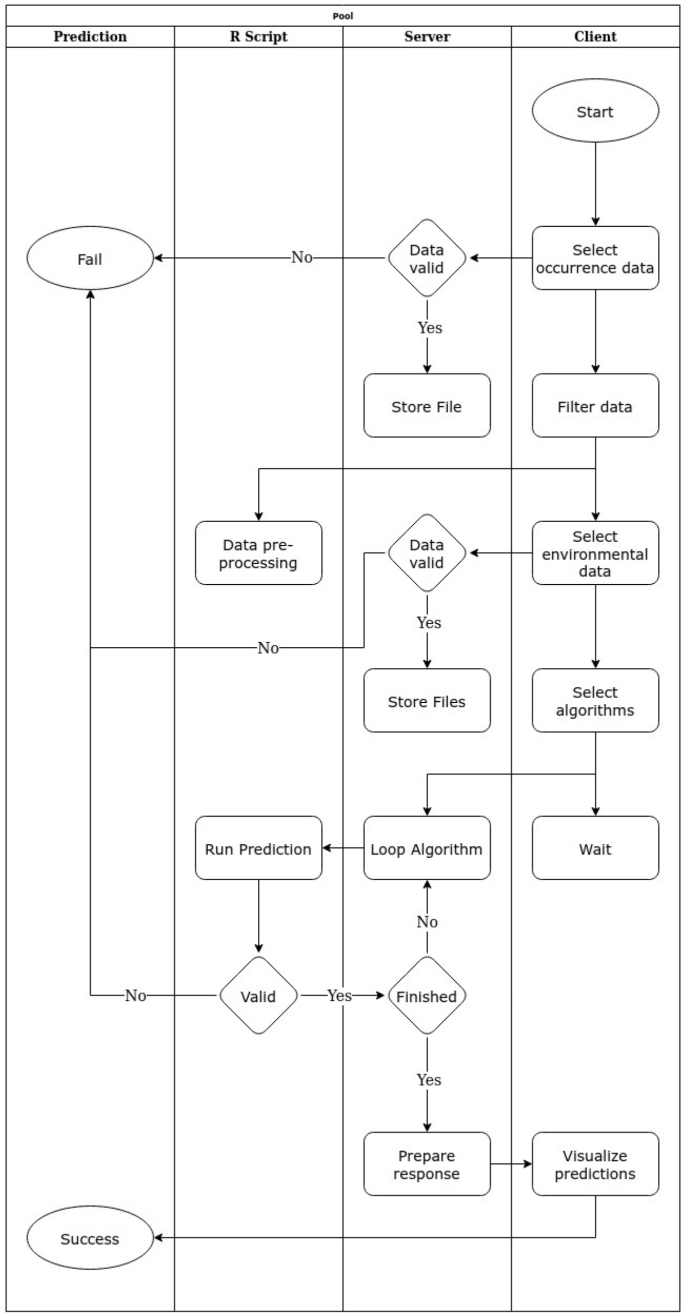

Fig. 10: Communication between R scripts, Server, and Client components of the architecture with an additional lane as process' status during the workflow.

3.3.1 Client - User Interface

The client consists of the web browser/application used to interact with Wave Labs. Here the user has all control regarding data input, formatting options, normalization filters, and data visualization.

The graphical UI's file structure is segmented into two sections, the application itself and a data store managed by redux.

The store is defined with "Actions" which execute Axios requests to get, update, create, and delete information from the back-end. The response from these requests directly impacts the values in the "State". To access this data or execute actions from the application, the system relies on reducers as a communication interface, reading or calling actions as needed.

However, not all application components should interact with the store, only one single parent state component has a connection established with the reducer. From these components, data is transferred as properties, or props as referred to in React's nomenclature. The remaining components are built as functional components, or simply JavaScript Functions being React Components which return JSX (React's Syntax).

Every one of these components possesses an internal and hidden store, also named State, which is accessible only from the component itself.

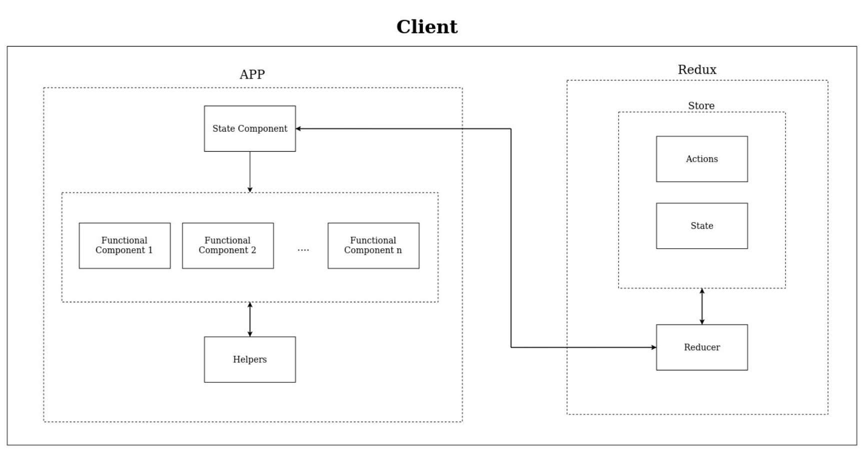

Fig. 11: Architecture - Client component

As aforementioned in section 3.1.1 and explained in the figure 4, the developed component is implemented in the Wave Labs dashboard and divided into four tasks. Following the established order of this dissertation, subsequent images represent (i) landing page and previous predictions table; (ii) Selection and filtering for occurrence data sources; (iii) Selection of environmental data source and covariates (iv) loading page; (v) and output visualization.

The landing page serves as an information provider regarding the dissertation workflow and data display mechanisms for previous predictions. Here the user may resume previous workflows and revisit already established predictions. Failed protocols can not be redone yet (fig. 12).

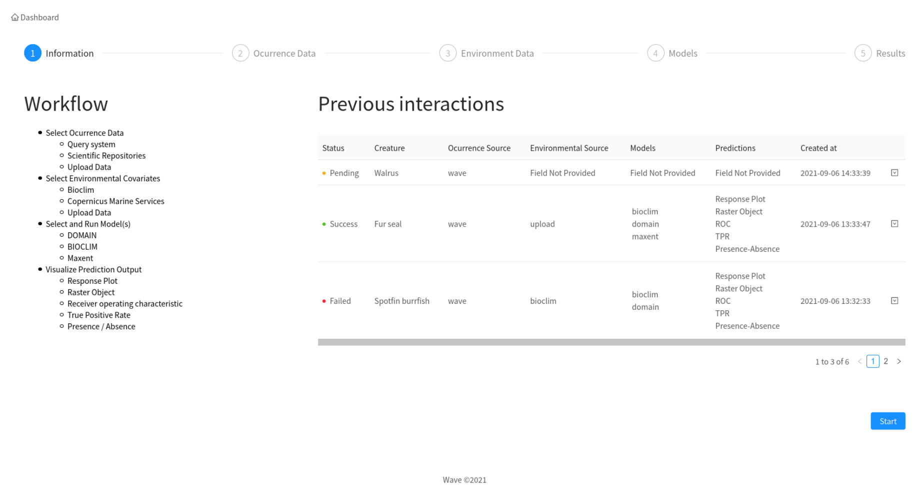

Fig. 12: Wave Labs Dashboard - Landing page

An occurrence data source can be selected from the options list on the left, while filters dependent on the options are presented on the right side menu. Below the selection box, there is a map view and table view with real-time consequences for the filters selected on top (fig. 14 A).

Likewise, environmental data sources, are displayed on the options list on the left, while the filters are displayed on the right side menu, considering consistency as an essential factor of the system. Below the selection box, in case of BIOCLIM data source, there is imagery for each of the variables for the current prediction. These are not the most recent data available due to inherited performance costs from generating these images for each new prediction process (fig. 14 B), however such imagery functions solely as feedback and the most recent values are always used. Additionally, option "Upload" grants a file upload component that supports directories of .nc, .asc, and .tiff files, and "Copernicus" supports the selection of six Copernicus products and a time filter for said product's data.

The model selection page only has a top filter component regarding the k-fold group size and which algorithms(s) to use on the prediction (fig. 14 C). The slider available is limited from two to ten and defaults to value five, and the user may choose how many of the algorithms as he so wishes to execute.

Hereinafter the system will compute the output of each algorithm chosen, thus the user needs to wait for output visualization. Consequently, how many more algorithms selected, greater the computation load and therefore time costs. Depending on the data and number of algorithms, the process may take up to 240 seconds, or 4 minutes for each of the algorithms, thus the user is presented with a loading page with Wave Labs' logo doing transverse waves until the process concludes (fig. 13). Finally, if the all algorithms have executed successfully, maps suitability scores are presented with performance metrics described previously as imagery plots (fig. 14 D). Some of metrics of evaluation include: (i) presence-absence maps; (ii) area under the ROC curve; (iii) true positive rate; (iv) and response plots based on environmental variables values range.

Fig. 13: Wave Labs Dashboard - Loading page

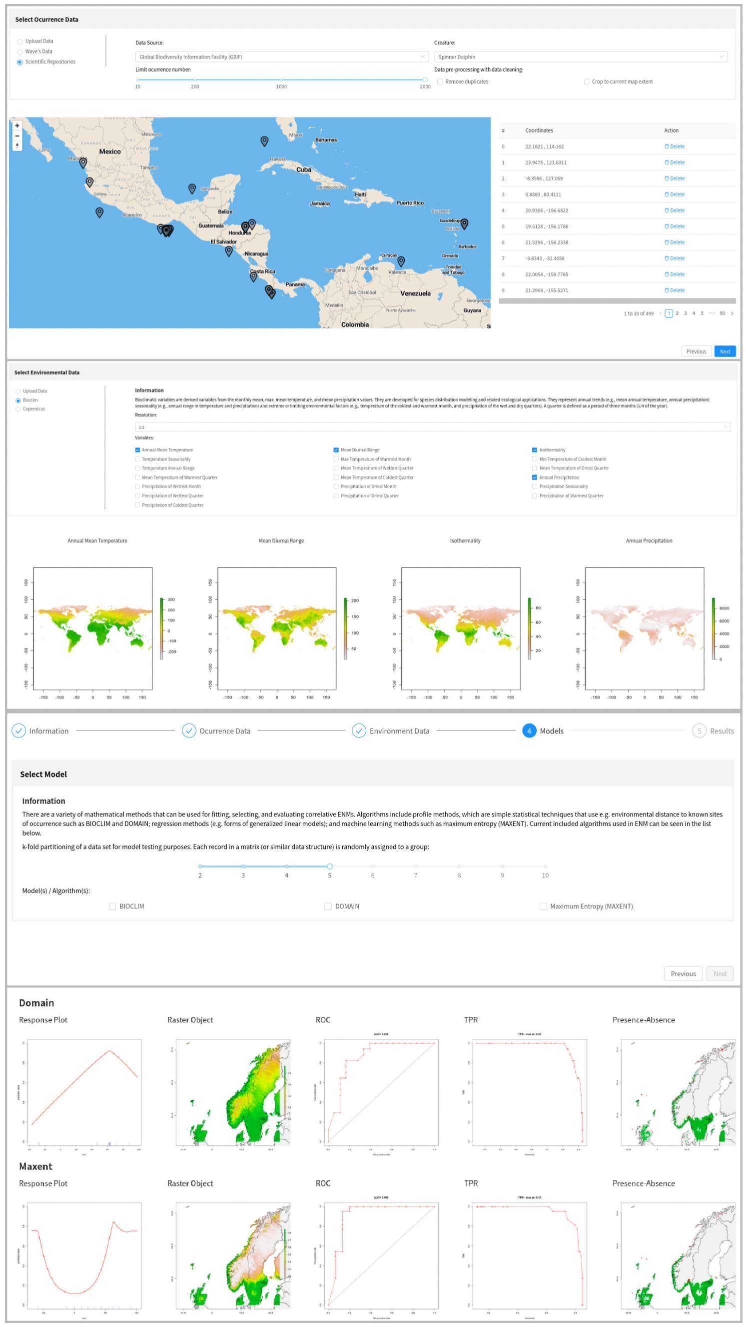

Fig. 14: From top to bottom: A) Wave Labs Dashboard - Occurrence data ; B) Wave Labs Dashboard - Environmental data; C) Wave Labs Dashboard - Model selection; D) Wave Labs Dashboard - Output visualization.

3.3.2 Server

The server is responsible for establishing the connection between remaing components by dispatching scripts and returning responses to the client. The aforementioned aspects are established through the API and database.

The Application programming interface (API). API is defined as a set of specifications, namely Hypertext Transfer Protocol (HTTP) request messages with JavaScript Object Notation (JSON) data structures and laravel eloquent models.

When the system receives a request, is loaded by the Route Service Provider that connects with the stateless API middleware group. Subsequently, the request information is redirected to the associated "Controller", where it delegates tasks to the correspondent dispatcher(ers).

The dispatcher may be an *Export* class handling the CSV files' data, it may be a *Filter* component filtering incoming and exporting data, a *Validator* managing incoming request parameters, a *Job* to wrap custom logic around the execution of queueable tasks which run asynchronously to the application, or a "*Model* that contains all of the Eloquent model classes with Active Record implementation for working with the database (referred in the model as $this).

Eloquent is an object-relational mapper (ORM) that facilitates interactions with the database. When using Eloquent, each database table has a corresponding model that is used to interact with objects belonging to it. In addition, Eloquent models provide interfaces for all CRUD (create, read, update, and delete) operations.

Additionally, there are custom methods/functions for the Active Record defined in the correspondent Model file, such as executing custom operations, obtaining computational parameters, and defining relationships between other models.

Resources are a transformation layer that sits between Eloquent models and the JSON responses that are returned to the application's uses.

In summary, the Client component sends requests to the API through the Routing middleware, then the Controller dispatches the payload to one or more dispatchers for the actions to be concluded. From here a Job may be dispatched executing a Model Process (controlled shell script), passing the validated request arguments and filtered values to an R script, as visible in figure 15, thus modifying/creating data information in the file system and/or database. Finally, the controller calls a Resource that sends a status message back to the Client finishing the request.

```
RunScript::dispatch('my-script-file', [$argument1, $argument2, $argument3]);
```

```
$command = array_merge([
    "Rscript",
    $script
], [$arguments]);

$process = new Process($command);
$process->setTimeout(240);
$process->setIdleTimeout(240);
$process->run();

return $process;
```

Fig. 15: Top image displays the action of dispatching a script job, while the bottom one demonstrates the contents of said job asynchronously.

ENM Database. In the dissertation only tables referring to ENM features are described, however, existing Wave Labs' tables that are affiliated with the current proposal are mentioned even if oversimplified for consistency in the structure.

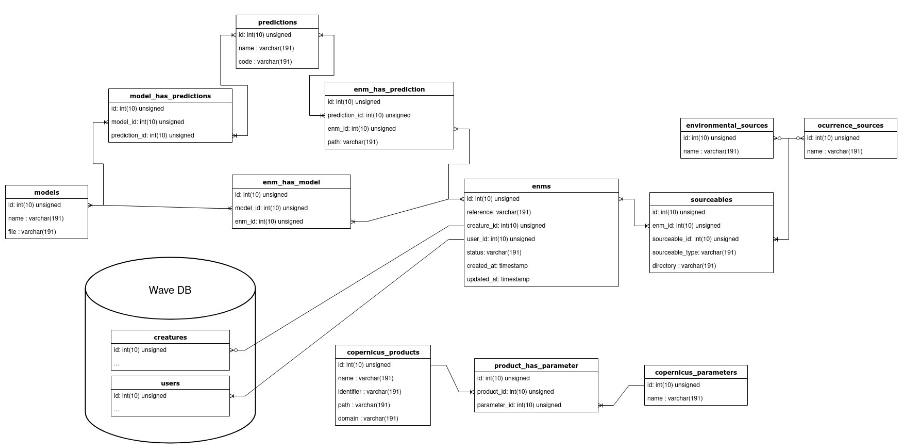

Fig. 16: Database structure

enms	"Reports" table that stores data for all references created by one protocol, such as creature being predicted, user, sources of data, current status of the prediction and timestamps.	
Fields	id	unique key
	reference	unique identifier for file system directories and files identification
	creature_id	foreign key for creatures
	user_id	foreign key for users
	status	current status of the report, may take "pending", "success", or "fail"
	created_at	timestamp column for date-time of creation
	updated_at	timestamp column for date-time of last update
occurrence_sources	Static table which holds current options regarding occurrence data input sources (e.g. wave, user upload, and scientific repositories).	
Fields	id	unique key
	name	identifier name of the input source
environmental_sources	Static table which holds current options regarding environmental data input sources (e.g. BIOCLIM, user upload, and Copernicus Marine Services).	
Fields	id	unique key
	name	identifier name of the input source
sourceables	Polymorphic pivot table between enms and both data input tables.	
Fields	id	unique key
	sourceable_id	polymorphic foreign key for occurrence and environmental sources
	sourceable_type	polymorphic foreign type for occurrence and environmental sources
	directory	null able identifier used for selection of Copernicus directory path when representing environmental data sources

Table – continued from previous page

models	Static table which holds current implemented algorithms (e.g. MaxEnt, Domain, and BIOCLIM).	
Fields	id	unique key
	name	identifier name of the algorithm
	file	name of the R script to execute this algorithm
enm_has_model	Pivot table for many to many relationship between enms and models	
Fields	id	unique key
	model_id	foreign key for models
	enm_id	foreign key for enms
predictions	Static table which holds available output formats (e.g. response plot, AUC, TPR, and map of suitability scores)	
Fields	id	unique key
	name	commercial name of the generated prediction (e.g. Response Plot, Presence-Absence, etc)
	code	identifier name to call during script's output process
model_has_prediction	Pivot table for many to many relationship between models and predictions	
Fields	id	unique key
	model_id	foreign key for models
	prediction_id	foreign key for predictions
model_has_prediction	Pivot table for many to many relationship between models and predictions	
Fields	id	unique key
	enm_id	foreign key for enms
	prediction_id	foreign key for predictions
	path	string code for the file name to access the output image for this prediction, composed by information of enm reference, model name, and prediction code

Table – continued from previous page

Wave Labs DB	Remaining database totalling over 120 tables of data.	
Fields	creatures	stores Wave Labs' fauna and flora
	users	table responsible for storing information regarding system's users
copernicus_products	Static table which holds currently available Copernicus' products.	
Fields	id	unique key
	name	commercial name of the product
	identifier	identifier name to call in back-end
	path	folder path to access files from this product when accessing Copernicus' FTP servers
	domain	geographic domain of data available for this product
copernicus_parameters	Static table which holds publicly available glossary regarding Copernicus Marine Services[37]	
Fields	id	unique key
	name	commercial name of the variable (e.g. Temperature, Salinity)
	identifier	unique identifier when accessing NetCDF variables
product_has_parameters	Pivot table for many to many relationship between Copernicus' products and parameters	
Fields	id	unique key
	product_id	foreign key for copernicus_products
	parameter_id	foreign key for copernicus_parameters

Table 2: Database tables descriptions and fields

Even if not currently required, some tables are designed for future features, enabling continuous growth of features. For example, Copernicus products parameters are not used, however, the system

[37]https://marine.copernicus.eu/glossary

is prepared for users to select which variables should be employed in the prediction, instead of using all available.

Static tables such as occurrence_sources, environmental_sources, predictions, and models exist to provide information dynamically to *enms*, meaning that if new data should be added to these tables, then the system adapts from new entries, therefore reducing costs of upgradability.

```
public function sourceables()
{
    return $this->hasMany('App\Sourceable');
}
```

```
public function sourceable()
{
    return $this->morphTo();
}

public function occurrenceData()
{
    return BelongsToMorph::build($this, OccurrenceDataSource::class, 'sourceable');
}

public function environmentalData()
{
    return BelongsToMorph::build($this, EnvironmnetalDataSource::class, 'sourceable');
}
```

Fig. 17: From left to right: a) Many to one relationship of *enms* to *sourceables*; (b) Polymorphic relation between *sourceables* and both data input source tables.

Most connections between tables come from simple many-to-one or many-to-many relations provided by foreign keys and pivot tables, as visible in table 2. However, we should highlight the many-to-many polymorphic relation between *enms* and environmental/occurrence sources, through the sourceables pivot table. This relation allows the ENM model to have more than one type of data source using a single association pivot table, as seen in figure 17 left). The polymorphic relationship between the pivot table and both data input sources is visible in figure 17 right), where the column sourceable_id represents the foreign key to one of the sources tables, while sourceable_type identifies which of the tables are we establishing relation with.

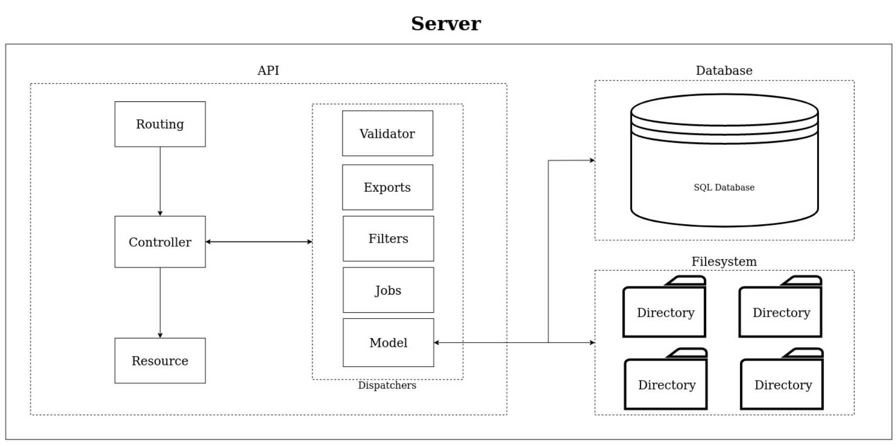

Fig. 18: Architecture - Server component

3.3.3 R Scripts

This component composed of various R modules, responsible for the protocol detailed in chapter 3.2.1. Each sub-component (occurrence data, environmental data, and prediction) share data between themselves through script arguments provided by the server, as demonstrated below:

$$\textbf{(1)}\ \textit{Rscript script.r argument}_1\ \ldots\ \textit{argument}_n$$

Then arguments are received as proceeding. When needed arguments are validated and/or converted to numerical values.

$$\textbf{(1)}\ \text{args} = \text{commandArgs(trailingOnly=TRUE)}$$
$$\textbf{(2)}\ argument_1 = \text{args[0]},\ argument_n = \text{args[n]}$$

As result of the web-based approach, every package associated in the protocol (e.g. raster, rgdal, maptools) needs to installed globally on the web server, thus end user requires no installation of software and no setup to run the predictions, truly achieving a "plug and play" experience, as presented in objectives of section 1.5.

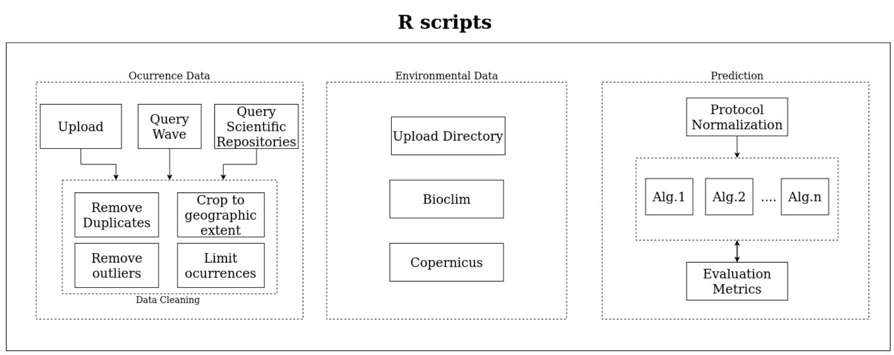

Fig. 19: Architecture - R scripts component

3.4 Data Pipelines

Not all data is currently loaded on the user's device or into the server, for instance, scientific repositories for occurrence data and environmental information from Copernicus Marine Services need to be fetched from external sources. Here the protocol of obtaining this data is explained, complementing the system's architecture.

3.4.1 Copernicus Marine Services Data

The Copernicus Marine Service provides several Application Programming Interfaces (APIs) to have access to its products and the related metadata[38]:

- HTTPs - MOTU API

[38] https://help.marine.copernicus.eu/en/articles/4794731-which-apis-are-provided#h_10ebc597b8

- HTTPs - WMS API
- HTTPs - WCS API
- HTTPs - ERDDAP API
- HTTPs - OPeNDAP API
- HTTPs - CSW API
- File Transfer Protocol (FTP)

Copernicus Marine products were downloaded through FTP servers, namely NearRealTime (NRT) for respectively ocean information from present and future. Any Copernicus Marine User can retrieve datasets by selecting files in a directory from Copernicus Marine File Transfer Protocol (FTP) servers. FTP data access has mechanisms to download at once the entire geospatial coverage and all variable(s) of a dataset, as produced by Copernicus Marine Producers, maximum concurrent download requests are set to 10 by IP-address, and performances of data retrieval depend on the server's disk read and bandwidth. As mentioned, we established an FTP connection with the NRT server for near real-time and forecast products, using available endpoint *ftp://nrt.cmems-du.eu/Core* and Copernicus Marine Services credentials to log in.

This data is fetched and stored locally on the server for fast access. The file system structure was replicated from the Copernicus FTP server, where files are stored in directories indexed by their year if contain monthly data or by month if contained daily data. Figure 20 demonstrates daily data on the left branch and monthly data on the right one.

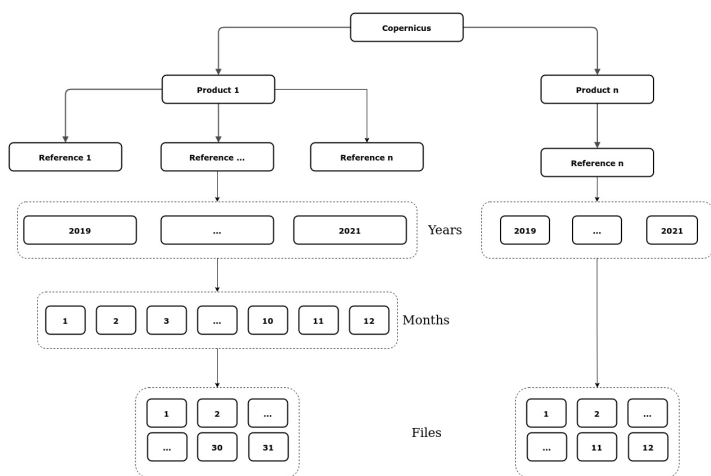

Fig. 20: Copernicus Marine Services FTP server file system structure

Data is obtained from an artisan command executed through a cron job, as a time-based job scheduler. Data is re-fetched every month because data currently being used from Copernicus FTP servers can take up to 500GB (All available data would be impossible, attending to current hardware limitations), creating bottlenecks for performance. Then the algorithm iterates over the years and months directories, storing them locally in the same structure, as presented in algorithm 3.

Algorithm 3 Copernicus Marine Services Data Fetch

```
Require: storagePath, products, files = []
  Copernicus ← Storage::server("copernicus")
  while products as product do
    years ← Copernicus::directories(product->path)
    while years as year do
      path ← path + year
      months ← Copernicus::directories(path)
      if months then
        while months as month do
          path ← path + year + month
          monthFiles ← Copernicus::files(path)
          array_push(files, monthFiles)
        end while
      else
        files ← Copernicus::files(path)
      end if
      while files as file do
        Copernicus::save(storagePath + path, file)
      end while
    end while
  end while
```

3.4.2 Scientific Repositories Occurrence Data

As mentioned multiples times during the protocol, one of the available environmental data sources is online scientific repositories. These were accessed from spocc (SPecies OCCurrence)[39], a package developed from the rOpenSci project, which aims to provide programmatic access to scientific data repositories on the web. The package delivers an API that normalizes the request process for nine major biodiversity repositories, namely:

- Global Biodiversity Information Facility
- iNaturalist
- VertNet
- Biodiversity Information Serving Our Nation
- eBird
- iDigBio
- Ocean Biogeographic Information System
- Atlas of Living Australia

Spocc has a function that takes species name and a repository from the previous list as arguments and returns an "occdat" object, however, this is not the data structure used for the remaining of the process. It's possible to combine results from occ calls to a single data.frame, using the function

[39] https://docs.ropensci.org/spocc/index.html

"occ2df", extracting following columns: (i) scientific name; (ii) longitude; (iii) latitude; (iv) data provide; (v) date; (vi) and unique key. Currently, the system will only store coordinates on the CSV file, discarding the rest.

Unlike Copernicus data, this one is fetched for every protocol, meaning that for each prediction data will be re-fetched without substantially affecting the performance of the system.

Algorithm 4 Scientific Repositories Data Retrieval

Require: $creature, repository$
libraty('spocc')
$data.occurrences \leftarrow spocc.occ(query = creature, from = repository)$
$data.dataframe \leftarrow spocc.occ2df(data.occurrences)$
return $data$

3.5 Study Setup

In this section, two distinct studies will be described, including their procedure. The first one inquiries how is the system perceived by users, while the latter the impact of system's outputs on the ecological awareness of regular sailors.

Both experiments' subjects are currently residing in Portugal, however, some inquiries have been conducted through remote user testing, as a consequence of current pandemic status, user preference, and/or geographic contexts, as a result of, for the first experiment, one user currently residing in Azores.

Disparities between remote and local inquiries have not been found, prevailing consistency between results from either methodology.

3.5.1 User Testing: How is the system perceived by users?

The first experiment inquiries the effectiveness of the system by employing a mixed analysis of usability through think-aloud assessments and verbal feedback given by subjects, followed by a System Usability Scale (SUS) questionnaire to users of different professional backgrounds.

Our sample size had two independent groups (N=16), namely: (i) expert group (N=4), providing experienced feedback for species distributions software and applications; (ii) experimental group (N=12), serving as a baseline evaluating current features from an interaction perspective.

Expert group (i) consisted of individuals that advised the development of the proposal, without ever interacting or viewing the system during its development. All subjects had extensive experience with ENM and multiple studies published with the said protocol.

From the review of subjects belonging to group (ii), inclusion criteria were developed which consisted of the following: (a) biology background; (b) basic knowledge on the concept of ENM; (c) no modelling-related projects published; (d) no computer science-based background.

As previously mentioned, the present study was carried during a pandemic. Minded that, remote user testing platforms were essential towards the evaluation of the dissertation, favoring some users' preferences. There are plenty of quantitative and qualitative methods for obtaining user feedback, so we employ methods for validating user flows during agile development cycles in three distinct phases:

(i) demographics profile; (ii) think-aloud assessments; (iii) post-survey (SUS). As a consequence of the current global status, the following web-based platforms were selected: Zoom, and Cisco Webex.

In total, each subject was exposed to the system in a single session with an average duration of 6 minutes and 53 seconds. Before each session, a profile tracing questionnaire was performed to obtain relevant data (age and technological expertise), which may produce disparity in information. The aforementioned survey is very important, considering it allows for data analysis that generates profile groups, indicating if some specific subject group is more affected by employed methods and detecting bias.

Users were then asked to explore the user interface while being guided to perform seven tasks considered relevant for the proposal. The task protocol consisted of (i) selecting some occurrence data from scientific repositories and then (ii) modifying data sources to select new data points, overriding the first step. Afterward, the user had to (iii) apply a minimum of two filters before (iv) selecting environmental data from BIOCLIM with 2.5km resolution and at least three variables. Next the protocol asks to (v) refresh the page and edit the previous entry to continue the protocol, (vi) select at least two algorithms to run and (vii) obtain a prediction output on the interface.

Simultaneously, an exploration section was recorded to create a key-term list regarding think-aloud assessments, generating a timeline of thoughts towards the system without affecting and/or influencing their actions.

After the exploration section, a rephrased version of SUS was presented as a ten (10) item questionnaire with a Likert scale from "Strongly disagree" (1) to "Strongly agree" (5). For item 2, the word "system" was replaced with "workflow", in item 4 the term "technical person" was specified as "developer", and item 5 the word "workflow" was also added for consistency reasons. Finally, for item 8, the word "awkward" was used rather than "cumbersome, due to reported confusion [109] (10% of participants) with non-English speakers regarding its wording.

Despite these differences, research on the SUS has shown, that slight changes to item wording most often lead to no detectable differences in reliability [110], therefore current adaptation is considered acceptable for present study. Moreover, future research should accomplish an European Portuguese validation.

Additionally, collected data from questionnaires was correlated with think-aloud assessments. Preceding assessments demonstrate reliance on the researcher's report. Thus, it is important to also collect other forms of data to track the validity of such conclusions. Minded prior reliance, an open opinion field was provided which we can equate with our conclusions and validate previous assessments.

3.5.2 Ecological Impact: Is the system able to motivate users towards ecological concerns?

The second experiment focuses on evaluating the social impact generated from the proposed system, promoting environmental awareness analysis, by alerting marine occurrences through publicly available environmental information.

The experiment inquiries the feasibility of using the proposed system as a fully automated systematic task, generating maps of suitability scores from near real-time satellite remote sensing and citizen scientists reports. These predictions will then be alerting end-users of abundance levels of specific taxa, providing ecological concern study proposals.

The evaluation of the experiment was performed through pre and post-questionnaires, directed towards two experimental groups (N=15; female=0; male=15; mean age=41.3) consisting of catamaran sailors (N=6; female=0; male=6; mean age=38.5) and fishing boats (N=9; female=0; male=9;

mean age=43.1). The sample was gathered from a total of four sailboats and modelling species include *Delphinus delphis* (common dolphin) and *Physeter macrocephalus* (sperm whale) in Funchal's Marina.

The pre questionnaire employed the New Ecological Paradigm (NEP) scale [111] to measure the endorsement of a "pro-ecological" world view. It is used extensively in environmental education, outdoor recreation, and other realms where differences in behavior or attitudes are believed to be explained by underlying values, a world view, or a paradigm. The scale is constructed from individual responses to fifteen Likert scale statements that measure agreement or disagreement.

The questionnaire was composed of 15 statements scored on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The deconstruction of the questionnaire allow for five sections [112], as follows: (i) growth of human societies (items 1, 6, and 11); (ii) anti-anthropocentric (items 2, 7 and 12); (iii) fragility of nature's balance (items 3, 8 and 13); (iv) rejection of human exceptionalism (items 4, 9 and 14); (v) and possibility of an ecological crisis (items 5, 10 and 15).

Then subjects were shown two prediction outputs for the species mentioned above, inquiring if such predictions could precipitate any ecological impact on the subject's behaviors. Images presented to both samples are visible as follows on figures 21 and 22.

The post questionnaire consists of an adaptation of the project "Raising Environmental Awareness in the Baltic Sea area" [113], here abbreviated as SPA, in October 1999 by the Finnish Environment Institute. This study provided guidelines for improving public awareness and environmental education in the Baltic Sea catchment area, therefore increasing the efforts to improve the state of this area. The study proved successful and the SPA project created a set of guidelines for improving public awareness and environmental education towards planning environmental information campaigns and activities. The experiment's pre questionnaire will be highly based on the questionnaire produced by this study and adapted according to the geographic area, modern requirements, sample demographics, and conclusions of the analysis, of same guidelines, provided by the "Analysis of the questionnaire on environmental awareness in the Baltic Sea" from Pemberton et al [114]. Additionally study shall assess values, attitudes, and threats towards marine biodiversity by assessing individual's general perceptions of marine biodiversity, as demonstrated in a study from Gkargkavouzi et al. [115] in Thessaloniki, a Greek coastal port city.

Questionnaire adaptations briefly mentioned before resulted in the following 17 item form, built from multiple choices, Likert scale items from one to five, and multiple-choice grids quantifying the impact of multiple factors per sentence. Further, sub-items of questions 9), 10) and 11) are divided into specific categories that are used towards worldviews analysis, additionally, each sub-item of question 11) was attributed a classification of evaluation. According to established methodology, the following questions were presented in the post-survey study:

1. What is your primary field of work?
2. How frequently do you usually go on boat trips?
3. Do you need some marine biodiversity knowledge in your field of work?
4. The amount of information I receive affects my ecological concern
5. The quality of the information I receive affects my ecological concern
6. I need more information about marine biodiversity problems.
7. Motivation towards marine biodiversity information
8. Which information channels do you mostly use?
9. Attitudes towards marine biodiversity (Items 1-4 represent an ecological view; items 5-8 depict an economical approach)

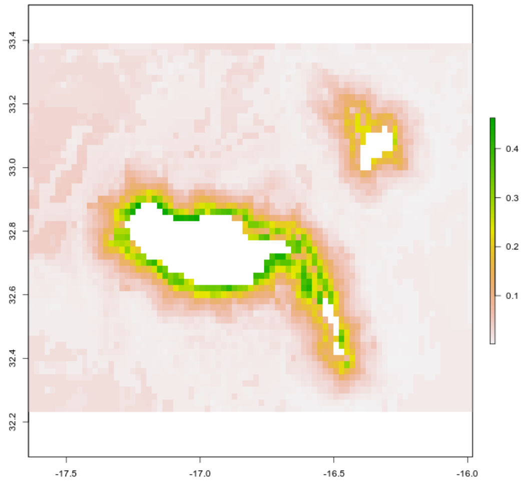

Fig. 21: Study model. Common dolphin

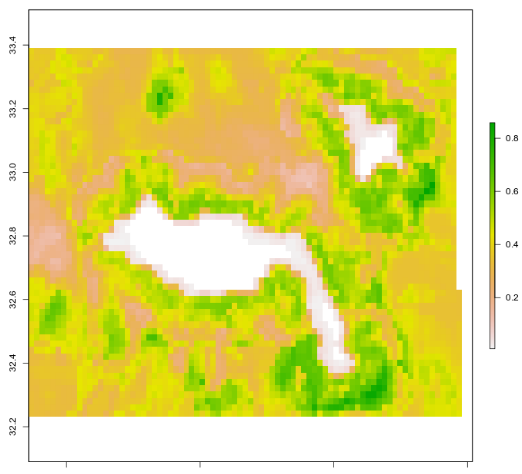

Fig. 22: Study model. Sperm whale

(a) All marine species have a right to exist in their natural environment

(b) I like learning about marine animals

(c) I enjoy watching animals in the ocean

(d) Every marine species plays an important role in the marine ecosystem

(e) I see little wrong with practices that negatively impact marine ecosystems if they have high economic benefits

(f) It is worth paying a little more for sea products to assist marine conservation

(g) I would support the use of fishing restrictions to assist in the conservation of marine species

(h) It is not humans' job to protect endangered marine animals

10. Threats towards marine biodiversity (Items 1-3 represent industrial threats; items 4-8 depict ocean exploration)

    (a) Products extraction (Oil and gas)

    (b) Pollution

    (c) Climate change

    (d) Human introduced species

    (e) Marine traffic (Waste and Noise Pollution)

    (f) Fisheries

    (g) Aquaculture

    (h) Marine renewable energy

11. Importance of marine biodiversity (Items 1-6 represent practical exploitation of nature; items 7-12 depict cultural and identity importance)

    (a) **Direct.** Provides resources (food and water)

    (b) **Recreation.** Place for outdoor recreation activities

    (c) **Scientific.** Opportunity for scientific observation or experimentation

    (d) **Pharmaceutical.** Provides materials to help healthcare

    (e) **Life Sustaining.** Produce, preserve, clean, and renew the air, soil, and water

    (f) **Economic.** Economic benefits and employment

    (g) **Therapeutic.** It makes me feel better

    (h) **Future.** Allows future generations to experience this place

    (i) **Intrisic.** Just because

    (j) **Spiritual.** Is significant to me

    (k) **Cultural.** Contributes to my cultural identity

    (l) **Perceveid Biodiversity.** Provides a variety of plants, wildlife, marine life, and other living organisms

Lastly, seven questions regarding the ecological impact of images 21 and 22 were employed. These questions aim to detect changes in the behaviors of sailboats when in high-density areas of marine taxa.

1. Did the prediction affect in any way your route?

2. If yes, in which way?
3. Would you slow down, keep velocity, or accelerate in high-density areas?
4. Can you explain why would that be?
5. Would you avoid, ignore, or go towards high-density areas?
6. Can you explain why would that be?

4 Results

Current chapter displays results retrieved when employing experiments described on section 3.5, evaluating how is the system perceived by users (section 3.5.1) and assessing the environmental impact of the prediction outputs (section 3.5.2).

4.1 User Testing

The experiment was initiated through a demographic evaluation of the sample, by gathering their age and technological expertise. Besides, the following questions consist of a series of statements towards systems usability inquiries and results distinguish the experimental group from the expert group.

– **DQ1.** Age (Figure 23)

– **DQ2.** Technical expertise (Figure 24)

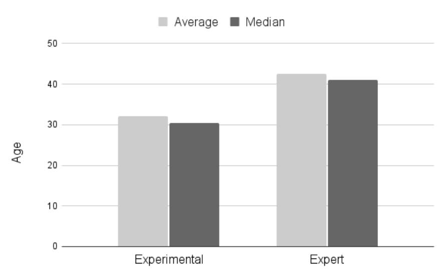

Fig. 23: Average and median age for both samples.

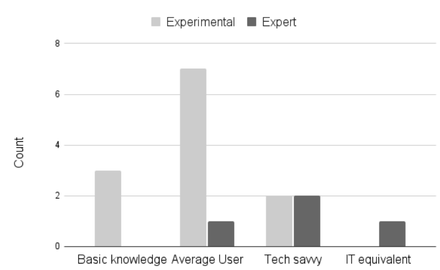

Fig. 24: Technological expertise of interviewees, divided in four categories.

– Think-aloud assessments were recorded from studies, fillerwords were removed (e.g. "I think") and finally similar words were grouped (e.g. learn, learned -> learn).

Fig. 25: Key-term diagram from think-aloud assessments.

– **SUS** System Usability Score results (Figures 26, 27, and 28)

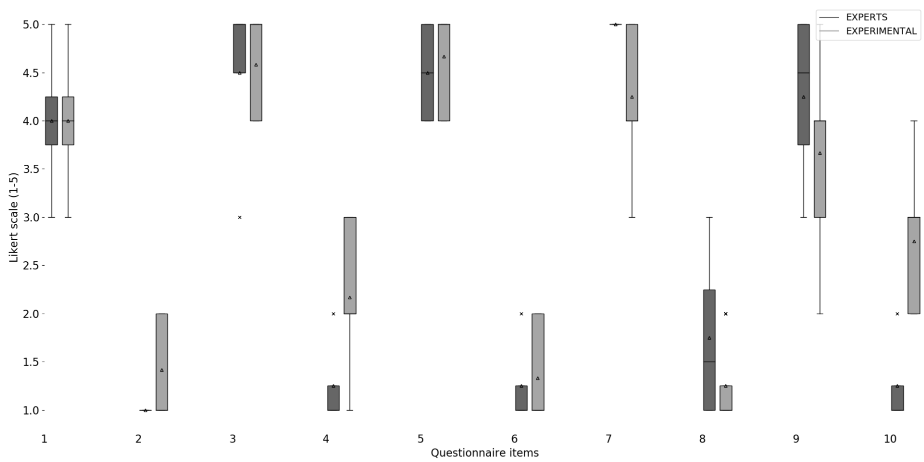

Fig. 26: SUS scores, where horizontal axis represents each item on the questionnaire and vertical axis represents the mean value of results.

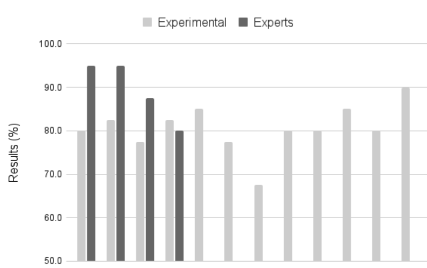

Fig. 27: SUS results in percentage. Each column displays an unique user.

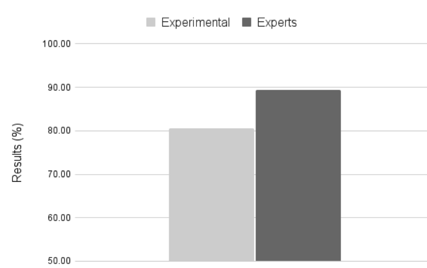

Fig. 28: SUS absolute results for both groups in percentage.

4.2 Environmental Impact

Following tables and charts demonstrate results gathered from experiment 2, pre (chapter 4.2.1) and post-questionnaires (chapter 4.2.2).

4.2.1 New Ecological Paradigm

Towards a better analysis of mean scores and deviations, the orientation of seven even-numbered items (representing anti-ecological world views) were reversed using formula 4, where a higher score indicates pro-NEP worldviews, hence all means have a tendency towards five (Likert scale maximum value). This operation was restricted solely to the data in Tables 3 and 4.

$$score(x) = max(x) + 1 - x, max(x) = 5 \tag{4}$$

Table 3, presents the results of the pre-questionnaire (NEP scale), distinguishing fisherman from catamaran sailors, to verify if some disparity between assessments is gathered. The aforementioned separation allows for a better understanding of beliefs and worldviews from both groups evaluating if field of work and cultural parameters impact the level of engagement to accept ecological awareness efforts.

Similarly, following tables adhere in accordance with the separation of both sample groups. Table 4 displays the deconstruction of the form as described in chapter 3.5.2, in which major averages are calculated from arrangements 1,6,11; 2,7,12; 3,8,13; 4,9,14; and 5,10,15.

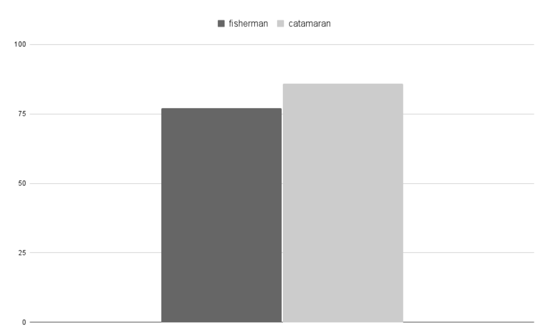

Fig. 29: NEP absolute results for both groups in percentage.

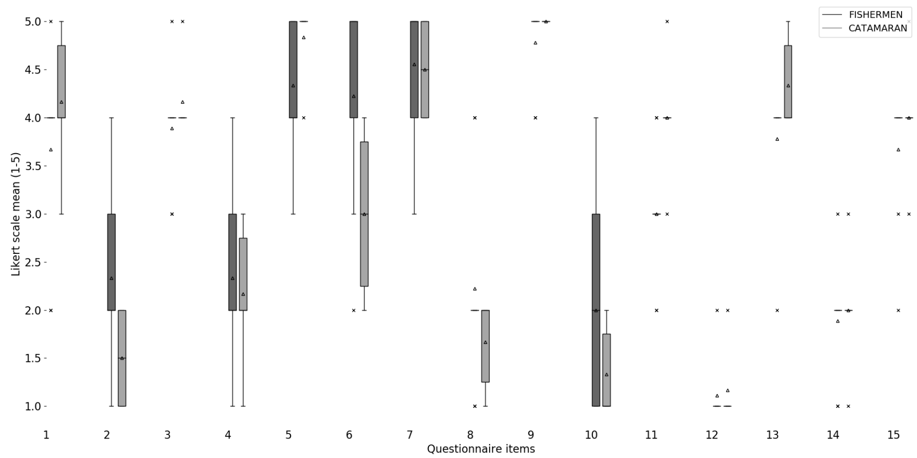

Fig. 30: NEP results

Item	Results (%)										Mean	
	Sd		D		N		A		Sa			
	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct
1	0	0	22	0	0	17	67	50	11	33	3.7	4.2
2[a]	22	50	33	50	33	0	11	0	0	0	3.7	4.5
3	0	0	0	0	22	0	67	83	11	17	3.9	4.2
4[a]	11	17	56	50	22	33	11	0	0	0	3.7	3.8
5	0	0	0	0	11	0	44	17	44	83	4.3	4.8
6[a]	0	0	11	33	11	33	22	33	56	0	1.8	3.0
7	0	0	0	0	11	0	22	50	67	50	4.6	4.5
8[a]	22	33	56	67	0	0	22	0	0	0	3.8	4.3
9	0	0	0	0	0	0	22	0	78	100	4.8	5.0
10[a]	44	67	22	33	22	0	11	0	0	0	4.0	4.7
11	0	0	22	0	56	17	22	67	0	17	3.0	4.0
12[a]	89	83	11	17	0	0	0	0	0	0	4.9	4.8
13	0	0	11	0	0	0	89	67	0	33	3.8	4.3
14[a]	22	17	67	67	11	17	0	0	0	0	4.1	4.0
15	0	0	11	0	11	17	78	67	0	17	3.7	4.0
NEP ecological views (%)											77[b]	86[b]

Table 3: New Ecological Paradigm questionnaire overview; Sd: strongly disagree; D: disagree; N: neutral; A: agree; Sa: strongly agree; Fh: fisherman; Ct: catamarans.

[a] Mean Likert scores after adjustment of direction with equation 4.

[b] Percentage regarding pro-NEP ecological views.

Section	Results (%)		Mean		Total	
	Fh	Ct	Fh	Ct	Results (%)	Mean
growth of human societies (items 1,6,11)	56.30	74.44	2.81	3.72	65.37	3.27
anti-anthropocentric (items 2,7,12)	87.41	92.22	4.37	4.61	89.81	4.49
fragility of nature's balance (items 3,8,13)	76.30	85.56	3.81	4.28	80.93	4.05
rejection of human exceptionalism (items 4,9,14)	83.70	85.56	4.19	4.28	84.63	4.23
possibility of an ecological crisis (items 5,10,15)	80.00	90.00	4.00	4.50	85.00	4.25

Table 4: New ecological paradigm from five different questions objectives, in which each of them is computed from the average of the respective items on chapter 3.5.2; Fh: fisherman; Ct: catamaran sailors.

4.2.2 Ecological Awareness

Following post-questionnaire results evaluate motivation, threats, attitudes, and values towards marine biodiversity from both study samples (expert and experimental groups). Following results were gathered after the exposure of the predictions (figures 21 and 22) and after subjects completed later sail trips, which in some cases took up to five days after the first questionnaire, regarding NEP results.

Q1. Marine biology information (Figure 31)

Q1.1. Do you need some marine biodiversity knowledge in your field of work?

Q1.2. The amount of information I receive affects my ecological concern

Q1.3. The quality of the information I receive affects my ecological concern

Q1.4. I need more information about marine biodiversity

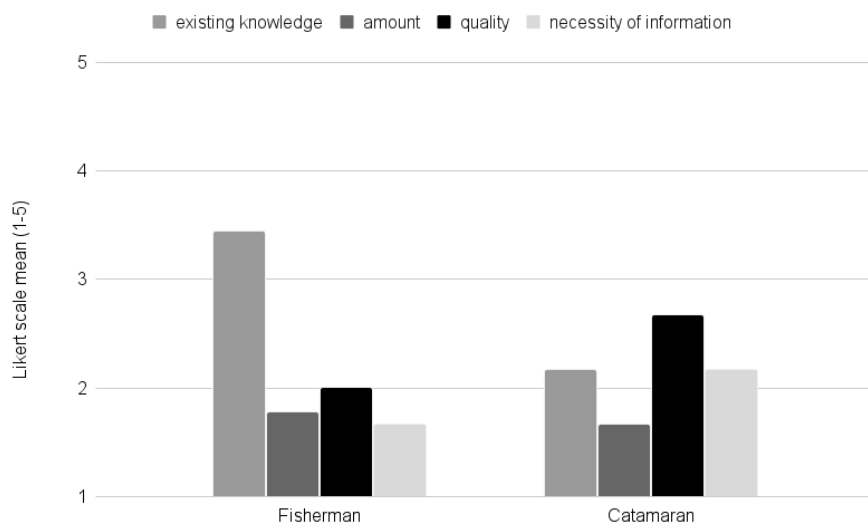

Fig. 31: Impact of the amount, quality and necessity of information from marine biodiversity problems. Regarding labels, Q1.1: existing knowledge; Q1.2: amount; Q1.3: quality; Q1.4: necessity of information.

Q2. Motivation towards marine biodiversity information (Figure 32).

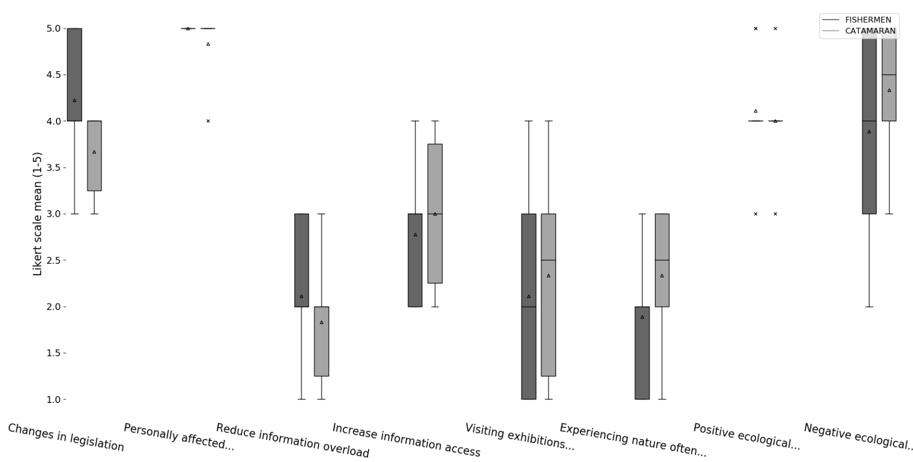

Fig. 32: Motivation towards marine biodiversity information. Horizontal axis contains eight motivation factors and vertical axis represents the mean value of results on the likert scale (1-5).

Q3. Which information channels do you mostly use? (Figure 33).

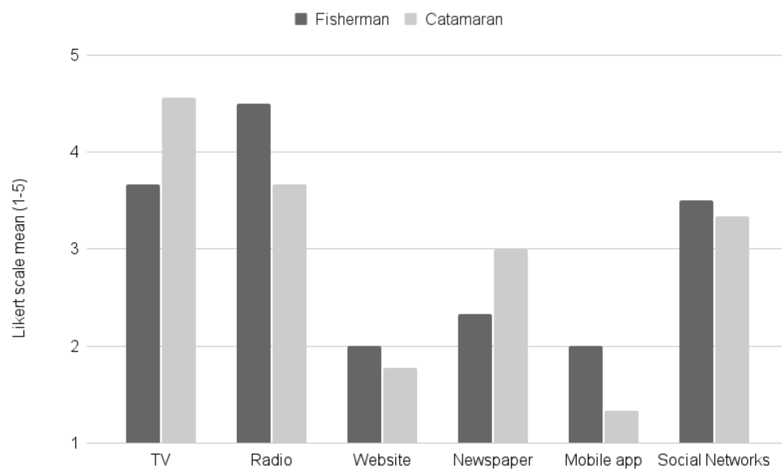

Fig. 33: Information channels are displayed on the horizontal axis and vertical axis contains mean values of results on the likert scale (1-5).

Q4. What makes biodiversity information useful and effective for you? [40]

Fig. 34: Key-term diagram from item Q4 in experiment 1.

Q5. Threats, attitudes, and importance of marine biodiversity (table 5 and figures 35, 36, 37, and 38)

[40]Key-terms were gathered in Portuguese, then translated into English, filler words/sentences were removed (e.g. "I think", "or", "In my opinion", etc) and finally similar words were grouped (e.g. learn, learned, learning -> learn)

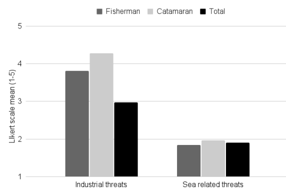

Fig. 35: Threats towards marine biodiversity.

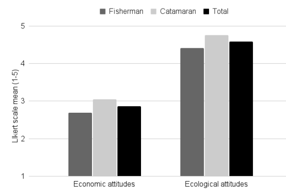

Fig. 36: Attitudes to marine biodiversity.

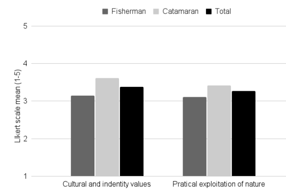

Fig. 37: Values of marine biodiversity.

Computed parameters	**Mean**		
	Fisherman	Catamaran	Total
Importance of marine biodiversity			
Practical exploitation	3.11	3.42	3.26
Cultural and identity	3.15	3.61	3.38
Threats to marine biodiversity			
Ocean exploration	1.84	1.97	1.91
Industrial	3.81	4.28	4.05
Attitudes towards marine biodiversity			
Economic view	2.69	3.04	2.87
Ecological view	4.42	4.75	4.58
Computed (Importance, threats, attitudes)			
Practical Exploitation; Industrial; Ecological	3.78	4.15	3.61
Practical Exploitation; Industrial; Economic	3.21	3.58	3.04
Practical Exploitation; Ocean exploration; Ecological	3.12	3.38	3.25
Practical Exploitation; Ocean exploration; Economic	2.55	2.81	2.68
Cultural; Industrial; Ecological	**3.79**	**4.21**	**3.65**
Cultural; Industrial; Economic	3.22	3.64	3.07
Cultural; Ocean exploration; Ecological	3.14	3.44	3.29
Cultural; Ocean exploration; Economic	2.56	2.87	2.72

Table 5: Threats, attitudes, and importance computed sub-item, obtained from the grand average of each of the classifications.

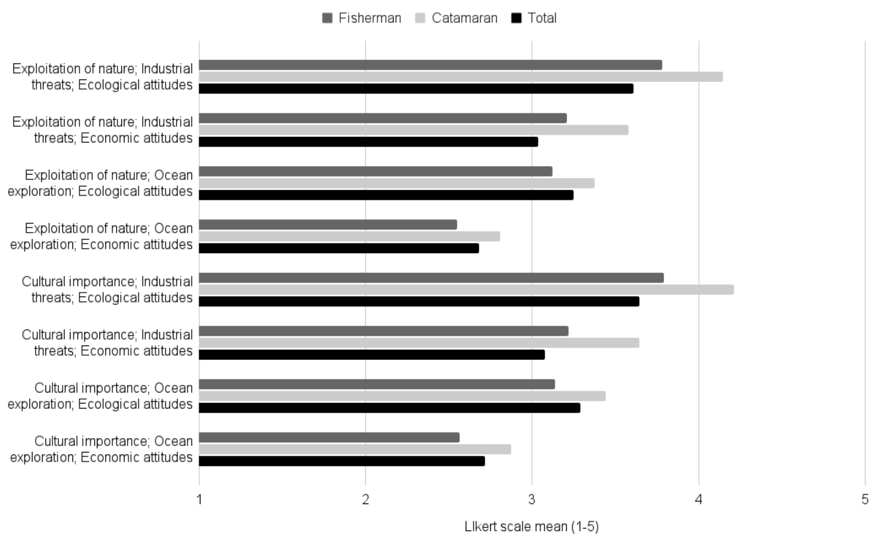

Fig. 38: Computed threats, attitudes, and importance of marine biodiversity, obtained from the grand average of each of the classifications.

Q5.1. Threats towards marine biodiversity (Table 6 and figure 39)

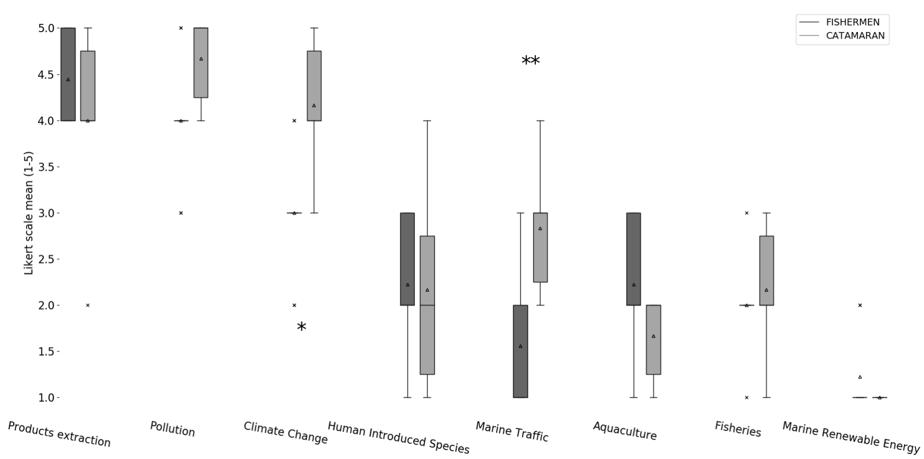

Fig. 39: Threats towards marine biodiversity. T tests adjusted with $*p < .05$, $**p < .01$, $***p < .001$.

Q5.2. Attitudes towards marine biodiversity (Table 7 and figure 40)

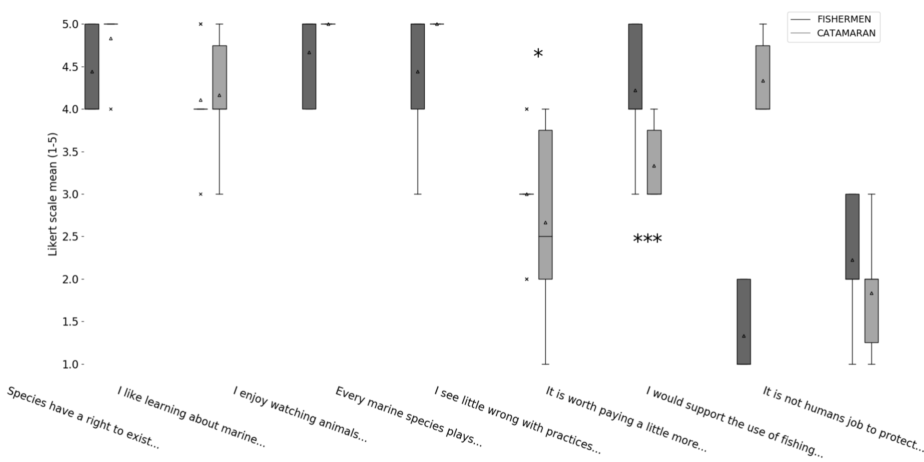

Fig. 40: Attitudes towards marine biodiversity. T tests adjusted with *p < .05, **p < .01, ***p < .001.

Q5.3. Values of marine biodiversity (Table 8 and figure 41)

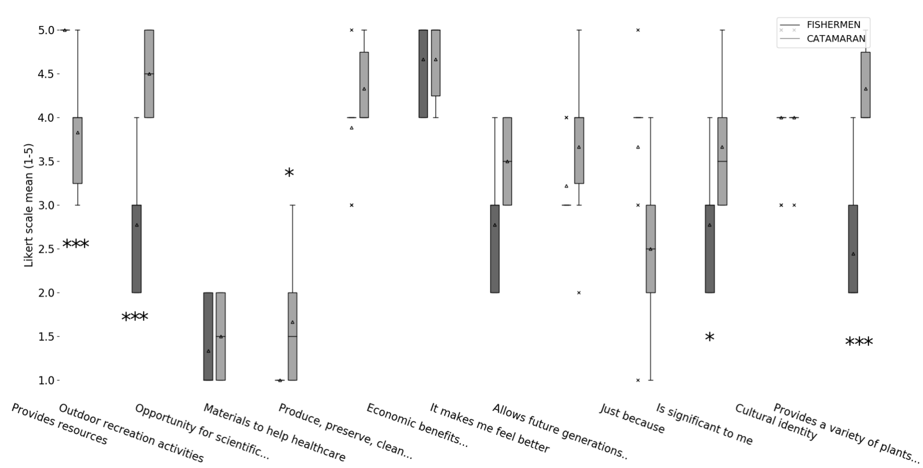

Fig. 41: Values of marine biodiversity. T tests adjusted with *p < .05, **p < .01, ***p < .001.

Threats	Results (%)										Mean		SD	
	Sd		D		N		A		Sa					
	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct		
Products extraction	0	0	0	17	0	0	56	50	44	33	**4.44**	**4.00**	0.53	1.10
Pollution	0	0	0	0	22	0	56	33	22	67	**4.00**	**4.67**	0.71	0.52
Climate change	0	0	22	0	56	17	22	50	0	33	**3.00**	**4.17**	0.71	0.75
Human introduced species	22	33	33	33	44	17	0	17	0	0	2.22	2.17	0.83	1.17
Marine traffic	56	0	33	33	11	50	0	17	0	0	1.56	2.83	0.73	0.75
Aquaculture	11	33	56	67	33	0	0	0	0	0	2.22	1.67	0.67	0.52
Fisheries	11	17	78	50	11	33	0	0	0	0	2.00	2.17	0.50	0.75
Marine renewable energy	78	100	22	0	0	0	0	0	0	0	1.22	1.00	0.44	0.00

Table 6: Threats towards marine biodiversity; Sd: strongly disagree; D: disagree; N: neutral; A: agree; Sa: strongly agree; All items are measures in a 5-point Liker scale; Mean of each factor is computed by averaging the items of the corresponding factor; SD: standard deviation.

Attitudes	**Results (%)**										**Mean**		**SD**	
	Sd		D		N		A		Sa					
	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct
All marine species have a right to exist in their natural environment	0	0	0	0	0	0	56	17	44	83	**4.44**	**4.83**	0.53	0.41
I like learning about marine animals	0	0	0	0	11	17	67	50	22	33	4.11	4.17	0.60	0.75
I enjoy watching animals in the ocean	0	0	0	0	0	0	33	0	67	100	**4.67**	**5.00**	0.50	0.00
Every marine species plays an important role in the marine ecosystem	0	0	0	0	11	0	33	0	56	100	**4.44**	**5.00**	0.73	0.00
I see little wrong with practices that negatively impact marine ecosystems if they have high economic benefits	0	17	22	33	56	17	22	33	0	0	3.00	2.67	0.71	1.21
It is worth paying a little more for sea products to assist marine conservation	0	0	0	0	11	67	56	33	33	0	4.22	3.33	0.67	0.52
I would support the use of fishing restrictions to assist in the conservation of marine species	67	0	33	0	0	0	0	67	0	33	1.33	4.33	0.50	0.52
It is not humans' job to protect endangered marine animals	11	33	56	50	33	17	0	0	0	0	2.22	1.83	0.67	0.75

Table 7: Attitudes towards marine biodiversity; Sd: strongly disagree; D: disagree; N: neutral; A: agree; Sa: strongly agree; All items are measures in a 5-point Liker scale; Mean of each factor is computed by averaging the items of the corresponding factor; SD: standard deviation.

Values	Results (%)										Mean		SD	
	Sd		D		N		A		Sa					
	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct	Fh	Ct
Provides resources (food and water)	0	0	0	0	0	33	0	50	100	17	**5.00**	3.83	0.00	0.75
Place for outdoor recreation activities	0	0	44	0	33	0	22	50	0	50	2.78	**4.50**	0.83	0.55
Opportunity for scientific observation or experimentation	67	50	33	50	0	0	0	0	0	0	1.33	1.50	0.50	0.55
Provides materials to help healthcare	100	50	0	33	0	17	0	0	0	0	1.00	1.67	0.00	0.82
Produce, preserve, clean, and renew air, soil, and water	0	0	0	0	22	0	67	67	11	33	3.89	**4.33**	0.60	0.52
Economic benefits and employment	0	0	0	0	0	0	33	33	67	67	**4.67**	**4.67**	0.50	0.52
It makes me feel better	0	0	44	0	33	50	22	50	0	0	2.78	3.50	0.83	0.55
Allows future generations to experience this place	0	0	0	17	78	17	22	50	0	17	3.22	3.67	0.44	1.03
Just because	11	17	0	33	11	33	67	17	11	0	3.67	2.50	1.12	1.05
Is significant to me	0	0	33	0	56	50	11	33	0	17	2.78	3.67	0.67	0.82
Contributes to my cultural identity	0	0	0	0	22	17	56	67	22	17	**4.00**	4.00	0.71	0.63
Provides a variety of plants, wildlife, marine life, and other living organisms	0	0	67	0	22	0	11	67	0	33	2.44	4.32	0.73	0.52

Table 8: Values of marine biodiversity; Sd: strongly disagree; D: disagree; N: neutral; A: agree; Sa: strongly agree; All items are measures in a 5-point Liker scale; Mean of each factor is computed by averaging the items of the corresponding factor; SD: standard deviation.

Q6. Did the prediction affect in any way your route? (Figure 42)

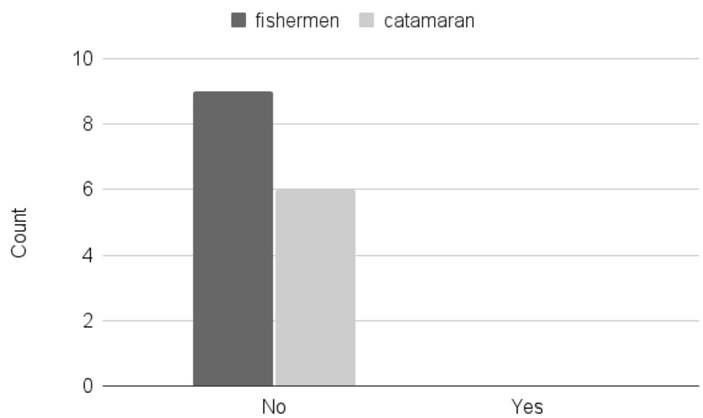

Fig. 42: Experiment 2. Q6

Q8. Would you slow down, keep velocity, or accelerate in high-density areas? (Figure 44)

Q9. Would you avoid, ignore, or go towards high-density areas? (Figure 44)

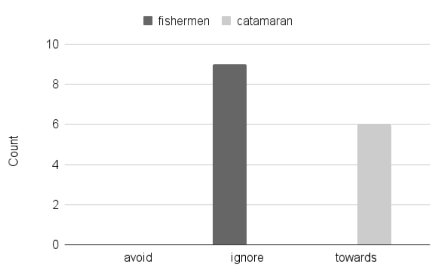

Fig. 43: Experiment 2. Q8

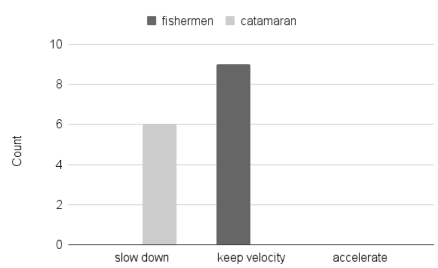

Fig. 44: Experiment 2. Q7

5 Discussion

In this chapter, results presented in chapter 4 are explained and conclusions are drawn, correlating research questions culminating with research contributions. Additionally, research challenges are described, providing a set of guidelines towards web-based automation of environmental niche modeling. Limitations for the proposed system are mentioned, moreover, all the constraints encountered in the process will be presented and restated for future developments.

5.1 Findings

As aforementioned, this section describes all previously presented results, section 4, and depicts in the following user studies presented in sections 3.4.1 and 3.4.2.

5.1.1 Experiment 1

The results of the questionnaire are summarized in Figure 27. Overall the results from the experiment are positive in terms of usability. It was calculated the overall SUS score according to [116] by determining each item's score contribution, from the range of 1-5 as follows: (i) for positive items (1, 3, 5, 7, and 9), the score contribution is the answered value minus 1; (ii) for negative items (2, 4, 6, 8 and 10), it is 5 minus the answered value; (iii) for missing values assign 3 (the center of the rating scale). To get the overall SUS score, multiplication of the sum of the item score contributions by 2.5 is applied, thus adjusting the overall SUS score to 0-100 with 2.5-point increments. Calculating the overall score for the experimental group, we reached 80.63/100, which fits on a grade "B" as acceptable, described by the grade rankings of SUS scores from the practical guide to the System Usability Scale [117] on image 45. Comparatively the expert group obtained a score of 89.38/100, which fits on a grade "B" as acceptable.

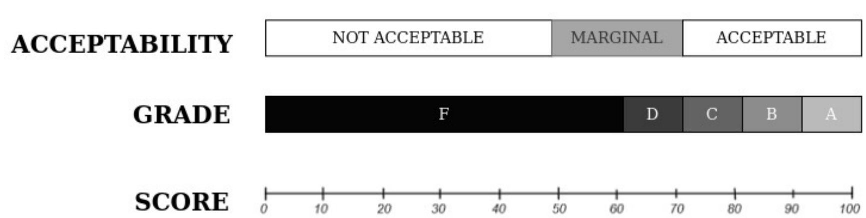

Fig. 45: Grade rankings of SUS scores from the practical guide to the System Usability Scale

From the experimental group results, 75% would use the system frequently, 100% agreed that the system was easy to use. Additionally 91.7% considered the workflow well integrated, correlating with 83.3% considering that system has a small learning curve. However, 33.3% of subjects required support to complete the procedure and only 58.3% felt confident when interacting with the system. In comparison, 75.0% of the expert sample considered the system easy to use and felt confident while exploring the tasks. All experts admitted the value of a web-based species distribution platform, however, experts acknowledge that in its current state, only 20.0% would use the system. None of the subjects considered the workflow unnecessarily complex and correspondingly there was a 100% acceptance regarding usability.

The system had positive results on users' perceptions. These are visible through the elicited positive mentions detected in the think-aloud inquiry of figure 34. Descriptions such as "this is so interesting" or "it is very easy to use" were expressed multiple times during the exploration task of the study. Added features were successful and important, resulting in a substantial amount of praise to the system.

Nonetheless, specifically, the expert group, thought the system was "too easy to use", which could create problems in future developments. The sample considered that workflow should be more technical towards experienced users. Subjects were active in suggestions and some demonstrated positive effects of engagement as was the case with "I could help you with additional data" and "Are you planning in adding more models?". In five distinct situations, subjects were interested in finding if the project would keep being developed, from which two offered technical support.

These findings showed that the subjects were interest in the proposal and were willing to engage in the activity again in the future, suggesting potential towards species distribution models in a web environment.

As predicted, different study groups demonstrated different approaches to evaluation. For instance, data inquiry showed that experienced subjects manifested a particular interest in further means of data source and additional filters over the data, while the experimental group had considerable interest in visualizing the data and prediction outputs. As a consequence of the context of the proposal and the small number of samples gathered, there was no correlation between demographics and feedback gathered.

5.1.2 Experiment 2

The final experiment benefits from pre-and post- questionnaires regarding the environmental impact of the proposal.

The environmental perception was determined by providing mean distribution and standard deviation for both sample groups, in tables 3 and **??** respectively.

The mean total, after reversal of orientation towards anti-NEP items, concluded as 3.8/5 or 77% environmental concerned for fisherman and 4.3/5 for catamarans' sailors with an environmental concern of 88%. These results suggest an overall positive underlying ecological worldview for both study groups, supporting the NEP statements to a varying degree.

Hereinafter, for the current chapter analysis, group i) will illustrate fisherman sample, while catamaran sailors will be mentioned as group ii).

Society Growth. There was a general disagreement around human population planning, with 78% of fishermen acknowledging that there are enough resources on Earth, with 22% considering these as limited. Comparatively, only 33% of catamaran sailors consider Earth's resources enough, consequently, 84% of subjects admit a very limited number of resources. However, there was a general consensus regarding the limit of people that Earth can support with 78% and 83% supporting the statement, for group i) and group ii) respectively.

Human-centric worldview. A positive environmental worldview contrasts with the idea that nature exists primarily for human use and that humans have the right to modify the natural environment to suit their needs. An examination of items 2, 7, and 12 evaluating anti-anthropocentric behaviors, confirm that 87% of group i) and 92% of group ii) refuse a human-centric worldview with the grant majority (89% and 100%, respectively) attributing the same importance to plants and animals as to humans to exist. Additionally, none of the subjects from either group sample, agree that humans are meant to rule over the rest of nature.

Nature's balance. Regarding the fragility of nature's balance, group (ii) imposes such statements 10% more often than group i), recognizing the balance of nature is very delicate and easily upset for an average 94% of both samples. Some mixed opinions were detected in group ii), respecting human interference and respective consequences, as 78% considered it often produces disastrous consequences.

Exemptionalism. The belief that the relationship between humans and the natural environment is unimportant because humans are "exempt" from environmental problems was analyzed from items 4, 9, and 14. 84% and 86% of group i) and ii) accept an anti-exemptionalist worldview. When insuring that human development will not make Earth unlivable, 33% of both groups remained neutral towards such statement. Ecological behaviors are detected for item 14 (Humans will eventually learn and be able to control nature), with 89% and 84% disagreeing with such statement. On item 9, there is a total consent that humans are still subject to laws of nature.

Ecologic crisis. The great majority of both samples (94%) agrees that human abuse of nature, however, 33% of fisherman consider that the ecological crisis has been greatly exaggerated. Regarding the probability of an ecological catastrophe group i) (78%) ii) (84%) admitted that current activities are unsustainable.

Generally, both groups have substantially pro-ecological worldviews, without significant disparities in results. Interestingly, although samples had mixed opinions considering the amount of resources on Earth and respective limits, there was a contradictory agreement with human population outgrowth. This finding suggests that both samples consider society growth as the main concern, independently of their beliefs on Earth resources. Such discovery relates to results regarding the ecologic crisis, in which the great majority of the sample agrees that humans abuse nature and human interference and consequently impact nature's balance.

Four key findings are identified as relevant from the post-survey: (i) Concern about the marine environment is mostly positive, confirming aforementioned conclusions of NEP; (ii) There is a disparity between fisherman and touristic perspectives on attitudes towards marine biodiversity; (iii) Field of work and cultural factors can influence the perception of marine issues more than the frequency of sailing activities; (iv) Individual's ecological concerns mostly depend on their beliefs, attitudes, and values directly impacting their perception of biodiversity issues and therefore influencing their level of engagement to accept ecological awareness efforts.

Regarding t tests' statistical relevance of results for both study samples, threats "Climate Change" and "Marine Traffic" were considered relevant and very relevant, respectively. Additionally, attitudes towards the increase on price of marine-related products were also considered relevant, with increase of fishing regulations achieving an extremely relevant status. Finally, values "Produce, preserve, clean, and renew air, soil, and water" and "Is significant to me" concluded as relevant, further "Provides resources (food and water)", "Place for outdoor recreation activities", and "Provides a variety of plants, wildlife, marine life, and other living organisms" were evaluated as extremely relevant.

Results indicated that people generally have ecological attitudes, indicating industrial threats as the main concern towards marine biodiversity. Regarding the importance of marine biodiversity, most people have a life-sustaining worldview, an identity connection, from cultural stimuli, and economic influence (employment). However, results differ in some items, with a significant portion of fishermen valuing direct benefits, such as resource access (food and water), while catamaran sailors highly worth recreational activities. Both samples indicated a slight tendency towards practical exploitation of nature over cultural and identity importance.

Information channels indicate that there is an opportunity for alerting mechanisms mostly from radio, social networks, and television. These should provide feedback on action-based consequences. Results also propose that changes in ecological legislation could benefit overall ecological awareness.

Even though environmental concern and ecological attitudes were observed from both pre and post-questionnaires, subjects admitted not changing their routes even if that meant a direct collision route with marine biodiversity, implying economic attitudes, and practical exploitation of nature. Furthermore samples admitted to keep velocity, ignoring high-density areas.

5.2 Research Contributions

This section connects the implemented work of this dissertation with the research questions presented in section 1.5, describing correlation from these questions and research contributions gathered from results.

As far as the literature review demonstrated, this dissertation presents the first web-based automated ecological niche modeling system. The system uses a modular script architecture that is executed according to a pre-established workflow in parallel with data manipulation, providing maps of suitability scores with statistical evaluation metrics.

- **RQ1 - Could the proposed system facilitate the process of employing ENMs, while retaining the flexibility existent with current solutions?**

  Ultimately, the system enables biologists to create solutions for their custom problems using simple protocols without the need for any third-party entities and extensive knowledge in programming contexts. Additionally, there are established standardized open-access pipelines with public available online repositories.

  Gathered from study inquiries, the proposed system allows fast and straightforward means by which to retrieve and manipulate occurrence and environmental data, which can then be supplied to modeling algorithms, indeed facilitating the process of employing ENMs.

  Model fitting remains a niche step that requires full adaptation based on given taxa and environmental data, it is therefore not suitable for the extensive normalization of models. Although flexibility should be affected, the overall accuracy of proposed solutions is promoted. Further studies regarding the balance between flexibility and accuracy of ENM solutions are necessary to maximize the performance of the system.

- **RQ2 - Can the ENMs be used as an alerting mechanism towards ecological problems?**

  The proposal employed an ecological evaluation on fisherman and catamaran sailors, with the intention to produce behaviors changes in high-density areas of taxa.

  Overall, predictions did not cause any significant impact on the ecological awareness of neither sample, suggesting that such procedure is not suitable for adopted study groups. The assessment of threats, attitudes, and importance of marine biodiversity has proven useful in identifying possible motives for the aforementioned findings. Although samples admit cultural and identity values with an ecological worldview, both groups predominantly consider industrial threats (products extraction, pollution, and climate change) as marine biodiversity's main concern. Further, both groups devalued marine traffic as a significant threat, withholding ocean exploration as a problem.

  Another conclusion may be drawn, in which information channel for studied groups, overly consisted on television, social networks, namely *Facebook*, and radio. The latter results from the frequent usage of such devices when on hold for sailing activities. Additionally, means of motivation towards ecological awareness suggest the most effective solutions as action-based feedback (positive and negative consequences of actions), and legislation changes, due to direct ecological impact and demand as a result of their occupation, respectively.

  These discoveries may be coupled, advocating that different approaches of alerting mechanisms and information channels may prove successful with considered demographics, towards increasing ecological awareness and therefore provoke any behaviors changes based on maps of suitability scores.

By showcasing this proposal to areas of interest, in an objective web application, the dissertation is enabling the potential of novel ideas that can be implemented from achieved guidelines.

From described contributions, the proposal addresses an important gap in the state of the art which is the scarcity of solutions for ecologists without technological backgrounds. Dissertation advances alerting solutions for marine biodiversity concerns by providing continuously acquired predictions, available for visualization, and further research by the scientific community.

5.3 Research Challenges

In this section, seven essential research challenges in the proposal are addressed. The goal of the current chapter is to provide high-level guidance for future researches in a web-based species distribution environment. These challenges will evolve as the requirements are better identified in the process of building complete distribution modeling systems. These problems could be addressed by researchers as independent research problems. Of course, some researchers may build complete systems by combining several of these.

1. **Data Identification.** Importing data is somewhat effortless, notwithstanding collecting, georeferencing, and cross-checking coordinate data. Although such validations are provided, there is no control over the number of records and bias roots. Further information collected from results suggests a need for additional data filtering processes, such as an automated outlier identification associated with a user manual review. Another limitation was observed, regarding the limitation of occurrence data, in which filtering data to N records, take the initial N rows, instead of a random N unordered subset, therefore creating a bias.

2. **Data Processing.** The emerging technological development of big data is evolving rapidly, driven in part by IoT environmental sensors [118]. Additionally, the volume of such data is analogously increasing at great rates. From this perspective, three congestion points are easily identified, particularly: (i) data fetching of Copernicus Marine Services; (ii) upload of occurrence data; (iii) and upload environmental files. While point (i) does not affect the wait-time of the workflow, the update of the data on the webserver lasts over 48 hours, due to a volume of data superior to 2 terabytes (considering the six Copernicus products mentioned in section 3.4.1). Points (ii) and (iii), however, affect said parameter immensely, leading to data uploads being limited to 300 megabytes as a direct result of payload crashes and/or timeouts. This is clearly a limitation that requires the evaluation of alternatives in further versions.

3. **Heterogeneous Data Quality.** Latest developments augment the amount of data available online, both for environmental and occurrence points. This enables a global provider loop and data collected with different measurements. However, owing to the historic development of systems, the exploitation of data, is often limited to the purpose for which the data was originally collected [119]. At the moment of deliverable, the system allows for a single data source, as a consequence of the inability to guarantee heterogeneous data quality between multiple sources.

4. **Decision Support and Recommendation System.** Although there were feedback messages being given to users, some of them could easily be avoided by decision support protocols. Failures such as insufficient occurrence points, environmental data points not overlapping with occurrence data, the abundance of *null* values should be verified at the appropriate step for each of these tasks. Currently, the system validates these parameters when modeling the algorithms, meaning that if have any of them occurs the prediction will fail, when it could have just warned the user when provided the data.

5. **Model Fitting.** Model fitting is, theoretically, rather similar between different algorithms. Most methods work as formulas identifying the dependent and independent variables. Proposal operates only with presence data for models implemented in package dismo, meaning that implementation is accessible, though prediction accuracy is impaired. There are plenty of studies [120–125] corroborating the superiority of presence-absence to estimate the current and potential distributions of biodiversity. As a consequence, current predictions should be used as comparative methods for external predictions, evaluating the congruence between results and adapted to enhance model calibration.

6. **Model Evaluation.** From all the steps described in the workflow, the most complex is undoubtedly model evaluation. It is much easier to create a model and make a prediction than to assess how good the model is and whether it is can be used for a specific purpose. Models have different statistical measures that provide feedback on how well the model fits the data, however, an ecological evaluation is as important. Proposal lacked on such metrics undermining predictions conceived from the system.

7. **System Usability.** Although the translation of the SUS questionnaire has been considered easy to understand and without any semantic or content problems [126], few items caused some confusion, and wording should be revised. The question "I think that I would be able to use this system frequently" was employed to an experimental group without previous experience in ENM, thus this item created a bias towards 1, as a result of people not using this type of system from the start. Additionally, the question "I found the system very awkward to use" caused some doubts in subjects, mostly from the word "awkward", suggesting that conclusions from Finstad et al. [109] should be verified in a European Portuguese validation.

5.3.1 Guidelines

Automating web-based species distribution protocols introduce additional limitations and challenges as described previously. However proposed system has the potential to simplify and accelerate the employment of ENM protocols due to feedback from impact, use cases, and scientific evaluations. The proposed system allows a fast and straightforward means by which to retrieve and manipulate occurrence and environmental data, which can then be implemented in alerting systems, projection, and evaluation for assessing distributions of species in geographic space and their corresponding environmental combinations.

The aforementioned challenges and feedback generated some guidelines for future implementations of similar systems.

1. **Environmental data is substantial.** Data files of environmental information are usually several gigabits in size. Uploading pipelines towards the webserver are good to have, avoiding limitations with HTTPS transfer rate and capacity. We propose ahead of time file uploads, through suitable file transfer protocols and only then the initiation of the workflow. You may also deconstruct the payload and upload them as a chunked transfer encoding, therefore reducing the bottleneck. Additionally, prepare your setup for this data with a minimum requirement of some terabytes.

2. **Geocentric evaluation of models is required.** Prediction output relies on ecology experts validation and most of the time statistical evaluation is not enough. Although the prediction may fit in multiple computational evaluations, if it fails in the geocentric context, it's not a good prediction. This is totally user-dependent, but you may help them by providing feedback of geographical context, time periods, and environmental ranges used in the prediction, which may help them identify inaccurate data, auto-correlation among the predictor variables, etc.

3. **The more reliable data sources, the better.** User usually feels the need to combine multiple data sources in the same modeling process. Although the predictions may focus on aquatic

biodiversity, it is still important to equate for extra factors, like human presence or predatory presence points. We might never be able to predict which variables the user may need for a specific species so it is better to just have the most.

4. **Introduce one a variable at a time.** As mentioned in the previous point, predicting what variables the user may need for a specific prediction is complex. Fitting the model for each species is a niche step, so you should focus on a subset of culturally relevant biodiversity and avoid generalizing every single step of the workflow. Only when a subset of species with similar conditions is incorporated, you focus on implementing the next subset. The same should be applied to environmental data, meaning that if you are working with biochemical parameters, you should adjust them to the models and then introduce new parameters.

5. **Give control over the workflow.** Essential features for 90% of users might be useless for the remaining. Instead of generalizing the entire workflow, give the user the choice to customize every step as an "Advanced" option. This might generate some code duplication, as a result of the deconstruction of the generalized code structure, however, the benefits overshadow such a problem.

6. **Flexibility is required.** You will always have to add new features to your system and may never know what the user might need. Build your system as flexible as possible so that new filters are as easy as possible to implement without disrupting the existing structure.

7. **Users need to restart the workflow.** Sometimes the only way to know if you chose the right filters is just to test it and see if it makes sense. Give the user's the option to start over as many times as they want and rerun workflows using different filters. It may take them 5 tries or 100, but they should have the option to retry.

8. **Cross-validation is almost always the solution.** Most of the time when you just want to validate the prediction, user's just want to compare it with confirmed abundances. if it takes the user 100 tries to get an accurate prediction, should he just start over when he wants to use new data? Of course not, they should be able to reuse workflows for new data without having to repeat the process every single time.

9. **Divide the workflow as small tasks.** Modeling the prediction of species takes several steps, from collecting species occurrence data to iterating the process to improve the model performance. Create these steps clearly defined in the UI, navigating the users through the steps of the workflow. This fragmentation of the workflow increase landing page conversions since they: reduce psychological friction, question overload, and help you avoid the Goldilocks Syndrome [127].

10. **Workflow as a collection of dispatchable jobs.** Steps of employing a modeling algorithm are time-consuming, resulting in some users leaving the application midway and continue it later. A dispatchable job approach is beneficial because for each task jobs will be added to a queue, meaning that no job will occur without previous dependencies are fulfilled and the job won't be affected by the user leaving the system.

5.4 Future Work

The proposed system will continue its development, however, the scope of the project will be slightly modified to overcome defined challenges and better follow guidelines provided.

The workflow employed and mentioned multiple times during the dissertation needs to be modified to incorporate the new requirements, by moving data input mechanisms from the middle of the workflow towards the beginning. Using FTP server access to upload environmental data to the web server system will overcome the significant bottleneck generated from HTTPS file transfer. This modification will be based on a new step in the workflow, named "Data Upload", resulting in the updated and improved structure from figure 46.

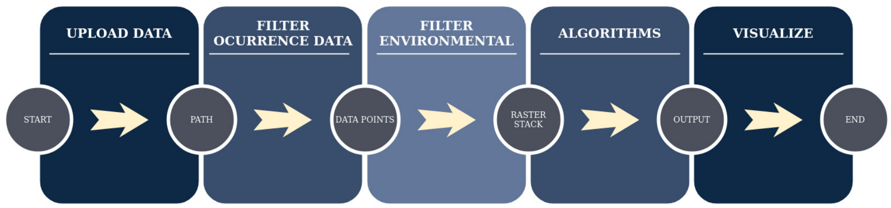

Fig. 46: Updated system workflow diagram

In the geographical context, data shall be filtered to fit a spatial polygon around Portugal's coastal area, including its islands. Hereinafter, predictions will not be affected by external cultural aspects, negatively impacting suitability maps. Additionally, data will be time-categorized, classifying it towards different times of the year, reducing bias related to the large disparity of occurrence points during the summer and winter periods.

Supplementary filters are required in environmental data treatment, such as the ability to choose which dimensions and variables to use from a single environmental file. Consequently, the possibility to use data from multiple sources in the same prediction is important for future versions of the system.

Three species distribution algorithms were implemented, while for the correlation of overall abundance levels, more algorithms should be added. Another concern from the users' point of view is the evaluation metrics displayed in the visualization step. While provided evaluation methods were insufficient, future works can try to improve them by providing a summary of initial parameters used in the modeling process as well as more relevant evaluation metrics.

From collected feedback, the ability to run the workflow with a percentage-based threshold will prove beneficial to the proposal. This option allows us to run the same prediction, but with the objective of deviating environmental values by N%, therefore we can compare how a specific species presence may be affected by a 5% increase or decrease in temperature.

Finally, future implementations will focus on further developing alerting mechanisms through automated modeling of species relevant to Madeira Island. As a result, a model using presence-absence data will be developed which will continuously update itself when data input is also updated. Further alerting protocols will be studied, towards presenting abundance alerts of marine biodiversity in the most efficient way towards boat owners.

6 Conclusion

Ecological processes have been a long-term focal point for researchers and ecology experts. In this dissertation, a new methodology is proposed, in which environmental niche modeling is web-based. The proposed system includes robust pipelines with established scientific repositories for large-scale studies.

Data was gathered from two experiments, one regarding the system's usability and the other regarding the ecological impact of the system's outputs. The former was collected from an adaptation of the settled SUS scale, while the latter benefited from pre and post-assessments through NEP and ecological awareness questionnaires.

The proposed system managed to simplify and accelerate the employment of ENM workflows, however, it also introduced state-of-the-art limitations. Seven major research challenges are addressed and discussed, from which ten guidelines towards future implementations were established.

The proposed system allows a fast and straightforward means by which to manipulate occurrence and environmental data, which can then be implemented in alerting systems, projection, and evaluation for assessing distributions of species in geographic space and their corresponding environmental combinations.

Although ENMs were not successful as an alerting mechanism towards ecological awareness of fisherman and catamaran' sailors, additional metrics, namely motivation, information channels, threats, attitudes, and values of marine biodiversity are considered for future attempts of such systems.

Furthermore, the proposal is meaningful when considering that there is an abundance of specialized software that supports the application of ENM protocols applications, however, entry-level systems are scarce, especially in educational systems, augmenting the barrier of initiation and discouraging future researchers to establish new studies. Given enormous progress in concept, there is an opportunity for a solution that supports both experts and beginners, without compromising simplicity and scientific validity.

References

[1] E. Sala, J. Mayorga, D. Bradley, R. B. Cabral, T. B. Atwood, A. Auber, W. Cheung, C. Costello, F. Ferretti, A. M. Friedlander *et al.*, “Protecting the global ocean for biodiversity, food and climate,” *Nature*, vol. 592, no. 7854, pp. 397–402, 2021.

[2] H. K. Lotze, M. Coll, and J. A. Dunne, “Historical changes in marine resources, food-web structure and ecosystem functioning in the adriatic sea, mediterranean,” *Ecosystems*, vol. 14, no. 2, pp. 198–222, 2011.

[3] M. J. Costello, M. Coll, R. Danovaro, P. Halpin, H. Ojaveer, and P. Miloslavich, “A census of marine biodiversity knowledge, resources, and future challenges,” *PloS one*, vol. 5, no. 8, p. e12110, 2010.

[4] S. D. Kraus, M. W. Brown, H. Caswell, C. W. Clark, M. Fujiwara, P. K. Hamilton, R. D. Kenney, A. R. Knowlton, S. Landry, C. A. Mayo *et al.*, “North atlantic right whales in crisis,” *Science*, vol. 309, no. 5734, pp. 561–562, 2005.

[5] R. M. Pace III, P. J. Corkeron, and S. D. Kraus, “State–space mark–recapture estimates reveal a recent decline in abundance of north atlantic right whales,” *Ecology and Evolution*, vol. 7, no. 21, pp. 8730–8741, 2017.

[6] W. C. Cook, “Notes on predicting the probable future distribution of introduced insects,” *Ecology*, vol. 12, no. 2, pp. 245–247, 1931.

[7] D. S. Maclagan, “The predictive method in animal ecology,” *Nature*, vol. 149, no. 3769, pp. 115–116, 1942.

[8] A. R. Wallace, “On the Zoological Geography of the Malay Archipelago,” *Zoological Journal of the Linnean Society*, vol. 4, no. 16, pp. 172–184, 06 2009. [Online]. Available: https://doi.org/10.1111/j.1096-3642.1860.tb00090.x

[9] J. M. Lobo, A. Jiménez-Valverde, and J. Hortal, “The uncertain nature of absences and their importance in species distribution modelling,” *Ecography*, vol. 33, no. 1, pp. 103–114, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0587.2009.06039.x

[10] D. L. Warren and S. N. Seifert, “Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria,” *Ecological Applications*, vol. 21, no. 2, pp. 335–342, 2011, _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/10-1171.1. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/10-1171.1

[11] W. Thuiller, B. Lafourcade, R. Engler, and M. B. Araújo, “Biomod – a platform for ensemble forecasting of species distributions,” *Ecography*, vol. 32, no. 3, pp. 369–373, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0587.2008.05742.x

[12] J. Elith and J. R. Leathwick, “Species Distribution Models: Ecological Explanation and Prediction Across Space and Time,” *Annual Review of Ecology, Evolution, and Systematics*, vol. 40, no. 1, pp. 677–697, Dec. 2009. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.110308.120159

[13] M. S. Wisz, R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan, “Effects of sample size on the performance of species distribution models,” *Diversity and Distributions*, vol. 14, no. 5, pp. 763–773, 2008, _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1472-4642.2008.00482.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1472-4642.2008.00482.x

[14] J. Lobo, "More complex distribution or more representative data?" *Biodiversity Informatics*, vol. 5, Jan. 2008.

[15] R. J. Hijmans and J. Elith, "Species distribution modeling with r," *R CRAN Project*, 2013.

[16] F. Fisheries, "Aquaculture department," *The state of world fisheries and aquaculture*, pp. 1–153, 2012.

[17] G. A. Polis, "Stability is woven by complex webs," *Nature*, vol. 395, no. 6704, pp. 744–745, 1998.

[18] K. Kaschner, D. P. Tittensor, J. Ready, T. Gerrodette, and B. Worm, "Current and future patterns of global marine mammal biodiversity," *PLoS one*, vol. 6, no. 5, p. e19653, 2011.

[19] D. Kachelriess, M. Wegmann, M. Gollock, and N. Pettorelli, "The application of remote sensing for marine protected area management," *Ecological Indicators*, vol. 36, pp. 169 – 177, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1470160X13002653

[20] A. Balmford, A. Bruner, P. Cooper, R. Costanza, S. Farber, R. E. Green, M. Jenkins, P. Jefferiss, V. Jessamy, J. Madden, K. Munro, N. Myers, S. Naeem, J. Paavola, M. Rayment, S. Rosendo, J. Roughgarden, K. Trumper, and R. K. Turner, "Economic reasons for conserving wild nature," *Science*, vol. 297, no. 5583, pp. 950–953, 2002. [Online]. Available: https://science.sciencemag.org/content/297/5583/950

[21] N. Beaumont, M. Austen, S. Mangi, and M. Townsend, "Economic valuation for the conservation of marine biodiversity," *Marine Pollution Bulletin*, vol. 56, no. 3, pp. 386 – 396, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0025326X07004535

[22] S. Wilson, A. Dolman, A. Cheal, M. Emslie, M. S. Pratchett, and H. P. Sweatman, "Maintenance of fish diversity on disturbed coral reefs," *Coral Reefs*, vol. 28, no. 1, pp. 3–14, 2009.

[23] T. P. Hughes, D. R. Bellwood, C. S. Folke, L. J. McCook, and J. M. Pandolfi, "No-take areas, herbivory and coral reef resilience," *Trends in Ecology & Evolution*, vol. 22, no. 1, pp. 1 – 3, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169534706003429

[24] C. Albaladejo, P. Sánchez, A. Iborra, F. Soto, J. A. López, and R. Torres, "Wireless sensor networks for oceanographic monitoring: A systematic review," *Sensors*, vol. 10, no. 7, pp. 6948–6968, 2010.

[25] C. E. Schroeder, D. A. Wilson, T. Radman, H. Scharfman, and P. Lakatos, "Dynamics of active sensing and perceptual selection," *Current Opinion in Neurobiology*, vol. 20, no. 2, pp. 172 – 176, 2010, cognitive neuroscience. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959438810000322

[26] D. A. Mouat, G. G. Mahin, and J. Lancaster, "Remote sensing techniques in the analysis of change detection," *Geocarto International*, vol. 8, no. 2, pp. 39–50, 1993. [Online]. Available: https://doi.org/10.1080/10106049309354407

[27] D. Lyzenga, "Passive remote-sensing techniques for mapping water depth and bottom features," *Applied optics*, vol. 17, pp. 379–83, 02 1978.

[28] J. Chen, K. Bittinger, E. S. Charlson, C. Hoffmann, J. Lewis, G. D. Wu, R. G. Collman, F. D. Bushman, and H. Li, "Associating microbiome composition with environmental covariates using generalized UniFrac distances," *Bioinformatics*, vol. 28, no. 16, pp. 2106–2113, 06 2012. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts342

[29] H.-C. Lee, A. Banerjee, Y.-M. Fang, B.-J. Lee, and C.-T. King, "Design of a multifunctional wireless sensor for in-situ monitoring of debris flows," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 11, pp. 2958–2967, 2010.

[30] C. A. Pérez, M. Jimenez, F. Soto, R. Torres, J. López, and A. Iborra, "A system for monitoring marine environments based on wireless sensor networks," in *OCEANS 2011 IEEE-Spain*. IEEE, 2011, pp. 1–6.

[31] P. Jiang, H. Xia, Z. He, and Z. Wang, "Design of a water environment monitoring system based on wireless sensor networks," *Sensors*, vol. 9, no. 8, pp. 6411–6434, 2009.

[32] G. Xu, W. Shen, and X. Wang, "Applications of wireless sensor networks in marine environment monitoring: A survey," *Sensors*, vol. 14, no. 9, pp. 16 932–16 954, 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/9/16932

[33] A. Bayo, D. Antolín, N. Medrano, B. Calvo, and S. Celma, "Early detection and monitoring of forest fire with a wireless sensor network system," *Procedia Engineering*, vol. 5, pp. 248–251, 2010.

[34] Y. E. Aslan, I. Korpeoglu, and Ö. Ulusoy, "A framework for use of wireless sensor networks in forest fire detection and monitoring," *Computers, Environment and Urban Systems*, vol. 36, no. 6, pp. 614–625, 2012.

[35] Y. Singh, S. Saha, U. Chugh, and C. Gupta, "Distributed event detection in wireless sensor networks for forest fires," in *2013 UKSim 15th International Conference on Computer Modelling and Simulation*. IEEE, 2013, pp. 634–639.

[36] I. Silva, L. A. Guedes, P. Portugal, and F. Vasques, "Reliability and availability evaluation of wireless sensor networks for industrial applications," *Sensors*, vol. 12, no. 1, pp. 806–838, 2012.

[37] G. Zhao *et al.*, "Wireless sensor networks for industrial process monitoring and control: A survey." *Netw. Protoc. Algorithms*, vol. 3, no. 1, pp. 46–63, 2011.

[38] R. Morais, S. G. Matos, M. A. Fernandes, A. L. Valente, S. F. Soares, P. Ferreira, and M. Reis, "Sun, wind and water flow as energy supply for small stationary data acquisition platforms," *Computers and electronics in agriculture*, vol. 64, no. 2, pp. 120–132, 2008.

[39] X. Li, Y. Deng, and L. Ding, "Study on precision agriculture monitoring framework based on wsn," in *2008 2nd International Conference on Anti-counterfeiting, Security and Identification*. IEEE, 2008, pp. 182–185.

[40] H. Qian, P. Sun, and Y. Rong, "Design proposal of self-powered wsn node for battle field surveillance," *Energy Procedia*, vol. 16, pp. 753–757, 2012.

[41] G. Padmavathi, D. Shanmugapriya, M. Kalaivani *et al.*, "A study on vehicle detection and tracking using wireless sensor networks," *Wireless Sensor Network*, vol. 2, no. 02, p. 173, 2010.

[42] D. Tacconi, D. Miorandi, I. Carreras, F. Chiti, and R. Fantacci, "Using wireless sensor networks to support intelligent transportation systems," *Ad Hoc Networks*, vol. 8, no. 5, pp. 462–473, 2010.

[43] M. Tubaishat, P. Zhuang, Q. Qi, and Y. Shang, "Wireless sensor networks in intelligent transportation systems," *Wireless communications and mobile computing*, vol. 9, no. 3, pp. 287–302, 2009.

[44] H. Lee, C. Wu, and H. Aghajan, "Vision-based user-centric light control for smart environments," *Pervasive and Mobile Computing*, vol. 7, no. 2, pp. 223–240, 2011.

[45] J. Bangali and A. Shaligram, "Energy efficient smart home based on wireless sensor network using labview," *American Journal of Engineering Research (AJER)*, vol. 2, no. 12, pp. 409–413, 2013.

[46] R. N. Handcock, D. L. Swain, G. J. Bishop-Hurley, K. P. Patison, T. Wark, P. Valencia, P. Corke, and C. J. O'Neill, "Monitoring animal behaviour and environmental interactions using wireless sensor networks, gps collars and satellite remote sensing," *Sensors*, vol. 9, no. 5, pp. 3586–3603, 2009.

[47] E. S. Nadimi, R. N. Jørgensen, V. Blanes-Vidal, and S. Christensen, "Monitoring and classifying animal behavior using zigbee-based mobile ad hoc wireless sensor networks and artificial neural networks," *Computers and Electronics in Agriculture*, vol. 82, pp. 44–54, 2012.

[48] M. Bahrepour, N. Meratnia, M. Poel, Z. Taghikhaki, and P. J. Havinga, "Distributed event detection in wireless sensor networks for disaster management," in *2010 international conference on intelligent networking and collaborative systems*. IEEE, 2010, pp. 507–512.

[49] M. Iacono, E. Romano, and S. Marrone, "Adaptive monitoring of marine disasters with intelligent mobile sensor networks," in *2010 IEEE Workshop on Environmental Energy and Structural Monitoring Systems*. IEEE, 2010, pp. 38–45.

[50] S. L. Ozesmi and M. E. Bauer, "Satellite remote sensing of wetlands," *Wetlands ecology and management*, vol. 10, no. 5, pp. 381–402, 2002.

[51] R. Sandau, "Status and trends of small satellite missions for earth observation," *Acta Astronautica*, vol. 66, no. 1, pp. 1 – 12, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094576509003488

[52] E. Guirado, S. Tabik, M. L. Rivas, D. Alcaraz-Segura, and F. Herrera, "Whale counting in satellite and aerial images with deep learning," *Sci Rep 9*, 2019. [Online]. Available: https://www.nature.com/articles/s41598-019-50795-9

[53] C. Bamford, N. Kelly, L. Dalla Rosa, D. Cade, H. Cubaynes, A. Mesquita, L. Gerrish, A. Friedlaender, and J. Jackson, "A comparison of baleen whale density estimates derived from overlapping satellite imagery and a shipborne survey," *Scientific Reports*, vol. 10, 07 2020.

[54] H. J. Lynch, R. White, A. D. Black, and R. Naveen, "Detection, differentiation, and abundance estimation of penguin species by high-resolution satellite imagery," *Polar Biology*, vol. 35, no. 6, pp. 963–968, 2012.

[55] A. S. Kennedy, A. N. Zerbini, O. Vásquez, N. Gandilhon, P. J. Clapham, and O. Adam, "Local and migratory movements of humpback whales (megaptera novaeangliae) satellite-tracked in the north atlantic ocean," *Canadian Journal of Zoology*, vol. 92, no. 1, pp. 9–18, 2014.

[56] P. T. Fretwell, I. J. Staniland, and J. Forcada, "Whales from space: counting southern right whales by satellite," *PloS one*, vol. 9, no. 2, p. e88655, 2014.

[57] H. C. Cubaynes, P. T. Fretwell, C. Bamford, L. Gerrish, and J. A. Jackson, "Whales from space: four mysticete species described using new vhr satellite imagery," *Marine Mammal Science*, vol. 35, no. 2, pp. 466–491, 2019.

[58] P. Wade, M. P. Heide-Jørgensen, K. Shelden, J. Barlow, J. Carretta, J. Durban, R. LeDuc, L. Munger, S. Rankin, A. Sauter *et al.*, "Acoustic detection and satellite-tracking leads to discovery of rare concentration of endangered north pacific right whales," *Biology letters*, vol. 2, no. 3, pp. 417–419, 2006.

[59] N. Platonov, I. Mordvintsev, and V. Rozhnov, "The possibility of using high resolution satellite images for detection of marine mammals," *Biology Bulletin*, vol. 40, no. 2, pp. 197–205, 2013.

[60] P. T. Fretwell, P. Scofield, and R. A. Phillips, "Using super-high resolution satellite imagery to census threatened albatrosses," *Ibis*, vol. 159, no. 3, pp. 481–490, 2017.

[61] M. A. LaRue, J. J. Rotella, R. A. Garrott, D. B. Siniff, D. G. Ainley, G. E. Stauffer, C. C. Porter, and P. J. Morin, "Satellite imagery can be used to detect variation in abundance of weddell seals (leptonychotes weddellii) in erebus bay, antarctica," *Polar Biology*, vol. 34, no. 11, p. 1727, 2011.

[62] C. R. McMahon, H. Howe, J. Van Den Hoff, R. Alderman, H. Brolsma, and M. A. Hindell, "Satellites, the all-seeing eyes in the sky: counting elephant seals from space," *PloS one*, vol. 9, no. 3, p. e92613, 2014.

[63] A. Borowicz, H. Le, G. Humphries, G. Nehls, C. Höschle, V. Kosarev, and H. J. Lynch, "Aerial-trained deep learning networks for surveying cetaceans from satellite imagery," *PloS one*, vol. 14, no. 10, p. e0212532, 2019.

[64] L. Dalla Rosa, E. Secchi, Y. Maia, A. Zerbini, and M. Heide-Jørgensen, "Movements of satellite-monitored humpback whales on their feeding ground along the antarctic peninsula," *Polar Biology*, vol. 31, no. 7, pp. 771–781, 2008.

[65] I. Duporge, O. Isupova, S. Reece, D. Macdonald, and T. Wang, "Using very high-resolution satellite imagery and deep learning to detect and count african elephants in heterogeneous landscapes," 09 2020.

[66] S. Stapleton, M. LaRue, N. Lecomte, S. Atkinson, D. Garshelis, C. Porter, and T. Atwood, "Polar bears from space: assessing satellite imagery as a tool to track arctic wildlife," *PLoS One*, vol. 9, no. 7, p. e101513, 2014.

[67] M. A. LaRue and S. Stapleton, "Estimating the abundance of polar bears on wrangel island during late summer using high-resolution satellite imagery: a pilot study," *Polar Biology*, vol. 41, no. 12, pp. 2621–2626, 2018.

[68] M. B. Araujo and A. Guisan, "Five (or so) challenges for species distribution modelling," *Journal of biogeography*, vol. 33, no. 10, pp. 1677–1688, 2006.

[69] A. T. Peterson, L. G. Ball, and K. P. Cohoon, "Predicting distributions of mexican birds using ecological niche modelling methods," *Ibis*, vol. 144, no. 1, pp. E27–E32, 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.0019-1019.2001.00031.x

[70] J. D. Lozier, P. Aniello, and M. J. Hickerson, "Predicting the distribution of sasquatch in western north america: anything goes with ecological niche modelling," *Journal of Biogeography*, vol. 36, no. 9, pp. 1623–1627, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2699.2009.02152.x

[71] K. L. Howell, R. Holt, I. P. Endrino, and H. Stewart, "When the species is also a habitat: Comparing the predictively modelled distributions of lophelia pertusa and the reef habitat it forms," *Biological Conservation*, vol. 144, no. 11, pp. 2656 – 2665, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006320711002886

[72] O. F. Anderson, J. M. Guinotte, A. A. Rowden, D. M. Tracey, K. A. Mackay, and M. R. Clark, "Habitat suitability models for predicting the occurrence of vulnerable marine ecosystems in the seas around new zealand," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 115, pp. 265 – 292, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967063716301637

[73] A. Bombosch, D. P. Zitterbart, I. Van Opzeeland, S. Frickenhaus, E. Burkhardt, M. S. Wisz, and O. Boebel, "Predictive habitat modelling of humpback (megaptera novaeangliae) and antarctic minke (balaenoptera bonaerensis) whales in the southern ocean as a planning tool for seismic surveys," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 91,

pp. 101 – 114, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967063714000995

[74] J. Guinan, C. Brown, M. F. Dolan, and A. J. Grehan, "Ecological niche modelling of the distribution of cold-water coral habitat using underwater remote sensing data," *Ecological Informatics*, vol. 4, no. 2, pp. 83 – 92, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574954109000053

[75] R. Muscarella, P. J. Galante, M. Soley-Guardia, R. A. Boria, J. M. Kass, M. Uriarte, and R. P. Anderson, "Enmeval: An r package for conducting spatially independent evaluations and estimating optimal model complexity for maxent ecological niche models," *Methods in Ecology and Evolution*, vol. 5, no. 11, pp. 1198–1205, 2014. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12261

[76] R. J. Hijmans, "Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model," *Ecology*, vol. 93, no. 3, pp. 679–688, 2012.

[77] S. D. Veloz, "Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models," *Journal of Biogeography*, vol. 36, no. 12, pp. 2290–2299, 2009.

[78] H. Qiao, A. T. Peterson, L. P. Campbell, J. Soberón, L. Ji, and L. E. Escobar, "Nichea: creating virtual species and ecological niches in multivariate environmental scenarios," *Ecography*, vol. 39, no. 8, pp. 805–813, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.01961

[79] B. Bentlage and M. Shcheglovitova, "Nichepy: modular tools for estimating the similarity of ecological niche and species distribution models," *Methods in Ecology and Evolution*, vol. 3, no. 3, pp. 484–489, 2012.

[80] M. E. Cobos, A. Peterson, N. Barve, and L. Osorio-Olvera, "Kuenm: An r package for detailed development of ecological niche models using maxent," *PeerJ*, vol. 7, p. e6281, 02 2019.

[81] W. Thuiller, D. Georges, and R. Engler, "biomod2: Ensemble platform for species distribution modelling," vol. 2, 01 2014.

[82] L. J. Hijmans R.J, Phillips S. and E. J., "dismo: Ensemble platform for species distribution modelling," 2015. [Online]. Available: https://cran.r-project.org/web/packages/dismo/dismo.pdf

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, 01 2012.

[84] N. Golding, T. A. August, T. C. D. Lucas, D. J. Gavaghan, E. E. van Loon, and G. McInerny, "The zoon r package for reproducible and shareable species distribution modelling," *Methods in Ecology and Evolution*, vol. 9, no. 2, pp. 260–268, 2018. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12858

[85] L. Osorio-Olvera, A. Lira-Noriega, J. Soberón, A. T. Peterson, M. Falconi, R. G. Contreras-Díaz, E. Martínez-Meyer, V. Barve, and N. Barve, "ntbox: An r package with graphical user interface for modelling and evaluating multidimensional ecological niches," *Methods in Ecology and Evolution*, vol. 11, no. 10, pp. 1199–1206, 2020. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13452

[86] S. E. Ahmed, G. McInerny, K. O'Hara, R. Harper, L. Salido, S. Emmott, and L. N. Joppa, "Scientists and software–surveying the species distribution modelling community," *Diversity and Distributions*, vol. 21, no. 3, pp. 258–267, 2015.

[87] R. K. Colwell and T. F. Rangel, "Hutchinson's duality: The once and future niche," *Proceedings of the National Academy of Sciences*, vol. 106, no. Supplement 2, pp. 19 651–19 658, 2009. [Online]. Available: https://www.pnas.org/content/106/Supplement_2/19651

[88] H. Qiao, C. Lin, L. Ji, and Z. Jiang, "mmweb - an online platform for employing multiple ecological niche modeling algorithms," *PLoS ONE*, vol. 7, 2012.

[89] J. Brown, J. Bennett, and C. French, "Sdmtoolbox 2.0: The next generation python-based gis toolkit for landscape genetic, biogeographic and species distribution model analyses," *PeerJ*, vol. 5, p. e4095, 12 2017.

[90] L. E. Escobar, A. Lira-Noriega, G. Medina-Vogel, and A. T. Peterson, "Potential for spread of the white-nose fungus (pseudogymnoascus destructans) in the americas: use of maxent and nichea to assure strict model transference," *Geospatial health*, pp. 221–229, 2014.

[91] B. Naimi and M. B. Araújo, "sdm: a reproducible and extensible r platform for species distribution modelling," *Ecography*, vol. 39, no. 4, pp. 368–375, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.01881

[92] S. Schmitt, R. Pouteau, D. Justeau, F. de Boissieu, and P. Birnbaum, "ssdm: An r package to predict distribution of species richness and composition based on stacked species distribution models," *Methods in Ecology and Evolution*, vol. 8, no. 12, pp. 1795–1803, 2017. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12841

[93] P. Verma, "Gracoli: A graphical command line user interface," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 3143–3146. [Online]. Available: https://doi.org/10.1145/2468356.2479631

[94] P. T. Groth and Y. Gil, "A scientific workflow construction command line," in *Proceedings of the 14th International Conference on Intelligent User Interfaces*, ser. IUI '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 445–450. [Online]. Available: https://doi.org/10.1145/1502650.1502716

[95] M. Sandal, F. Benedetti, M. Brucale, A. Gomez-Casado, and B. Samorì, "Hooke: an open software platform for force spectroscopy," *Bioinformatics*, vol. 25, no. 11, pp. 1428–1430, 03 2009. [Online]. Available: https://doi.org/10.1093/bioinformatics/btp180

[96] S. Hultstrand and R. Olofsson, "Git - cli or gui: Which is most widely used and why?" Ph.D. dissertation, 2015. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-10539

[97] J. Kass, B. Vilela, M. Aiello-Lammens, R. Muscarella, C. Merow, and R. Anderson, "Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion," *Methods in Ecology and Evolution*, vol. 9, 01 2018.

[98] L. Citi, "Development of a neural interface for the controlof a robotic hand," 2009.

[99] B. Myers and M. B. Rosson, "Survey on user interface programming," 01 1992, pp. 195–202.

[100] U. Andriolo, G. Gonçalves, F. Bessa, and P. Sobral, "Mapping marine litter on coastal dunes with unmanned aerial systems: A showcase on the atlantic coast," *Science of The Total Environment*, vol. 736, p. 139632, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0048969720331521

[101] J. R. Jambeck and K. Johnsen, "Citizen-based litter and marine debris data collection and mapping," *Computing in Science Engineering*, vol. 17, no. 4, pp. 20–26, 2015.

[102] A. Thanacoody, G. Bekaroo, A. Santokhee, and S. Juddoo, "Analyzing the prospects and acceptance of mobile-based marine debris tracking," in *Smart and Sustainable Engineering for*

Next Generation Applications, P. Fleming, B. M. Lacquet, S. Sanei, K. Deb, and A. Jakobsson, Eds. Cham: Springer International Publishing, 2019, pp. 256–267.

[103] I. Almarashdeh, K. E. Eldaw, M. AlSmadi, U. Badawi, F. Haddad, O. A. Abdelkader, G. Jaradat, A. Alkhaldi, and Y. Qawqzeh, "Search convenience and access convenience: The difference between website shopping and mobile shopping," in *International Conference on Soft Computing and Pattern Recognition*. Springer, 2018, pp. 33–42.

[104] Z. Tupikovskaja-Omovie, D. Tyler, S. Dhanapala, and S. Hayes, "Mobile app versus website: a comparative eye-tracking case study of topshop," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, vol. 9, 2015.

[105] L. J. Beaumont, L. Hughes, and M. Poulsen, "Predicting species distributions: use of climatic parameters in bioclim and its impact on predictions of species' current and future distributions," *Ecological Modelling*, vol. 186, no. 2, pp. 251–270, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304380005000530

[106] T. H. Booth, H. A. Nix, J. R. Busby, and M. F. Hutchinson, "bioclim: the first species distribution modelling package, its early applications and relevance to most current maxent studies," *Diversity and Distributions*, vol. 20, no. 1, pp. 1–9, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.12144

[107] J. Elith*, C. H. Graham*, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. M. Overton, A. Townsend Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann, "Novel methods improve prediction of species' distributions from occurrence data," *Ecography*, vol. 29, no. 2, pp. 129–151, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2006.0906-7590.04596.x

[108] G. Carpenter, A. Gillison, and J. Winter, "Domain: a flexible modelling procedure for mapping potential distributions of plants and animals," *Biodiversity & Conservation*, vol. 2, no. 6, pp. 667–680, 1993.

[109] K. Finstad, "The system usability scale and non-native english speakers," *Journal of usability studies*, vol. 1, no. 4, pp. 185–188, 2006.

[110] J. R. Lewis, "Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995.

[111] M. Anderson, "New ecological paradigm (nep) scale," *Berkshire Encyclopedia of Sustainability*, vol. 6, pp. 260–262, 01 2012.

[112] H. D. Denis and L. N. Pereira, "Measuring the level of endorsement of the new environmental paradigm: a transnational study," *Dos Algarves: A Multidisciplinary e-Journal*, no. 23, pp. 4–26, 2017.

[113] M. Partanen-Hertell, P. Harju-Autti, K. Kreft-Burman, and D. Pemberton, "Raising environmental awareness in the baltic sea area," 1999.

[114] D. Pemberton, M. Partanen-Hertell, and P. Harju-Autti, *Analysis of the questionnaire on environmental awareness in the Baltic Sea. Analysis of the answers received to the questionnaire targeted on several key expert groups within the 14 countries of the Baltic Sea catchment area.* Suomen ympäristökeskus, 1999.

[115] A. Gkargkavouzi, G. Halkos, and S. Matsiori, "Assessing values, attitudes and threats towards marine biodiversity in a greek coastal port city and their interrelationships," *Ocean & Coastal Management*, vol. 167, pp. 115–126, 2019.

[116] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[117] J. Sauro, "A practical guide to the system usability scale: Background," *Benchmarks & best practices*, 2011.

[118] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Business horizons*, vol. 60, no. 3, pp. 293–303, 2017.

[119] N. Micic, D. Neagu, F. Campean, and E. H. Zadeh, "Towards a data quality framework for heterogeneous data," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 2017, pp. 155–162.

[120] R. Kent and Y. Carmel, "Presence-only versus presence–absence data in species composition determinant analyses," *Diversity and Distributions*, vol. 17, no. 3, pp. 474–479, 2011.

[121] J. Elith, C. Graham, R. Valavi, M. Abegg, C. Bruce, A. Ford, A. Guisan, R. J. Hijmans, F. Huettmann, L. Lohmann *et al.*, "Presence-only and presence-absence data for comparing species distribution modeling methods," *Biodiversity informatics*, vol. 15, no. 2, pp. 69–80, 2020.

[122] A. E. Gelfand and S. Shirota, "Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data," *Ecological Monographs*, vol. 89, no. 3, p. e01372, 2019.

[123] P. C. Fiedler, J. V. Redfern, K. A. Forney, D. M. Palacios, C. Sheredy, K. Rasmussen, I. García-Godos, L. Santillán, M. J. Tetley, F. Félix *et al.*, "Prediction of large whale distributions: a comparison of presence–absence and presence-only modeling techniques," *Frontiers in Marine Science*, vol. 5, p. 419, 2018.

[124] A.-X. Zhu, Y. Miao, L. Yang, S. Bai, J. Liu, and H. Hong, "Comparison of the presence-only method and presence-absence method in landslide susceptibility mapping," *Catena*, vol. 171, pp. 222–233, 2018.

[125] A. M. Gormley, D. M. Forsyth, P. Griffioen, M. Lindeman, D. S. Ramsey, M. P. Scroggie, and L. Woodford, "Using presence-only and presence–absence data to estimate the current and potential distributions of established invasive species," *Journal of Applied Ecology*, vol. 48, no. 1, pp. 25–34, 2011.

[126] A. I. Martins, A. F. Rosa, A. Queirós, A. Silva, and N. P. Rocha, "European portuguese validation of the system usability scale (sus)," *Procedia Computer Science*, vol. 67, pp. 293–300, 2015.

[127] B. Bozkurt, "Enkephalins and the opioid system of the heart: Too much of a good thing or the goldilocks syndrome?" 2019.