

Towards an integrated ecological modelling framework to assess the risk of Alien Species spread in the Mediterranean Sea: *Caulerpa cylindracea* as a case study

Abstract

Predicting the future spread of Alien Invasive Species (AIS) in the Mediterranean Sea is a critical challenge for marine risk assessment due to their significant ecological and economic impact.

The aim of this thesis is to provide a unified theoretical framework for AIS risk assessment that integrates both correlative (Machine Learning) and mechanistic (process-based) models to predict suitable areas and simulate the invasion process.

This thesis addresses the development of the three key components to achieve this goal:

The first component is the preparation of a solid data basis. The dataset contains current and future key bioclimatic variables for the Mediterranean Sea. It covers the period 2005–2099 under the high emissions scenario RCP8.5 at high spatial and temporal resolution.

The second component is a robust Machine Learning Ecological Niche Modelling (ML-ENM) method. It includes three major methodological advances: (i) an improved model tuning phase, (ii) a novel site weighting scheme to improve the validity of performance metrics, and

(iii) a multi-criteria decision-making framework for unbiased model selection. These innovations address critical challenges of ML-ENM modelling related to: maximising model transferability, mitigating data bias and ensuring ecological validity.

The third component is a spatial interaction model that explicitly simulates vessel-mediated invasions from known presence sites. The probability of invasion is modelled as the strength of the connection between “infected” and “exposed” (susceptible) areas, using a production-constrained gravity model. The model incorporates data on vessel traffic, suitability and substrate type.

The framework is presented through an applied case study of *Caulerpa cylindracea*, one of the most dangerous invasive alien species in the Mediterranean. The results consist of annual suitability maps of *C. cylindracea* for the period 2000-2050 and the maps of simulated annual invasions for the period 2020-2023. Despite a significant negative trend in suitability, the species continues to expand its range due to increasing shipping activities, highlighting the detrimental impact of human activities on the basin.

The framework is modular and easily transferable to allow for future extensions and applications to other target species, significantly expanding the reliability and depth of the AIS risk assessment. The dataset and scripts to reproduce the analysis are publicly available.

Table of Contents

ABSTRACT	1
THESIS SUMMARY	6
INTRODUCTION	12
ALIEN INVASIVE SPECIES	12
<i>The invasion process</i>	13
<i>Invasion kinetics</i>	15
<i>Corridors, vectors, pathways</i>	17
CAULERPA CYLINDRACEA AS A CASE STUDY	19
<i>Physiological Characteristics</i>	19
<i>Biological Traits</i>	20
<i>Ecological Impacts and Interactions</i>	21
<i>A model species for biological invasion studies</i>	22
THE MEDITERRANEAN SEA.....	23
<i>Geological History and Oceanographic Features</i>	23
<i>Biodiversity and Ecological Richness</i>	25
Human Impacts on the Mediterranean.....	25
<i>Alien Invasive Species in the Mediterranean Sea</i>	26
<i>Vectors and Pathways of Introduction</i>	27
The Suez Canal (Lessepsian Migration)	27
Shipping Activities.....	27
Aquaculture and Fisheries	28
<i>Impacts of AIS</i>	29
Ecological Impacts.....	29
Socio-Economic Impacts	29
THE MARINE STRATEGY FRAMEWORK DIRECTIVE (MSFD)	30
<i>MSFD Focus on Alien Invasive Species</i>	30
<i>Strategies and Measures to Combat AIS</i>	32
Prevention.....	32
Early Detection and Rapid Response	32
Challenges from Climate Change and Human Activities.....	32
TOWARD AN INTEGRATED MODELLING FRAMEWORK	33
THEORETICAL FOUNDATION	33
OVERVIEW OF ECOLOGICAL NICHE MODELLING	34
<i>Mechanistic and Correlative Approaches</i>	35
Mechanistic Models	35
Correlative Models	36
SPECIES' OCCURRENCE DATA	37
<i>Presence-Absence Data</i>	38
Modelling assumptions	38
Limitations in interpretation	38
<i>Presence-Only Data</i>	39
Modelling assumptions	39
Limitations in interpretation	40
MAXENT	41
<i>How it works</i>	41
Ecological perspective	41

Machine Learning perspective	42
<i>Data input</i>	44
<i>Output interpretation</i>	44
OVERVIEW OF SPATIAL INTERACTIONS MODELS.....	45
<i>Gravity models</i>	47
Gravity models to estimate global biological invasions hotspots	48
Gravity models and probability of invasion	49
Limitations	50

2005-2099 HIGH RESOLUTION BIOCLIMATIC VARIABLES FOR THE SURFACE AND BOTTOM OF THE MEDITERRANEAN SEA. 51

AN EASY-TO-USE, HIGH-QUALITY DATASET FOR RISK ASSESSMENT	51
ARTICLE: DOWNSCALED AND BIAS-CORRECTED BIOCLIMATIC DATASET FOR THE MEDITERRANEAN SEA (2005–2099)	52
Authors	52
Affiliations.....	52
Corresponding author’s email address and Twitter handle	52
Keywords	52
<i>Abstract</i>	52
<i>SPECIFICATIONS TABLE</i>	53
<i>VALUE OF THE DATA</i>	55
<i>BACKGROUND</i>	56
<i>DATA DESCRIPTION</i>	56
<i>EXPERIMENTAL DESIGN, MATERIALS AND METHODS</i>	60
COMPUTATIONAL ENVIRONMENT:	60
MATERIALS:	60
GIS MODELLING:	62
EXPERIMENTAL DESIGN and METHODS:	62
<i>LIMITATIONS</i>	63

TOWARD A RELIABLE ENM PROCESS 65

ABSTRACT	66
DEFINING THE SCOPE.....	67
DATA PROCESSING	68
<i>Occurrence data</i>	68
<i>Occurrence data cleaning</i>	69
<i>Environmental data</i>	70
<i>Variable preparation and selection</i>	72
<i>Calibration region</i>	73
<i>Spatial autocorrelation</i>	75
MAXENT LEARNING AND CALIBRATION	76
<i>Calibration phase</i>	77
MODEL EVALUATION	77
<i>Improving the informative value of performance metrics</i>	80
Site-specific weighting	81
<i>Weighted and unweighted metrics</i>	82
<i>Delta metrics</i>	84
MODEL SELECTION	85
<i>Multi-Criteria Decision Problem approach</i>	86
<i>Biological realism</i>	88
SPATIAL AND TEMPORAL PROJECTIONS	91
<i>From continuous probability to suitability classes</i>	91
Uncertainty in classification	91

<i>Time series analysis</i>	92
Mann-Kendall Test.....	92
<i>Extrapolation Risk Analysis</i>	93
<i>Analysis of the results</i>	96
CONCLUSION	99
<i>A robust ENM process</i>	99
<i>Insights on invasion spread</i>	100
<i>Limitations</i>	100
A MECHANISTIC MODEL TO EXPLICITLY ASSESS EXPOSURE TO AIS SECONDARY INVASION	101
INTRODUCTION	101
<i>An improved gravity model to estimate probability of invasion mediated by vessel traffic</i>	102
<i>Data and methods</i>	104
Data.....	104
Methods	106
<i>Results and discussion</i>	109
Deterrence function	109
Per-ship probability of initiating an invasion	110
Expected invasions per year	113
Spatial patterns of invasion	115
Substrate-specific invasions	116
<i>Conclusion</i>	117
<i>Limitations</i>	118
CONCLUSIONS AND FINAL REMARKS.....	120
FUTURE DIRECTIONS	123
APPENDIX A	125
BIBLIOGRAPHY	129

Thesis summary

Advances in the availability of data and Machine Learning (ML) are revolutionising the way we live and think. This revolution has permeated every scientific field over the past decade, and it is going to increase in the future (Malde et al., 2020).

ML applications hold immense potential for the management and protection of marine ecosystems, which are increasingly threatened by human activities and climate change.

However, realising this potential requires not only the best available data and techniques (Goodwin et al., 2022), but also a holistic view and approach, a clear definition of the scope, system components and boundaries, transparent methods and an iterative review of the process.

Biological invasions are among the greatest and growing threats to biodiversity and human well-being (Brondízio et al., 2021). Formally, Alien Invasive Species (AIS) are defined as:

“Non-Indigenous Species (NIS)¹ that display invasive characteristics and threaten or adversely impact biodiversity and related ecosystem services” (Regulation 1143/2014/EU, 2014).

It is worth reflecting on how much information this definition entails. The term “*alien*” refers to the concept of *Non-Indigenous Species (NIS)*¹: Species that have been introduced outside their native range due to human activities and can survive and reproduce. The term “*Invasive*”

¹ Non-indigenous species (NIS; synonyms: alien, exotic, non-native, allochthonous) are species, subspecies or lower taxa introduced outside of their natural range (past or present) and outside of their natural dispersal potential. This includes any part, gamete or propagule of such species that might survive and subsequently reproduce. Their presence in the given region is due to intentional or unintentional introduction resulting from human activities. Natural shifts in distribution ranges (e.g. due to climate change or dispersal by ocean currents) do not qualify a species as a NIS. However, secondary introductions of NIS from the area(s) of their first arrival could occur without human involvement due to spread by natural means (Regulation 1143/2014/EU, 2014).

refers to a complex concept that can be seen as a “judgment”. It represents how the “community²”, to the best of its knowledge³, data and vision, evaluates the *effects*⁴ of the alien species as significantly negative. An alien species has the potential (high *risk*⁵) to spread uncontrollably and alter the functioning and services of the ecosystem at any level by directly (or indirectly) damaging one or more key species and services.

AIS risk assessment is an objective process of estimating *risk*, broadly defined as the product of the probability of events and the severity of their impacts (Leung et al., 2012).

High-resolution, regional, AIS risk assessments are mandatory under the Marine Strategy Framework Directive (MSFD) (Zenetos et al., 2012), and the application of ML approaches is the most feasible way to achieve this goal. However, there are still concerns about the reliability of ML models, especially when applied to ecological data, which are often incomplete and uncertain (Yackulic et al., 2013).

The reasons for this are partly due to a lack of integrated knowledge of the scientific problem: *It's not just about biology*. The advent of machine learning has provided researchers with extremely powerful modelling tools but has also imposed a complexity that requires multidimensional skills — including mathematics, statistics and coding.

It's not just about statistics and coding. The technical aspects (e.g. model assumptions) are necessary to respect the "datum" nature of the data and produce numerical results, but they should also provide ecologically meaningful conclusions.

² Community here is a ‘lazy’ term for society.

³ Scientific and non-scientific.

⁴ quantification of changes in the environment and in the recipient's ecosystem based on a reference scenario.

⁵ The potential for adverse consequences for human or ecological systems, recognising the diversity of values and objectives associated with such systems.

The first goal of this work is to promote a more meaningful use of ML techniques for Ecological Niche Modelling (ML-ENMs) to predict the current and future spread of AIS.

The novel framework is based on and builds upon three theoretical concepts:

- *Data can inform, but not explain*

ML is currently excellent for identifying complex patterns and correlations but lacks interpretability and causality. In this research, ML is used as a hypothesis generator. The centrepiece of the research is the development of a protocol for post-hoc (exploratory) hypothesis testing.

- *Machine learns what we teach it*

ML models operate within the “universe” defined by the context-rich data provided. The contextual richness and diversity of the data shape the model's capabilities, biases and limitations. Context is the most trivial aspect of ML as it determines both the understanding and performance of the model, making it a “*wicked concept*”⁶ (Kutor et al., 2021).

- *Risk decision making approach*

The scope is to identify, assess, and prioritise potential risks in the ML-ENM process to inform and guide modelling decisions and actions. The main purpose is to develop strategies to maximise the validity⁷ of the context and mitigate the impact of bias in the data. There is no right or wrong, only better or worse.

⁶ An idea that is inherently complex, difficult to define universally, and continuously evolving. It resists clear-cut answers and incorporates diverse, often contradictory elements that prevent it from being resolved or universally understood. This concept is applied to areas with high ambiguity and multidimensionality, such as wisdom, where ongoing debates and reinterpretations shape its understanding.

⁷ Degree of informativeness with respect of the process being modelled.

The second goal is to extend the framework with a component that makes it possible to *simulate* the human-mediated invasion process. Mechanistic approaches such as spatial interaction models are better suited for this purpose. They can be used to investigate invasion dynamics by modelling both the species dispersal mechanism and the pathway of invasion.

Spatial interaction models are based on the premise that the interaction between two locations is influenced by three key factors:

- *The properties of the origin location*
- *The properties of the destination location*
- *The distance between locations*

Often used to study transport and migration, they can model the likelihood of an AIS spreading from one area to another, taking into account factors such as the intensity of shipping traffic (a major vector for marine invasions), habitat and substrate suitability.

By integrating spatial interaction into the framework, we can simulate the process of invasion and determine not only where an AIS could potentially establish itself, but also how and when it could reach these areas. This allows for a more comprehensive risk assessment that includes both the potential for establishment and the mechanisms of spread.

The structure of the thesis follows the logical order of the research journey.

The first chapter sets out the background and basic concepts relevant to the AIS risk assessment:

It begins with a characterisation of the invasion process and a description of the species selected for the case study: *Caulerpa cylindracea*. The study area, the Mediterranean Sea, and its oceanographic and ecological features are presented, and a general overview of the

region-specific AIS knowledge is given. The chapter ends with a synthesis of the legal framework for AIS risk assessment as provided by the MSFD.

The second chapter defines the theoretical foundations and modelling components of the framework. Ecological niche models and spatial interaction models are presented from a general perspective. Then the specific models adopted in this framework, as well as data types used for the assessment, are thoroughly explained and contextualised. Both models, *MaxEnt* and the *production-constrained gravity model* are presented from a mathematical and ecological perspective.

The third chapter is dedicated to the Mediterranean Sea and the bioclimatic dataset created for the purpose of this study. Dataset creation led to the publication of a data article on *Data in Brief Journal* and it is proposed here in its published form. It presents the dataset and describes the procedure adopted to produce it, providing a template script to reproduce the analysis.

The fourth chapter contains the contributions to improving the reliability of ENMs. These advances include the use of a novel weighting system to reduce spatial bias and ensure more ecologically meaningful evaluation metrics. In addition, a multi-criteria decision-making process is proposed to provide an objective ranking of model performance. The methodological innovations and potential applications are presented using *Caulerpa cylindracea* as case study and *MaxEnt*. This chapter is the centrepiece of the research in this thesis. It led to an article that is currently under revision but is presented here in an extended version to provide a step-by-step critical analysis of the theoretical ML-ENM process, the proposed new methods and the challenges of practical implementation. Results obtained for *C. cylindracea* are presented and interpreted along the chapter.

The fifth chapter presents the improved *production-constrained gravity model* and describes its role in simulating vessel-mediated probability of invasion for the Mediterranean Sea. Using known presence sites as origins and the number of vessels crossing each origin as the *production-constrain*, the model estimates the number of expected invasions. Model behaviour, important variables and hotspots of invasion are analysed, as well as a characterisation of the invasion probability per substrate type.

The thesis ends with a conclusion and final remarks chapter, that summarise the findings, insights and future directions of framework development.

Introduction

This chapter presents the basic concepts for assessing AIS risk, beginning with an overview of the invasion process and a profile of *Caulerpa cylindracea*, the invasive species under study. The unique oceanographic and ecological characteristics of the Mediterranean are then described. An overview of the AIS phenomenon in the Mediterranean, highlighting most dangerous species, impacts and pathways is provided. The chapter concludes with a summary of the legal framework for AIS risk assessment under the MSFD.

Alien Invasive Species

“Non-Indigenous Species (NIS)⁸ that display invasive characteristics and threaten or adversely impact biodiversity and related ecosystem services” (Regulation 1143/2014/EU, 2014).

Alien Invasive Species (AIS) have emerged as one of the greatest threats to the biodiversity and the ecological (and economical) balance of our seas (Haubrock et al., 2021).

An alien species has the potential (high *risk*⁹) to spread uncontrollably and alter the functioning and services of the ecosystem at any level by directly (or indirectly) damaging one or more key species and services.

⁸ Non-indigenous species (NIS; synonyms: alien, exotic, non-native, allochthonous) are species, subspecies or lower taxa introduced outside of their natural range (past or present) and outside of their natural dispersal potential. This includes any part, gamete or propagule of such species that might survive and subsequently reproduce. Their presence in the given region is due to intentional or unintentional introduction resulting from human activities. Natural shifts in distribution ranges (e.g. due to climate change or dispersal by ocean currents) do not qualify a species as a NIS. However, secondary introductions of NIS from the area(s) of their first arrival could occur without human involvement due to spread by natural means.

⁹ The potential for adverse consequences for human or ecological systems, recognising the diversity of values and objectives associated with such systems (Regulation 1143/2014/EU, 2014).

The effects of invasions can be diverse and unpredictable (Katsanevakis et al., 2014, 2016; Mazza et al., 2014). Of particular concern are AIS that can be defined as transformers, i.e. invasive species that are ecosystem engineers (sensu Lawton, 1994) of different ecosystems. These species can upend existing communities and trophic relationships, alter biogeochemical cycles, reshape the habitat landscape and act as a catalyst for further invasions.

The assessment of biological invasions is one of the most pressing challenges in analysing cumulative impacts and risks (Vantarová et al., 2023). The risk of invasion is closely related to the extent to which the recipient ecosystem is impacted, and invasions have multiple impacts on the recipient region.

AIS risk assessment is an objective process of estimating *risk*, broadly defined as the product of the probability of events and the severity of their impacts (Leung et al., 2012).

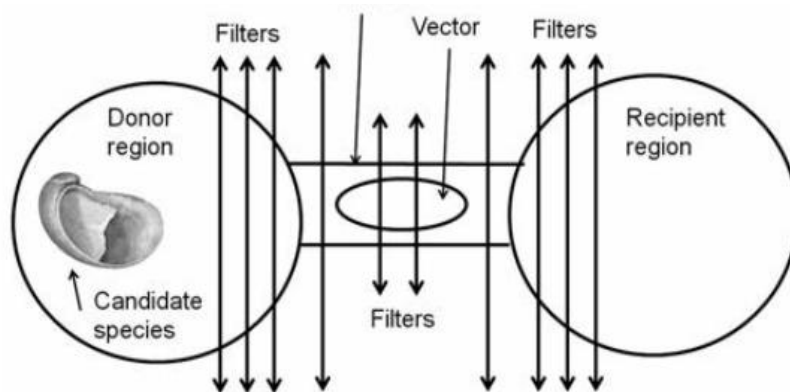
The invasion process

Even if a conspicuous number of NIS were deliberately introduced for cultural, economic or ecological reasons, most introductions are accidental.

The process of invasion is a complex and dynamic phenomenon characterised by a series of interconnected events influenced by both ecological and anthropogenic factors (Figure 1). It is in fact a continuum of complex interactions between the invading species and the recipient ecosystem, mediated by various biotic and abiotic factors (C. Boudouresque & Verlaque, 2012).

The first phase of this process involves the transport of a species from its native range (donor region) to a new environment (recipient region).

Figure 1 Schematic representation of the introduction process. Taken from Bouderesque & Verlaque. 2012)



During this journey, invading species encounter various filters - ecological, geographical or anthropogenic barriers that make it difficult for them to survive and colonise new sites. These filters act as selective barrier, allowing only those species or individuals with suitable traits to survive. Environmental conditions such as temperature, salinity and pH can act as abiotic filters, while biotic filters can include competition with native species, predators or the absence of necessary mutualistic partners. Human activities can also create, remove or alter these filters for example through pollution, habitat modification or climate change.

This movement is facilitated by vectors and corridors, which serve as means or pathways for the transfer of species.

After arriving in a new environment, the success of the introduction process depends on the species' ability to overcome these filters and establish a self-sustaining population, a phase known as *naturalisation*. This critical phase is influenced by several factors, including the number and genetic diversity of founding individuals (propagule pressure), the species' traits, and the ecological characteristics of the recipient ecosystem (Simberloff, 2009).

There is a consensus that healthy, resilient ecosystems are more resistant to biological invasions, as they often offer limited niche availability and have stronger defence mechanisms, while artificial or disturbed habitats are more susceptible to invasions.

If *naturalisation* is successful, the species can enter an *expansion* phase characterised by population growth and spread to adjacent areas. This phase is often characterised by rapid increases in abundance and range expansion as the species exploits available resources and potentially displaces native species. The speed and extent of this spread can vary greatly depending on the characteristics of the species, the characteristics of the invaded ecosystem and external factors such as disturbance regimes and climate change.

This is followed by the phase of *persistence*, in which the population of the invasive species integrates into the new ecosystem and maintains a self-sustaining presence over time.

Invasion kinetics

The long-term persistence of an invasive species can have profound ecological and evolutionary impacts on both the invading organism and the recipient ecosystem. These effects are largely determined by the kinetics of invasion, a concept related to chemical kinetics in that it focuses on dynamic processes. Specifically, invasion kinetics analyses the growth rate of invasive populations, the dynamics of their spread, their dependence on initial conditions and the existence of equilibrium states. Two major models of invasion kinetics have been proposed to describe the long-term dynamics of invasive species populations (C. Boudouresque & Verlaque, 2012).

- "*Natural fluctuation*" model (Figure 2a). Characterised by an initial expansion phase followed by a persistence phase in which the invader's abundance shows a plateau-like behaviour with more or less abrupt fluctuations around an equilibrium level (C. F.

Boudouresque, 1999). According to this model, the invader integrates into the recipient ecosystem and often achieves dominance but coexists with native species.

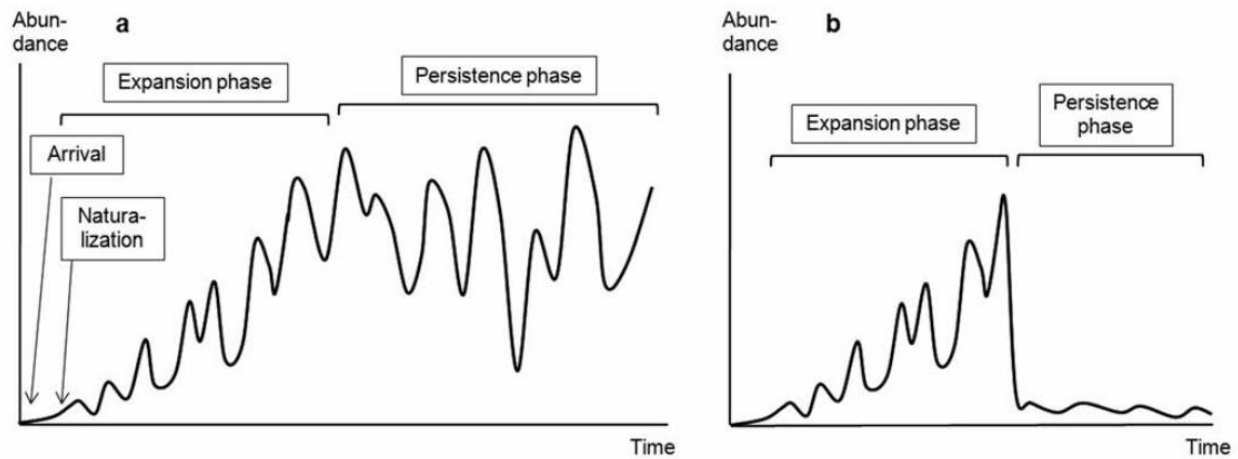
Fluctuations in abundance are similar to those of native species and are determined by predator-prey relationships, parasite-host interactions, recruitment variability, and climatic episodes. This model implies lasting ecological impacts on the invaded community, including potential displacement of native species, habitat modification and altered ecosystem processes. Over time, both invasive and native species may undergo evolutionary changes as they adapt to the new biotic interactions.

- "*Boom-and-bust*" model (Figure 2b). Characterised by a dramatic population explosion during the expansion phase, followed by a rapid decline to a low level of abundance (Simberloff & Gibbons, 2004). This collapse can be triggered by various mechanisms: (1) overexploitation of resources by the invasive species, (2) accumulation of natural enemies or pathogens, (3) loss of genetic diversity in the founder population, or (4) environmental changes that reduce habitat suitability. The boom phase often leads to an acute, severe disruption of the invaded ecosystem. However, the subsequent bust phase may allow the community to recover, albeit in an altered state. In some cases, the invader may persist at a low level and possibly retain the capacity for future outbreaks.

These contrasting models have important implications for both the trajectory of invasion and its consequences for ecosystems. The natural fluctuation model suggests more predictable, chronic effects and the potential for long-term coexistence and co-evolution between invaders and native species. The *boom-and-bust* model implies less predictable dynamics with the possibility of transient but intense ecosystem impacts followed by at least partial

recovery.

Figure 2 a) "Natural fluctuation" model; b) "Boom and bust" model. Taken from Bouderesque & Verlaque (2012).



Understanding which model applies to a given invasion is crucial for forecasting its long-term effects and developing appropriate management strategies.

Corridors, vectors, pathways

The most important factor for an invasion is the presence of a corridor and something that acts as a vector (Lieurance et al., 2023).

A "corridor" in the context of AIS refers to any human-related route that facilitates or enables the movement and dispersal of organisms between ecologically unconnected areas. The term can be meant both geographically (e.g. Suez Canal) and conceptually (e.g. "shipping route", "trade route").

A "vector" is any means that directly carries and transmits an invasive specimen from one place to another. Vectors include various means of transport such as ships, aeroplanes and vehicles, as well as the goods they carry, e.g. agricultural produce, landscaping materials and ornamental plants. Vectors play a crucial role in the initial introduction of invasive species

into new environments, as they not only transport the invasive species, but often allow them to overcome physical, chemical and biological filters (Saul et al., 2017).

In the context of AIS risk assessment, it is common to refer to the concept of *pathways* of introduction.

Pathways describe the processes and routes by which invasive species are transported from their native range to a new location. These pathways can be intentional or unintentional and often involve human activities that facilitate the movement of species (Hulme et al., 2008).

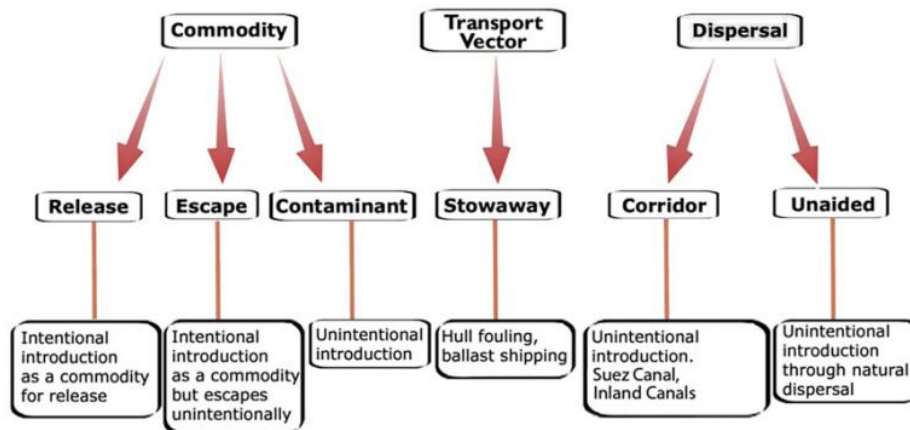
See Figure 3 for a visual representation of the pathway's categorisation.

They can be broadly categorised as follows:

- *Release*: the intentional release of an organism into the wild, e.g. for hunting or fishing.
- *Escape*: When organisms unintentionally escape from captivity or confinement, such as pets or plants from gardens and aquaculture.
- *Contamination*: Organisms that arrive as unintended passengers with imported goods, such as insects or seeds that have been mixed with agricultural products.
- *Stowaway*: Organisms travelling on means of transport without the intention of humans to transport them, such as organisms in ballast water or on ship hulls.
- *Corridor*: Routes that facilitate the dispersal of species across landscapes and are often enhanced by human intervention such as canals or roads.
- *Unsupported*: Natural spread from one area to another without direct human intervention, but often following a primary human-mediated introduction.

It is crucial to assess and describe the introduction pathways as they strongly determine species spread in the recipient region.

Figure 3 Simplified framework to categorise pathways of introduction of AIS. Taken from Hulme et al., (2008).



Caulerpa cylindracea as a case study

Caulerpa cylindracea (Sonder, 1845), originally from south-west Australia, was first discovered in the Mediterranean off the coast of Libya in 1990 (Nizamuddin, 1991).

In less than 20 years, *C. cylindracea* colonised almost every coast of the basin and even reached the North Atlantic (Klein & Verlaque, 2008), becoming the most widespread invasive species in the Mediterranean (Katsanevakis et al., 2016; Streftaris & Zenetos, 2006).

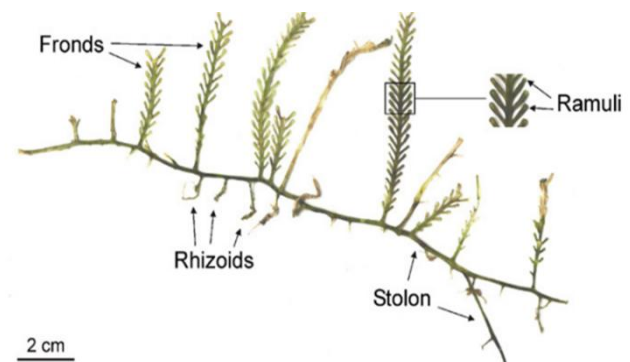
Physiological Characteristics

Caulerpa cylindracea exhibits a unique uniaxial siphonous thallus structure (Figure 4), which is predominantly divided into a creeping axis (stolon) and erect shoots known as fronds (Verlaque et al., 2003). These fronds can grow up to 19 cm high and are characterized by radially or distichously arranged vesiculate branchlets that facilitate nutrient absorption and growth (Klein & Verlaque, 2008).

The thallus's morphology, combined with its ability to produce thin rhizoids for substrate anchorage, enables the algae to establish itself firmly even in varying marine substrates, from rocky bottoms to seagrass meadows and artificial structures (Piazzi et al., 2016).

A key aspect of its physiology is its remarkable growth rate and its high plasticity in photoacclimation. The alga can adjust its photosynthetic activity to low-light conditions, which allows it to colonize deep subtidal zones as well as surface environments exposed to intense sunlight (Bernardeau-Esteller et al., 2015).

Figure 4 Thallus of the invasive *Caulerpa cylindracea* from the Gulf of Marseille (-30 m). Herbarium specimen, J. Klein (Klein & Verlaque, 2008).



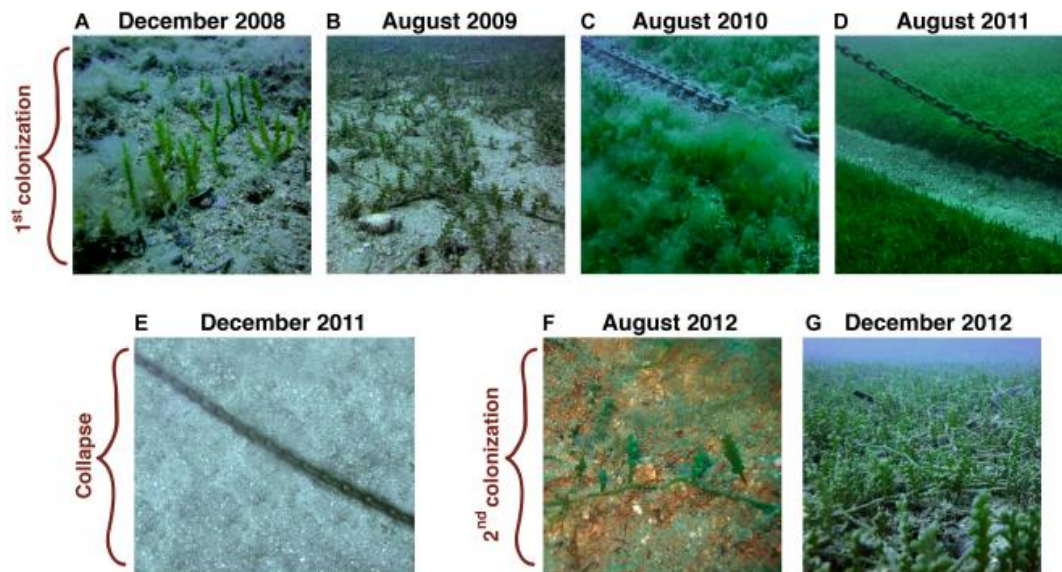
Biological Traits

The invasive success of *C. cylindracea* in the Mediterranean is primarily due to its impressive biological properties. It can spread rapidly by fragmentation, with small parts able to establish new colonies on any kind of substrate (Bulleri et al., 2018). Its stolons can elongate up to 2 cm per day, allowing rapid colonisation of large areas and the formation of dense mats (Piazzi et al., 2016). See Figure 5 for the time lapse of an invaded site at Vilanova i la Geltrú (Spain), taken from Santamaria et al. (2021).

Its exceptional morphological plasticity allows it to thrive in different environments and even transform into a resistant form under stress, enabling it to survive for long periods in unfavourable conditions (Santamaría et al., 2021). Furthermore, thanks to the remarkable ability to absorb and store nutrients, the alga can thrive in both eutrophic and nutrient-poor environments (Gennaro & Piazzi, 2014).

Finally, *C. cylindracea* shows a high resistance to herbivory, thanks to strong defence mechanisms (production of toxic alkaloids, e.g. *caulerpin*) and rapid regeneration of damaged tissue (Bulleri & Malquori, 2015).

Figure 5 Evolution of the abundance of *Caulerpa cylindracea* over a period of 4 years in a population under study, close to Vilanova i la Geltrú (Spain). Taken from Santamaria et al. (2021)



Ecological Impacts and Interactions

The ecological impacts of *C. cylindracea* on invaded ecosystems are profound. The species actively competes with native canopy-forming macroalgae and seagrasses such as *Posidonia oceanica*, altering marine community structure, trophic relationships and favouring the introduction of new AIS (Alomar et al., 2016; Bachir Bouiadjra et al., 2021; Bulleri et al., 2010).

It reduces both alpha and beta diversity by forming dense, monospecific mats that prevent the recovery of native populations, leading to a decline in species richness and habitat complexity (Piazzi et al., 2016).

Its preference for urbanised coasts and disturbed habitats is well documented. In these environments, the alga benefits from reduced competition with native species that are less adapted to the altered conditions, exerting further pressure. (Houngnandan et al., 2022).

A model species for biological invasion studies

Caulerpa cylindracea can be considered as a model species for biological invasion studies based on presence-only data and machine learning.

Due to its long invasion history, morphological characteristics and extensive experimental research, the available data basis and knowledge of the species closely match the assumptions required for modelling.

Firstly, the long invasion history and wide distribution of *C. cylindracea* in the Mediterranean indicate that *the species is in equilibrium in this region*. This means that the available occurrences reflect most, if not all, of the areas where the species is present as the species has likely saturated its potential distribution range (Jarnevich et al., 2015).

Thanks to the great attention dedicated to the species and numerous professional and citizen science monitoring projects, *C. cylindracea* is probably *one of the most sampled species* in the basin (Piazzi et al., 2016; Turicchia et al., 2021).

In addition, the distinct and easily recognisable morphology of *C. cylindracea* greatly *reduces the risk of misidentification* during field surveys. This ensures that the data used for modelling is robust and reliable (Wang & Jackson, 2023).

Finally, the extensive experimental research conducted on *C. cylindracea* in both controlled mesocosms and in situ experiments provides a *solid knowledge of the ecophysiology and biology* of the species and provides *insights to validate* the model predictions (Bulleri et al., 2011; Santamaría et al., 2021).

The Mediterranean Sea

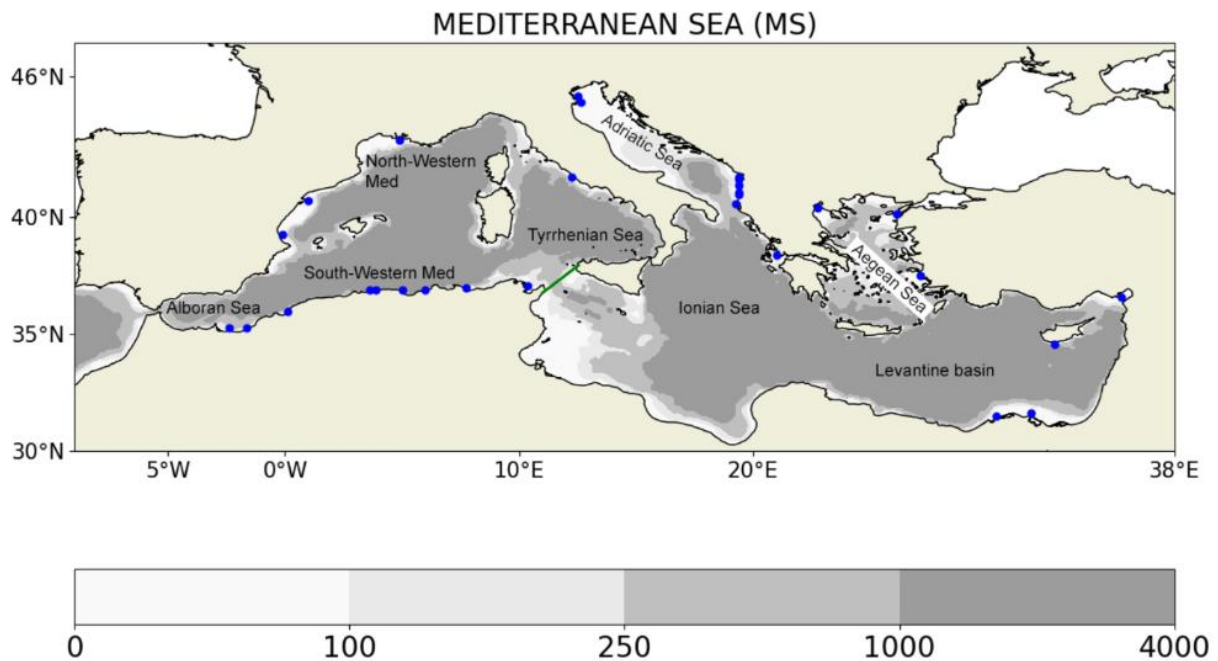
The Mediterranean Sea, a semi-enclosed basin bordering three continents, is a region of immense ecological, historical and geopolitical importance (Lo Brutto, 2021). Its unique geographical location and complex oceanographic features have favoured a rich biodiversity that has made it one of the world's most important hotspots of marine biodiversity (Coll et al., 2010). However, this ecological richness is increasingly threatened by human activities, climate change and the spread of alien invasive species (AIS)(Canu et al., 2022; Solidoro et al., 2013; Streftaris & Zenetos, 2006). The introduction and spread of AIS disrupt native ecosystems, alter habitats and pose significant socio-economic challenges (Katsanevakis et al., 2014; Zenetos & Galanidi, 2020).

This chapter provides a comprehensive overview of the ecological characteristics of the Mediterranean, key human activities and their impacts, the pressing problem of AIS and the legal framework that exists to address these challenges. It lays the foundation for the need for high quality, ready-to-use data to support risk assessments and predictive modelling efforts.

Geological History and Oceanographic Features

The Mediterranean Sea is the remnant of the ancient Tethys Sea with a rich and complex geological history shaped by tectonic processes. In the Mesozoic era, rifting, spreading and eventual closure of the Tethys Ocean occurred, driven by the collision of the African and Eurasian tectonic plates, creating the modern Mediterranean basin (Picotti et al., 2014). Major geological events such as the Messinian Salinity Crisis (ca. 5.96 to 5.33 million years ago), during which the sea almost completely dried up before being replenished by the Strait of Gibraltar, drastically altered sedimentation patterns and biodiversity (Krijgsman et al., 1999).

Figure 6 Mediterranean Sea bathymetry. Blue dots indicate major rivers and green line separates Western and Eastern basin. Taken from Solidoro et al. (2022).



Today, the Mediterranean is a semi-enclosed basin in which a unique combination of oceanographic processes takes place, leading to a high degree of endemism (Agiadi et al., 2024). The separation from the Atlantic Ocean has led to the isolation of marine species and contributed to the region being considered a biodiversity hotspot (Coll et al., 2010).

The Mediterranean is characterised by highly oligotrophic conditions, especially in its central and eastern basins, which are among the most nutrient-poor marine regions in the world (Agusti et al., 2017; Di Biagio et al., 2022). The thermohaline circulation is driven by opposing salinity and temperature gradients between the surface and deep waters (Menna et al., 2021).

This dynamic system enables the formation of deep-water masses such as the Eastern Mediterranean Deep Water (EMDW) and the Western Mediterranean Deep Water (WMDW), which play a crucial role in nutrient cycling and carbon sequestration (Solidoro et al., 2022).

Seasonal stratification has a significant influence on nutrient distribution. In productive coastal areas, such as *Posidonia oceanica* meadows, the availability of nutrients supports

high primary production (Gobert et al., 2002). In contrast, deep-sea environments such as seamounts and submarine canyons are characterised by lower nutrient levels and host unique communities adapted to these oligotrophic conditions (Bonanno et al., 2018). These habitats are further impacted by climate change, as the most recent projections indicate increasing ocean warming and acidification, particularly in the eastern Mediterranean (Solidoro et al., 2022; Zunino et al., 2019).

Biodiversity and Ecological Richness

Although the Mediterranean represents less than 1% of the global marine area, it harbours between 4% and 18% of the world's marine species, of which about 25% are endemic (Mannino et al., 2017). Its diverse habitats include coastal wetlands, estuaries that are important for migratory birds, and deep-sea hydrothermal vents that harbour unique organisms (Coll et al., 2011). Iconic species such as the Mediterranean monk seal (*Monachus monachus*) and various cetaceans, including the short-beaked common dolphin (*Delphinus delphis*), are an integral part of the region's biodiversity and cultural heritage (Gabrie et al., 2012).

The *Posidonia oceanica* meadows, often referred to as the "lungs of the Mediterranean," play a vital role in carbon sequestration and are important nurseries for numerous fish species (Martínez-Abraín et al., 2022). These habitats are critical for maintaining the ecological balance and supporting the complex marine food web (Campagne et al., 2015).

Human Impacts on the Mediterranean

The Mediterranean Sea has been heavily impacted by human activities in the past as well as today (Canu et al., 2022). Overfishing has drastically reduced populations of commercially important species, leading to the collapse of some fisheries and causing significant

ecological shifts (Fiorentino & Vitale, 2021; Micheli et al., 2013). Coastal development driven by rapid urbanisation and tourism has led to habitat destruction, particularly in sensitive areas such as wetlands and seagrass meadows (Hidalgo et al., 2022; Newton et al., 2014).

The construction of the Suez Canal in 1869 created a direct link between the Red Sea and the Mediterranean and changed the biogeography of the region by facilitating the introduction of non-native species — a phenomenon known as lessepsian migration (Galil, 2023).

Climate change is exacerbating these existing pressures by causing sea level rise, ocean acidification and increasing sea surface temperatures. These changes lead to species shifts, alter reproductive cycles and can lead to mass extinctions of vulnerable marine populations (Bednarsek et al., 2023; Lejeusne et al., 2010; Solidoro et al., 2013).

Alien Invasive Species in the Mediterranean Sea

The introduction of AIS has become one of the most pressing environmental problems in the Mediterranean (Katsanevakis et al., 2013). These species disrupt food webs, alter habitats and displace native species, leading to significant ecological and socio-economic impacts (Goren et al., 2016). AIS in the Mediterranean are diverse and range from macroalgae to invertebrates and fishes. Many of these species are thermophilic and thrive in the warming waters of the Mediterranean, especially as sea surface temperatures rise due to climate change (Galil et al., 2018).

The eastern Mediterranean has become a gateway for numerous species from the Red Sea through the Suez Canal. In addition, vectors such as shipping, aquaculture and recreational boating contribute to the introduction and spread of AIS, which represents a significant ecological pressure on the native ecosystems (Carreño & Lloret, 2021).

Vectors and Pathways of Introduction

The Suez Canal (Lessepsian Migration)

The Suez Canal is one of the most important pathways of AIS introduction. Since its opening, over 300 species have migrated from the Red Sea, many of which have established permanent populations (Galil et al., 2018). Lessepsian species, often thermophilic find the increasingly warmer conditions in the Mediterranean favourable, allowing them to expand their range. Notable invaders include the rabbitfishes *Siganus luridus* and *Siganus rivulatus*, which have impacted local algal communities and competed with native herbivorous fishes (Galil, 2023).

Shipping Activities

Shipping is another critical vector for the introduction of AIS in the Mediterranean. The busy maritime traffic in the Mediterranean, including the major shipping routes between Europe, Asia and Africa, increases the vulnerability of the region to bioinvasions despite increasing safety protocols (Baquero et al., 2021; Seebens et al., 2013).

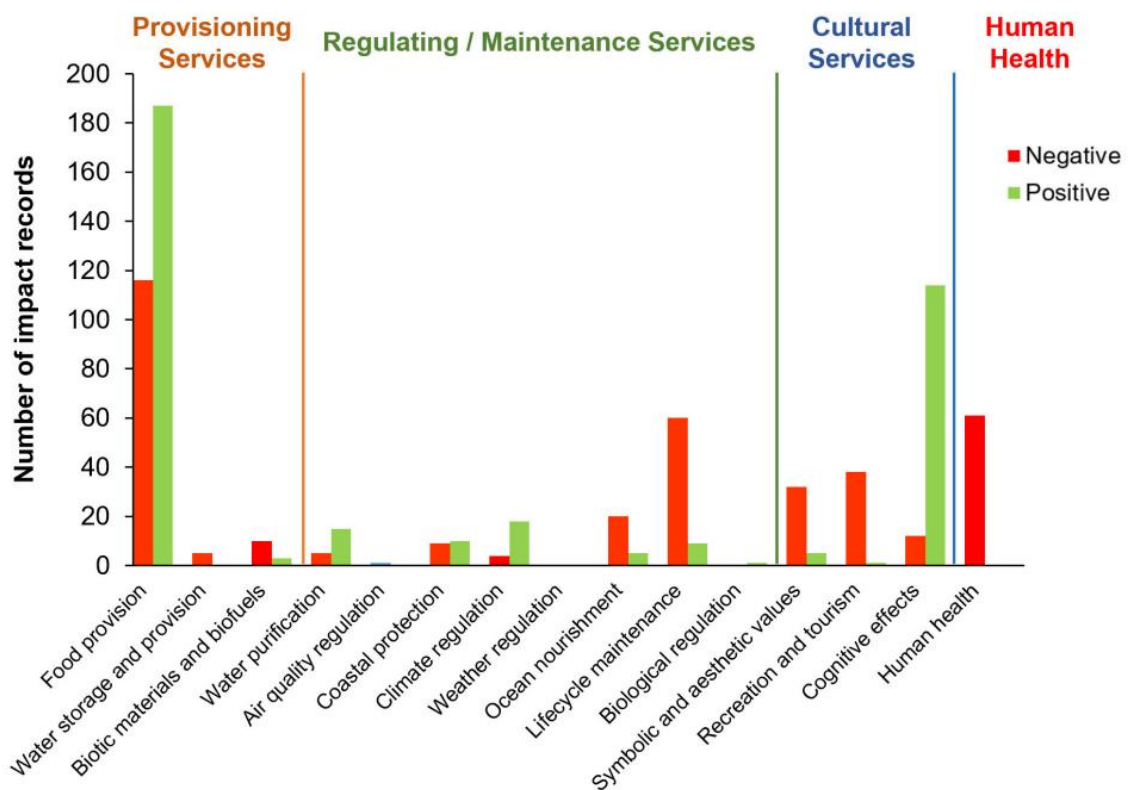
Of particular concern is that recreational boating and tourism, which bring millions of visitors to the coasts of the Mediterranean every year, also play a role in the introduction of AIS (Ashton et al., 2022). Small vessels can transport invasive species, particularly through hull fouling and the accidental transport of organisms. The high volume of recreational traffic, especially during the tourist season, increases the likelihood of AIS spreading to different parts of the Mediterranean. *Caulerpa cylindracea*, which has rapidly colonised the Mediterranean, is a notable example of the potential impact of such vectors (Cantasano et al., 2017).

Aquaculture and Fisheries

Aquaculture, a fast-growing industry in the Mediterranean, has inadvertently contributed to the introduction of AIS. Non-native species introduced for farming, such as the Pacific oyster (*Crassostrea gigas*), have escaped into the wild. In addition, associated species, including parasites and pathogens, can be accidentally introduced during aquaculture operations (Herbert et al., 2016).

Fisheries have also played a role, especially when non-native species are deliberately introduced to augment fish stocks or accidentally released. Such introductions can have unforeseen ecological consequences, such as competition with or predation on native species, disrupting the equilibrium of marine ecosystems (Tarkan et al., 2021).

Figure 7 Overview of AIS negative and positive impacts on ecosystem services and human health for the Mediterranean Sea. Taken from Tsirintanis et al. (2022).



Impacts of AIS

Ecological Impacts

AIS can severely disrupt native ecosystems by competing with native species for resources, altering habitat structures and introducing new predators (Tsirintanis et al., 2022). *Caulerpa cylindracea* displaces native seagrass species such as *Posidonia oceanica* (Piazzi et al., 2016) provoking a cascading effect on biodiversity, as these habitats provide shelter and nursery areas for numerous marine species (Bachir Bouiadjra et al., 2021).

The comb jellyfish *Mnemiopsis leidyi* is another notorious AIS that has disrupted local food webs by preying on zooplankton, a crucial food source for fish larvae and other marine organisms. This predation leads to a decline in fish populations and affects the overall ecological balance (Piccardi et al., 2024).

Socio-Economic Impacts

The presence of AIS poses a major economic challenge, particularly for industries dependent on marine resources. The decline in fish stocks due to competition and predation by AIS has a direct impact on the profitability of fisheries and threatens the livelihoods of coastal communities. The tourism sector is also at risk: AIS, in particular invasive algae, can completely alter the coastal landscape (Haubrock et al., 2021; Marbuah et al., 2014).

Finally, AIS can pose a risk to public health (Figure 7). Invasive jellyfish and poisonous fishes increase the frequency of stings, which can lead to serious injuries and put additional strain on local health systems. The spread of harmful algal blooms associated with some AIS can lead to an accumulation of toxins in seafood, posing a risk to human consumers (Bédry et al., 2021).

The Marine Strategy Framework Directive (MSFD)

The Marine Strategy Framework Directive (MSFD), which was introduced by the European Union in 2008, is a pivotal legislative framework for the protection of the marine environment in Europe (Directive 2008/56/EC). Its primary objective is to achieve good environmental status (GES) of EU marine waters, maintain the sustainability of marine ecosystems and at the same time balance the socio-economic activities dependent on marine resources (Commission Decision (EU) 2017/848). The GES is defined by 11 descriptors, including biodiversity, fisheries, eutrophication and in particular the prevention of the introduction of non-indigenous species (NIS), including AIS.

The MSFD emphasises an ecosystem-based management approach that combines environmental protection with sustainable use. It mandates EU Member States to develop and implement marine strategies aimed at assessing, monitoring and achieving the GES. These strategies are regularly updated to address new challenges such as climate change and the increase in human activities.

MSFD Focus on Alien Invasive Species

The MSFD specifically addresses AIS under Descriptor 2 (D2), which focuses on the management of NIS to prevent adverse effects on marine ecosystems (Regulation (EU) No 1143/2014; Commission Decision (EU) 2017/848). D2 outlines specific criteria:

- *Primary criterion (D2C1):* The number of new NIS introduced by human activities must be minimised and ideally reduced to zero. Member States are responsible for setting thresholds for the introduction of new introductions, in coordination at regional or sub-regional level.

- *Secondary criteria (D2C2 and D2C3):* These assess the abundance and distribution of established NIS and their impacts. D2C2 measures the spatial distribution and population sizes of established NIS, especially invasive species, while D2C3 analyses the proportion of ecosystems or species groups that are negatively affected by these invasions.

To achieve the objectives of the MSFD in relation to AIS, Member States must develop comprehensive marine strategies that include:

- *Monitoring and reporting:*

Ongoing monitoring of NIS to detect new introductions, assess the abundance and distribution of established species and monitor ecological impacts. The data collected must be reported at national and EU level to enable coordinated regional strategies.

- *Thresholds setting:*

Establish thresholds for the introduction of NIS in each six-year assessment period in coordination with neighbouring countries. These thresholds will serve as benchmarks for assessing progress towards GES under Descriptor 2.

- *Implementation of preventive and management measures:*

Development and implementation of measures to prevent the introduction of new AIS, including the control of introduction pathways such as ballast water from shipping and aquaculture. Strategies to manage established AIS populations are also needed to mitigate their impact on marine ecosystems.

Strategies and Measures to Combat AIS

Prevention

Prevention is a key focus of the MSFD's approach to combating AIS, which aims to reduce human introduction through strict controls on the main pathways of introduction. This includes the implementation of the International Convention for the Control and Management of Ships' Ballast Water and Sediments, which mandates treatment of ballast water to remove or destroy organisms. In addition, biosecurity protocols and public awareness campaigns aim to limit accidental releases from aquaculture and tourism.

Early Detection and Rapid Response

Early detection and rapid response mechanisms are critical components. Member States must have robust monitoring systems in place to promptly detect new NIS introductions and take immediate action to control or eradicate species before they become established and cause significant damage.

Challenges from Climate Change and Human Activities

The MSFD recognises the exacerbating effects of climate change on the spread of AIS. Rising sea temperatures, ocean acidification and changing current patterns facilitate the spread of species to new areas, including the Mediterranean. The Directive emphasises the inclusion of climate change projections in marine management strategies and highlights proactive measures to prevent climate-induced displacement of NIS.

Increasing human activities, such as shipping, tourism and resource extraction, exacerbate the challenges of AIS management. The MSFD calls for increased cooperation between countries and sectors and emphasises that an integrated, transboundary approach is crucial for successful AIS management.

Toward an integrated modelling framework

Predicting the current and future spread of Alien Invasive Species (AIS) in the Mediterranean Sea is a major challenge due to the complex interplay of ecological factors and human activities. Conventional approaches often rely on isolated assessments— focussing either on habitat suitability models or on the risk of introduction through shipping. There is an urgent need to develop holistic and transferable frameworks that assess the cumulative impacts of climate change and human activities on the spread of AIS based on limited data (Pyšek et al., 2020).

This chapter provides the theoretical basis for the framework, defines the intended outcomes and provides the general context and the specific formulation for each of the modelling techniques implemented.

Theoretical foundation

This research represents a step towards a more comprehensive AIS risk assessment process by drawing direct inspiration from the Biotic-Abiotic-Movement (BAM) concept proposed by Soberon & Peterson (2005). Soberón and Peterson conceptualises the distribution of species as a function of three core components:

- *Biotic interactions (B)*: Relationships with other species, including competition, predation and mutualism, that may facilitate or hinder the establishment of populations.
- *Abiotic factors (A)*: Non-living environmental conditions such as climate and physical parameters that set the physiological limits for species survival.
- *Movement constraints (M)*: The accessibility of geographic regions to the species, which is influenced by dispersal opportunities and barriers to movement.

Building on this conceptual foundation, the proposed integrated modelling framework combines correlative Ecological Niche Models (ENMs) with mechanistic spatial interaction models to represent and solve the BAM components:

The correlative ENM addresses the A and, to some extent, the B component by modelling the ecological niche of the species, while the mechanistic model addresses the M component by modelling the invasion pathways and simulating vessel-mediated dispersal.

To ensure the coherence and realism of the assessment, the entire framework must be based on a unique and comprehensive dataset tailored to the specific relevant aspects of the species being assessed, including climate change projections and human activities.

The intended outcomes are:

- 1) *Habitat suitability maps* expressed as probability of occurrence of the species per time unit, given the known sites in which the species is present.
- 2) *Invasion likelihood maps* expressed as estimated number of invasions per time unit.

Overview of Ecological Niche Modelling

Ecological Niche Models (ENMs) are mathematical representations that predict where species may exist based on their ecological niches—abstract spaces defined by environmental conditions that sustain a species (Peterson et al., 2011). An ecological niche is a multi-dimensional hypervolume that encompasses all possible conditions under which a species can survive and reproduce (Hutchinson, 1957). This niche concept is then mapped onto geographic space, reflecting the complex interplay between species and their environments, guided by both ecological and geographical principles.

ENMs are conceptual frameworks that bridges ecological theory and spatial modeling, to provide insights into the spatial dynamics of species' habitats (Franklin, 2010).

The predictive power of ENMs is crucial in forecasting species' responses to future climatic changes, making them indispensable for ecological risk analysis and the assessment of invasive species (AIS) threats.

Mechanistic and Correlative Approaches

ENM can be roughly divided into two categories: mechanistic models and correlative models. Although these approaches differ in their methods and principles, they both aim to estimate the niches of species and predict their distribution under different environmental conditions (Peterson et al., 2015).

Mechanistic Models

Mechanistic models, also known as process-based models, focus on defining a species' fundamental niche- the full range of environmental conditions (e.g. temperature, salinity, nutrient concentration and light availability) under which a species can survive, grow and reproduce in the absence of limiting factors such as competition and predation. These models establish a direct link between environmental conditions and the physiological and biophysical processes that regulate survival and reproduction (Kearney & Porter, 2004).

One of the key strengths of mechanistic models is that their ability does not depend on geographic location. This predictive ability is particularly valuable in scenarios involving climate change or the introduction of non-native species into marine systems, as the models can estimate potentially suitable areas based on physiological principles (Evans et al., 2015).

Despite their precision, mechanistic models reach their limits due to their high data requirements and the computational complexity needed for accurate parameterisation. For many species, and for AIS this is often the case, as detailed physiological data may be lacking. And even when such data are available, the sensitivity of these models to small variations in parameter values can significantly affect their results. (Riddell et al., 2023)

Correlative Models

Correlative models, also known as empirical or statistical models, rely on associations between species occurrence data and environmental variables to infer the realised niche of a species (Peterson et al., 2015). The realised niche represents the subset of the fundamental niche that a species actually occupies, taking into account not only the abiotic conditions but also the constraints imposed by biotic interactions (e.g. competition, predation) and the dispersal limitations of the species (Hutchinson, 1957). In other words: While the fundamental niche describes where a species could live under ideal conditions, the realised niche reflects where it actually lives due to the complex interplay of ecological and geographical factors.

These models rely on statistical associations between the observed occurrence of species and environmental variables to infer the conditions under which species persist (Elith et al., 2006). In the marine context, this could mean correlating species distribution data with variables such as sea surface temperature, salinity, nutrient levels and ocean currents. By overlaying species occurrence data with these environmental layers, correlative models can identify suitable habitats and predict potential distribution patterns without prior knowledge of the species (Franklin, 2010).

Due to the complexity of the statistical relationships to be derived and the huge amounts of data that need to be analysed, the most common approach to this task is to implement Machine Learning or Artificial Intelligence algorithms (ML-ENMs) for pattern recognition (Beery et al., 2021; Phillips et al., 2006).

However, their main limitations arise from several key assumptions: They assume that the current distribution is in equilibrium with existing environmental conditions and that occurrence data fully represent the species' ecological range. In particular, these assumptions are often not fulfilled in the case of AIS, as only distorted and incomplete information is usually available (Jarnevich et al., 2015).

One of the main contributions of this work is the development of a conceptual framework (ENM process) specifically tailored to the study of the current and future diffusion of AIS and entirely aimed at mitigating the impact of bias and distorted information on the correlative modelling process.

Species' Occurrence Data

Species occurrence data are a fundamental element in the study of biogeography and the modelling of ecological niche. They provide essential insights into the spatial and temporal distribution of species.

Species occurrences usually come from various sources such as field observations, museum specimens, contributions from citizen scientists and remote sensing technologies. The usefulness of these data is significantly influenced by their quality, which depends on factors such as spatial accuracy, temporal consistency and the thoroughness of sampling efforts (Peterson et al., 2011).

Occurrence data play a central role in both correlative and mechanistic ecological niche models (ENMs). Correlative models postulate their entire foundation to observed occurrences. Mechanistic models use occurrence data to calibrate model parameters (Evans et al., 2015).

Presence-Absence Data

Presence-absence data explicitly document where a species is found (presence) and where it was actively searched for but not found (absence). This dual aspect of the data provides a balanced view that is particularly valuable in ecological modelling, as it allows researchers to distinguish between suitable and unsuitable habitats (Peterson et al., 2011).

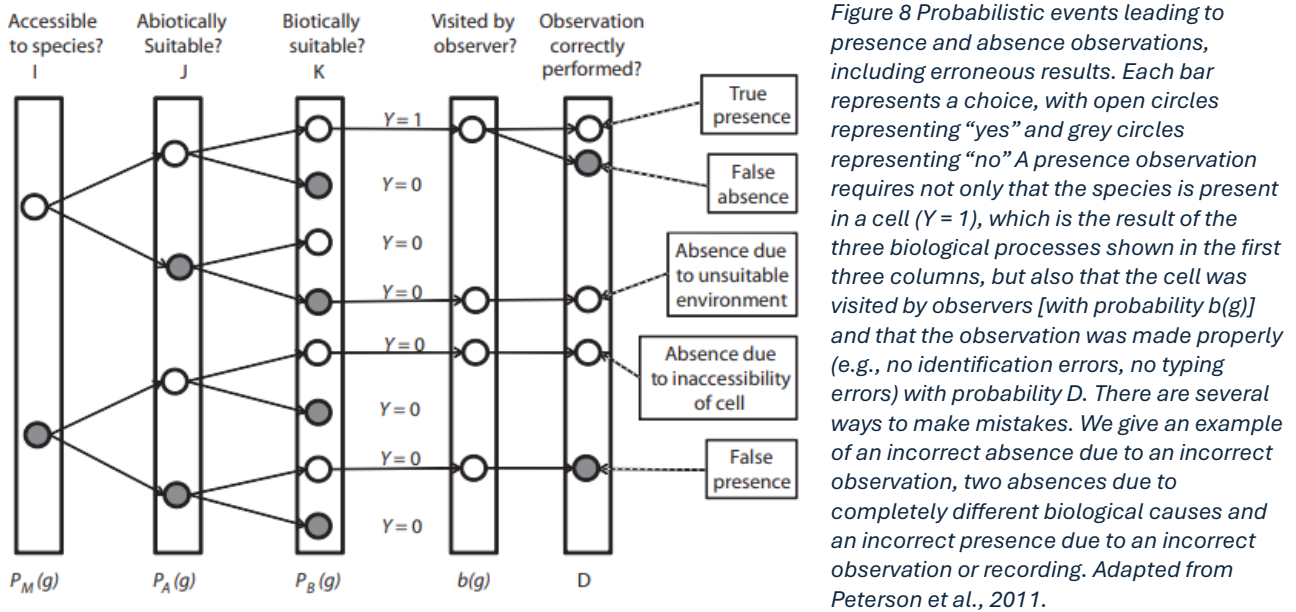
Modelling assumptions

The use of presence-absence data assumes that the absence data reflect unsuitable habitats, meaning that the species are not present due to unsuitable environmental conditions. See Figure 8 for a visualisation of the sampling process as a series of probabilistic events leading to a record. Another important assumption is that the sampling effort is the same at all sites, i.e. that there is an equal chance of detecting the species if it is present in each area sampled. This uniformity is important to avoid biases that could distort the species distribution models (Jarnevich et al., 2015; Peterson et al., 2011).

Limitations in interpretation

One of the main limitations of presence-absence data is the problem of false absences, where a species is present but not detected due to factors such as insufficient survey efforts or seasonal variation. This can lead to incorrect conclusions regarding the ecological preferences and distribution limits of the species. In addition, the absence of species may be due to factors unrelated to habitat suitability, such as recent local extinctions or barriers to

species dispersal, which complicates the interpretation of these data in modelling (Elith et al., 2020).



Presence-Only Data

Presence-only data consist of records indicating locations where a species has been observed without providing information on sites where the species has not been detected. This type of data usually comes from opportunistic observations, museum collections or reports from citizen scientists. Although not as informative as absence information, presence-only data are invaluable, especially when systematic surveys data are not available (Peterson et al., 2011).

Almost the totality of occurrence records available on AIS belongs to this class.

Modelling assumptions

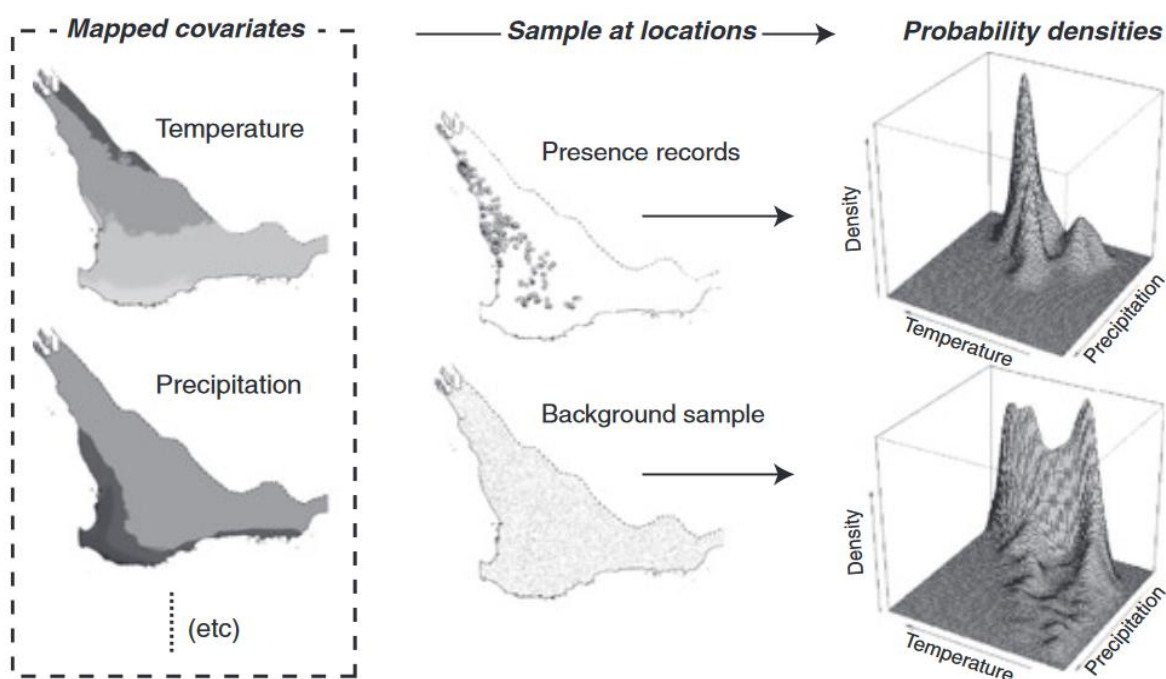
Models based on presence-only data make stronger assumptions than models that use presence-absence data. These models typically assume that the occurrences represent the environmental conditions suitable for the species. Since true absence data are not available,

these models employ background or pseudo-absence data to simulate areas where the species may not occur. Randomly selected background sites, work as contrasts for the presence point (Allouche et al., 2006; Coro et al., 2015).

Limitations in interpretation

The main drawback of presence-only data lies in the lack of direct absence records, which can lead to an overestimation of suitable habitat (Saupe et al., 2018). The challenge with this type of data is that it is difficult to distinguish between truly unsuitable areas and those for which data are simply not available or lacking (Yackulic et al., 2013). Presence-only data also tend to be biased toward more accessible or frequently surveyed locations, which can distort the perceived distribution of the species if this is not taken into account in the analysis (Fourcade et al., 2014). Careful statistical adjustments and critical interpretations are needed to mitigate these biases in species distribution models (Warren & Seifert, 2011).

Figure 9 A schematic representation of the probability densities. The maps on the left show two examples of mapped covariates. In the centre are the locations of the presence and background samples. The density estimates on the right are not in geographic space (map), but show the distributions of values in covariate space for the presence (top right) and background (bottom right) samples. These could represent the densities $f_1(z)$ and $f(z)$ for a simple model with linear characteristics. Taken from Elith et al., (2011).



MaxEnt

MaxEnt is a well-established machine learning algorithm for simulating the geographical distribution of species using presence-only data (Phillips & Dudík, 2008).

Considered as one of the best performing algorithm on presence-only data (Elith & Franklin, 2013), it is widely used and successful applications can be found in estimating AIS distribution (Karuppaiah et al., 2023; Sayit et al., 2019; A. M. West et al., 2016; Zhu et al., 2014).

It is based on the principle of maximum entropy, which is rooted in information theory. This principle states that, given limited information, the probability distribution that best represents the current state of knowledge is the one with maximum entropy, subject to the constraints imposed by the available data (Phillips et al., 2006).

How it works¹⁰

Ecological perspective

MaxEnt makes its predictions by comparing the environmental conditions at the locations where the species is known to occur with those of the entire landscape represented by the background data.

From a statistical perspective, *MaxEnt* estimates the relative suitability of each site by comparing the environmental conditions at the presence sites with those of the entire study area represented by the background data. Specifically, *MaxEnt* models the probability distribution of the environmental conditions under which the species occurs ($f_1(z)$) relative to

¹⁰ For a rigorous mathematical description of the algorithm, see Appendix A.

the probability distribution of the same conditions in the background ($f(z)$) (Elith et al., 2011).

See Figure 9 for a visual representation of $f_1(z)$ and $f(z)$ in a simple case.

This approach quantifies the extent to which the environmental conditions at the sites differ from the environmental conditions observed in the study area. By estimating this ratio, *MaxEnt* identifies which ones are disproportionately represented in the presence data compared to the background, highlighting areas with a higher likelihood of suitable habitat (Merow et al., 2013).

The comparison between the presence data and the background data allows *MaxEnt* to "learn" from the available information. In other words, it distinguishes between suitable and less suitable habitats by determining which environmental conditions are most strongly associated with the occurrence of the species compared to those generally present in the landscape (Elith & Franklin, 2013).

Machine Learning perspective

MaxEnt attempts to define an optimised formula that connects the probability of occurrence to environmental variables and (1) best explains the known presence patterns with (2) the least assumptions.

From a mathematical point of view, this translates in the definition of an *objective function*¹¹ to optimise. This function has two terms: *gain* and *regularisation*, a penalty term.

Gain corresponds to the task (1): it measures how much better the model predicts the species' distribution compared to a null model that assumes the species is equally likely to be found anywhere (a uniform distribution).

¹¹ The "numbers" that MaxEnt seeks to maximise during model training.

Regularisation (2) prevents the model from relying heavily on complex features unless their inclusion is justified by a significant improvement in fit (*gain*) by applying a penalty that increases with the complexity of the model.

The *feature classes* in *MaxEnt* determine the mathematical relationships between the environmental variables and the occurrence of the species. They define how the model can use the input variables to construct the probability distribution (Phillips & Dudík, 2008).

MaxEnt feature classes include:

- *Linear features*: Capture direct, proportional relationships between the variables and the occurrence of species.
- *Quadratic features*: Allow for parabolic relationships and model scenarios in which probability increases up to a certain point and then decreases.
- *Product functions*: Represent interactions between variables by multiplying them together.
- *Threshold functions*: Capture abrupt transients in species' suitability as a piecewise constant response (0 – 1) to environmental variables.
- *Hinge functions*: Expand threshold functions to include piecewise linear response to environmental variables.

Complex features involve higher order terms that allow the model to capture intricate patterns in the data, which is referred to as flexibility (James et al., 2021b). This is a desirable property, but it should be borne in mind that complex features (e.g. quadratic or hinge features) with a high weighting (importance) lead to extremely (often excessively) specific models. Flexibility makes the model very susceptible and vulnerable to training data-

specific¹² or random¹³ spatial patterns that are misinterpreted as determinants of presence.

The so-called *overfitting*.

Users can operate on these aspects by predefining a set of *feature classes (fc)* and the amount of the penalty applied to complex features via the *regulation multiplier (rm)*.

Data input

MaxEnt requires two primary types of input: *presence-only* data and *environmental predictors*.

Environmental predictors can be continuous or categorical variables that are used as covariates by the model.

In addition to these inputs, *MaxEnt* generates or uses user-entered background points to represent the range of environmental conditions in the study area. Careful selection of the background is critical as it provides the context through which *MaxEnt* “perceives” what constitutes a suitable habitat for the species (Elith et al., 2011).

Output interpretation

The main output from *MaxEnt* is the *Relative Occurrence Rate (ROR)*, which is a measure indicating the relative probability of occurrence in each cell of the study area, given the environmental conditions. ROR represents the likelihood of a cell being similar to known occurrence locations in terms of its environmental characteristics. It is essentially a measure of how strongly the environmental conditions in each grid cell align with those where the species has been observed (Merow et al., 2013).

¹² Bias

¹³ Noise

The logistic output is a transformed version of the ROR that estimates the probability of occurrence of a species in each cell, assuming that the presence data represent a random sample from the landscape. While this transformation can make the output more intuitive as it scales values between 0 and 1, it also introduces a strong assumption that is often violated in real-world scenarios: It assumes that species presence data are evenly distributed across the landscape, which is often not the case due to biases in observations and uneven sampling effort (Fourcade et al., 2014).

MaxEnt outputs, when assumptions are met, can be interpreted as an index of relative habitat suitability, highlighting areas that resemble the known habitat of the species (Yackulic et al., 2013).

Overview of Spatial Interactions Models

Spatial interaction models are a fundamental tool in geography, urban planning, and science of transports. They provide a structured approach to understanding and predicting the flows of people, goods, information, and even services between locations (K. Haynes & Fotheringham, 1985). The underlying principle of these models is that spatial interactions (such as migration, trade and tourism) are not random but are influenced by the spatial structure of places and the characteristics of the entities involved. In essence, spatial interaction models allow to quantify and examine how distance, population size and connectivity affect the movement and exchange between places, providing insights into complex spatial relationships and the forces that drive them (Fotheringham & O'Kelly, 1988).

In the last decades, these kind of models were applied to model environmental and ecological connectivity with interesting results (McRae et al., 2008; Van Etten & Hijmans, 2010).

In its most general form, a spatial interaction model can be written as:

$$T_{ij} = f(V_i, W_j, S_{ij}) \quad (1)$$

Where:

- T_{ij} represents *interaction* between i and j locations (*flow*)
- V_i represents a vector of *origins attributes* (*propulsiveness*)
- W_j represents a vector of *destination attributes* (*attractiveness*)
- S_{ij} represents a vector of *separation attributes* (*deterrence*)

This expansion allows for extreme flexibility in modelling complex spatial interactions that makes this family of models particularly interesting for the purpose of *exposure* assessment and even more when applied to AIS. For example, V_i may be interpreted as the *invasion potential* (e.g. propagule pressure) and W_j as the *attractive potential* (e.g. suitability).

Moreover, S_{ij} can be augmented by introducing ecological barriers and route preferences, making it an appropriate solution to model the movement across the oceans.

In the context of biological invasions, these models have been applied to simulate the dispersal of species across landscapes and illustrate the interplay between environmental characteristics, habitat suitability and the spatial structure of ecosystems (Falaschi et al., 2018).

One of the most widely used spatial interaction model is the so-called *gravity model*.

Gravity models

The concept of gravity models has its roots in Newton's law of universal gravitation, where the force between two objects is proportional to the product of their masses and inversely proportional to the square of the distance between them. These models quantify how interactions (e.g. trade, migration or biological dispersal) take place between different locations based on factors that influence movement from origin to destination (K. Haynes & Fotheringham, 1985). In the most basic form, gravity models follow the Newton's law formula:

$$T_{ij} = k \frac{M_i M_j}{d_{ij}^2} \quad (2)$$

Where:

- T_{ij} represents *interaction* between i and j locations
- M_i, M_j represents “masses” of the interacting locations
- d_{ij} represents the *distance*
- k is a constant that ensures mass balance

This model is based on the analogy that places with high “quantity” of a feature (e.g. population) interact more, while the cost or difficulty of interaction increases with distance.

In order to capture the complexity of spatial interactions in the real world, gravity models have undergone several generalizations and expansions. Gravity models exist to estimate flows

relying on information of total interactions, total outflows and total inflows both for origin and destination specific formulations (K. E. Haynes et al., 1980).

An example of gravity models applied to the AIS invasion risk mediated by shipping traffic is provided, as it is the foundation and the starting point for the development of the third component of this framework.

Gravity models to estimate global biological invasions hotspots

The foundation for the application of gravity models to biological invasions is the work of Drake & Lodge (2004). The study addresses the threats posed by invasive species transported by ballast water via international maritime trade and contributing to the increasing biotic homogenisation of ecosystems. Prior to this study, ballast water management was recognised as important, but there was a lack of comprehensive models to quantify invasion risks and prioritise management strategies.

The study had a substantial influence on subsequent research by highlighting the importance of reducing the propagule pressure (the rate at which invasive species are introduced) across all ship visits rather than focusing on eliminating key ports from the network. This shifted the focus in marine invasion science from managing high-risk locations to addressing broader systemic risks, contributing to international regulatory discussions such as the International Maritime Organization's (IMO) Ballast Water Management Convention, which was adopted to control invasive species spread through maritime pathways (Srèbaliené et al., 2019).

The work uses a multidisciplinary approach that combines gravity modelling of spatial interactions with network flow analysis to identify key pathways for biological invasions. Drake and Lodge adapted this framework to predict invasion risks from port to port, drawing on empirical data from 243 harbours worldwide.

Gravity models and probability of invasion

They define the probability of invasion as:

$$\pi = 1 - (1 - p)^n \quad (3)$$

Where:

- p is the per-ship probability of initiating an invasion
- n is the number of ships

p was estimated empirically from Ricciardi (2001) while n was estimated using a double constrained gravity model with this form:

$$F_{ij} = A_i B_j O_i D_j f(d_{ij}) \quad (4)$$

Where:

$$- A_i = \left(\sum_z B_j D_j f(d_{ij}) \right)^{-1} \quad (5)$$

$$- B_j = \left(\sum_z A_i O_i f(d_{ij}) \right)^{-1} \quad (6)$$

- O_i is the total flow from origins
- D_j is the total flow to destinations
- $f(d_{ij})$ is the deterrence function

A_i and B_j are the constrain terms that ensure the balance of all outflows and inflows, and are estimated iteratively (Fotheringham & O'Kelly, 1988).

Their model took into account the volume of shipping traffic and the distance between ports to assess the likelihood of species introductions. By applying a distance-dependent *deterrence* function, they identified "hot spots" of invasion where high shipping traffic increases the risk of species introduction.

A central finding of the paper is that reducing the overall probability of species introduction per ship visit is far more effective in slowing global biotic homogenization than eliminating high-risk ports. The study demonstrated that global invasion risk is diffuse and systemic rather than being concentrated in a few high-traffic areas. By using stochastic simulations to compare different management strategies, Drake and Lodge showed that strategies focusing on reducing per-ship invasion probability would lead to a greater reduction in the cumulative number of invasions compared to the elimination of key pathways or ports.

Limitations

While the paper presents a robust and innovative approach, there are key assumptions and limitations that affect its generalizability. Shipping patterns are dynamic, and without continuous updating, the model may not fully account for changes in global trade routes or emerging markets. Additionally, the per-ship invasion probability is based on generalized estimates, which may not capture species-specific or port-specific differences in invasion risks. This creates uncertainty in predicting real-world outcomes, particularly in understudied regions where invasive species dynamics are less understood.

Moreover, there is a lack of ecological characterisation of the invasion process. Per-ship probability of initiating an invasion greatly varies due to environmental suitability of both the donor and recipient site.

2005-2099 High resolution bioclimatic variables for the surface and bottom of the Mediterranean Sea.

An easy-to-use, high-quality dataset for Risk Assessment

Effective management measures, as emphasised in the MSFD, are highly dependent on accurate and comprehensive data. Data is crucial for the development of forecasting models, for policy decisions and for the implementation of targeted management strategies (Omodei et al., 2024).

In recent years, considerable efforts in ecological modelling have led to the development of highly reliable global and regional general circulation models (GCMs) and biogeochemical models (BGCs). These models simulate physical and biological processes with increasing accuracy and provide valuable insights into the dynamics of marine ecosystems under current and future climate scenarios (Giorgi, 2019).

However, the complexity and scale of the data generated by these models pose a major challenge for broad practical application. The high dimensionality and large volume of the data, combined with the complex technical procedures (e.g. downscaling, bias correction) required to make these data interoperable, can be an obstacle for users who are not experts in geospatial data analysis (Fortin & Dale, 2005).

For this reason, one of the key scientific outcomes of this research is the creation and publication of a high-resolution, ready-to-use dataset of the main physical and biogeochemical variables of the Mediterranean Sea for the 21st century, considering the RCP8.5 scenario. The dataset is presented as in the associated article “Downscaled and bias-

corrected bioclimatic dataset for the Mediterranean Sea (2005–2099)” published on Data in Brief Journal as open access (Fianchini et al., 2024b).

Article: Downscaled and bias-corrected bioclimatic dataset for the Mediterranean Sea (2005–2099)

Authors

Marco Fianchini^{1,2,3}, Cosimo Solidoro¹, Donata Canu¹.*

Affiliations

¹National Institute of Oceanography and Applied Geophysics - OGS Via Beirut 2, 34151 Trieste - Trst,

Italy

² Department of Life Sciences, University of Trieste, 34127, Trieste, Italy

³ Department of Environmental Life Sciences, University of Udine, 33100, Udine, Italy

Corresponding author's email address and Twitter handle

mfianchini@ogs.it

Keywords

Biogeochemical model; RCP8.5; Nutrients; Climate Change; Bioclimatic variables

Abstract

This dataset provides annual statistical descriptors (mean, minimum, maximum, range and standard deviation) of key biogeochemical and physical variables for the Mediterranean Sea.

It covers the period 2005-2099 under a high emissions scenario (CMIP5-RCP8.5), with a

spatial resolution of 1/24 degree (~4km²). The CMIP5-RCP8.5 scenario considers the current trajectory of increasing greenhouse gas emissions and population growth to the end of the century with nominal policies to reduce emissions. Variables include temperature, salinity, pH, water velocity, nutrients (NO₃, PO₄, NH₄), dissolved inorganic carbon, oxygen, and net primary production. Data are available for both surface and at bathymetry level. The original projections were generated using OGSTM-BFM and MFS16 models at daily time and 1/16 degree grid resolution. These were downscaled to 1/24 degree resolution and bias corrected by quantile delta mapping using CMEMS reanalysis products from 2005 to 2020. The dataset is provided in a user-friendly format so that it is accessible for various ecological and environmental modelling applications.

SPECIFICATIONS TABLE

Subject	Ecological Modelling.
Specific subject area	Mediterranean Sea Physics and Biogeochemistry annual statistical descriptors for the 21th century in the light of RCP8.5 scenario.
Type of data	Processed, netcdf format.
Data collection	All the processing was performed in Python 3.12.3 using the packages xarray, dask scipy, numpy, pandas, cmethods, copernicusmarine, ddsapi, on a High Performance Computing system (<i>Galileo100</i> , Cineca-Italy) using a total of 24 cores and 256GB RAM. Surface and bottom Reanalysis data at monthly resolution were obtained by

directly downloading them through *copernicusmarine* API service (A detailed method to obtain bathymetry level data is provided at <https://help.marine.copernicus.eu/en/articles/6049335-how-to-retrieve-a-variable-at-bathymetry-level-in-python>). Future projected data for *RCP8.5* scenario at daily resolution were downloaded through *ddsapi*, averaged to monthly data and homogenised in terms of dimensions and variables names. Trilinear interpolation was performed to downscale projected data to Reanalysis spatial resolution and missing values were filled using *Nearest Neighbour* method, in a conservative approach. The same procedure used to extract bathymetry level variables as for Reanalysis data was used. Quantile Delta Mapping algorithm was applied to minimise distributional biases between projections and reanalysis time-series data using the 2005-2020 overlapping period as control.

Data source

Raw data are stored and freely available (Creative Commons

location

Attribution 4.0 International (CC BY 4.0) license) at:

CMCC DDS API:

<https://dds.cmcc.it/#/dataset/medsea-cmip5-projections-biogeochemistry/>

<https://dds.cmcc.it/#/dataset/medsea-cmip5-projections-physics/>

E.U. Copernicus Marine Service Information:

https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_PHY_006_004

[E3R1](#)

	https://doi.org/10.25423/cmcc/medsea_multiyear_bgc_006_008_med_bfm3
Data accessibility	<p>Data are publicly available at:</p> <p>Repository name: 2005-2099 High resolution bioclimatic variables for the surface and bottom of the Mediterranean Sea</p> <p>Data identification number: 10.5281/zenodo.12780160</p> <p>Direct URL to data: https://zenodo.org/doi/10.5281/zenodo.12780160</p> <p>Download requires no registration or any other procedure.</p>
Related research article	<i>none.</i>

VALUE OF THE DATA

- The data provide an extended set of annual projected key bioclimatic variables for the surface and bottom of the Mediterranean Sea for the 21th century at high spatial resolution.
- Bioclimatic variables at regional scale can be useful for a variety of ecological and environmental research applications, ranging from environmental management and risk assessment to biogeographic and species distribution modelling studies.
- The data follows the World Geodetic System standard (WGS84) and are ready-to-use.

- The OGSTM-BFM model provides accurate simulations of plankton productivity and biogeochemical cycles for the Mediterranean Sea.

BACKGROUND

The Mediterranean Sea is characterised by its unique marine ecosystems and significant economic activities, including fishing and tourism. It has complex biogeochemical dynamics influenced by unique oceanographic and climatic conditions. Climate change is exacerbating these dynamics and affecting the availability and distribution of nutrients. The 1/16° resolution Mediterranean Forecasting System and the *OGSTM-BFM* model provide detailed insights into the biogeochemical present and future state of the basin under different emission scenarios (Reale et al., 2022; Solidoro et al., 2022). However, the complexity and scale of the data hinder their widespread, practical application. This dataset addresses these challenges by providing annual statistical descriptors of key climate change indicators and nutrients. The dataset is obtained by harmonising, downscaling and bias-correcting the model simulations outputs (Reale et al., 2022). The decision to focus on the *RCP8.5* scenario is because it can be considered a fairly realistic scenario compared to *RCP4.5* (Schwalm et al., 2020). The result is a ready-to-use dataset containing variables for the surface and bathymetric levels (Tong et al., 2021). With improved accessibility and interpretability, this derived dataset represents an optimised resource for biologists, ecological modellers and policy makers involved in marine conservation and climate resilience planning.

DATA DESCRIPTION

The dataset is composed of 95 files in a single folder (Fianchini et al., 2024a). Names are structured as ‘*{statistical_indicator}_{varname}_{layer}.nc*’ (e.g. *mean_thetao_bottom.nc*).

Note that for Net Primary Production, there are layers reporting the vertical integration of 0-

220m depth layers named '{*statistical_indicator*}_{*nppv*}_integrated.nc'. See *Table 1* for a complete list of files and variables.

Table 1 Description of the dataset files. Variable names, units, statistics and file names describe the content of the repository.

Long name	Variable name	Unit	Statistic	Filenames	
Temperature	<i>thetao</i>	°C	<i>min</i>	<i>min_thetao_surface.nc</i>	<i>min_thetao_bottom.nc</i>
			<i>max</i>	<i>max_thetao_surface.nc</i>	<i>max_thetao_bottom.nc</i>
			<i>range</i>	<i>range_thetao_surface.nc</i>	<i>range_thetao_bottom.nc</i>
			<i>std.dev</i>	<i>Std.dev_thetao_surface.nc</i>	<i>std.dev_thetao_bottom.nc</i>
			<i>mean</i>	<i>mean_thetao_surface.nc</i>	<i>mean_thetao_bottom.nc</i>
Salinity	<i>so</i>	PSU	<i>min</i>	<i>min_so_surface.nc</i>	<i>min_so_bottom.nc</i>
			<i>max</i>	<i>max_so_surface.nc</i>	<i>max_so_bottom.nc</i>
			<i>range</i>	<i>range_so_surface.nc</i>	<i>range_so_bottom.nc</i>
			<i>std.dev</i>	<i>std.dev_so_surface.nc</i>	<i>std.dev_so_bottom.nc</i>
			<i>mean</i>	<i>mean_so_surface.nc</i>	<i>mean_so_bottom.nc</i>
Water velocity	<i>Wv</i>	m/s	<i>min</i>	<i>min_WV_surface.nc</i>	<i>min_WV_bottom.nc</i>
			<i>max</i>	<i>max_WV_surface.nc</i>	<i>max_WV_bottom.nc</i>
			<i>range</i>	<i>range_WV_surface.nc</i>	<i>range_WV_bottom.nc</i>
			<i>std.dev</i>	<i>std.dev_WV_surface.nc</i>	<i>std.dev_WV_bottom.nc</i>
			<i>mean</i>	<i>mean_WV_surface.nc</i>	<i>mean_WV_bottom.nc</i>

pH	<i>ph</i>	.	<i>min</i>	<i>min_ph_surface.nc</i>	<i>min_ph_bottom.nc</i>
			<i>max</i>	<i>max_ph_surface.nc</i>	<i>max_ph_bottom.nc</i>
			<i>range</i>	<i>range_ph_surface.nc</i>	<i>range_ph_bottom.nc</i>
			<i>std.dev</i>	<i>std.dev_ph_surface.nc</i>	<i>std.dev_ph_bottom.nc</i>
			<i>mean</i>	<i>mean_ph_surface.nc</i>	<i>mean_ph_bottom.nc</i>
Ammonium	<i>nh4</i>	[mmol <i>m-3</i>]	<i>min</i>	<i>min_NH4_surface.nc</i>	<i>min_NH4_bottom.nc</i>
			<i>max</i>	<i>max_NH4_surface.nc</i>	<i>max_NH4_bottom.nc</i>
			<i>range</i>	<i>range_NH4_surface.nc</i>	<i>range_NH4_bottom.nc</i>
			<i>std.dev</i>	<i>std.dev_NH4_surface.nc</i>	<i>std.dev_NH4_bottom.nc</i>
			<i>mean</i>	<i>mean_NH4_surface.nc</i>	<i>mean_NH4_bottom.nc</i>
Nitrate	<i>no3</i>	[mmol <i>m-3</i>]	<i>min</i>	<i>min_NO3_surface.nc</i>	<i>min_NO3_bottom.nc</i>
			<i>max</i>	<i>max_NO3_surface.nc</i>	<i>max_NO3_bottom.nc</i>
			<i>range</i>	<i>range_NO3_surface.nc</i>	<i>range_NO3_bottom.nc</i>
			<i>std.dev</i>	<i>std.dev_NO3_surface.nc</i>	<i>std.dev_NO3_bottom.nc</i>
			<i>mean</i>	<i>mean_NO3_surface.nc</i>	<i>mean_NO3_bottom.nc</i>
Phosphate	<i>po4</i>	[mmol <i>m-3</i>]	<i>min</i>	<i>min_PO4_surface.nc</i>	<i>min_PO4_bottom.nc</i>
			<i>max</i>	<i>max_PO4_surface.nc</i>	<i>max_PO4_bottom.nc</i>
			<i>range</i>	<i>range_PO4_surface.nc</i>	<i>range_PO4_bottom.nc</i>
			<i>std.dev</i>	<i>std.dev_PO4_surface.nc</i>	<i>std.dev_PO4_bottom.nc</i>
			<i>mean</i>	<i>mean_PO4_surface.nc</i>	<i>mean_PO4_bottom.nc</i>

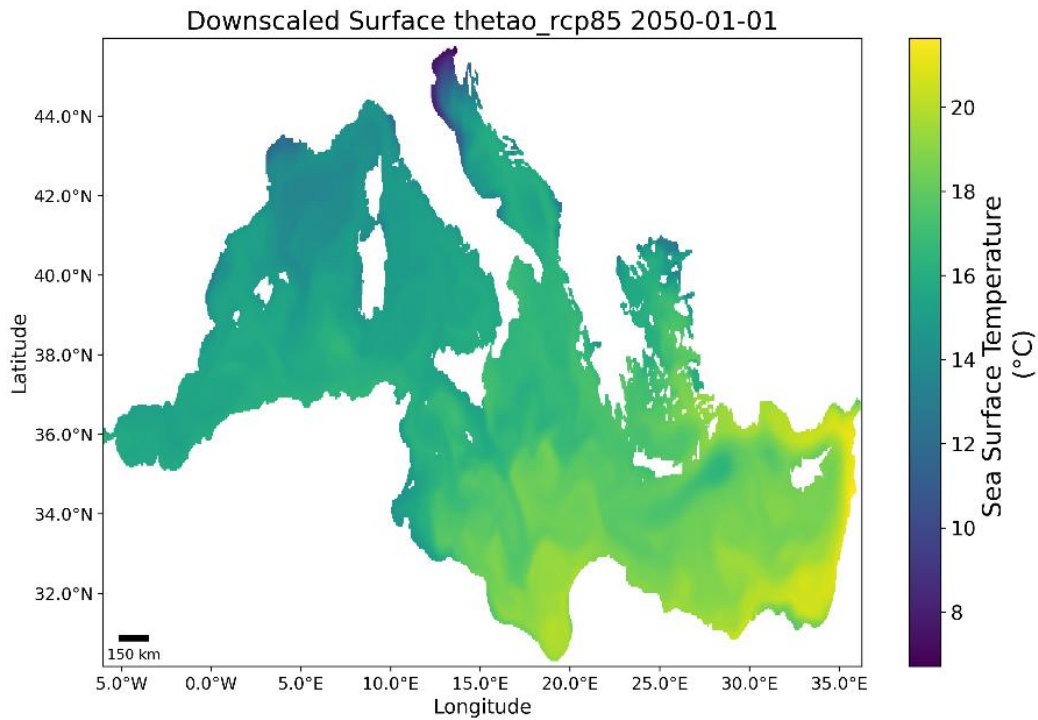
Dissolved Oxygen	o2	[mmol m-3]	min	min_o2_surface.nc	min_o2_bottom.nc
			max	max_o2_surface.nc	max_o2_bottom.nc
			range	range_o2_surface.nc	range_o2_bottom.nc
			std.dev	std.dev_o2_surface.nc	std.dev_o2_bottom.nc
			mean	mean_o2_surface.nc	mean_o2_bottom.nc
Net Primary Production	nppv	[mg m-2 day-1]	min	min_nppv_integrated.nc	
			max	max_nppv_integrated.nc	
			range	range_nppv_integrated.nc	
			std.dev	std.dev_nppv_integrated.nc	
			mean	mean_nppv_integrated.nc	
Dissolved Inorganic Carbon	dissic	[mol m-3]	min	min_DIC_surface.nc	min_DIC_bottom.nc
			max	max_DIC_surface.nc	max_DIC_bottom.nc
			range	range_DIC_surface.nc	range_DIC_bottom.nc
			std.dev	std.dev_DIC_surface.nc	std.dev_DIC_bottom.nc
			mean	mean_DIC_surface.nc	mean_DIC_bottom.nc

Each file covers the Mediterranean Sea (BBOX: *lon.min* = -6.00; *lon.max* = 36.28;

lat.min=30.21; *lat.max* = 45.97, CRS = EPSG:4326) for each year from 2005 to 2099. See

Figure 10 for an example.

Figure 10 Projection of the mean surface temperature for the Mediterranean Sea in January 2050.



EXPERIMENTAL DESIGN, MATERIALS AND METHODS

COMPUTATIONAL ENVIRONMENT:

Processing was performed on *Galileo 100* HPC system (Cineca-Italy) using 24 cores and 256GB RAM using Python language. Dependencies are: *copernicusmarine*, *ddsapi*, *xarray*, *dask*, *pandas*, *numpy*, *scipy*, *matplotlib*, *cmethods*.

MATERIALS:

The dataset was created using data from two primary sources: the European Union's Copernicus Marine Service Information (*CMEMS*) and the Centro Euro-Mediterraneo sui Cambiamenti Climatici (*CMCC*). The specific products used were:

1. Mediterranean Sea Biogeochemistry Reanalysis (Cossarini et al., 2021). E.U. Copernicus Marine Service Information (*CMEMS*). Marine Data Store (*MDS*).

2. Mediterranean Sea Physical Reanalysis (Escudier et al., 2021). E.U. Copernicus Marine Service Information (*CMEMS*). Marine Data Store (*MDS*).

3. MEDSEA CMIP5 Projections Physics (Lovato et al., 2013). CMCC Data Delivery System (*DDS*).

4. MEDSEA CMIP5 Projections Biogeochemistry (Reale et al., 2022). CMCC Data Delivery System (*DDS*).

THE FUTURE OF THE MEDITERRANEAN IN THE LIGHT OF THE RCP 8.5 SCENARIO:

Simulations following the *RCP8.5* scenario predict significant changes in the Mediterranean by the end of the *21st* century:

1. Increased water temperature and stratification, reduction in vertical mixing and nutrient upwelling.

2. Reduced availability of phosphorus and nitrogen in the upper layers, limiting primary productivity and phytoplankton populations. The eastern Mediterranean could develop into oligotrophic conditions.

3. Reduced oxygen solubility and increased acidification due to higher *CO2* uptake, affecting marine life, especially calcifying organisms.

4. Altered ecosystem composition and functionality with possible northward migration of species adapted to warm conditions and decline of less adaptable species.

5. Variable regional impacts, with potentially greater changes in the eastern Mediterranean compared to the western part.

6. Fluctuations in salinity due to changes in precipitation, river flows and evaporation rates, affecting water density and circulation patterns.

GIS MODELLING:

All data processing was done using the native coordinate system World Geodetic System 1984 (WGS84) (*EPSG:4326*) to transform the regular 1/16 degree grid of the modelled data into the regular 1/24 degree grid of the reanalysis data. The distances are calculated “on the fly” using the *Haversine* formula. This approach minimises the risk of introducing integration and numerical errors associated with the planar projection and preserves the latitude/longitude trends as produced by the model simulations.

EXPERIMENTAL DESIGN and METHODS:

Daily raw *RCP8.5* data from the *OGSTM-BFM* model for the period 2005-2100 with a grid resolution of 1/16 degree were processed to obtain bias-corrected statistical descriptors with a grid resolution of 1/24 degree and an annual time scale. Data processing included the following steps (package.function in brackets):

1. Monthly aggregation: the daily data were averaged monthly to reduce the data size and conform to the *CMEMS* reanalysis format.
2. Spatial downscaling: The data were downscaled to 1/24 degree using trilinear interpolation (*scipy.interpolate.interpn*) and masked (*numpy.where*) to exactly match the spatial grid of the *CMEMS* reanalysis.
3. Imputation of missing values: Missing values in the downscaled data set were filled using a nearest neighbour algorithm (*xarray.Dataset.ffill*), following a conservative approach. This ensured that the resulting dataset had the same number of values and the same spatial structure as the *CMEMS* reanalysis dataset while limiting extrapolation artifacts.
4. Bias correction: A Quantile Delta Mapping algorithm (*cmethods.adjust*, method = *quantile_delta_mapping*, quantiles = 10000) was used to align projections to the trends in the

CMEMS reanalysis time series (Tong et al., 2021). The overlapping 2005-2020 period was used as a reference to fit a transfer function that maps the model's distribution to the reanalysis distribution while maintaining the relative changes projected by the model.

5. Calculation of statistical descriptors: Bias-corrected data were aggregated annually and five statistical indices were calculated: Minimum(*numpy.min*), Maximum(*numpy.max*), Mean(*numpy.mean*), Range(*numpy.range*) and Standard Deviation(*numpy.std*). The data were exported as *netcdf* files.

An example/template script for reproducing the entire process, starting with the download of the raw data, can be found in the Supplementary Material.

LIMITATIONS

It is important to recognise that all model projections are subject to uncertainties. These can come from various sources, such as: Model structure, model parameterisation, uncertainties in initial conditions, uncertainties in boundary conditions and forcing functions (scenarios), numerical approximations. Usually the uncertainties in the initialisation dominate on a shorter time scale, the uncertainties in the scenarios are generally more important on a longer time scale, while the uncertainty in the choice of model parameterisation is always important. However, the results of a systematic global sensitivity analysis on the parameter uncertainties in a complex marine ecosystem model show that while the details of the time course of many variables are sensitive to the choice of parameter values, the integrated indicators are much more robust.

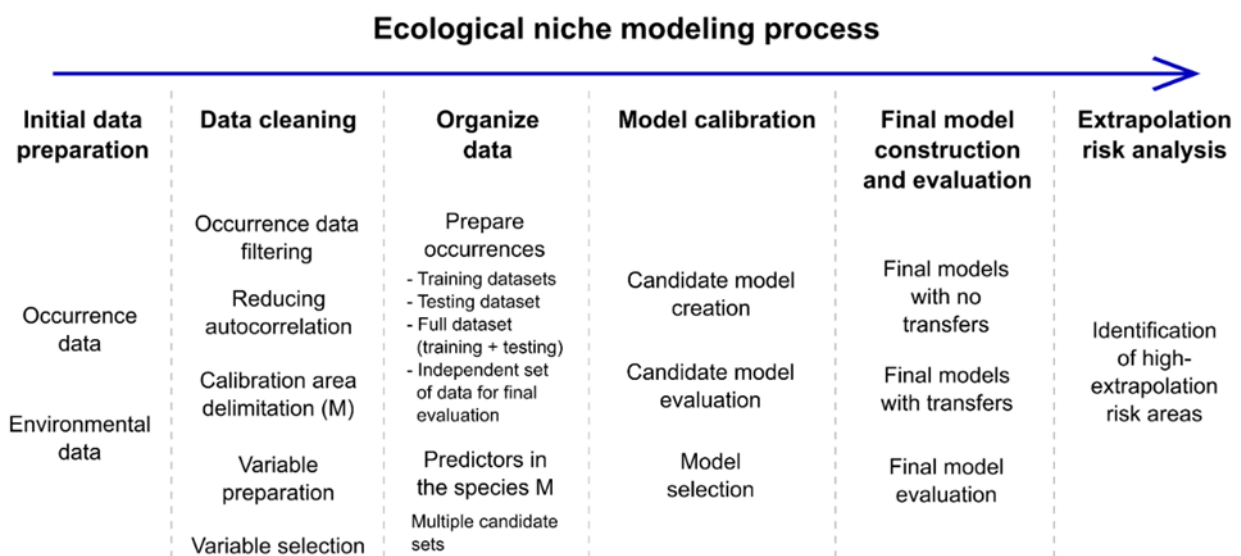
In this case case, the importance of the atmospheric forcing and boundary conditions (i.e. the scenarios) should be emphasised, which can strongly influence the dynamics of the Mediterranean Sea, including vertical mixing, and thus the nutrient distribution and the

dynamics of the lower trophic levels. Another point to consider is the spatial horizontal and vertical discretisation of the model. The *OGSTM-BFM* model focuses on resolving the dynamics of the low trophic levels in the euphotic zone at high resolution, while the representation of the processes in deep waters and near the bottom is less detailed. A final consideration concerns the downscaling and bias correction processes, which adjust and harmonise the dataset to the trends observed in the reanalysis but may also introduce artefacts and smooth environmental signals. Considering all these factors, the dataset is well suited for studies at regional and sub-regional scales but should be used with caution for applications at scales below grid resolution.

Toward a reliable ENM process

This chapter represents the central scientific contribution of this research and describes a comprehensive re-evaluation and improvement of the Ecological Niche Modelling (ENM) workflow (see Figure 11). This approach was specifically developed to address the complex challenges of modelling AIS.

Figure 11 Schematic description of the ecological niche modeling process. Adapted from Cobos et al. (2019).



The content of this chapter is an extended version of the article: “*Improving the reliability of MaxEnt predictions through multi-criteria decision analysis and site-weighted metrics: Caulerpa cylindracea case study*” (Fianchini et al., 2024. Under review).

The aim is to demonstrate a transparent and reproducible pathway that integrates best practises in data processing, model calibration, evaluation and risk assessment.

Each section covers a stage of the modelling process and addresses the underlying assumptions, constraints and methods. This facilitates a comprehensive understanding of how each element contributes to improving the reliability of ML-ENM.

Although we will address specific MaxEnt aspects and assumptions throughout the study, the framework is transferable to other ML-ENM algorithms.

Abstract

The increasing availability of data with fine spatial and temporal resolution has popularised the use of correlative Ecological Niche Models (ENMs) in the field of biological invasions.

However, to predict the spread of Alien Invasive Species (AIS), it is important to maximise the transferability of the models. Ensuring a balance between model complexity and ecological validity is essential for reliable results. In this study, we propose a comprehensive framework to deal with biased data, assess extrapolation risk, evaluate the impact of threshold uncertainties on suitability classification, and identify models with minimal "artificially enhanced" performance. We use MaxEnt to model the current and future distribution of *Caulerpa cylindracea*, one of the most dangerous AIS in the Mediterranean.

We evaluated 1240 model configurations under a "present scenario" to identify the model configuration that most accurately reflects the pattern of species abundance in a withheld dataset. To do this, we used a set of site-weighted performance metrics and applied a multi-criteria decision analysis to rank the candidate models. The annual projected probability of occurrence was then categorised into four suitability classes. We analysed the impact of threshold uncertainty on the classification results. In addition, a time series analysis of the annually predicted areas for each suitability class was performed to identify trends.

The success of the training process is confirmed by the fact that the response curves are consistent with the knowledge of the ecophysiology of the species. All models show a general shift towards a lower probability of occurrence in the future scenario. The time series analysis

of the high suitability areas showed a significant negative trend, indicating that the spread of the species is likely to be limited to the currently suitable areas.

The use of site-weighted metrics refines the models evaluation and increases confidence in the results. It is of particular value in highlighting overfitting and prioritising the selection of models with desirable, specific properties and scope. Confidence in model reliability allows the projection to be taken further to create and investigate suitability time series, opening up a new class of analyses in the process of Ecological Niche Modelling.

Defining the scope

Before beginning, we need to define what we expect from the model and clearly define the spatial and temporal boundaries of the study.

For *C. cylindracea*, we want to train a robust *MaxEnt* model to estimate the present and future suitability in the Mediterranean Sea considering the RCP8.5 climate projections. To support reliability of the projections, we want to characterise model uncertainty.

The “present” we consider, i.e. “where” in time the model is trained, are the average environmental conditions for the year 2000-2020, a choice dictated by the time span in which most of the occurrences were collected.

The “future” refers to the average environmental conditions for 2030-2050 and can be considered an arbitrary choice.

In addition, the projection is driven by utilising the high temporal resolution of the environmental data to obtain annual maps for the entire period, i.e. 2000-2050, to investigate whether trends exist or are likely to exist in suitable area.

Data processing

The first practical step of the modelling process is the collection of input data, namely *occurrences* and *environmental predictors*.

Occurrence data

To make *MaxEnt* “learn”, we need to provide positive examples of species presence in the geographical space. These examples are commonly called “*occurrences*”. Thanks to the joint efforts of scientific community and international organisations in the last two decades, *occurrences* data are easily available in a standardised form through a diverse ecosystem of data provider, e.g. the Ocean Biodiversity Information System (OBIS, <https://obis.org/>), the Global Biodiversity Information Facility (GBIF, <https://www.gbif.org/>) (Gaiji et al., 2013).

Specific information systems were developed to gather occurrence data of AIS, such as the European Alien Species Information Network (EASIN, <https://easin.jrc.ec.europa.eu/easin>).

The herbarium specimen digitalisation and the emergence of citizen science projects has augmented the size and number of datasets worldwide (Sandahl & Tøttrup, 2020), with positive effects on the ENM practice (Smith et al., 2023).

Such an abundance of diverse data sources mitigates the numerical demand for data, but poses a trivial challenge to data quality (Fourcade et al., 2014), as these systems bring together opportunistic sampling, machine observations, scientific sampling campaigns, herbarium collections and elaborated datasets from monitoring projects. This has required the development and adoption of standard formats such as the Darwin Core (Wieczorek et al., 2012), that provide end users with a wide range of metadata to perform quality checks and filtering on the raw datasets.

For the purposes of the study, *C. cylindracea* occurrences up to 2020 were derived from GBIF and EASIN (GBIF.org, 2024; Trombetti et al., 2013).

One of the retrieved datasets, was retained and elaborated separately to be used as a test. *The Reef Check Med Dataset on Key Mediterranean Marine Species 2001–2020 (RCMED)* is the result of a long-term campaign conducted by trained citizen scientists, aimed to estimate presence (with semi-quantitative abundance) and absence of 43 key taxa, among which *C. cylindracea* (Turicchia et al., 2021).

Occurrence data cleaning

In this phase, the task is to obtain a valid list of cells¹⁴ that are labelled as presences.

The first operation consists in filtering duplicate entries and occurrence records that lacks metadata required by the user to be reliable. This usually imply removal of records that has no (or clearly wrong) coordinates, coordinates uncertainty, date of the sampling.

The second process consists in the labelling of sites that are considered representative of the presence ($f_1(z)$, presented in chapter). This stage implies subtle (often overlooked) assumptions about the representativeness of the samples (Franklin, 2010). A quick comparison of the spatial scale of an opportunistic sampling (order of square metres) and the cell of a typical grid (order of square kilometres) makes this issue clearer.

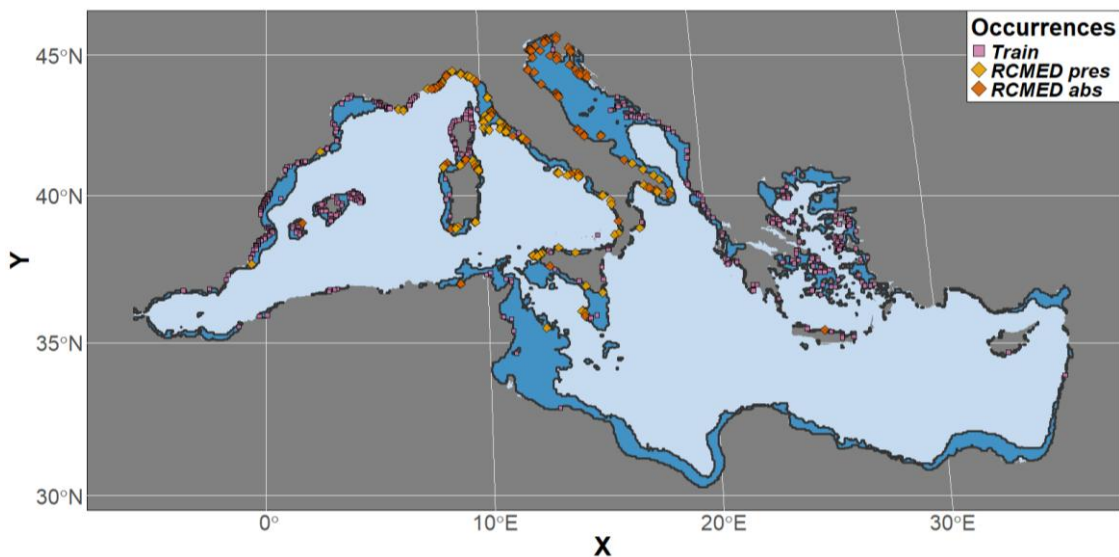
In contrast to the usual approach of considering each occurrence as a valid reason for labelling a site as a presence, in this study a site was labelled as presence if it was supported

¹⁴ Site is used as a synonym for cell and is used interchangeably. The reason for this is to emphasise, depending on the case, the biological value of the grid unit or the spatial nature of the datum.

by at least three records with non-identical coordinates. The cleaning process led to the selection of 1667 sites to be used as examples to train the model.

The RCMED dataset was also processed: The lowest frequency among records falling into a cell was kept and a cell were only labelled as absence if at least 5 records with absence and no presence fell into a cell. The process led to identify 81 presence sites and 85 absence sites. See Figure 12 for a map reporting the train and test sites.

Figure 12 Presence (and absence) sites identified after the cleaning process. Dark blue areas identify the domain of the study (Mediterranean waters with depths < 220 m).



Environmental data

The backbone of realised niche derivation is the assignment of occurrences to the environmental conditions under which they were recorded. This is often achieved using geospatial analysis, that represents the environmental conditions as cells (or pixels) with a defined spatial and temporal resolution that form a grid with a predefined extent (the study area). The environmental conditions usually comprise several environmental variables (predictors) that form the levels (layers) of environmental information that characterise each cell (Fortin & Dale, 2005).

Occurrences are represented as points with geographic coordinates that allow to directly link them to the environmental conditions represented in the cells in which they fall.

The set of predictors usually comprise physical, chemical or biological related variables or indexes that are assumed to exert an influence on the species presence (Peterson et al., 2011).

Common sources of predictors for regional scale studies are remote sensing data, General Circulation Models (GCMs) and Bio-Geo-Chemical Model (BGCMs). The reason is that they provide a continuous and consistent estimation of the variables for the whole study area, that would be impossible to achieve by direct observations (Franklin, 2010).

Our attention is focused on GCMs and BGCMs as these models not only estimates past and present conditions but provide also future projections according to pre-defined scenarios (Lovato et al., 2013; Solidoro et al., 2022).

For the purposes of this study, we used as initial *predictors* all the available environmental datasets contained in Fianchini et al., (2024), already presented chapter 3. To enrich the characterisation of the Biotic component of the environment, two categorical *predictors*, namely *biozone* and *substrate* (Vasquez et al., 2021; available through EMODnet service: <https://emodnet.ec.europa.eu/>), were integrated.(Bulleri et al., 2018; Caronni et al., 2021; Piazzini et al., 2016)

The term *biozone* refers to a specific vertical section or depth zone in a habitat. It defines the environmental layers based on physical factors such as light availability, depth or wave action that influence species and communities.

Substrate describes the surface or material on which organisms live. It can include various natural and artificial surfaces such as rocks, sand, mud or biogenic structures (e.g. coral

reefs). Substrate type plays a crucial role in the classification of habitats as it influences the communities and ecosystems that develop.

See Table 2 for a list of the biozone and substrate types considered in the analysis.

Table 2 EUNIS Biozone and Substrate types.

Layer	ID	Description
<i>Biozone</i>	1	Infralittoral
	2	Shallow circalittoral
	3	Deep circalittoral
<i>Substrate</i>	1	Seabed
	2	Sand
	3	Rock or other hard substrata
	4	Sandy mud
	5	Coarse & mixed sediment
	6	Fine mud
	7	Muddy sand
	8	<i>Posidonia oceanica</i> meadows
	9	Dead mattes of <i>Posidonia oceanica</i>
	10	Coralligenous platforms
	12	Fine mud or Sandy mud or Muddy sand

Variable preparation and selection

To ensure consistent spatial analysis, the environmental layers must be precisely aligned and have a uniform spatial resolution. In addition, these layers should be projected onto an appropriate planar space using a Coordinate Reference System (CRS) that is specifically suited to the type of spatial analysis being performed. When projecting three-dimensional data onto a two-dimensional space, distortions in area, shape, distance, direction or scale are inevitable (Fortin & Dale, 2005). Therefore, selecting a projection that best preserves the properties most important to the analysis is critical. An often overlooked assumption with *MaxEnt* is that it is area-dependent (Renner & Warton, 2013). Therefore, the use of an equal-area projection is required to maintain analytical accuracy.

In this work, Eckert IV projection and an area preserving resampling method (Johnson & Clarke, 2021) were used to prepare the environmental layers.

The main purpose of this phase is to identify a set of environmental variables (predictors) that is able to statistically characterise the environment in a biological meaningful space for the species (Cobos et al., 2019).

A number of biologically important variables (min/max T, max salinity, mean phosphate concentration, min current velocity) were identified through literature review (see Chapter 2.4) and stepwise Variance Inflation Factor (VIF) (correlation threshold = 0.5, VIF threshold = 2.5; James et al., 2021). The VIF analysis helps to mitigate collinearity problems that could artificially inflate the result of the analysis (Zuur & Ieno, 2016). From the original dataset, 8 continuous variables were identified, covering different aspects of the species, from mechanical tolerance to nutrient exploitation and habitat preference. Biozone and substrate, were retained due to their ecological importance.

Table 3 Layers selected after VIF analysis and used in the modelling process.

Name	Var	unit	Statistic	Values	VIF score
Temperature	thetao	°C	max	surface	1.40
			min	surface	1.36
Salinity	so	PSU	max	bottom	1.20
Water velocity	KE	m/s	min	bottom	1.01
Phosphate	po4	[mmolm ⁻³]	mean	bottom	1.69
Net Primary Production	nppv	[mgm ⁻² day ⁻¹]	mean	water column sum	1.49
Biozone	-	-	-	categorical	-
Substrate	-	-	-	categorical	-

Calibration region

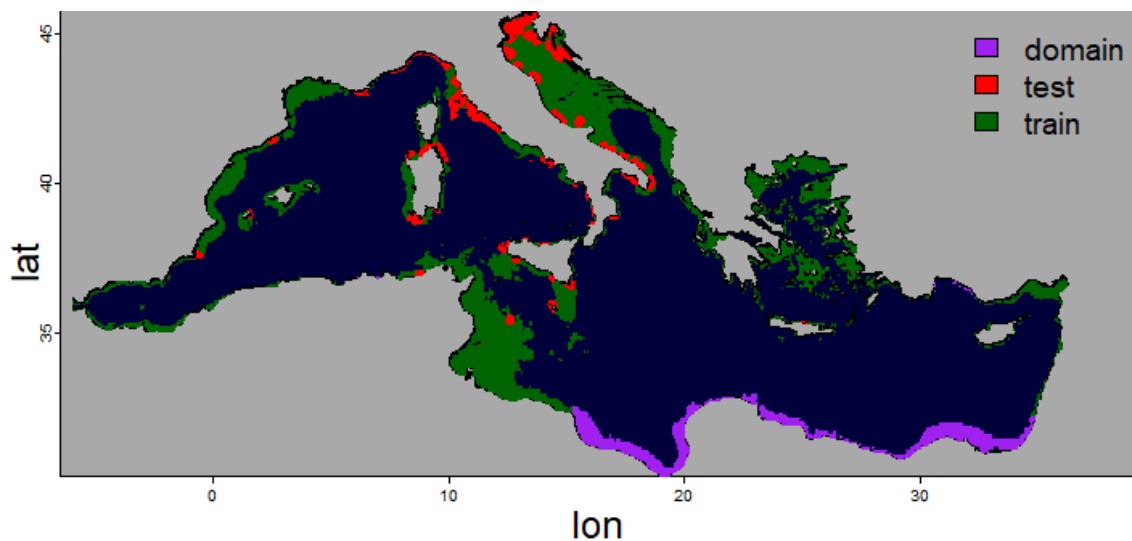
The definition of the calibration region is probably the most important step in ML-ENM (Luna et al., 2024; Owens et al., 2013).

In *MaxEnt*, the calibration region refers to the geographical region in which the environmental variables are sampled, and which serves as the context for the model (its universe) to characterise the regions in the environmental space (See Figure 13). The algorithm estimates the statistical relationship that connects the probability of occurrence in this context. In other words: We ask the model to "draw" the (realised) niche of the species on a "plot" of the (realised) theoretical domain of possible environmental conditions.

Ecologically, it refers to the geographic region accessible via dispersal over relevant time periods, where the environmental conditions were "experienced" by the species in question¹⁵.

A consequence is the removal of all cells that shows bathymetry depths greater than 200 m, well outside the euphotic zone.

Figure 13 Geographical representation of the calibration ("train") and testing ("test") regions. "Domain" indicate cells present in the study but not used for the inference.



For the purposes of this analysis, background is defined as the set of valid cells within 200 km from any presence cell, but distant at least 20 km¹⁶ from the cells used as test (RCMED). This ensures that the validation happens on *extrapolation* capabilities rather than *interpolation*. Cells used as background are 25194.

¹⁵ In practice, it is the process that selects the points that are to be used as (meaningful) "background" $f(z)$.

¹⁶ Subjective distance, related to the need of not excluding too many points from training and the need to break spatial autocorrelation between train and test data.

Spatial autocorrelation

Accounting for spatial or environmental stratification in contextual data is essential to reduce the likelihood of models interpreting exogenous factors (e.g., unequal sampling effort shaping geographic gradients of presence cells) or random spatial patterns as (over)explanatory variables (Radosavljevic & Anderson, 2014). Spatial autocorrelation, i.e. the non-random arrangement of data points in space, can lead to biased model evaluations if it is not adequately considered.

A basic approach for dealing with spatial autocorrelation is to estimate empirical variograms for each environmental variable. Variograms quantify the degree of spatial dependence by measuring the similarity between observations as a function of the distance separating them (Fortin & Dale, 2005). Analysing variograms can indicate a critical distance beyond which spatial autocorrelation becomes negligible. This distance serves as a threshold to ensure the spatial independence of the observations used in training and validating the model (Roberts et al., 2017).

Conventional cross-validation methods often inadvertently split geographically close observations between training and test sets, leading to overly optimistic performance estimates due to spatial autocorrelation.

To mitigate this problem, the *blockCV R* package (Valavi et al., 2019), specifically designed to create spatially disjoint subsets for cross-validation, was implemented.

The function *cv_spatial_autocor* was used to analyse the spatial autocorrelation of the environmental variables throughout the study area. This made it possible to estimate a distance threshold (*drange*) at which autocorrelation becomes minimal. To counteract the "edge effect", where observations near the boundaries of blocks may exhibit residual spatial

dependence, the distance range was increased by 50% (Roberts et al., 2017). A *drange* of 399106.5 meters was used to create the blocks.

Five spatially disjoint samples (with *drange* as minimum distance) within the calibration region are created using the 'random' method in the *cv_spatial* function.

This approach allowed for robust model evaluation by systematically partitioning the data into spatially independent subsets, thereby enhancing the generalizability and reliability of the predictive models.

MaxEnt learning and calibration

Data are now ready to be “fed” into *MaxEnt*.

The goal is to train a model using calibration data and evaluate its ability to fulfil a specific task using a set of known outputs (*test*), outside the calibration space.

To train a model, *MaxEnt* requires the user to provide two settings¹⁷: response types (*feature combination, fc*) and a penalty for complex features (*regulation multiplier, rm*).

Each combination of *fc* and *rm* is assimilable to a *hypothesis* on the complexity of the relationship between *predictors* and *presence*. On top of this hypothesis, the algorithm tries to estimate predictors' importance and species response to each variable (calibration phase).

The output is a formula that “predict” the species probability of occurrence as a combination of the environmental variables.

¹⁷ Hyperparameters, in ML terms.

Calibration phase

Hyperparameters' choice strongly affects model results and is the primary task of the model calibration phase (Low et al., 2020; Morales et al., 2017).

The “philosophy” supported by this research work is that calibration should be considered as a *ML* optimisation task and involve an extensive exploration of hyperparameters' space to ensure a robust model selection (Low et al., 2020).

In this work, all possible combinations of regularisation multipliers (*rm*: from 0.5 to 20, 0.5 each) and *MaxEnt* feature classes (*fc*: linear, quadratic, products, threshold, hinge, categorical) are tested, resulting in 1240 model configurations. Calibration was performed with no clamping (assumptions on response curves outside the training range), as suggested by Guisan et al. (2017).

Model evaluation

The most important task in ML-ENM is the evaluation of model performance. Performance refers to a quantitative measure of the model's capabilities, typically assessed on a set of reliable data not used in the calibration process.

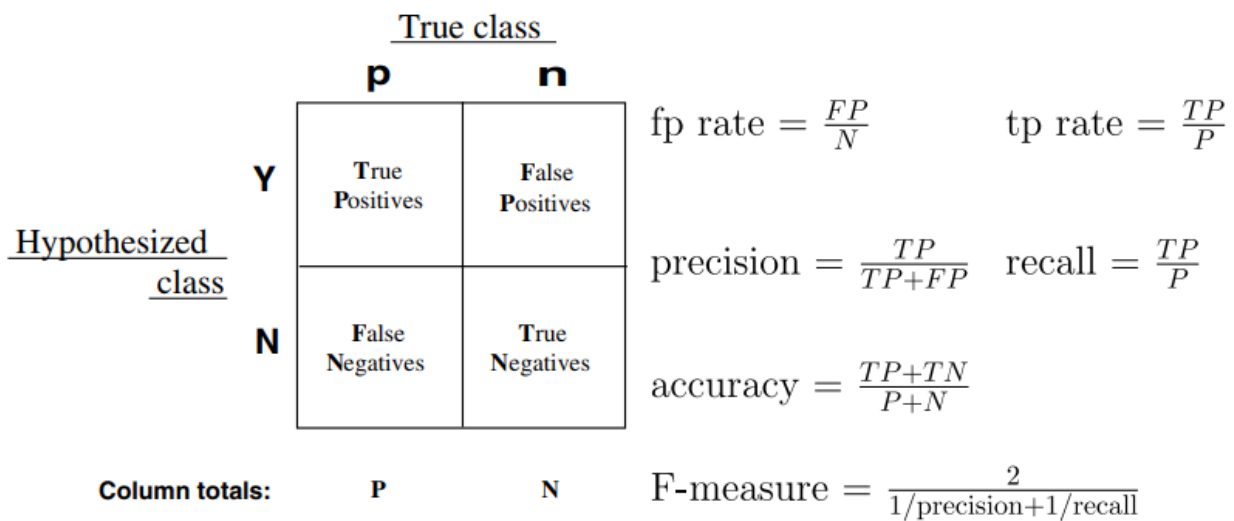
The key capabilities that are measured are *predictive ability*, *transferability* and *complexity*.

Predictive ability assesses how well the model can distinguish between presence and absence.

Transferability measures how well the model generalizes to new, unseen data. It is crucial for ensuring the model's predictions can be reliably applied in other context than the calibration region¹⁸.

Complexity refers to the level of sophistication in the model structure. It depends on the number of the features included, therefore is related to model flexibility.

Figure 14 Confusion matrix and common performance metrics. Taken from Fawcett (2006).



The selection of appropriate evaluation metrics is non-trivial, as different metrics provide different interpretations of model performance. Researchers need to carefully select metrics that are aligned with their specific research questions and the ecological context of their study (Elith et al., 2010).

Evaluation metrics can be divided into two types:

1. *Threshold-dependent metrics.*

These metrics require the selection of a specific probability threshold to classify predictions into presence or absence. The central element is the construction of a confusion matrix

¹⁸ In the optic of AIS distribution modelling, this is the most desirable model capability.

(Fawcett, 2006). See Figure 14 for a visual representation and the definition of key performance metrics based on it.

They are particularly important because they evaluate the performance of the model on tasks that are very close to practical, real-world applications. The downside is that the choice of threshold becomes the most important factor in interpreting performance (and results). Despite the great attention it has received in the scientific community, there are still no optimal strategies for selecting meaningful thresholds, so the choice is to some extent subjective (Elith & Franklin, 2013; Rajbahadur et al., 2019).

2. *Threshold-independent metrics.*

These metrics test the model's performance across the full range of possible thresholds and provide an unbiased assessment as they focus on the model's ability to rank positive cases higher than negative cases, rather than its ability to be "right". This means that they are more robust when it comes to highly imbalanced data sets, making them the most important tools for evaluating the performance of ENMs (Fawcett, 2006; Peterson et al., 2008).

There is a rich corpus of critical reviews of the performance assessment process and comparisons of evaluation metrics in the scientific literature (Fawcett, 2006; Low et al., 2020; Warren & Seifert, 2011; Wunderlich et al., 2019). Despite differing views on which metrics to use in which case, there is a consensus on the need to evaluate model performance using a mixture of threshold-dependent and independent metrics (Guisan et al., 2017).

In Table 4, the set of performance metrics used to evaluate models of *C.cylindracea* are presented.

Table 4 Summary of the performance metrics used in this work. SEDI and OR10p are threshold-dependent metrics, while AUC, pROC and CBI are threshold-independent metrics.

Metric	Rationale	Description	Reference
AUC	Predictive ability	Measures the probability that a randomly selected positive example will be ranked higher than a randomly selected negative one	Fawcett, 2006
pROC	Statistical significance	A modified ROC analysis that adjusts the AUC to focus on the proportion of area correctly predicted as present, accounting for model-specific prediction spans and prioritizing omission errors.	Peterson et al., 2008
SEDI	Predictive ability	Measures the accuracy of deterministic forecasts of rare binary events, robust to low prevalence	Wunderlich et al., 2019
OR.10p	Predictive ability	Percentage of test sites where presence was not predicted, using the 10th percentile of training presence values as the threshold	Muscarella et al., 2014
AUCdiff	Overfitting	Measures the risk of model overfitting by comparing AUC values between training and test datasets	Warren and Seifert, 2011
AICc	Complexity	Relative measure of model fit that considers the number of parameters and penalizes overfitting	Burnham and Anderson, 2003
CBI	Transferability	Measures the accuracy and the reliability in predicting presence based on the Spearman rank correlation between predicted suitability and observed presence	Hirzel et al., 2006

Improving the informative value of performance metrics

This work introduces a novel weighting system that aims to make the results of the metrics more robust against spatial biases in the observations and more meaningful for the ecological signal.

Ecological significance is increased by using additional data (if available) related to occurrences to account for site-specific ecological relevance in the metric calculation.

Weighted metrics are obtained by applying weights to account for different importance levels in the data. This adjustment ensures that certain observations (e.g., more critical locations or underrepresented regions) are given more emphasis during evaluation. Weighted metrics

provide a more balanced and realistic assessment of model performance when data is imbalanced or certain predictions are prioritized.

Site-specific weighting

Occurrences and background test sites in cross-validation are weighted according to their reciprocal distance to reduce the impact of spatial autocorrelation on the metric (Fourcade et al., 2014).

For the RCMED dataset, the weighting of presence sites is based on reciprocal distance and abundance, while the weighting of absence sites is based on reciprocal distance and proximity to presence sites. It is assumed that sites with a higher abundance are more representative of the realized niche and that absences close to presence points are more informative for the "edge" of the distribution.

Points play a decisive role in the demarcation of suitable and unsuitable areas. However, the effects of clustered points can distort interpretations due to spatial autocorrelation as well as sampling and detection biases (Phillips et al., 2009). To counteract these divergent effects, a distance weighting system was developed with respect to a specific set of points (occurrence and presence points).

This distance weighting system reduces the influence of spatially contiguous points by applying decreasing weights with increasing point density. This influence decreases with increasing distance up to a threshold value (*drange*), above which the points receive the maximum weight.

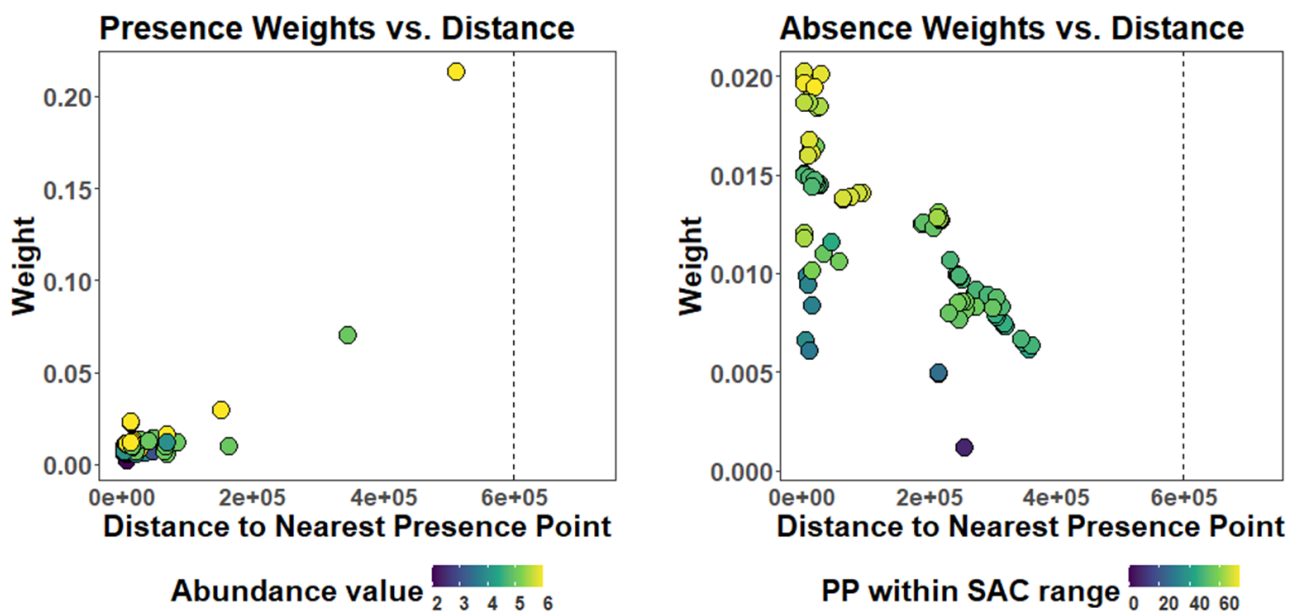
This system is used for background, occurrence and RCMED presence points (Figure 15, left), where the weighting for the latter is multiplied by the abundance index to weight the abundance.

For RCMED absence points (Figure 15, right), proximity to presence points leads to a higher weighting, indicating their greater importance in representing suitability gradients.

This mechanism ensures that the weighting for points further away from presence points decreases until a maximum distance (*drange*) is reached, after which a fixed weighting is applied.

It is worth to notice that this method allows to use basically any auxiliary information as weighting factor (e.g. time) and is customizable based on species-specific knowledge (i.e. a *drange* meaningful for the species biology).

Figure 15 Presence (left) and absence (right) weights for the RCMED test set. The dashed line indicates the value assigned to each point the Spatial Auto Correlation range from absence sites (right). PP= Presence Points.



Weighted and unweighted metrics

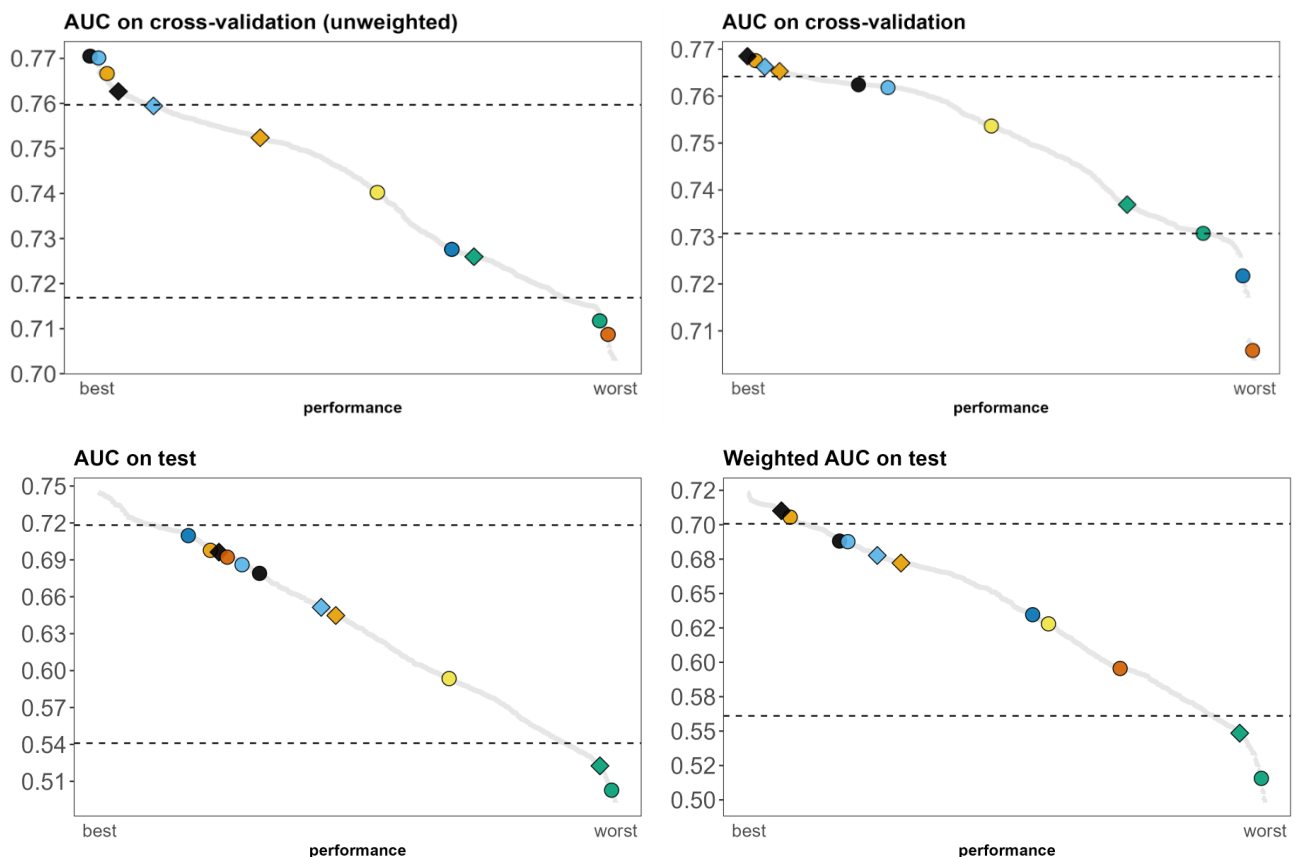
Using abundance values in the weighting somehow aligns the model performance measured with the weighted metrics with its ability to reproduce abundance patterns at presence sites and mimic the known suitability gradient.

The introduction of weighted metrics led to unexpected results that illustrate the complexity of interpreting model performance in the context of machine learning. The performances of most models change, in some cases drastically, between cross-validation and RCMED dataset, and between weighted and unweighted metrics. This change is significant as it alters the ranking of the models and thus the model selected. These inconsistencies are interpreted

Figure 16 A comparison of the AUC, evaluated on cross-validation (top) and RCMED datasets (bottom). In each plot, all 1240 model configurations are sorted from left to right, from best to worst. Candidate models are highlighted in colour (see legend to the right), while non-candidates are shown as grey dots. The dashed lines show the 10th and 90th percentile performance values calculated for the entire group of models. The plots on the left show the “unweighted” AUC values, while the weighted values are shown on the right. Note how the ranking changes between datasets and after applying the weights. This effect is particularly important when we evaluate and rank the performance of the models using the metrics. Since it is not possible to evaluate the “true” values of the metrics, we consider them all as somehow (simultaneously) informative and biased. The proposed approach is to identify an ideal model with perfect scores (PIS) and use it as a reference for ranking the candidates.

Criteria and model selected

- AICc - fc.QT_rm.1
- AUC - fc.T_rm.4.5
- AUCdiff - fc.Q_rm.20
- CBI - fc.HT_rm.6.5
- OR10p - fc.T_rm.16
- pROC - fc.HPT_rm.0.5
- SEDI - fc.T_rm.4
- ◆ wAUC - fc.H_rm.7
- ◆ wAUCdiff - fc.Q_rm.17
- ◆ wCBI - fc.H_rm.13.5
- ◆ wSEDI - fc.HT_rm.12.5



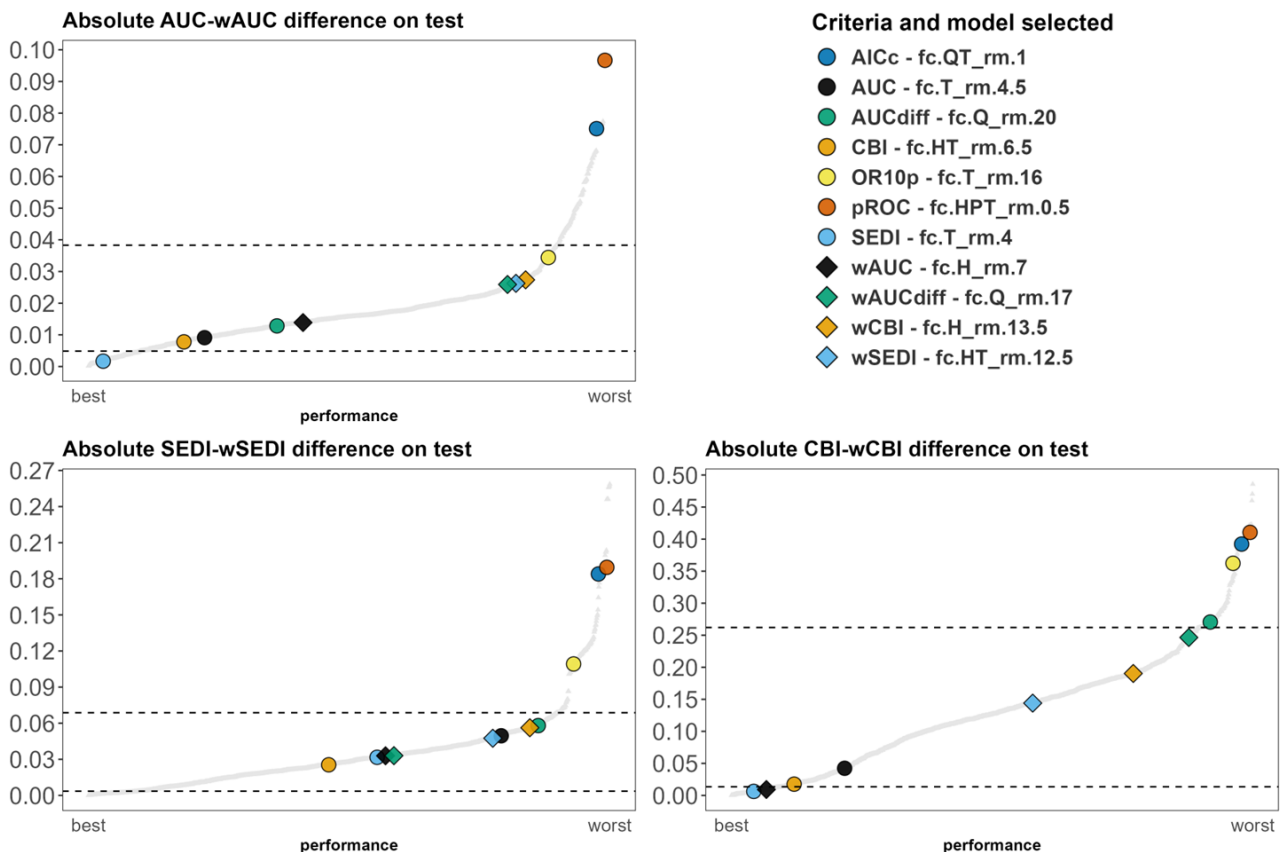
as an artificial distortion of metrics' scores due to biases in the data¹⁹. Figure 16 presents an example of the magnitude and impact of this change.

Delta metrics

As can be seen in Figures 16-17, the performance of the models in the weighted and unweighted versions of the metrics differs considerably in some cases.

From a theoretical perspective, a good model is expected to perform well on both versions, and the difference (METRIC – wMETRIC, hereafter referred to as the *delta*-metric) is an extremely informative indicator of artificially boosted performances (overfitting) and/or the strength of spatial bias.

Figure 17 Delta metrics for AUC (top-left), SEDI (bottom-left) and CBI (bottom-right). Some candidates' performance are more affected than others when considering site-weighting.



¹⁹ Here the data refer to all data, as the biases can (and probably do) affect the training and test datasets during cross-validation as well as the independent test data set.

As a matter of fact, a positive *delta* can be interpreted as a measure of overfitting, as the information we obtain is that the model performs better on the least important points (i.e. sites that are spatially clustered and/or have low abundance) than on the most informative points (i.e., sites that are spatially isolated and/or have high abundance).

Indeed, a negative *delta* can be interpreted as the model's ability to perform reasonably well on the most important test points in the set.

For the purpose of model selection, an objective way to evaluate *delta* in absolute values, avoiding strong interpretations and limiting the information to the fact that *delta* should ideally be zero.

Model selection

The main purpose of the evaluation metrics is to support model selection. During the calibration phase, a variety of models are trained and evaluated using the chosen methods to identify one or a few that are best suited to the scope of the research (Muscarella et al., 2014).

Using a single metric is not advisable as it usually measures a single model capability and, more importantly, each metric has some drawbacks and limitations (Dorji et al., 2020; Fawcett, 2006; Peterson et al., 2008; Velasco & González-Salazar, 2019).

For this reason, it is common practice to define *subjective* selection criteria that involve multiple metrics (Cobos et al., 2019; Radosavljevic & Anderson, 2014; Warren & Seifert, 2011).

A set of 13 metrics was considered in this study: the (unweighted) metrics shown in Table 4, wAUC, wCBI, wSEDI and the respective *delta* metrics.

Multi-Criteria Decision Problem approach

One advance of this research is to interpret the task of model selection as a Multi-Criteria Decision Problem (MCDP). MCDP is a theoretical framework developed in the field of *operations research* to evaluate and prioritise a limited number of alternatives based on their performance with respect to multiple, often conflicting, criteria. Typically, these problems involve identifying relevant factors, evaluating their relative importance and integrating them to select the best or a set of optimal alternatives (Taherdoost & Madanchian, 2023).

One of the most widely used approaches for solving MCDP is the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). TOPSIS is a MCDP method that ranks alternatives based on their distance from a Positive Ideal Solution (PIS), the most desirable, and a Negative Ideal Solution (NIS), the least desirable. The best alternative has the shortest distance to the ideal solution and the furthest distance to the anti-ideal solution (Hwang et al., 1993).

TOPSIS provides a coherent and mathematically sound framework through which all evaluation metrics can be interpreted together in respect to PIS but does not solve the problem of subjectivity (Chakraborty, 2022).

For this purpose, a different *operations research* method is used to estimate the objective criteria weights: CRiteria Importance Through Inter-criteria Correlation or CRITIC (Krishnan et al., 2021)

CRITIC weights the criteria on the basis of their variability (standard deviation) and the degree of conflict (correlation) between them. Criteria with higher variability and lower correlation with others are weighted higher.

To objectively evaluate the best model among the 1240 configurations tested for *C. cylindracea*, a TOPSIS ranking was performed for the subset of models (candidates)

maximising a single metric, using the objective weights estimated by CRITIC analysis. The

ranking considered both the cross-validation and the RCMED dataset.

PIS is defined here as an ideal model that has the highest score for each performance metric

in both datasets and the lowest score for the delta metrics. The reason for choosing the

smallest differences as a desirable model feature is that we prefer transferability (consistent

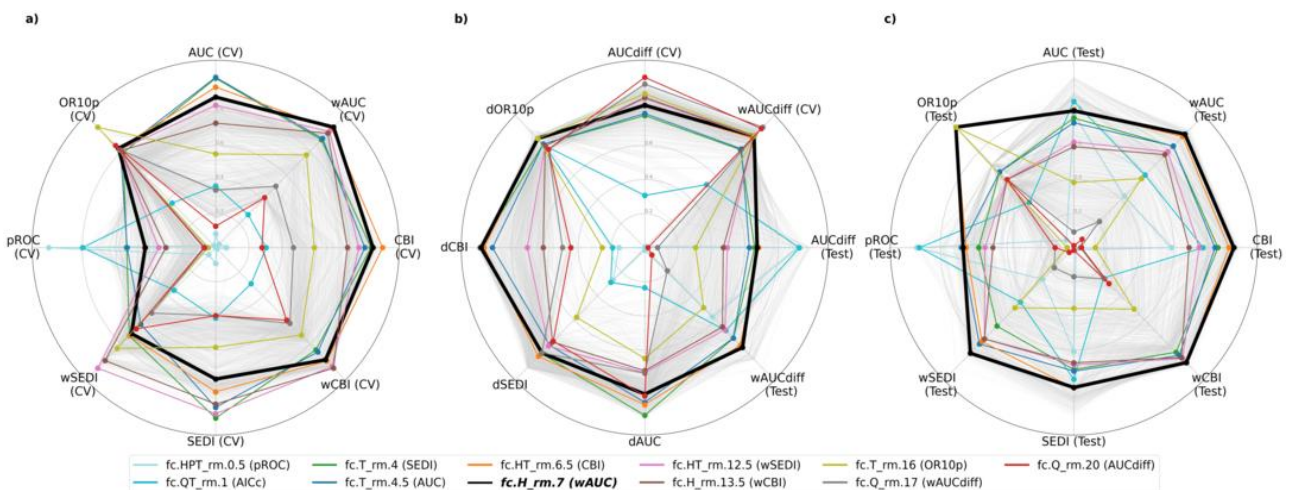
performance regardless of different weighting systems) over specialisation.

performance regardless of different weighting systems) over specialisation.

Table 5 Candidates scores for both cross-validation and RCMED test, ranked by TOPSIS score.

Name	Tune Args	TOPSIS	wAUC		pROC			wSEDI		wAUCdiff		OR10.p		AICc	wCBI	
			cv	test	cv	test	pvalue	cv	test	cv	test	cv	test		cv	test
wAUC	<i>fc.H_rm.7</i>	0.88	0.77	0.71	1.40	1.19	0.00	0.52	0.51	0.19	0.10	0.01	0.06	8413.49	0.88	0.76
CBI	<i>fc.HT_rm.6.5</i>	0.85	0.77	0.71	1.40	1.19	0.00	0.52	0.45	0.19	0.10	0.02	0.06	8359.82	0.88	0.76
AUC	<i>fc.T_rm.4.5</i>	0.74	0.76	0.69	1.45	1.19	0.00	0.50	0.46	0.20	0.12	0.03	0.07	8263.02	0.85	0.71
SEDI	<i>fc.T_rm.4</i>	0.73	0.76	0.69	1.46	1.17	0.00	0.53	0.38	0.21	0.12	0.04	0.07	8244.15	0.84	0.69
wSEDI	<i>fc.HT_rm.12.5</i>	0.68	0.77	0.68	1.36	1.16	0.00	0.61	0.44	0.20	0.14	0.01	0.09	8467.64	0.90	0.73
wCBI	<i>fc.H_rm.13.5</i>	0.60	0.77	0.67	1.36	1.15	0.00	0.59	0.44	0.20	0.14	0.01	0.09	8497.15	0.90	0.73
OR10p	<i>fc.T_rm.16</i>	0.41	0.75	0.63	1.22	1.05	0.00	0.56	0.29	0.12	0.10	0.01	0.13	8595.95	0.79	0.42
AICc	<i>fc.QT_rm.1</i>	0.38	0.72	0.63	1.58	1.25	0.00	0.41	0.26	0.37	0.17	0.09	0.09	8114.82	0.61	0.22
wAUCdiff	<i>fc.Q_rm.17</i>	0.38	0.74	0.55	1.23	1.07	0.00	0.47	0.09	0.20	0.14	0.00	0.19	8676.61	0.75	0.22
AUCdiff	<i>fc.Q_rm.20</i>	0.22	0.73	0.52	1.23	1.07	0.00	0.51	0.02	0.18	0.14	0.00	0.22	8701.19	0.74	0.25
pROC	<i>fc.HPT_rm.0.5</i>	0.21	0.71	0.60	1.68	1.25	0.00	0.32	0.15	0.50	0.23	0.19	0.11	8303.27	0.48	0.04

Figure 18 Kiviati diagram of model rankings for unweighted (a), weighted (c), AICc and delta (b) metrics. The delta metrics are calculated as follows: $dMETRIC = |METRIC_{unw} - METRIC_w|$. Coloured lines represent model candidates, while grey lines represent all other model configurations. The model selected by TOPSIS, highlighted in bold, shows the most consistent overall performance.

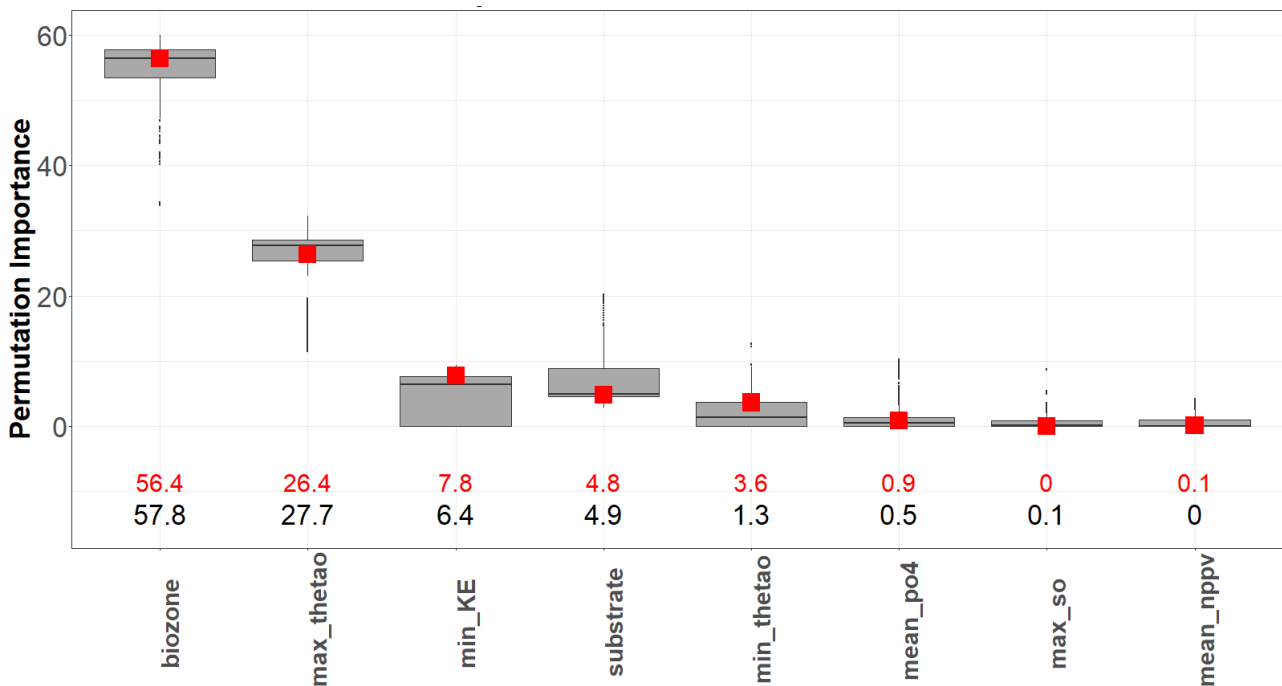


Interestingly, the model closest to the PIS, *fc.H_rm.7*, was “identified” by a weighted metric (*wAUC*). For a visual presentation of model performance, see Figure 18. A summary of the candidates’ evaluation is provided in Table 5.

Biological realism

In the context of ML-ENM, a "reality check" is essential: the statistical relationships derived from the model between the probability of occurrence and the environmental variables should be consistent with existing knowledge about the biology of the species.

Figure 19 Boxplot showing the influence of the individual variables on MaxEnt’s ability to correctly identify presence sites, as estimated from the 1240 trained models. Red squares highlight variables importance according to the best model. Best model and mean importance values are reported at the bottom of the plot. See the main text for an explanation of how the permutation importance is calculated by the MaxEnt software.



The “reality check” consist in the evaluation of *variables importance* and *response curves*.

MaxEnt estimates the importance of the variables by randomly shuffling the values for each variable in the training dataset, which contains both presence and background data. The performance of the model, as measured by AUC, is then re-evaluated using this permuted

data set. The decrease in AUC resulting from the permutation of each variable reflects the importance of the variable to the accuracy of the model. This decrease is normalised and expressed as a percentage to provide a clear measure of the contribution of each variable. Variables that cause a significant decrease in AUC when permuted are considered critical to the model and indicate strong dependence (Phillips & Dudík, 2008).

Boxplots of variables permutation's importance computed for all the 1240 fitted models are shown in Figure 19.

Biozone and (max) temperature are the most important variables, followed by currents velocity and substrate. This is reasonable, according to different authors (Bulleri et al., 2018; Caronni et al., 2021; Piazzini et al., 2016).

Response curves illustrate how the predicted probability of presence (or suitability) changes as a function of an environmental variable, with all other variables remaining at their average or fixed values. They help to visualise and interpret the influence of each environmental variable on species distribution by showing species preferences or tolerances for each environmental condition (Elith & Franklin, 2013).

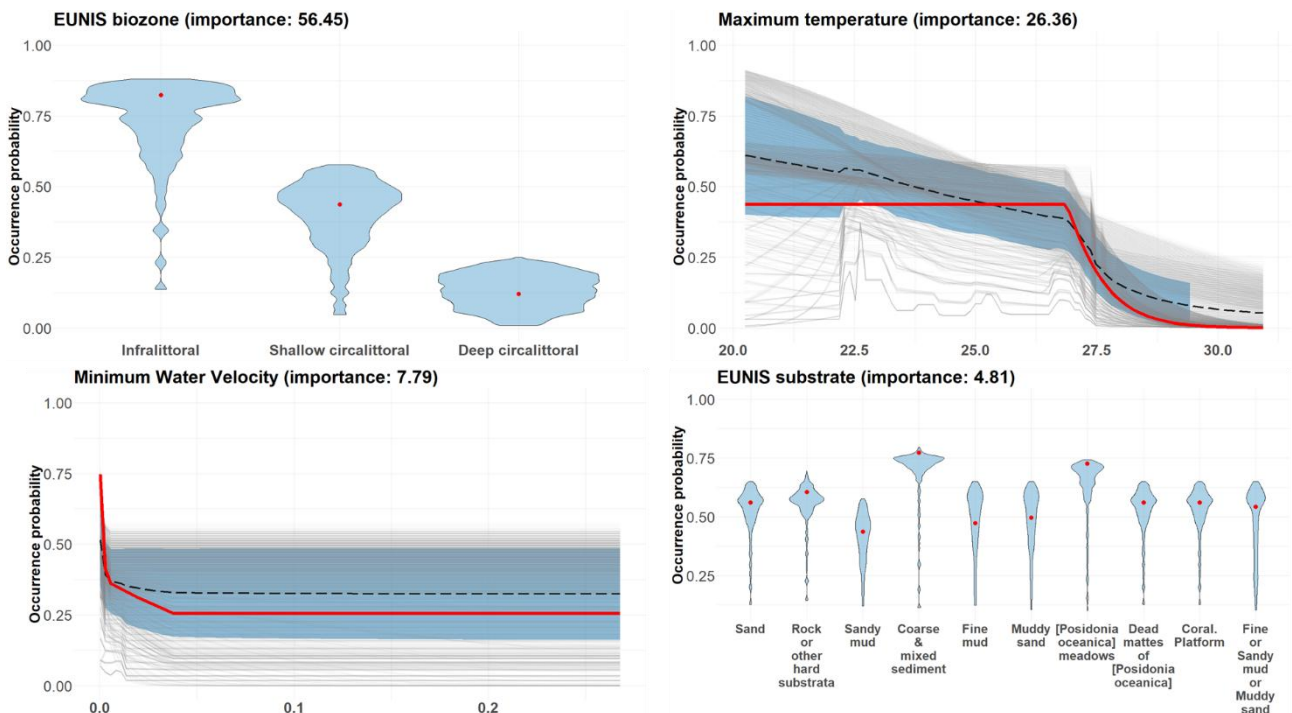
The response curves of *C. cylindracea* for the four most important variables, as estimated by the 1240 fitted models, are shown in Figure 20.

The preference for shallow waters, but the ability to survive and thrive at greater depths, is consistent with observations (Cantasano et al., 2017). The relationship with the substrate reflects the species' ability to colonise all types of sediments, with a preference for complex substrates that facilitate anchoring (Bulleri et al., 2011; Piazzini et al., 2016).

Caronni et al. (2021) investigated the influence of water motion and temperature on gametogenesis and spawning of *C. cylindracea*. The number of fertile thalli (FT) and releasing

thalli (RT) is already disturbed at a low water velocity, with high success rates only occurring at "no water movement". The study found an optimum temperature of 24 °C and a strong negative effect on the RT/FT ratio at 30 °C.

Figure 20 Response curves for biozone (topleft), maximum temperature (topright), minimum water velocity (bottomleft) and substrate (bottomright). The red lines/points signal the selected model response, grey lines are the other models. The dashed line is the "mean" response curve, light-blue area shows $\pm 1.s.d.$



It is worth noting that analysing the response curves is also useful to identify potential problems in the data and therefore in training. As discovered by Santamaría et al. (2021), a cryptic, filamentous form of *C. cylindracea* emerges at temperatures above 28 °C, which can survive and return to its original form when the stressful conditions disappear. The cryptic form has a completely different thallus, indicating a strong correlation between the probability of detection and temperature, which could undermine the validity of the estimated response and thus the predictions.

Considering the high ecological realism of the models and TOPSIS results, *fc.H_rm.7* was selected as the best model and used for the projection phase.

Spatial and temporal projections

The last phase of ENMs is to generate maps of presence-absence or suitability (Cobos et al., 2019).

The selected model was projected for the entire geographical area according to the "present" and "future" scenarios (see Defining the scope section) and for each year available in the dataset, assuming that species do not adapt to new conditions.

From continuous probability to suitability classes

MaxEnt creates maps that represent the probability of occurrence as continuous values between 0 and 1. To create presence-absence maps or suitability maps from these results, these continuous values must be classified into discrete categories. This classification usually involves the selection of threshold values. *MaxEnt* offers 11 methods for estimating thresholds and alternative, more complex methods exist in literature (Phillips et al., 2017).

For *C. cylindracea*, the threshold that maximises the sum of sensitivity (true positive rate) and specificity (true negative rate) (maxSSS; Liu et al., 2016) was used to classify the probability of occurrence into suitability.

In particular, four suitability classes are defined on the basis of the *maxSSS* value (*th*): *high suitability* (HS: $P > th$), *medium suitability* (MS: $\frac{2th}{3} < P \leq th$), *low suitability* (LS: $th \frac{th}{3} < P \leq \frac{2th}{3}$), and *no suitability* (NS: $0 < P \leq \frac{th}{3}$).

Uncertainty in classification

The suitability classification depends heavily on the threshold value (Merow et al., 2013).

MaxEnt specifies a single *maxSSS* threshold for the final model, but actually determines *k* thresholds, one for each partition used in the cross-validation process (Muscarella et al., 2014).

By extracting the *maxSSS* thresholds estimated by the fitted models for each partition, we can evaluate how the classification results change depending on the threshold variability, which provides an additional measure of the uncertainty of the predictions.

Time series analysis

Annual time series for the areas classified in each suitability category are compiled. *Mann-Kendall Trend (M-K) Test*, as implemented in Collaud Coen et al. (2020) is used to determine the presence of significant monotonic trends within the time series.

Mann-Kendall Test

The Mann-Kendall test is a non-parametric method to detect trends in a time series without assuming a specific distribution.

Starting with the null hypothesis of no trend, the test evaluates the ranks of data points to determine if there's a consistent upward or downward movement. A significant Mann-Kendall statistic indicates the presence of a trend. The test provides two outputs:

- *Kendall's Tau*: This coefficient represents the strength and direction of the trend. A positive value indicates an increasing trend, while a negative value suggests a decreasing trend.
- *Pvalue*: The probability of observing the given result if the null hypothesis of no trend is true. A small p-value (<0.05) suggests that the observed trend is statistically significant.

The Mann-Kendall test primarily discerns the direction of a trend. Kendall's tau in the test output measures both the strength and direction of the trend. A positive tau indicates an upward trend, while a negative value suggests a downward one. The closer tau is to 1 or -1, the stronger and more consistent the trend. However, while tau indicates trend consistency, it doesn't quantify the actual magnitude or rate of change in the data (Collaud Coen et al., 2020). Results are shown in Figure 25.

Extrapolation Risk Analysis

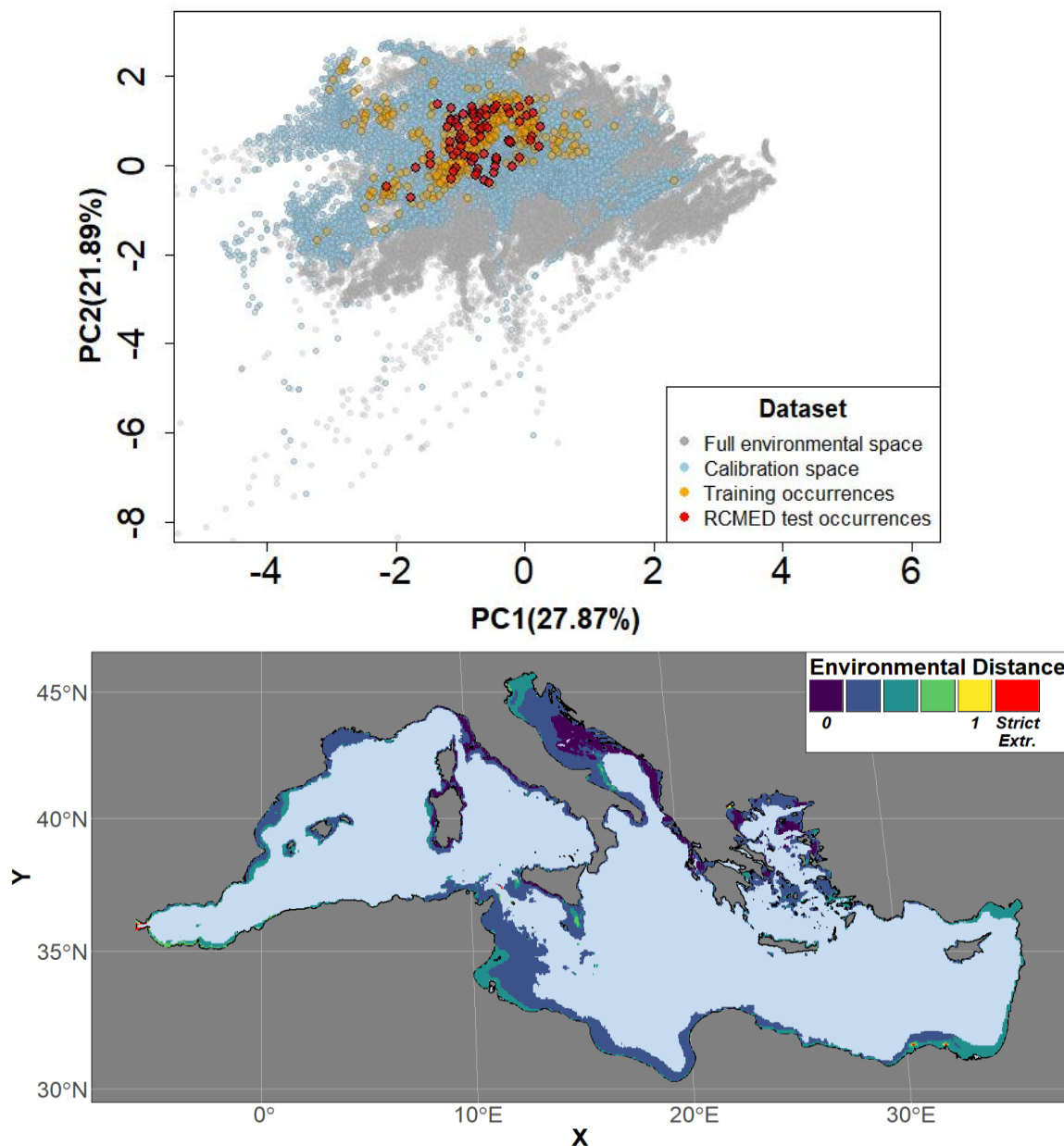
The reliability of *MaxEnt* projections depends on a fundamental assumption: the consistency of the relationships between environmental variables and species responses in both calibration and projection space. The model learns these relationships from the calibration space and applies them to make predictions. Projection involves either *interpolation* or *extrapolation*, each of which has a different degree of reliability (Cobos et al., 2019).

Interpolation takes place when the projection space contains environmental conditions that are *analogous* to those in the calibration space, so that predictions can be made within the bounds of the observed data. This scenario provides more reliable predictions as they are based on patterns that the model was trained on (Elith et al., 2006).

Conversely, extrapolation occurs when predictions need to be made for environmental conditions that are *non-analogous* to the calibration space. Extrapolation estimates species responses beyond the original range of observation based on relationships with other variables, without empirical evidence of these relationships in the new conditions (Fitzpatrick & Hargrove, 2009).

Figure 21 (right) This PCA diagram visually represents the entire environmental space (in grey) of the study area for the entire study period (2000-2050). Notice how the calibration space, shown in light blue, “surrounds” the (most) training occurrences (orange) and the entirety of the RCMED test occurrences (red). This ensures that the model has had the opportunity to explore and evaluate the surrounding space of all presence points during training. We can be confident that 1) the evaluation metrics in the RCMED dataset were obtained by interpolation and that 2) ecological gradients controlling the presence probability were available during training. In fact, only a few training occurrences are located at the edge of the calibration space. Presences at the edge of the calibration space leads to some uncertainty in the response curves as the models do not “know” what happens beyond these ecological coordinates.

NOTE: Visually, the strict extrapolation can be thought of as predicting the probability of presence for a point outside the light blue area: The greater the distance, the less confidence we can have in the quality of the prediction, as we expect the statistical relationships between the variables to be increasingly different from those estimated by the model.



To assess whether the predictions are interpolation or extrapolation, researchers use techniques such as Multivariate Environmental Similarity Surfaces (MESS; Phillips & Dudík, 2008) or Principal Component Analysis.

These methods compare the distributions of environmental variables between calibration and projection spaces and identify areas where predictions may be less reliable due to novel conditions. This understanding is critical for assessing model reliability and determining appropriate applications for the projections (Elith & Franklin, 2013).

The pivotal research of Owens et al. (2013) established the conditions for the projection of *MaxEnt* onto a new environmental space by introducing the concept of *Mobility Oriented Parity* (MOP).

MOP is designed to quantify the environmental similarity between calibration and projection regions and helps to identify areas of *strict extrapolation*²⁰. It improves MESS by using multivariate distances (e.g. Mahalanobis or Euclidean) to assess how closely the environmental conditions in the projection space resemble the closest part of the environmental conditions in the calibration space.

In practise, the environmental conditions represented in the calibration region must cover the full range of conditions observed in the occurrences to reduce *strict extrapolation*.

In the methodology presented in this study, MOP analysis entails a classification of the Euclidean distance between calibration and projection space, which serves as an index of extrapolation risk (Figure 21bottom). The thresholds for classification were determined using the percentiles of a chi-squared distribution fitted to the degrees of freedom of the environmental variables. In particular, the environmental distances were categorised into

²⁰ Strict extrapolation refers to the process of applying an ENM to environmental conditions that are completely outside the range of values used in the model calibration. This type of extrapolation occurs when models are forced to predict species suitability in environments that were not part of the original training data, often resulting in biologically unrealistic predictions.

quintiles, with all distances exceeding the 90th percentile of the chi-squared distribution being considered as *strict extrapolation*.

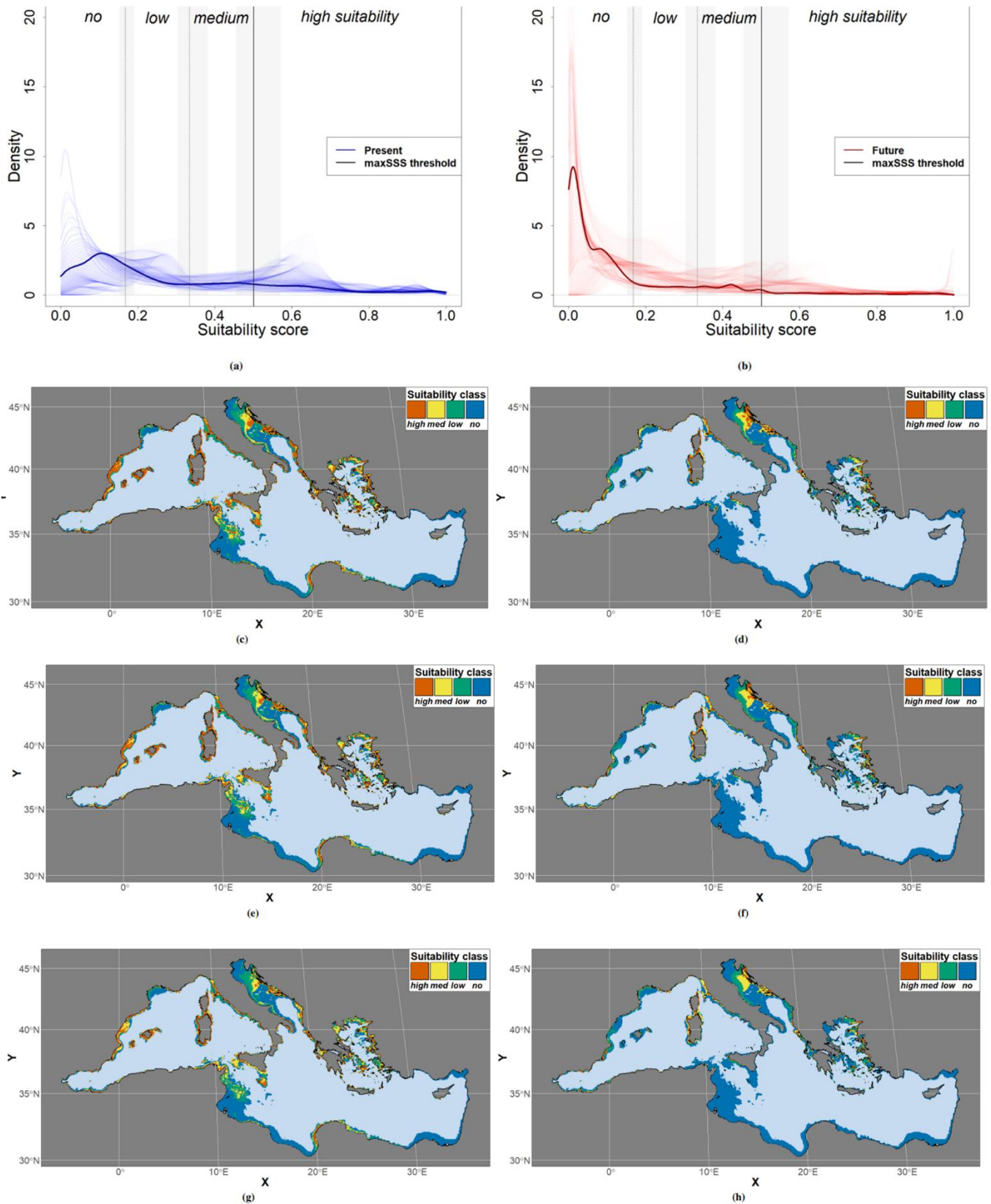
The environmental conditions in the cross-validation scheme must envelop the environmental conditions of RCMED dataset. By taking this precaution, we can ensure that the results of the model for the RCMED dataset result from interpolation of known conditions rather than extrapolation into unknown environmental spaces. This ensures that the model recognizes reasonable environmental combinations and avoids strict extrapolations when projecting to the test sites, thereby avoiding artificial results. See Figure 21top for a visual representation of calibration and projection spaces.

Analysis of the results

The training process provided the models a sufficiently large calibration and validation context to evaluate the independent dataset mainly by interpolation rather than extrapolation, which strengthens the validity of the results. The proportion of cells classified as 'strict extrapolation' is 0.21% of the total. See Figure 21 for a visual representation of the calibration, validation and test environmental space and a semi-quantitative map of uncertainty of prediction based on MOP analysis.

The probability density functions (PDFs) for predicted occurrence probability for both present (Figure 22a) and future (Figure 22b) scenarios, as determined by each of the 1240 fitted models, show a general shift towards an increase in lower probabilities of occurrence at the expense of medium to high probability values in the future. Figure 22c-h illustrates the effect of *maxSSS* threshold uncertainty. The maps show the different suitability classifications for the present (left) and the future (right) according to *maxSSS_{min}* (c-d), *maxSSS* (e-f) and *maxSSS_{max}* (g-h) thresholds.

Figure 22 Probability Density Functions for the predicted probability of occurrence by the 1240 different MaxEnt models for present (a) and future (b) conditions. Thick lines highlight the selected model. Vertical lines represent maxSSS based thresholds used to classify probability of occurrence into suitability. Grey shaded areas indicate threshold range. Suitability maps show how introducing thresholds' uncertainty affects the present(left) and future (right) predictions. "worst scenario" (c-d), using: maxSSSmin; "normal scenario" (e-f), using: maxSSS; "best scenario" (g-h), using: maxSSSmax.

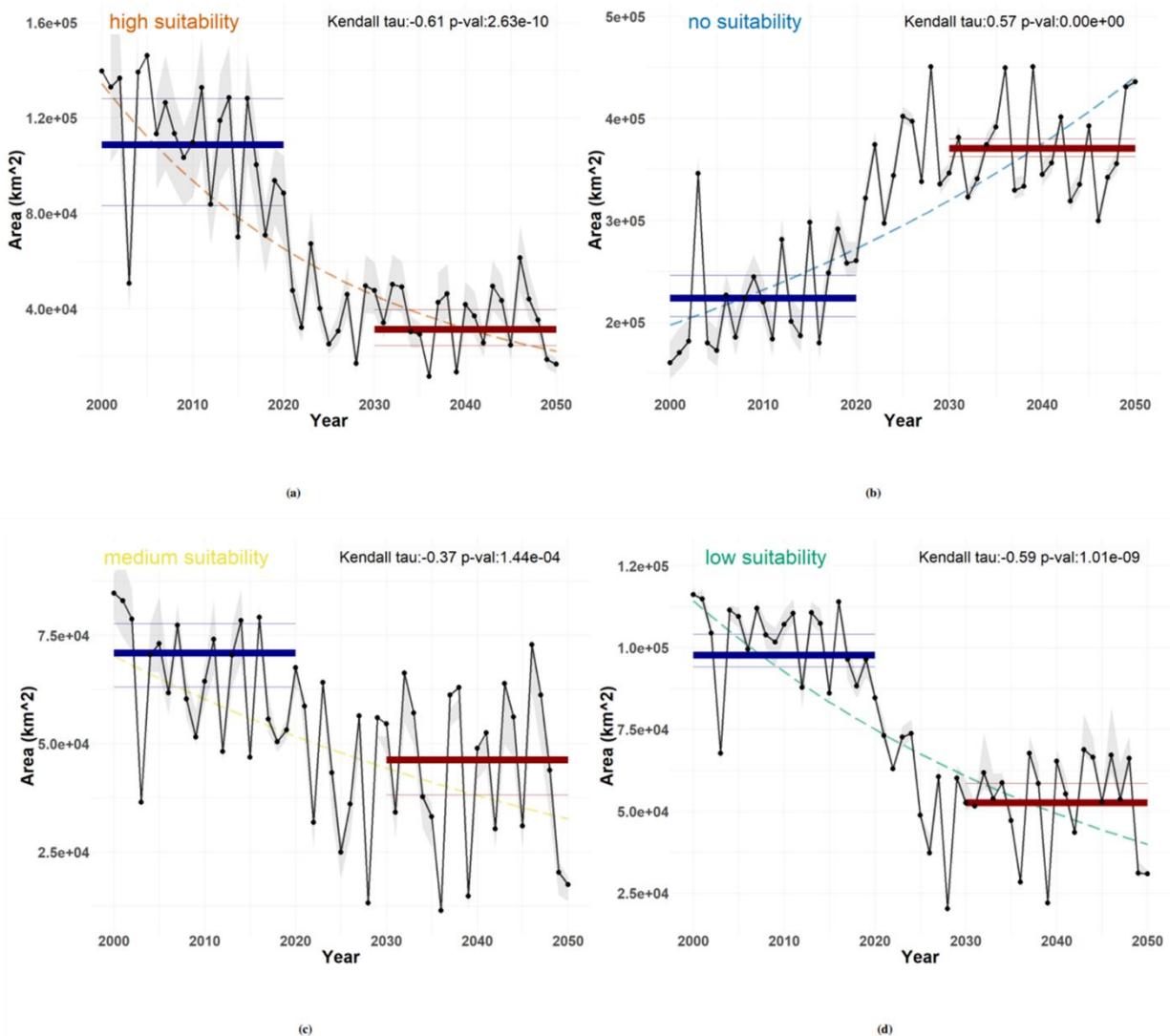


The predicted suitable areas (expressed in km^2) decrease in the future both in the scenarios and in the annual projections regardless of the threshold used.

The time series analysis of the annual projected suitability area shows a significant negative trend for *HS*, *MS* and *LS* while *NS* shows a significant positive trend. See figure 23.

Changes in the threshold can have a significant impact on the classification results if the peaks of the probability density have similar values or if the threshold has a high degree of uncertainty (Figure 22a-c and Figure 23a). This effect can be seen in Figure 22 when comparing c,g suitability maps. This underlines the need for a clear assessment of

Figure 23 Area (km^2) for each class as predicted by projecting the model to each available year. The grey area shows the variability associated with threshold uncertainty. Red and blue horizontal lines indicate the area (thick) predicted for present and future.



uncertainty in the classification process, to provide additional insights for a better informed decision-making process.

These results indicate that further spread of the species is likely to be limited, despite some fluctuations in the currently suitable areas. The smooth annual oscillation is reminiscent of the "persistence phase" for the "natural fluctuation" invasion model, as defined in C. F. Boudouresque (1999).

Conclusion

A robust ENM process

In this study, an optimised *MaxEnt* model was used to predict the current and future habitat suitability of *C. cylindracea* in the Mediterranean Sea using IPCC-AR5-RCP8.5 projections with a fine spatial and temporal resolution. In line with the work of Moreno-Amat et al. (2015) and Radosavljevic & Anderson (2014), our approach emphasises the treatment of overfitting, sampling bias and spurious results over pure model performance. The use of using site weights in the calculation of performance metrics should become a standard analysis in the ENM practice to amplify the ecological signal and better evaluate the models.

The availability of abundance measures for the independent test set enabled a further step in model ranking, revealing the triviality of performance metrics in the context of presence-only regressive methods. The TOPSIS analysis provided a robust and unbiased method for determining the best performing model among the candidates. Model selection should be treated as a Multi-Criteria Decision Problem with a clear definition of the ideal solution, as individual metrics capture only partial and biased information.

Sensitivity analysis to threshold variability provides a nuanced measure of uncertainty in the results and underscores the critical importance of careful threshold evaluation and selection for management and restoration purposes.

The software to reproduce this study (based on R programming language) is available on request and will be freely available on publication.

Insights on invasion spread

In our study, the annual changes in predicted suitability for *C. cylindracea* in the Mediterranean were examined to gain insights into the spread of the invasion.

Although suitability shows a negative trend due to the adverse effects of the rising temperature, sexual reproduction can give rise to new genotypes that favour the invasion profile, as noted by Caronni et al. (2021).

The analysis showed that the Adriatic Sea is a new hotspot for the species. This observation is consistent with the results of Iveša et al. (2015), which investigated the *C. cylindracea* spread in this region between 2004 and 2014.

In summary, our dynamic suitability predictions shed valuable light on the potential spread of *C. cylindracea* and expand the possibilities for utilising ENM results. The comprehensive methodological analysis sets a standard for future research in model evaluation and selection in the context of machine learning-based ENM techniques.

Limitations

Invasion dynamics in the real world often involve complex relationships that go beyond mere suitability and are often neglected in correlative ENMs. Factors such as the adaptability of invasive species, evolutionary changes and the intricate role of human activities not only alter

the invasion trajectory, but also distort the patterns in the data we use to train our models to an unknown degree. An example was presented in “*Biological Realism*” section, regarding the cryptic form of *C. cylindracea*.

Further research is needed to identify and manage the nature and strength of biases in the data and to find better ways to select models that are less prone to overfitting problems. In addition, the critical role of thresholds in shaping outcomes requires comprehensive and refined procedures for selecting thresholds and assessing uncertainty so that decision makers can take more informed action to limit species invasion risk.

A mechanistic model to explicitly assess exposure to AIS secondary invasion

Introduction

The Mediterranean Sea is increasingly threatened by AIS, with *Caulerpa cylindracea* proving to be one of the most aggressive and ecologically disruptive invaders (Piazzini et al., 2016).

Predicting the current and future spread of AIS species is a multifaceted challenge that goes beyond identifying suitable habitats. It requires a comprehensive understanding of both the ecological factors and the mechanisms by which these species spread, especially in the context of human activities. ML-ENMs have greatly improved our ability to map potential distribution areas by focussing on abiotic and biotic factors. However, they often overlook the likelihood of reaching these areas, highlighting a critical gap in our predictive capabilities.

This gap emphasises the importance of incorporating movement constraints, the "M" component of the Biotic-Abiotic Movement (BAM) framework (Soberon & Peterson, 2005). The BAM framework states that the distribution of a species is determined not only by the abiotic environment (A) and biotic interactions (B), but also by its ability to reach suitable regions (M). In this work, we have investigated the intricate vectors and pathways of AIS invasions, and highlighted how *C. cylindracea* may spread due to human-mediated dispersal linked to shipping traffic using a particular spatial interactions class of models, namely *gravity models*.

C. cylindracea spreads naturally by vegetative fragmentation, allowing it to colonise neighbouring areas efficiently (Bulleri et al., 2018). However, its rapid spread in the Mediterranean region has been greatly facilitated by human activities (Cantasano et al., 2017). Ships not only inadvertently transport fragments of the alga attached to hulls or anchors, but also directly contribute to fragmentation when travelling over *C. cylindracea* meadows (Creese et al., 2004).

Through this complexity, the link between our refined ENM process and mechanistic models provides a more holistic and accurate assessment of invasion risks. This integration is crucial for species such as *C. cylindracea*, where the interplay between ecological suitability and dispersal mechanisms determines the invasion trajectory (Piazzi et al., 2016).

An improved gravity model to estimate probability of invasion mediated by vessel traffic

The framework presented here builds on the gravity model developed by Drake & Lodge (2004) and introduces several important advances that address the limitations of the original approach by focussing modelling efforts on the AIS-specific biological aspects and ecological dynamics of vessel-mediated invasions.

As the authors note in their conclusions, the likelihood of invasion per vessel is the true driver of invasion risk, and there is a large gap in characterising the destinations' *exposure* to invasion.

To address these gaps and investigate in depth the role of invasion probability per vessel, the gravity model was revised and enriched to integrate biological and anthropic factors to characterise the attributes of origin (V_i), destination (W_j), hereafter defined as *Origin Infective Potential*, *Destination Establishment Potential*, respectively.

In this study, the factors shaping origin and destination attributes are *substrate type* and *habitat suitability*.

Substrate type is one of the main factors that regulates fragment production. Sandy substrates are able to generate an order of magnitude more fragments than rocky substrates (Creese et al., 2004), and the geometrical complexity of the substrate interacts with the capacity of anchors and hulls to harvest fragments (A. M. West et al., 2016).

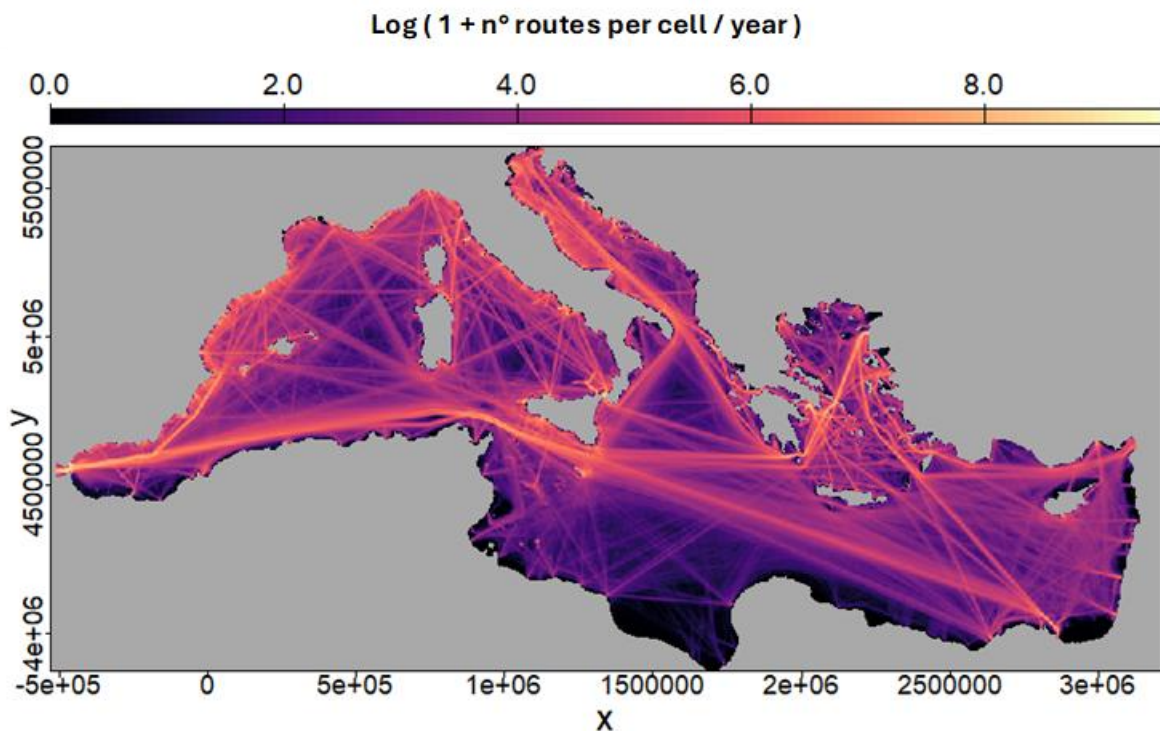
Habitat suitability indicates the likelihood that a species can survive and thrive in a particular location based on environmental conditions. At the place of origin, a higher habitat suitability correlates with greater species fitness and the availability of reproductive material (Uyà et al., 2018). At the place of destination, it influences the potential for survival and the successful establishment (Bulleri et al., 2018).

By integrating these mechanistic models with the correlative ENMs, we comprehensively address each component of the BAM framework. This ensures that our predictions reflect not only where the species can survive and thrive, but more importantly, where (and how) it can physically spread. This comprehensive framework allows us to produce detailed maps of

exposure as a likelihood of invasion through human-mediated dispersal (Soberon & Peterson, 2005).

The importance of this integrated approach lies in its ability to inform management and mitigation strategies. The identification of high-risk areas for vessel-mediated introduction can, for example, lead to stricter regulations for hull cleaning and anchoring in these areas (E. J. West et al., 2007).

Figure 24 Visual representation of average annual route density for 2020 retrieved from EMODnet. Logarithm values to enhance contrasts.



Data and methods

In this section, we will describe the data used as inputs and the methodology implemented to estimate the expected invasions per year.

Data

Traffic data for the period 2020-2023, namely *route density* and *vessel density* yearly averages for all vessel types were retrieved from EMODnet Data Services

(<https://emodnet.ec.europa.eu/en>) at a spatial resolution of 1km x 1km. See Figure 24 for a visual representation of the route density.

Route density data are expressed as raster data in which each cell's value represent the number of routes that crosses the cell in a year (EMSA, 2019). They are used to estimate realistic vessel paths across the network.

Vessel density data are expressed as raster data in which each cell's value represent the monthly average hours spent by vessels for a determinate year (European Commission. Directorate General for Maritime Affairs and Fisheries., 2023).

The *Species occurrence points*, the *Habitat suitability* maps for the period (2020-203) and *EUNIS substrate* map are taken from the results of the work presented in the previous chapter. These are two categorical rasters. Table 6 provides a comprehensive overview of the classes, and the weights assigned to each class.

Table 6 Suitability and Substrate classes. We assigned a numerical value to each, based on relevant studies on *C. cylindracea* fragmentation and colonisation processes. A brief explanation of the rationale is provided.

Classification	Category	Weight	Rationale
Habitat Suitability Classes	No Suitability	0.05	Accounts for uncertainty in the suitability estimation
	Low Suitability	0.33	–
	Medium Suitability	0.66	–
	High Suitability	0.95	Accounts for uncertainty in the suitability estimation
Substrate Types (estimated from: Creese et al., 2004; West et al., 2007; Bulleri et al., 2017)	Seabed	0.5	Base value when no information is available on the substrate
	Sand	1.0	Higher fragment generation in sand anchors
	Rock or other hard substrata	0.2	Lower fragment generation, even at high stress
	Sandy mud	0.7	Moderate impact combining sand and mud properties
	Coarse & mixed sediment	0.65	Discrete potential for fragment generation
	Fine mud	0.4	Reduced generation due to fragment trapping
	Muddy sand	0.55	Slightly higher fragment capacity than fine mud
	Posidonia oceanica meadows	0.1	Isolated specimen in the meadows
	Dead mattes of P. oceanica	0.8	Suitable place for the species to thrive
	Coralligenous platforms	0	Low production and reduced fragments creation from hard structures
Fine mud/Sandy mud/Muddy sand	0.75	Mixed positive effects	

The data processing involved the homogenisation of raster data, obtained by resampling traffic data, and converting *Vessel Density* to *Vessels Per Cell*. This is an approximation of the number of ships passing through the cell in a year. The methodology is provided in the technical report of the product (EMODnet, 2023).

Based on findings from literature on *C. cylindracea*, *habitat suitability* and *substrate* were converted into numerical values, symbolising the *importance* of each class in the process of fragments formation and colonisation.

A synthetic list of weights, parameters used as input and related references is provided in Table 6.

Methods

Despite the similarity in the approach, the network we are trying to model is completely different from that of Drake and Lodge (2004). In this case, we want to estimate the *directed flow from the origins* (cells in which the species is present) to *destinations* (valid cells of our domain in which vessels passes through).

F_{ij} takes the form of the so called *Production-Constrained Gravity Model* (K. Haynes & Fotheringham, 1985):

$$F_{ij} = A_i O_i D_j f(d_{ij}) \quad (7)$$

Where:

- $A_i = \left(\sum_j D_j f(d_{ij}) \right)^{-1}$ (8)

- O_i is the total *production* from origins (= *Vessels Per Cell*)

- D_j is the *attractiveness* factor of destinations (= *Vessels per Cell*)
- $f(d_{ij})$ is the *deterrence* function

A_i is the production constraint term ensuring that the sum of the inflows is equal to the sum of all outflows (Fotheringham & O’Kelly, 1988).

The *deterrence* function $f(d_{ij})$, following Drake & Lodge (2004) is built by fitting a multimodal gamma distribution (Young et al., 2019) to mean route lengths.

To estimate the route lengths, local maxima route density along the coast are defined as *ports*. The least distance between each pair of *ports* is computed using the inverse of the route density as “resistance” surface according to the methodology defined by Etten (2017). In this way, realistic routes are “simulated” and measured.

The formula for the probability of invasion is modified to consider ecological aspects of the invasion process:

$$\pi_{ij} = 1 - (1 - p_{ij})^{F_{ij}} \quad (9)$$

- p_{ij} is the per-ship probability of initiating an invasion, depending on *origin* and *destination* characteristics
- F_{ij} is the number of estimated *ij flow*

In particular:

$$p_{ij} = p_0 * OIP_i * (1 + DEP_j) \quad (10)$$

$$OIP_i = H_i^\alpha * S_i^\beta * \omega_i \quad (11)$$

$$DEP_j = H_j^\gamma * S_j^\delta * \omega_j \quad (12)$$

Where:

- OIP_i is the *Origin Invasive Potential*
- DEP_j is the *Destination Establishment Potential*
- H is the *Habitat Suitability*
- S is the *Substrate Type*
- $\alpha, \beta, \gamma, \delta$ regulate the impact
- ω are an *efficacy factors*

ω assumes a particular importance from the point of view of *risk* assessment because it could be interpreted as a vessel-type specific coefficient and/or the efficacy of biosecurity protocols.

A sensitivity analysis is performed on $\alpha, \beta, \gamma, \delta, \omega$ to explore the impact of each parameter.

Values tested for $\alpha, \beta, \gamma, \delta$ are 1.0 and 1.5 while values for ω are 0.5 and 1.0, resulting in 64 simulations for each year.

Getis-Ord Gi statistics are used to spatially identify (and cluster) hotspots of probability of invasion and test whether they change across the different simulations and years (Bivand & Wong, 2018).

Results and discussion

Deterrence function

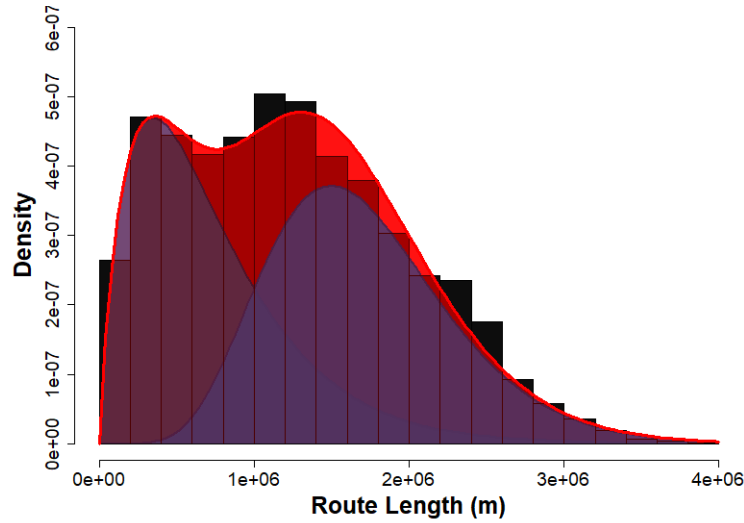
The analysis of route data identified 2188 ports. To estimate a representative distribution of distances, the pairwise shortest path between the identified harbours was calculated. The probability density of the distance from port to port and the fitted bimodal distribution used as the deterrence function are shown in Figure 25. Parameters estimated for the 2-modes gamma distribution: *Shape* parameters: 1.89 and 8.14; *Scale* parameters: 387606 and 209673; *Mixture* coefficient: 0.47 and 0.53. For details on the parametrisation and fitting process, see Young et al. (2019).

The deterrence function enables a modulation of the traffic flow to the origin-arrival distance that depends on two types of routes. In this formulation, it shapes the response of the model so that short and medium distances (up to 1000 km) and long distances (>1000 km) are taken into account.

The reason is to incorporate the effect of the transport vector in “shielding” the fragments/propagules from ecological barriers, including distance. Fragments and propagules of *C. cylindracea* can survive and remain viable for days, allowing them to travel hundreds if not thousands of kilometres (Cabrini et al., 2019; Creese et al., 2004).

Since the estimation of the gamma distribution is based on the route density, it implicitly includes the traffic volume and offers an improved risk-oriented definition of shipping traffic. The disadvantage is that vessel types are not taken into account, which is likely to lead to an underrepresentation of important vectors such as pleasure crafts, which account for only a small proportion of total vessel traffic but may play an important role in the spread of *Caulerpaceae* (E. J. West et al., 2007).

Figure 25 Probability density of port-to-port distance (route length, in meters). Red area shows the fitted multimodal gamma distribution ($k=2$). Blue areas represent the mixed gamma distributions. Shape parameters: 1.89 and 8.14; Scale parameters: 387606 and 209673; Mixture coefficient: 0.47 and 0.53. For details on the parametrisation and fitting process, see Young et al. (2019). The distribution is used for the deterrence function in the production-constrained model.



Per-ship probability of initiating an invasion

The analysis of variance (ANOVA) was performed to determine the influence of various ecological and anthropogenic factors on the per-ship probability of initiating an invasion p_{ij} .

The response variable p_{ij} , was log-transformed to reduce heteroscedasticity.

Table 7 contains the results of the ANOVA analysis, in which all the significant terms are listed, sorted by F-values.

The per-vessel efficacy factors at the origin (ω_i) and destination (ω_j) had the largest effects on the probability of invasion, with F-values of 185,194.41 and 12,876.94, respectively.

The strong effect of vessel efficacy factors emphasizes the impact that biosecurity protocols can have in managing invasion risk. A higher efficacy in transporting fragments is the most important determinant of invasion probability, which is consistent with the work of Drake & Lodge (2004).

Table 7 ANOVA results for the significant terms, using mean p_{ij} as response and testing all single factors and their interactions.

	Sum Sq	Df	F-value	Pr(F)
ω_i	3039.00	1.00	185194.41	0.00
ω_j	211.31	1.00	12876.94	0.00
Year	147.05	3.00	2987.04	0.00
β	20.22	1.00	1232.22	0.00
α	3.87	1.00	235.78	0.00
δ	3.00	1.00	182.99	0.00
γ	1.74	1.00	105.98	0.00
$\alpha : \beta$	1.41	1.00	86.03	0.00
$\alpha : \text{Year}$	4.09	3.00	83.07	0.00
$\beta : \delta$	0.19	1.00	11.44	0.00
$\alpha : \delta$	0.11	1.00	6.69	0.01
$\delta : \text{Year}$	0.26	3.00	5.21	0.00
$\alpha : \gamma$	0.08	1.00	4.68	0.03
$\gamma : \text{Year}$	0.15	3.00	2.95	0.03

Among the ecological parameters, the coefficients for substrate type at origin (β) and destination (δ) as well as habitat suitability at origin (α) and destination (γ) were significant terms. In particular, β had a pronounced effect ($F = 1,232.22$), emphasising the role of substrate in fragment production and dispersal potential.

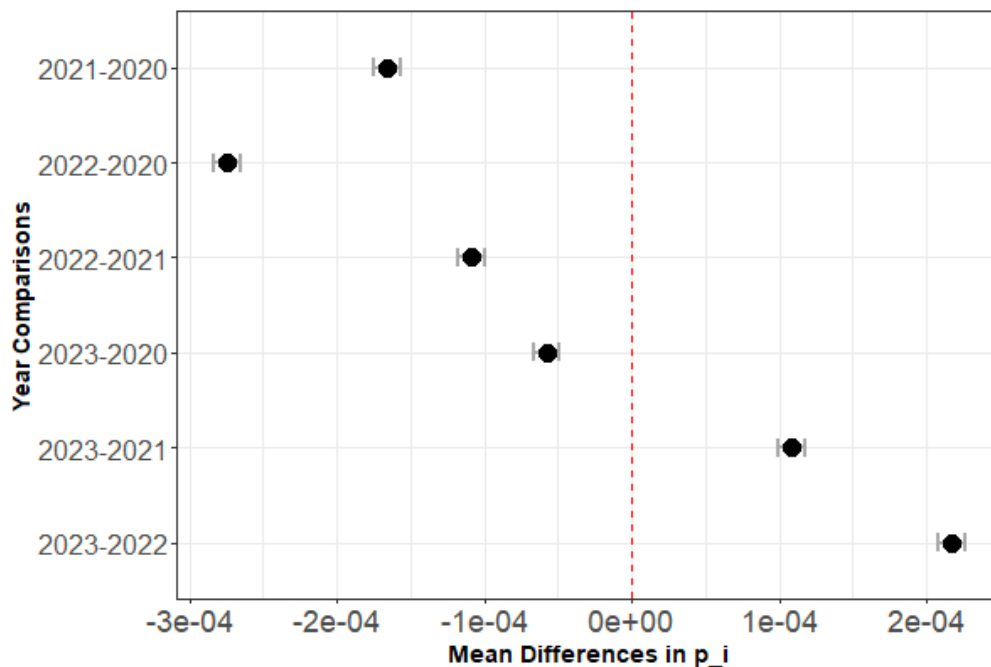
Several two-way interactions were significant, suggesting that the effects of ecological parameters are not completely independent. The interaction between habitat suitability and substrate type at origin was significant, indicating that the influence ($F = 86.03$) of habitat suitability on p_{ij} is moderated by substrate type. The interaction between habitat suitability at origin and year was also significant, indicating temporal changes ($F = 83.07$) in how habitat suitability shapes invasion risk.

The significant influence of substrate type at origin underscores the importance of suitable substrates in fragment production. Sandy substrates, which are associated with more fragments than rocky ones (Creese et al., 2004), contribute to higher invasion probabilities.

The interaction between habitat suitability and substrate type at origin suggests that optimal environmental conditions combined with favourable substrates increase p_{ij} .

The effect of year was also significant ($F = 2,987.04$). The temporal variation captured by the significant year effect and its interactions with habitat suitability and substrate type suggest that changes over time, possibly due to environmental fluctuations or increased vessel activity, shape the invasion dynamics.

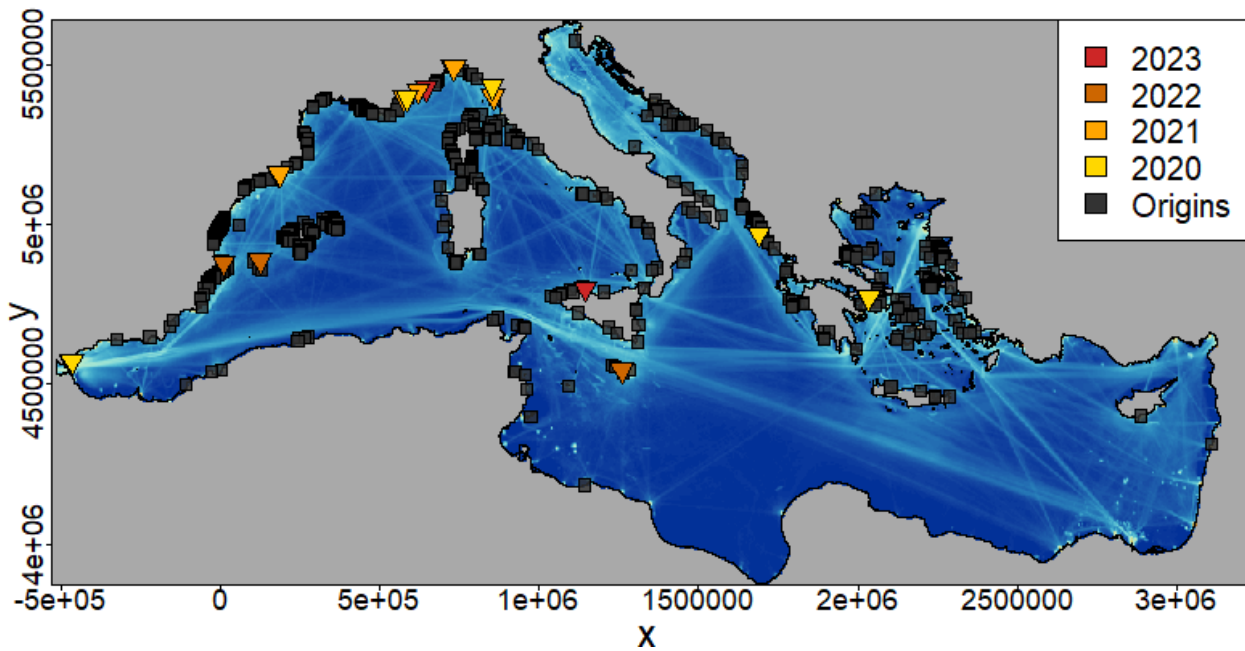
Figure 26 Tukey's Honest Significant Difference plot. Grey bars report the 95% confidence interval. 2020 is used as the baseline.



To further investigate the temporal dynamics, post-hoc pairwise comparisons were performed using the Tukey Honest Significant Difference (HSD) test. The results (Figure 26) show that all pairwise year comparisons are statistically significant ($p < 0.001$) and reveal differences in p_{ij} across the years analysed.

These findings indicate that average p_{ij} of *C. cylindracea* initially decreased after 2020 but gradually recover in 2023. The pattern likely mirrors the temporal trend in habitat suitability (Figure 23a).

Figure 27 Maps reporting *C. cylindracea* presence points used as origins (black), and cells showing $\pi > 1$ (coloured). Colors indicate the first year in which $\pi > 1$ occurs. The layer underneath highlights major routes.



Expected invasions per year

For each year, 64 independent simulations were carried out to test the effects of the different parameter combinations. The invasion probability π_j , as the route density, has an extremely skewed distribution (see Figure 28left) with 99% of the values below than 0.12 and a very long tail (max value is 10.36).

The expected number of invasions per year seems to generally increase with time, regardless of the parameterisations. This pattern is interesting since p_{ij} is lower than the 2020 baseline, showing that the impact of increasing vessel density (from the 2020 density: +2.3%, +6.5%,

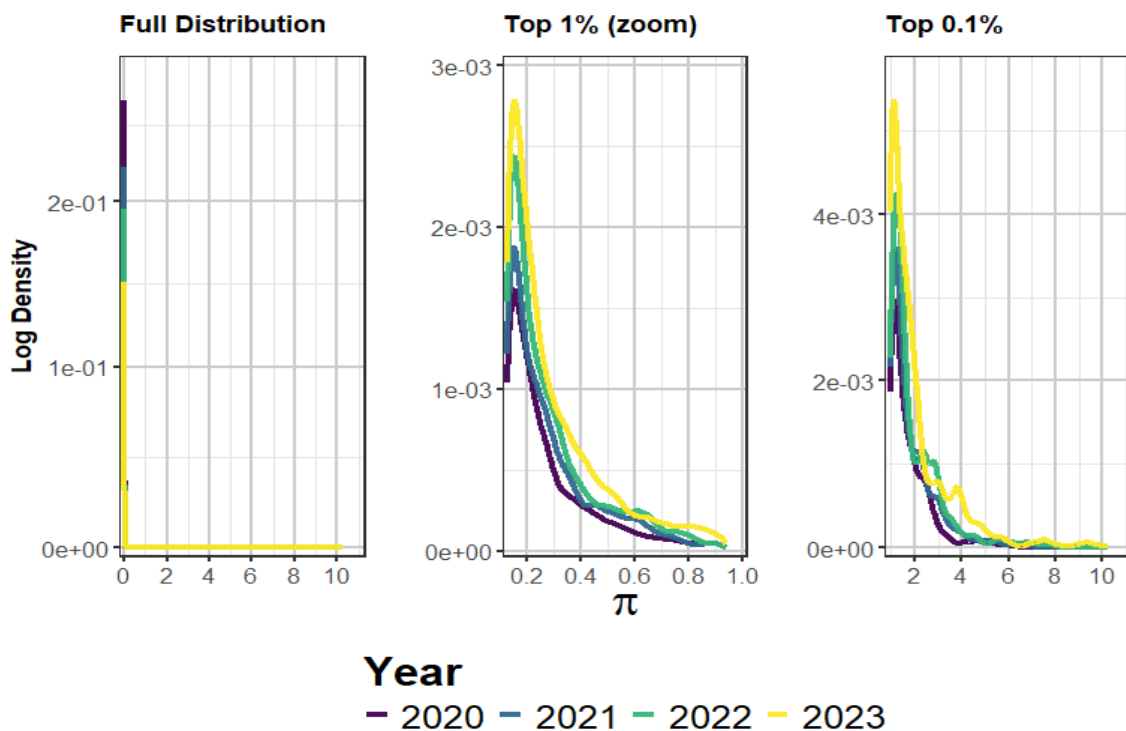
+6.7%, respectively) can overcome the impact of decreasing suitability (see Figure 23a) over the years.

Year	Mean \pm SD	Minimum	Maximum
2020	21.34 \pm 2.61	6	39
2021	27.47 \pm 2.81	12	45
2022	33.67 \pm 3.29	18	51
2023	46.16 \pm 5.46	21	88

Table 8 Summary statistics for the estimated number of cells with $\pi > 1$ per year. The simulations show great variability but all agrees in a general increase in the invasion probability.

Table 8 provides a statistical overview of how the number of cells with $\pi > 1$ varies through simulations and with time while Figure 27 shows the position of origins and cells that obtained $\pi > 1$ in every simulation. Interesting insights comes from observing the probability density function of π in relation with time:

Figure 28 Estimated density of invasions per year. Observing the full distribution is not feasible. 99th percentile value is 0.12 and max value 10.26. Plot in the middle represents highlights the range of top 1%-0.1% values, while the plot on the right shows top 0.1% values. At high π -values, the peak values of invasion density per year appear to shift to the right, indicating an increasing risk of invasion with time.



Higher values increase more than lower ones. In Figure 28centre and Figure 28right density peaks progressively shift rightward indicating a higher incidence of higher values.

Spatial patterns of invasion

To ensure reliable identification of areas acting as hotspots of invasion, only those cells identified as hotspots by the Gi statistics in all 64 simulations within the same year are considered. We then overlaid the annual hotspot maps to analyse temporal fluctuations and trends in invasion patterns. This enabled the identification of persistent invasion hotspots and provided valuable insights into the spatio-temporal dynamics of invasion processes by highlighting areas of persistently high invasion risk and their changes over time. The fact that the hotspots are becoming more evenly distributed over time and the risk of invasions is spreading to an ever larger region is a major cause for concern. The species, through subsequent successful invasions, can naturally expand and invade coastlines in between two hotspots from multiple sides.

See Figure 29 for a detailed visualisation of the hotspots in the Strait of Sicily, representing hotspots in time and locations of estimated invasions events.

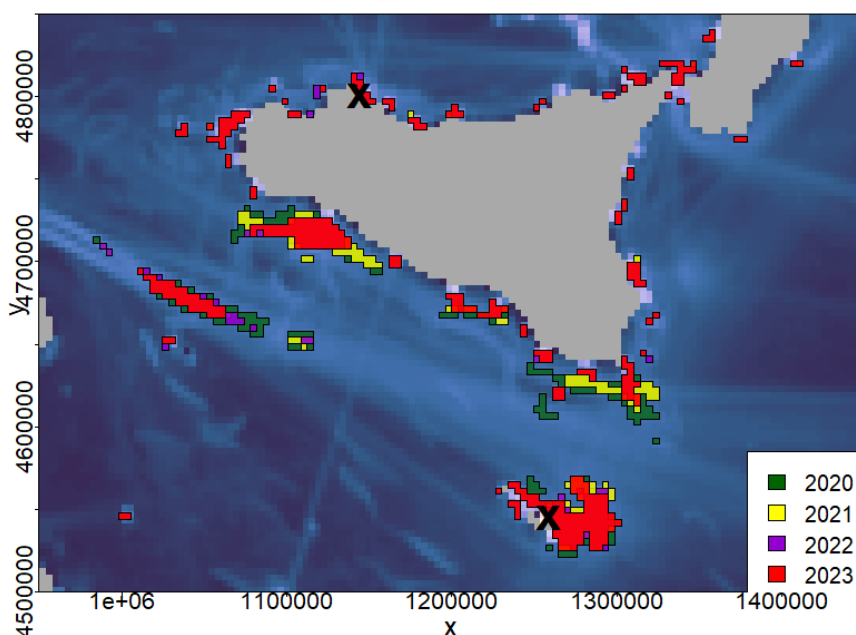


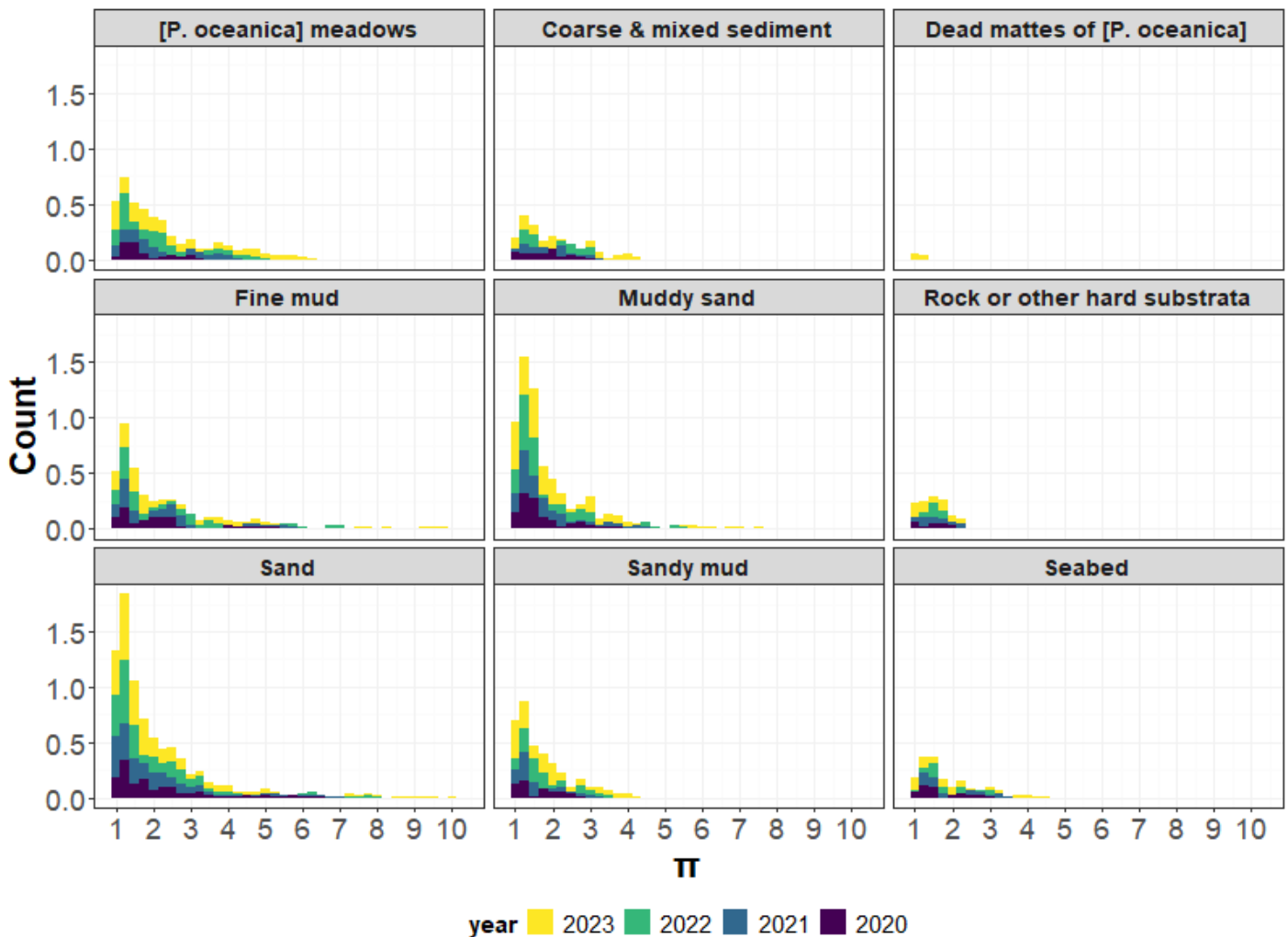
Figure 29 Hotspot areas, as identified by Gi cluster analysis. Since more than 99% of cells have an invasion probability of zero or very close to zero, a regional visualisation of the annual changes in hotspot patterns is difficult. A detailed look at the Strait of Sicily, one of the most important regions in terms of ship-borne invasion risk, is presented as a visual example. The increase in ship traffic from 2020 to 2023 exerts a double effect on probability of invasion as it both increase in absolute values through time, and relatively through space, resulting in the appearance of new invasions hotspots (black crosses). The underneath layer highlights major routes.*

Substrate-specific invasions

Figure 30 displays the distribution of $\pi > 1$ across the different substrate types over four years (2020-2023). The complex interactions between substrate exposure to invasion, vessel intensity and route density have non-linear effects on π .

In particular, the model is able to downweigh likelihood of invasion on the base of the interactions mediated by vessels: i.e *Dead mattes of P. oceanica* and *rock or other hard substrata* have few invasion events despite their high establishment potential (see Table 6). This is likely due to the uneven distribution of substrate types and correlation between route

Figure 30 Estimated count distribution of $\pi > 1$ per substrate type and year. Substrate types with no $\pi > 1$ values are not shown. Likelihood of invasion increase both in magnitude and frequency, as the peaks are higher and the right tail becomes longer. Notably, *Dead mattes of P.oceanica* reach values of $\pi > 1$ only in 2023.



density and substrate type. Sandy and muddy substrates are the most represented types in coastal areas, particularly those interested by high vessel density, and this manifests in a higher number of invasion events. The positive trend in vessel density and changes in route density likely result in more concentrated traffic that in turns produce higher π values despite the general decrease in suitability values.

Conclusion

This study introduces a refined gravity model to quantify the secondary invasion risk of *C. cylindracea* in the Mediterranean Sea, providing important insights into vessel-mediated spread of the species between 2020 and 2023.

The sensitivity analysis shows that vessel spreading efficacy is the most important factor for invasion probability at both origin and destination, while substrate type at origin emerges as the overall most important ecological factor. In addition, several two-way interactions emerged as significant, highlighting dependencies among ecological parameters. Notably, habitat suitability's influence on p_{ij} was moderated by substrate type, and temporal shifts were observed in the impact of habitat suitability on invasion risk across years.

The temporal variation captured by the significant year effect and its interactions with habitat suitability and substrate type suggest that changes over time, possibly due to environmental fluctuations or increased vessel activity, shape the invasion dynamics. For the simulated years, the temporal analysis shows an intriguing trend in which probability of invasion constantly increases despite lower per-ship probability of starting an invasion. This suggests a prevailing effect of increasing vessel and route densities on invasion risk.

This effect also shapes the *exposure* of substrate type to invasion risk. Sandy and muddy substrates, for instance, have a higher invasion rate, which is likely due to their prevalence in

coastal regions characterised by high vessel density. Conversely, substrates with higher intrinsic establishment potential, such as dead *P. oceanica* mattes and rocky and hard substrates appear to be less frequently invaded. This discrepancy suggests that biological compatibility of the substrate is less important than the spatial distribution of vessel traffic in determining the outcome of invasion.

Spatially, invasion hotspots are becoming more evenly distributed across the Mediterranean over time, increasing the area at risk and highlighting the potential for new invasion hotspots. This suggests that *C. cylindracea* is likely expanding its range thank to vessel activity.

These results provide a quantitative basis for targeted monitoring and intervention strategies in the Mediterranean: i.e new invasions seem likely to already strongly impacted areas (e.g., Sicily, Ligurian Sea), highlighting the importance of increased monitoring and specific containment measures in these areas.

The fine spatial resolution of the model allows for locally tailored management measures, while the temporal trend remarks the importance of improved biosecurity protocols to address increasing invasive pressure from maritime transport. Although future validation with new occurrence data could improve prediction accuracy, this framework already provides a solid basis for assessing regional exposure to secondary invasion by *C. cylindracea* at a scale relevant to Maritime Spatial Planning.

Limitations

The challenge of identifying meaningful parameters is critical, especially when applying local estimates (e.g. p_0) or when parameters' interactions are very uncertain.

An example is the impact of habitat suitability at of the origins: some authors have found that *C. cylindracea* increases fragments and propagule production when it is under stress (Santamaría et al., 2021; Uyà et al., 2018). This implies that α depends on H_i and behaves non-monotonically. At low suitability values, α may have high values due to the species biological response to the stressful situation, potentially contributing to p_i rather than hampering it.

Another major limitation lies in the simplified representation of the fleet and the characterisation of the *deterrence* function and vessel efficacy. More effort should be made to estimate *vessel-type* parameterisation to enhance a more realistic estimate of invasion rate. This could, for example, help to give a higher weighting to the impact of pleasure crafts on local spread and a lower weighting to the impact of cargo and tanks, that follow strict biosecurity protocols during operation (David et al., 2015).

Conclusions and final remarks

The increasing complexity of environmental challenges demands innovative approaches that can effectively integrate multiple layers of analysis while maintaining scientific rigor.

This thesis has developed and demonstrated a comprehensive framework for assessing the risk of Alien Invasive Species (AIS) spread in the Mediterranean Sea, with particular focus on *Caulerpa cylindracea* as a case study. The appropriate use of ML methods can provide powerful tools for understanding and predicting the dynamics of marine ecosystems, even with fragmented or scarce data. This is crucial for an effective conservation and sustainable management of marine resources.

The framework's strength lies in its multi-layered perspective on understanding and predicting AIS spread. The groundwork is a robust dataset that ensures that subsequent analyses are built upon reliable and scientifically sound information. The high-resolution dataset for the Mediterranean Sea spanning 2005-2099 featuring key biogeochemical and physical variables under the RCP8.5 scenario provides the essential environmental context for both current and future predictions allowing the possibility to quantitatively consider the impacts of climate change onto species future distribution.

Significant methodological progress has been made in the field of Machine Learning Ecological Niche Modeling (ML-ENM). The cornerstone is the development of a Multi-Criteria Decision Making (MCDM) framework for performing model selection, based on the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) and CRITIC (CRiteria Importance Through Intercriteria Correlation) methods. The centrepiece of TOPSIS is the definition of the Positive Ideal Solution (PIS) in order to rank the models. Defining PIS is the same as setting the research objective (e.g. maximising transferability or predictive power) in

the ML realm. This ensure that the best model is selected meaningfully. On the other hand, CRITIC determines the weighting of the criteria based on the variability of the criteria and the intercorrelations. It provides an objective system that efficiently organises the performance distance measured by TOPSIS according to the nature of the data, mitigating the impact of biased data on metrics.

Another important achievement is the introduction of the site weighting scheme, which allows additional information on occurrences to be used to embed ecological significance into the evaluation metrics. The weighting scheme is the basis for the introduction of delta metrics, which make it possible to recognise artificially inflated (overfitted) model performance.

Particularly important advances in the context of AIS are the refined methods for assessing threshold uncertainty and extrapolation risk. Threshold variability is relevant for scenario generation, while analysing extrapolation risk is a diagnostic tool that is essential for quantifying the reliability of projections.

The result of applying this robust ENM process is a model that project a general, significant decrease of suitable areas for the Mediterranean Sea, that is a quite interesting finding, considering the relevance of the species.

The integration of the model outcomes into a mechanistic framework that simulates vessel mediated dispersal provided even more interesting insights about *C. cylindracea* invasion dynamics in the Mediterranean Sea:

Despite projections shows a general decline in environmental suitability under future climate scenarios, the species continues to expand its range. This apparent paradox highlights the dominant role of human activities, particularly maritime transport, in AIS spread. Through

explicit *flow* representation, it's shown that vessel density and route patterns have a greater influence on invasion outcomes than substrate and suitability. This is indeed dependent on the *deterrence* function and, more specifically, on its bimodal gamma representation. The gamma distribution roughly approximate aggregated traffic data and is certainly biased towards medium-long route lengths. This means that the dominance of vessel density on estimated number of invasions holds on average, but not in certain cases, such as the *flows* to neighbouring and very close cells or in the case of vessel-specific flows. This imply that the outcomes are relevant to regional scale assessments but must be taken with caution in local assessments.

The temporal and spatial analyses conducted through our framework revealed emerging invasion hotspots across the Mediterranean, with particular concern for already impacted areas like Sicily and the Ligurian Sea. The increasingly even distribution of invasion probability over time suggests a broadening of risk across the basin, pointing to the need for widespread rather than localized monitoring efforts.

The framework developed in this thesis makes substantial contributions to the implementation of the Marine Strategy Framework Directive (MSFD). By providing a standardized, reproducible approach for assessing AIS risk at the required spatial and temporal resolutions, it directly supports evidence-based decision-making in marine management. The integration of correlative and mechanistic models can enable to the possibility to distinguish between climate-driven and anthropogenic factors in species spread, allowing for more targeted and effective interventions.

Future directions

This work provides a solid foundation for AIS risk assessment and pave the way to even more integrated risk characterisation with the addition of vulnerability layers, biotic resistance and socio-economic factors.

Furthermore, the development and integration of other pathways (e.g. aquaculture release) or processes (e.g. natural expansions and gene flows between populations Van Etten & Hijmans, 2010) could be straightforward as they could be included in the spatial interaction model.

From a technical and methodological perspective, there is an ample opportunity of improvement:

Refining the modelling of vessel-mediated spread represents an important direction for future research. The excessive importance of shipping in invasion dynamics points to the need for detailed vessel-specific parameterisation and per-ship type assessments. This would involve developing risk profiles for different vessel types and evaluating the effectiveness of current biosecurity protocols. Such refinements would improve our ability to predict and manage invasion risks associated with maritime activities.

In addition, to make the mechanistic model a useful tool for predicting invasions, a robust model validation system needs to be developed.

The framework presented in this thesis represents a significant advancement in AIS risk assessment. Its most notable feature, however, lies in its adaptability: it is designed to be easily transferable across different contexts and prepared to accommodate future developments. For this reason, the software infrastructure was designed from the outset to be as modular and flexible as possible. A fundamental aim was to define “blocks” of

operations (with standard input-output data formats) that would allow the implementation of the hypothetical user's preferred method with little effort. (e.g. one could opt for a random forest or GAM model instead of *MaxEnt*).

As marine ecosystems face increasing pressures from both climate change and human activities, the need for robust, integrated approaches to environmental risk assessment becomes ever greater. This framework provides a new set of tools and, more importantly, a spacious toolbox.

Appendix A

Marco Fianchini

October 31, 2024

A Mathematical Foundations of MaxEnt

This appendix provides a mathematical treatment of the Maximum Entropy (MaxEnt) in the context of machine learning and statistical inference.

A.1 Problem Formulation

Let $X \subseteq \mathbb{R}^d$ be the feature space representing environmental variables, and $Y = \{0, 1\}$ be the binary label space indicating species presence (1) or absence (0). We aim to estimate the conditional probability distribution $p(y|x)$, where $x \in X$ and $y \in Y$.

Given:

- A set of n presence locations: $\{x_1, \dots, x_n\} \subset X$
- A set of m background points: $\{z_1, \dots, z_m\} \subset X$
- A set of k feature functions: $f_j : X \rightarrow \mathbb{R}$, for $j = 1, \dots, k$

A.2 Maximum Entropy Principle

The MaxEnt principle states that we should choose the probability distribution that maximises entropy while satisfying the constraints imposed by our empirical observations. Mathematically, we are trying to maximise:

$$H(p) = - \int_X p(x) \log p(x) dx \quad (1)$$

subject to the constraints:

$$\mathbb{E}_p[f_j] = \frac{1}{n} \sum_{i=1}^n f_j(x_i) \quad \text{for } j = 1, \dots, k \quad (2)$$

where $\mathbb{E}_p[f_j]$ denotes the expected value of feature f_j under distribution p .

A.3 Duality and Exponential Form

Using the method of Lagrange multipliers, we can show that the solution to this constrained optimisation problem has exponential form:

$$p_\lambda(x) = \frac{\exp(\sum_{j=1}^k \lambda_j f_j(x))}{Z(\lambda)} \quad (3)$$

where $\lambda = (\lambda_1, \dots, \lambda_k)$ are the Lagrange multipliers and $Z(\lambda)$ is the partition function:

$$Z(\lambda) = \int_X \exp(\sum_{j=1}^k \lambda_j f_j(x)) dx \quad (4)$$

A.4 Optimization

The dual problem is to find λ^* that maximises the log-likelihood:

$$\lambda^* = \lambda \left\{ \sum_{i=1}^n \sum_{j=1}^k \lambda_j f_j(x_i) - n \log Z(\lambda) \right\} \quad (5)$$

This optimization is typically performed using numerical methods such as L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm).

A.5 Regularization

To prevent overfitting, MaxEnt employs L1 and L2 regularization. The regularised optimization problem becomes:

$$\lambda^* = \lambda \left\{ \sum_{i=1}^n \sum_{j=1}^k \lambda_j f_j(x_i) - n \log Z(\lambda) - \sum_{j=1}^k (\beta_1 |\lambda_j| + \beta_2 \lambda_j^2) \right\} \quad (6)$$

where β_1 and β_2 are the L1 and L2 regularization parameters, respectively.

A.6 Feature Engineering

MaxEnt uses several types of features derived from the environmental variables:

- Linear: $f_j(x) = x_j$
- Quadratic: $f_j(x) = x_j^2$
- Product: $f_{jk}(x) = x_j x_k$
- Threshold: $f_j(x) = \mathbb{I}(x_j > t)$, where \mathbb{I} is the indicator function and t is a threshold
- Hinge: $f_j(x) = \max(0, x_j - t)$

A.7 Prediction

For a new point x^* , the raw prediction is given by:

$$p_{\lambda^*}(x^*) = \frac{\exp(\sum_{j=1}^k \lambda_j^* f_j(x^*))}{Z(\lambda^*)} \quad (7)$$

The logistic output, which can be interpreted as the probability of species presence, is calculated as:

$$P(y = 1|x^*) = \frac{c \cdot p_{\lambda^*}(x^*)}{1 - p_{\lambda^*}(x^*) + c \cdot p_{\lambda^*}(x^*)} \quad (8)$$

where c is the prevalence of the species in the study area.

B Model Evaluation

Performance metrics used in the thesis:

B.1 Area Under the Receiver Operating Characteristic Curve (AUC)

AUC measures the probability that a randomly selected positive example will be ranked higher than a randomly selected negative one.

$$\text{AUC} = \frac{1}{n_+ n_-} \sum_{i: y_i=1} \sum_{j: y_j=0} \mathbb{I}(\hat{p}_i > \hat{p}_j) \quad (9)$$

where n_+ and n_- are the numbers of positive and negative samples, respectively, and \mathbb{I} is the indicator function.

B.2 Partial ROC (pROC)

pROC is a modified ROC analysis that adjusts the AUC to focus on the proportion of area correctly predicted as present, taking into account model-specific prediction margins and prioritising omission errors.

$$\text{pROC} = \frac{\int_{E_1}^{E_2} \text{ROC}(E) dE}{(E_2 - E_1) \cdot (E_2 - E_1)/2} \quad (10)$$

where E_1 and E_2 are the lower and upper bounds of the partial area, respectively.

B.3 Symmetric Extremal Dependence Index (SEDI)

SEDI measures the accuracy of deterministic forecasts of rare binary events, robust to low prevalence.

$$\text{SEDI} = \frac{\ln(F) - \ln(H) + \ln(1 - H) - \ln(1 - F)}{\ln(F) + \ln(H) + \ln(1 - H) + \ln(1 - F)} \quad (11)$$

where F is the false alarm rate and H is the hit rate.

B.4 OR.10p

OR.10p represents the percentage of test sites where presence was not predicted, using the 10th percentile of training presence values as the threshold.

$$\text{OR.10p} = \frac{\text{Number of test presences below threshold}}{\text{Total number of test presences}} \times 100\% \quad (12)$$

B.5 AUCdiff

AUCdiff measures the risk of model overfitting by comparing AUC values between training and test datasets.

$$\text{AUCdiff} = \text{AUC}_{\text{training}} - \text{AUC}_{\text{test}} \quad (13)$$

B.6 Akaike Information Criterion corrected (AICc)

AICc is a relative measure of model fit that takes into account the number of parameters and penalises complexity.

$$\text{AICc} = 2k - 2 \ln(\hat{L}) + \frac{2k(k+1)}{n-k-1} \quad (14)$$

where k is the number of parameters, \hat{L} is the maximum likelihood estimate, and n is the sample size.

B.7 Continuous Boyce Index (CBI)

CBI measures the accuracy and reliability in predicting presence based on the Spearman rank correlation between predicted suitability and observed presence.

$$\text{CBI} = \frac{1}{n} \sum_{i=1}^n \frac{P_i - E_i}{P_i + E_i} \quad (15)$$

where P_i and E_i are the predicted and expected frequencies in bin i , respectively.

Bibliography

- Agiadi, K., Hohmann, N., Gliozzi, E., Thivaïou, D., Bosellini, F. R., Taviani, M., Bianucci, G., Collareta, A., Londeix, L., Faranda, C., Bulian, F., Koskeridou, E., Lozar, F., Mancini, A. M., Dominici, S., Moissette, P., Campos, I. B., Borghi, E., Iliopoulos, G., ... García-Castellanos, D. (2024). The marine biodiversity impact of the Late Miocene Mediterranean salinity crisis. *Science*, 385(6712), 986–991. <https://doi.org/10.1126/science.adp3703>
- Agusti, S., Martinez-Ayala, J., Regaudie-de-Gioux, A., & Duarte, C. M. (2017). Oligotrophication and Metabolic Slowing-Down of a NW Mediterranean Coastal Ecosystem. *Frontiers in Marine Science*, 4. <https://doi.org/10.3389/fmars.2017.00432>
- Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Ashton, G. V., Zabin, C. J., Davidson, I. C., & Ruiz, G. M. (2022). Recreational boats routinely transfer organisms and promote marine bioinvasions. *Biological Invasions*, 24(4), 1083–1096. <https://doi.org/10.1007/s10530-021-02699-x>
- Bachir Bouiadjra, B., Ghellai, M., Daoudi, M., Behmene, I. E., & Bachir Bouiadjra, M. E. A. (2021). Impacts of the invasive species *Caulerpa cylindracea* Sonder, 1845 on the algae flora of the west coast of Algeria. *Biodiversity Data Journal*, 9, e64535. <https://doi.org/10.3897/BDJ.9.e64535>
- Baquero, R. A., Ayllón, D., & Nicola, G. G. (2021). Are the EU biosecurity legislative frameworks sufficiently effective to prevent biological invasions in the Natura 2000 network?—A case study in Mediterranean Europe. *Environmental Science & Policy*, 120, 21–28.
- Bednarsek, N., Guilloux, B., Canu, D. M., Galdies, C., Guerra, R., Simoncelli, S., Feely, R. A., Pelletier, G., Gašparović, B., Godrijan, J., Malej, A., Solidoro, C., Turk, V., & Zunino, S. (2023). Ocean Acidification as a Governance Challenge in the Mediterranean Sea: Impacts from Aquaculture and Fisheries. In S. Partelow, M. Hadjimichael, & A.-K. Hornidge (A c. Di), *Ocean Governance: Knowledge Systems, Policy Foundations and Thematic Analyses* (pp. 403–432). Springer International Publishing. https://doi.org/10.1007/978-3-031-20740-2_18
- Bédry, R., de Haro, L., Bentur, Y., Senechal, N., & Galil, B. S. (2021). Toxicological risks on the human health of populations living around the Mediterranean Sea linked to the invasion of non-indigenous marine species from the Red Sea: A review. *Toxicon*, 191, 69–82. <https://doi.org/10.1016/j.toxicon.2020.12.012>
- Bernardeau-Esteller, J., Ruiz, J. M., Tomas, F., Sandoval-Gil, J. M., & Marín-Guirao, L. (2015). Photoacclimation of *Caulerpa cylindracea*: Light as a limiting factor in the invasion of native Mediterranean seagrass meadows. *Journal of Experimental Marine Biology and Ecology*, 465, 130–141. <https://doi.org/10.1016/j.jembe.2014.11.012>
- Bivand, R. S., & Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, 27(3), 716–748. <https://doi.org/10.1007/s11749-018-0599-x>
- Bonanno, A., Zgozi, S. W., Jarboui, O., Mifsud, R., Ceriola, L., Basilone, G., & Arneri, E. (2018). Marine ecosystems and living resources in the Central Mediterranean Sea: An introduction. *Hydrobiologia*, 821(1), 1–10. <https://doi.org/10.1007/s10750-018-3692-y>

Boudouresque, C. F. (1999). *INTRODUCED SPECIES IN THE MEDITERRANEAN : ROUTES, KINETICS AND CONSEQUENCES* (MAP Technical Report Series, pp. 51–72) [Proceedings of the workshop on invasive *Caulerpa* species in the Mediterranean]. UNEP.

Boudouresque, C., & Verlaque, M. (2012). An overview of species introduction and invasion processes in marine and coastal lagoon habitats. *Cahiers de Biologie Marine*, 53.

Brondízio, E., Settele, J., Diaz, S., Ngo, H. T., Experts, G., & Mohamed, A. (2021). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. <https://doi.org/10.5281/zenodo.383188>

Bulleri, F., Alestra, T., Ceccherelli, G., Tamburello, L., Pinna, S., Sechi, N., & Benedetti-Cecchi, L. (2011). Determinants of *Caulerpa racemosa* distribution in the north-western Mediterranean. *Marine Ecology Progress Series*, 431, 55–67. <https://doi.org/10.3354/meps09137>

Bulleri, F., & Malquori, F. (2015). High tolerance to simulated herbivory in the clonal seaweed, *Caulerpa cylindracea*. *Marine Environmental Research*, 107, 61–65. <https://doi.org/10.1016/j.marenvres.2015.04.004>

Bulleri, F., Tamburello, L., Pusceddu, A., Bonechi, L., Cau, A., Moccia, D., & Gribben, P. E. (2018). Fragment quality and sediment organic loading regulate the survival of an invasive, clonal seaweed. *Biological Invasions*, 20(8), 1953–1959. <https://doi.org/10.1007/s10530-018-1685-4>

Cabrini, M., Cerino, F., De Olazabal, A., Di Poi, E., Fabbro, C., Fornasaro, D., Goruppi, A., Flander-Putrlle, V., Francé, J., & Gollasch, S. (2019). Potential transfer of aquatic organisms via ballast water with a particular focus on harmful and non-indigenous species: A survey from Adriatic ports. *Marine pollution bulletin*, 147, 16–35.

Campagne, C. S., Salles, J.-M., Boissery, P., & Deter, J. (2015). The seagrass *Posidonia oceanica*: Ecosystem services identification and economic evaluation of goods and benefits. *Marine Pollution Bulletin*, 97(1), 391–400. <https://doi.org/10.1016/j.marpolbul.2015.05.061>

Cantasano, N., Pellicone, G., & Di Martino, V. (2017). The spread of *Caulerpa cylindracea* in Calabria (Italy) and the effects of shipping activities. *Ocean & Coastal Management*, 144, 51–58. <https://doi.org/10.1016/j.ocecoaman.2017.04.014>

Canu, D. M., Zunino, S., Bondoni, M., Brandini, C., Čermelj, B., Drago, A., Faure, V., Galea, A., Galiana, S., Garcia, X., Grenz, C., Lapucci, C., Ličer, M., Lipizer, M., Mavrič, B., Perna, M., Previati, M., Queirós, L., Ruitton, S., & Viure, L. (2022). *Mapping human impacts to support sustainable uses of marine ecosystems in the Mediterranean sea* (EGU22-9926). EGU22. Copernicus Meetings. <https://doi.org/10.5194/egusphere-egu22-9926>

Caronni, S., Delaria, M. A., Gentili, R., Montagnani, C., Navone, A., Panzalis, P., & Citterio, S. (2021). First Report of Gametogenesis and Spawning for the Invasive Alga *Caulerpa cylindracea* in the Tyrrhenian Sea: The Key Role of Water Motion and Temperature. *Frontiers in Marine Science*, 8. <https://www.frontiersin.org/articles/10.3389/fmars.2021.774274>

Carreño, A., & Lloret, J. (2021). Environmental impacts of increasing leisure boating activity in Mediterranean coastal waters. *Ocean & Coastal Management*, 209, 105693. <https://doi.org/10.1016/j.ocecoaman.2021.105693>

Chakraborty, S. (2022). TOPSIS and Modified TOPSIS: A comparative analysis. *Decision Analytics Journal*, 2, 100021. <https://doi.org/10.1016/j.dajour.2021.100021>

- Cobos, M. E., Peterson, A. T., Barve, N., & Osorio-Olvera, L. (2019). kuenm: An R package for detailed development of ecological niche models using Maxent. *PeerJ*, 7, e6281. <https://doi.org/10.7717/peerj.6281>
- Coll, M., Piroddi, C., Albouy, C., Lasram, F., Cheung, W., & Christensen, V. (2011). The Mediterranean Sea under siege: Spatial overlap between marine biodiversity, cumulative threats and marine reserves. *Global Ecology and Biogeography*, 20, 1–16.
- Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Ben Rais Lasram, F., Aguzzi, J., Ballesteros, E., Bianchi, C. N., Corbera, J., Dailianis, T., Danovaro, R., Estrada, M., Frogliani, C., Galil, B. S., Gasol, J. M., Gertwagen, R., Gil, J., Guilhaumon, F., Kesner-Reyes, K., ... Voultsiadou, E. (2010). The biodiversity of the Mediterranean Sea: Estimates, patterns, and threats. *PLoS One*, 5(8), e11842. <https://doi.org/10.1371/journal.pone.0011842>
- Collaud Coen, M., Andrews, E., Bigi, A., Martucci, G., Romanens, G., Vogt, F. P. A., & Vuilleumier, L. (2020). Effects of the prewhitening method, the time granularity, and the time segmentation on the Mann–Kendall trend detection and the associated Sen’s slope. *Atmospheric Measurement Techniques*, 13(12), 6945–6964. <https://doi.org/10.5194/amt-13-6945-2020>
- Commission Decision (EU) 2017/848 of 17 May 2017 Laying down Criteria and Methodological Standards on Good Environmental Status of Marine Waters and Specifications and Standardised Methods for Monitoring and Assessment, and Repealing Decision 2010/477/EU (Text with EEA Relevance.), 125 OJ L (2017). <http://data.europa.eu/eli/dec/2017/848/oj/eng>
- Coro, G., Magliozzi, C., Vanden Berghe, E., Bailly, N., Ellenbroek, A., & Pagano, P. (2015). Estimating absence locations of marine species from data of scientific surveys in OBIS. *Ecological Modelling*, 323, 61–76. <https://doi.org/10.1016/j.ecolmodel.2015.12.008>
- Cossarini, G., Feudale, L., Teruzzi, A., Bolzon, G., Coidessa, G., Solidoro, C., Di Biagio, V., Amadio, C., Lazzari, P., Brosich, A., & Salon, S. (2021). High-Resolution Reanalysis of the Mediterranean Sea Biogeochemistry (1999–2019). *Frontiers in Marine Science*, 8. <https://doi.org/10.3389/fmars.2021.741486>
- Creese, R., Glasby, T. M., & Davis, A. (2004). *Eradicating and preventing the spread of the invasive alga Caulerpa taxifolia in NSW*. NSW Fisheries.
- David, M., Gollasch, S., Elliott, B., & Wiley, C. (2015). Ballast water management under the ballast water management convention. *Global Maritime Transport and Ballast Water Management: Issues and Solutions*, 89–108.
- Di Biagio, V., Salon, S., Feudale, L., & Cossarini, G. (2022). Subsurface oxygen maximum in oligotrophic marine ecosystems: Mapping the interaction between physical and biogeochemical processes. *Biogeosciences*, 19(23), 5553–5574. <https://doi.org/10.5194/bg-19-5553-2022>
- Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 Establishing a Framework for Community Action in the Field of Marine Environmental Policy (Marine Strategy Framework Directive) (Text with EEA Relevance), EP, CONSIL, 164 OJ L (2008). <http://data.europa.eu/eli/dir/2008/56/oj/eng>
- Drake, J. M., & Lodge, D. M. (2004). Global Hot Spots of Biological Invasions: Evaluating Options for Ballast-Water Management. *Proceedings: Biological Sciences*, 271(1539), 575–580.
- Elith, J., & Franklin, J. (2013). Species distribution modeling. In *Encyclopedia of Biodiversity: Second Edition* (pp. 692–705). Elsevier Inc.

- Elith, J., Graham, C., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A., Li, J., Lohmann, L., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J., Peterson, A., & Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ferrier, S., Ford, A., Guisan, A., Hijmans, R. J., Huettmann, F., Lohmann, L., Loiselle, B., Moritz, C., Overton, J., Peterson, A. T., Phillips, S., Richardson, K., Williams, S., Wiser, S. K., ... Zimmermann, N. E. (2020). Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics*, *15*(2), Articolo 2. <https://doi.org/10.17161/bi.v15i2.13384>
- Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. <https://doi.org/10.1111/j.1472-4642.2010.00725.x>
- EMSA. (2019). *Promotion of interoperability between industry and competent authorities in the European Maritime Single Window (EMSW) environment under the CISE Process*. <https://emsa.europa.eu/damage-stability-study/items.html?cid=14&id=4806>
- Escudier, R., Clementi, E., Cipollone, A., Pistoia, J., Drudi, M., Grandi, A., Lyubartsev, V., Lecci, R., Aydogdu, A., Delrosso, D., Omar, M., Masina, S., Coppini, G., & Pinardi, N. (2021). A High Resolution Reanalysis for the Mediterranean Sea. *Frontiers in Earth Science*, *9*, undefined-undefined. <https://doi.org/10.3389/feart.2021.702285>
- Etten, J. van. (2017). R Package gdistance: Distances and Routes on Geographical Grids. *Journal of Statistical Software*, *76*, 1–21. <https://doi.org/10.18637/jss.v076.i13>
- European Commission. Directorate General for Maritime Affairs and Fisheries. (2023). *EU vessel density map: Detailed method*. Publications Office. <https://data.europa.eu/doi/10.2771/518143>
- Evans, T. G., Diamond, S. E., & Kelly, M. W. (2015). Mechanistic species distribution modelling as a link between physiology and conservation. *Conservation Physiology*, *3*(1), cov056. <https://doi.org/10.1093/conphys/cov056>
- Falaschi, M., Mangiacotti, M., Sacchi, R., Scali, S., & Razzetti, E. (2018). Electric circuit theory applied to alien invasions: A connectivity model predicting the Balkan frog expansion in Northern Italy. *Acta Herpetologica*, *13*, 33–42. https://doi.org/10.13128/Acta_Herpetol-20871
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Fianchini, M., Solidoro, C., & Canu, D. (2024a). *2005-2099 High resolution bioclimatic variables for the surface and bottom of the Mediterranean Sea*. (Versione 1.0) [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.12780161>
- Fianchini, M., Solidoro, C., & Canu, D. (2024b). Downscaled and bias-corrected bioclimatic dataset for the Mediterranean Sea (2005–2099). *Data in Brief*, *57*, 110846. <https://doi.org/10.1016/j.dib.2024.110846>
- Fiorentino, F., & Vitale, S. (2021). How Can We Reduce the Overexploitation of the Mediterranean Resources? *Frontiers in Marine Science*, *8*. <https://doi.org/10.3389/fmars.2021.674633>

- Fitzpatrick, M. C., & Hargrove, W. W. (2009). The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, 18(8), 2255–2261. <https://doi.org/10.1007/s10531-009-9584-8>
- Fortin, M.-J., & Dale, M. R. T. (2005). *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press.
- Fotheringham, A., & O’Kelly, M. (1988, dicembre 5). *Spatial Interaction Models: Formulations and Applications*. <https://www.semanticscholar.org/paper/Spatial-Interaction-Models:Formulations-and-Fotheringham-O%E2%80%99Kelly/3d7be4c3830fd92a5cf0c79534d25e5658355dd2>
- Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS One*, 9(5), e97122. <https://doi.org/10.1371/journal.pone.0097122>
- Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction* (1^a ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511810602>
- Gabrie, C., Lagabrielle, E., Bissery, C., Crochelet, E., Meola, B., Webster, C., Claudet, J., Chassanite, A., Marinesque, S., Robert, P., & Goutx, M. (2012). *The Status of Marine Protected Areas in the Mediterranean Sea 2012*.
- Gaiji, S., Chavan, V., Ariño, A. H., Otegui, J., Hobern, D., Sood, R., & Robles, E. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodiversity Informatics*, 8(2), Articolo 2. <https://doi.org/10.17161/bi.v8i2.4124>
- Galil, B. S. (2023). A Sea, a Canal, a Disaster: The Suez Canal and the Transformation of the Mediterranean Biota. In C. Lutmar & Z. Rubinovitz (A c. Di), *The Suez Canal: Past Lessons and Future Challenges* (pp. 199–215). Springer International Publishing. https://doi.org/10.1007/978-3-031-15670-0_10
- Galil, B. S., Marchini, A., & Occhipinti-Ambrogi, A. (2018). East is east and West is west? Management of marine bioinvasions in the Mediterranean Sea. *Estuarine, Coastal and Shelf Science*, 201, 7–16. <https://doi.org/10.1016/j.ecss.2015.12.021>
- GBIF.org. (2024). *GBIF Occurrence Download* [Dataset]. <https://doi.org/10.15468/dl.ctg7ht>
- Gennaro, P., & Piazzini, L. (2014). The indirect role of nutrients in enhancing the invasion of *Caulerpa racemosa* var *cylindracea*. *Biological Invasions*, 16(8), 1709–1717. <https://doi.org/10.1007/s10530-013-0620-y>
- Giorgi, F. (2019). Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next? *Journal of Geophysical Research: Atmospheres*, 124(11), 5696–5723. <https://doi.org/10.1029/2018JD030094>
- Gobert, S., Laumont, N., & Bouquegneau, J.-M. (2002). *Posidonia oceanica* meadow: A low nutrient high chlorophyll (LNHC) system? *BMC Ecology*, 2(1), 9. <https://doi.org/10.1186/1472-6785-2-9>
- Goodwin, M., Halvorsen, K. T., Jiao, L., Knausgård, K. M., Martin, A. H., Moyano, M., Oomen, R. A., Rasmussen, J. H., Sørvalen, T. K., & Thorbjørnsen, S. H. (2022). Unlocking the potential of deep learning for marine ecology: Overview, applications, and outlook†. *ICES Journal of Marine Science*, 79(2), 319–336. <https://doi.org/10.1093/icesjms/fsab255>

- Goren, M., Galil, B. S., Diamant, A., Stern, N., & Levitt-Barmats, Y. (2016). Invading up the food web? Invasive fish in the southeastern Mediterranean Sea. *Marine Biology*, 163(8), 180. <https://doi.org/10.1007/s00227-016-2950-7>
- Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models: With Applications in R*. <https://doi.org/10.1017/9781139028271>
- Haubrock, P. J., Turbelin, A. J., Cuthbert, R. N., Novoa, A., Taylor, N. G., Angulo, E., Ballesteros-Mejia, L., Bodey, T. W., Capinha, C., Diagne, C., Essl, F., Golivets, M., Kirichenko, N., Kourantidou, M., Leroy, B., Renault, D., Verbrugge, L., & Courchamp, F. (2021). Economic costs of invasive alien species across Europe. *NeoBiota*, 67, 153–190. <https://doi.org/10.3897/neobiota.67.58196>
- Haynes, K. E., Phillips, F. Y., & Mohrfeld, J. W. (1980). The entropies: Some roots of ambiguity. *Socio-Economic Planning Sciences*, 14(3), 137–145. [https://doi.org/10.1016/0038-0121\(80\)90017-8](https://doi.org/10.1016/0038-0121(80)90017-8)
- Haynes, K., & Fotheringham, A. (1985). *Gravity and Spatial Interaction Models*. Grant Ian Thrall. <https://www.semanticscholar.org/paper/Gravity-and-Spatial-Interaction-Models-Haynes-Fotheringham/9d34750038f9835a38fe403c01e273ce62de103d>
- Herbert, R. J. H., Humphreys, J., Davies, Clare. J., Roberts, C., Fletcher, S., & Crowe, Tasman. P. (2016). Ecological impacts of non-native Pacific oysters (*Crassostrea gigas*) and management measures for protected areas in Europe. *Biodiversity and Conservation*, 25(14), 2835–2865. <https://doi.org/10.1007/s10531-016-1209-4>
- Hidalgo, M., El-Haweet, A. E., Tsikliras, A. C., Tirasin, E. M., Fortibuoni, T., Ronchi, F., Lauria, V., Ben Abdallah, O., Arneri, E., Ceriola, L., Milone, N., Lelli, S., Hernández, P., Bernal, M., & Vasconcellos, M. (2022). Risks and adaptation options for the Mediterranean fisheries in the face of multiple climate change drivers and impacts. *ICES Journal of Marine Science*, 79(9), 2473–2488. <https://doi.org/10.1093/icesjms/fsac185>
- Hulme, P. E., Bacher, S., Kenis, M., Klotz, S., Kühn, I., Minchin, D., Nentwig, W., Olenin, S., Panov, V., Pergl, J., Pyšek, P., Roques, A., Sol, D., Solarz, W., & Vilà, M. (2008). Grasping at the routes of biological invasions: A framework for integrating pathways into policy. *Journal of Applied Ecology*, 45(2), 403–414. <https://doi.org/10.1111/j.1365-2664.2007.01442.x>
- Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22, 415–427. <https://doi.org/10.1101/SQB.1957.022.01.039>
- Hwang, C.-L., Lai, Y.-J., & Liu, T.-Y. (1993). A new approach for multiple objective decision making. *Computers & Operations Research*, 20(8), 889–899. [https://doi.org/10.1016/0305-0548\(93\)90109-V](https://doi.org/10.1016/0305-0548(93)90109-V)
- Iveša, L., Djakovac, T., & Devescovi, M. (2015). Spreading patterns of the invasive *Caulerpa cylindracea* Sonder along the west Istrian Coast (northern Adriatic Sea, Croatia). *Marine Environmental Research*, 107, 1–7. <https://doi.org/10.1016/j.marenvres.2015.03.008>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. <https://doi.org/10.1007/978-1-0716-1418-1>
- Jarnevich, C. S., Stohlgren, T. J., Kumar, S., Morissette, J. T., & Holcombe, T. R. (2015). Caveats for correlative species distribution modeling. *Ecological Informatics*, 29(1), 6–15. <https://doi.org/10.1016/j.ecoinf.2015.06.007>
- Johnson, J. M., & Clarke, K. C. (2021). An area preserving method for improved categorical raster resampling. *Cartography and Geographic Information Science*, 48(4), 292–304. <https://doi.org/10.1080/15230406.2021.1892531>

- Karuppaiah, V., Maruthadurai, R., Das, B., Soumia, P. S., Gadge, A. S., Thangasamy, A., Ramesh, S. V., Shirsat, D. V., Mahajan, V., Krishna, H., & Singh, M. (2023). Predicting the potential geographical distribution of onion thrips, *Thrips tabaci* in India based on climate change projections using MaxEnt. *Scientific Reports*, *13*(1), 7934. <https://doi.org/10.1038/s41598-023-35012-y>
- Katsanevakis, S., Tempera, F., & Teixeira, H. (2016). Mapping the impact of alien species on marine ecosystems: The Mediterranean Sea case study. *Diversity and Distributions*, *22*(6), 694–707. <https://doi.org/10.1111/ddi.12429>
- Katsanevakis, S., Wallentinus, I., Zenetos, A., Leppäkoski, E., Çinar, M., Oztürk, B., Grabowski, M., Golani, D., & Cardoso, A. (2014, agosto 5). *Species-specific review of the impacts of invasive alien species on ecosystem services and biodiversity in the European Seas*.
- Katsanevakis, S., Zenetos, A., Belchior, C., & Cardoso, A. C. (2013). Invading European Seas: Assessing pathways of introduction of marine aliens. *Ocean & Coastal Management*, *76*, 64–74. <https://doi.org/10.1016/j.ocecoaman.2013.02.024>
- Kearney, M., & Porter, W. P. (2004). Mapping the Fundamental Niche: Physiology, Climate, and the Distribution of a Nocturnal Lizard. *Ecology*, *85*(11), 3119–3131. <https://doi.org/10.1890/03-0820>
- Klein, J., & Verlaque, M. (2008). The *Caulerpa racemosa* invasion: A critical review. *Marine Pollution Bulletin*, *56*(2), 205–225. <https://doi.org/10.1016/j.marpolbul.2007.09.043>
- Krijgsman, W., Hilgen, F., Raffi, I., Sierro, F., & Wilsonk, D. (1999). Chronology, causes and progression of the Messinian Salinity Crisis. *Nature*, *400*. <https://doi.org/10.1038/23231>
- Krishnan, A. R., Kasim, M. M., Hamid, R., & Ghazali, M. F. (2021). A Modified CRITIC Method to Estimate the Objective Weights of Decision Criteria. *Symmetry*, *13*(6), Articolo 6. <https://doi.org/10.3390/sym13060973>
- Kutor, S. K., Amoak, D., Owusu, B., & Kyeremeh, E. (2021). Theorizing “Wicked Concept” and Reconceptualizing Wisdom as Wicked. *The Professional Geographer*, *73*(4), 632–640. <https://doi.org/10.1080/00330124.2021.1933553>
- Lawton, J. H. (1994). What Do Species Do in Ecosystems? *Oikos*, *71*(3), 367–374. <https://doi.org/10.2307/3545824>
- Lejeusne, C., Chevaldonné, P., Pergent-Martini, C., Boudouresque, C. F., & Pérez, T. (2010). Climate change effects on a miniature ocean: The highly diverse, highly impacted Mediterranean Sea. *Trends in ecology & evolution*, *25*(4), 250–260.
- Leung, B., Roura-Pascual, N., Bacher, S., Heikkilä, J., Brotons, L., Burgman, M. A., Dehnen-Schmutz, K., Essl, F., Hulme, P. E., Richardson, D. M., Sol, D., & Vilà, M. (2012). TEASIng apart alien species risk assessments: A framework for best practices. *Ecology Letters*, *15*(12), 1475–1493. <https://doi.org/10.1111/ele.12003>
- Lieurance, D., Canavan, S., Behringer, D. C., Kendig, A. E., Minter, C. R., Reisinger, L. S., Romagosa, C. M., Flory, S. L., Lockwood, J. L., Anderson, P. J., Baker, S. M., Bojko, J., Bowers, K. E., Canavan, K., Carruthers, K., Daniel, W. M., Gordon, D. R., Hill, J. E., Howeth, J. G., ... Wanamaker, C. (2023). Identifying invasive species threats, pathways, and impacts to improve biosecurity. *Ecosphere*, *14*(12), e4711. <https://doi.org/10.1002/ecs2.4711>
- Liu, C., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, *6*(1), 337–348. <https://doi.org/10.1002/ece3.1878>

- Lo Brutto, S. (2021). Historical and Current Diversity Patterns of Mediterranean Marine Species. *Diversity*, 13(4), Articolo 4. <https://doi.org/10.3390/d13040156>
- Lovato, T., Vichi, M., & Oddo, P. (2013). *High-Resolution Simulations of Mediterranean Sea Physical Oceanography Under Current and Scenario Climate Conditions: Model Description, Assessment and Scenario Analysis*. <https://doi.org/10.2139/ssrn.2637861>
- Low, B. W., Yiwen, Z., Tan, H., & Yeo, D. (2020). Predictor complexity and feature selection affect Maxent model transferability: Evidence from global freshwater invasive species. *Diversity and Distributions*, 27. <https://doi.org/10.1111/ddi.13211>
- Luna, S., Peña-Peniche, A., & Mendoza-Alfaro, R. (2024). Species distribution model accuracy is strongly influenced by the choice of calibration area. *Biodiversity Informatics*, 18. <https://doi.org/10.17161/bi.v18i.22655>
- Malde, K., Handegard, N. O., Eikvil, L., & Salberg, A.-B. (2020). Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77(4), 1274–1285. <https://doi.org/10.1093/icesjms/fsz057>
- Mannino, A. M., Balistreri, P., & Deidun, A. (2017). The marine biodiversity of the Mediterranean Sea in a changing climate: The impact of biological invasions. *Mediterranean identities-environment, society, culture*, 101–127.
- Marbuah, G., Gren, I.-M., & McKie, B. (2014). Economics of Harmful Invasive Species: A Review. *Diversity*, 6(3), Articolo 3. <https://doi.org/10.3390/d6030500>
- Martínez-Abraín, A., Castejón-Silvo, I., & Roiloa, S. (2022). Foreseeing the future of *Posidonia oceanica* meadows by accounting for the past evolution of the Mediterranean Sea. *ICES Journal of Marine Science*, 79(10), 2597–2599. <https://doi.org/10.1093/icesjms/fsac212>
- Mazza, G., Tricarico, E., Genovesi, P., & Gherardi, F. (2014). Biological invaders are threats to human health: An overview. *Ethology Ecology & Evolution*, 26(2–3), 112–129.
- McRae, B. H., Dickson, B. G., Keitt, T. H., & Shah, V. B. (2008). USING CIRCUIT THEORY TO MODEL CONNECTIVITY IN ECOLOGY, EVOLUTION, AND CONSERVATION. *Ecology*, 89(10), 2712–2724. <https://doi.org/10.1890/07-1861.1>
- Menna, M., Gerin, R., Notarstefano, G., Mauri, E., Bussani, A., Pacciaroni, M., & Poulain, P.-M. (2021). On the Circulation and Thermohaline Properties of the Eastern Mediterranean Sea. *Frontiers in Marine Science*, 8. <https://doi.org/10.3389/fmars.2021.671469>
- Micheli, F., Halpern, B. S., Walbridge, S., Ciriaco, S., Ferretti, F., Frascchetti, S., Lewison, R., Nykjaer, L., & Rosenberg, A. A. (2013). Cumulative Human Impacts on Mediterranean and Black Sea Marine Ecosystems: Assessing Current Pressures and Opportunities. *PLOS ONE*, 8(12), e79889. <https://doi.org/10.1371/journal.pone.0079889>
- Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: Are we paying attention to recommendations? A systematic review. *PeerJ*, 5, e3093. <https://doi.org/10.7717/peerj.3093>
- Moreno-Amat, E., Mateo, R. G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J.-C., & García-Amorena, I. (2015). Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, 312, 308–317. <https://doi.org/10.1016/j.ecolmodel.2015.05.035>

- Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution*, 5(11), 1198–1205. <https://doi.org/10.1111/2041-210X.12261>
- Newton, A., Icely, J., Cristina, S., Brito, A., Cardoso, A. C., Colijn, F., Riva, S. D., Gertz, F., Hansen, J. W., Holmer, M., Ivanova, K., Leppäkoski, E., Canu, D. M., Mocenni, C., Mudge, S., Murray, N., Pejrup, M., Razinkovas, A., Reizopoulou, S., ... Zaldívar, J.-M. (2014). An overview of ecological status, vulnerability and future perspectives of European large shallow, semi-enclosed coastal systems, lagoons and transitional waters. *Estuarine, Coastal and Shelf Science*, 140, 95–122. <https://doi.org/10.1016/j.ecss.2013.05.023>
- Omodei, E., Grilli, J., Marsili, M., & Sanguinetti, G. (2024). Quantitative Human Ecology: Data, Models and Challenges for Sustainability. In S. Fantoni, N. Casagli, C. Solidoro, & M. Cobal (A c. Di), *Quantitative Sustainability: Interdisciplinary Research for Sustainable Development Goals* (pp. 79–89). Springer International Publishing. https://doi.org/10.1007/978-3-031-39311-2_6
- Owens, H. L., Campbell, L. P., Dornak, L. L., Saupe, E. E., Barve, N., Soberón, J., Ingenloff, K., Lira-Noriega, A., Hensz, C. M., Myers, C. E., & Peterson, A. T. (2013). Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling*, 263, 10–18. <https://doi.org/10.1016/j.ecolmodel.2013.04.011>
- Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, 213(1), 63–72. <https://doi.org/10.1016/j.ecolmodel.2007.11.008>
- Peterson, A. T., Papeş, M., & Soberón, J. (2015). Mechanistic and Correlative Models of Ecological Niches. *European Journal of Ecology*, 1(2), Articolo 2. <https://doi.org/10.1515/eje-2015-0014>
- Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press. <https://doi.org/10.1515/9781400840670>
- Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, 40(7), 887–893. <https://doi.org/10.1111/ecog.03049>
- Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, 31(2), 161–175. <https://doi.org/10.1111/j.0906-7590.2008.5203.x>
- Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological applications*, 19(1), 181–197.
- Piazzi, L., Balata, D., Bulleri, F., Gennaro, P., & Ceccherelli, G. (2016). The invasion of *Caulerpa cylindracea* in the Mediterranean: The known, the unknown and the knowable. *Marine Biology*, 163, 163:161. <https://doi.org/10.1007/s00227-016-2937-4>
- Piccardi, F., Poli, F., Sguotti, C., Tirelli, V., Borme, D., Mazzoldi, C., & Barausse, A. (2024). Assessing the impact of the invasive ctenophore *Mnemiopsis leidyi* on artisanal fisheries in the Venice Lagoon: An interdisciplinary approach. *Hydrobiologia*. <https://doi.org/10.1007/s10750-024-05505-6>
- Picotti, V., Negri, A., & Capaccioni, B. (2014). The Geological Origins and Paleoceanographic History of the Mediterranean Region: Tethys to Present. In S. Goffredo & Z. Dubinsky (A c. Di), *The Mediterranean*

- Sea: Its history and present challenges* (pp. 3–10). Springer Netherlands. https://doi.org/10.1007/978-94-007-6704-1_1
- Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643. <https://doi.org/10.1111/jbi.12227>
- Reale, M., Cossarini, G., Lazzari, P., Lovato, T., Bolzon, G., Masina, S., Solidoro, C., & Salon, S. (2022). Acidification, deoxygenation, and nutrient and biomass declines in a warming Mediterranean Sea. *Biogeosciences*, *19*(17), 4035–4065. <https://doi.org/10.5194/bg-19-4035-2022>
- Regulation (EU) No 1143/2014 of the European Parliament and of the Council of 22 October 2014 on the Prevention and Management of the Introduction and Spread of Invasive Alien Species, 317 OJ L (2014). <http://data.europa.eu/eli/reg/2014/1143/oj/eng>
- Renner, I. W., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics*, *69*(1), 274–281. <https://doi.org/10.1111/j.1541-0420.2012.01824.x>
- Ricciardi, A. (2001). Facilitative interactions among aquatic invaders: Is an «invasional meltdown» occurring in the Great Lakes? *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(12), 2513–2525. <https://doi.org/10.1139/f01-178>
- Riddell, E., Burger, I., Tyner-Swanson, T., Biggerstaff, J., Muñoz, M., Levy, O., & Porter, C. (2023). Parameterizing mechanistic niche models in biophysical ecology: A review of empirical approaches. *The Journal of experimental biology*, *226*. <https://doi.org/10.1242/jeb.245543>
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillerá-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. <https://doi.org/10.1111/ecog.02881>
- Roy, H. E., Pauchard, A., Stoett, P., & Renard Truong, T. (2024). *IPBES Invasive Alien Species Assessment: Full report*. Zenodo. <https://doi.org/10.5281/zenodo.11629357>
- Sandahl, A., & Tøttrup, A. P. (2020). Marine Citizen Science: Recent Developments and Future Recommendations. *Citizen Science: Theory and Practice*, *5*(1), Articolo 1. <https://doi.org/10.5334/cstp.270>
- Santamaría, J., Golo, R., Cebrian, E., García, M., & Vergés, A. (2021). Stressful Conditions Give Rise to a Novel and Cryptic Filamentous Form of *Caulerpa cylindracea*. *Frontiers in Marine Science*, *8*. <https://www.frontiersin.org/articles/10.3389/fmars.2021.548679>
- Saul, W.-C., Roy, H. E., Booy, O., Carnevali, L., Chen, H.-J., Genovesi, P., Harrower, C. A., Hulme, P. E., Pagad, S., Pergl, J., & Jeschke, J. M. (2017). Assessing patterns in introduction pathways of alien species by linking major invasion data bases. *Journal of Applied Ecology*, *54*(2), 657–669. <https://doi.org/10.1111/1365-2664.12819>
- Saupe, E. E., Barve, N., Owens, H. L., Cooper, J. C., Hosner, P. A., & Peterson, A. T. (2018). Reconstructing Ecological Niche Evolution When Niches Are Incompletely Characterized. *Systematic Biology*, *67*(3), 428–438. <https://doi.org/10.1093/sysbio/syx084>
- Sayit, H., Nurbay, A., Xu, Z., Arman, J., Shao, H., & Vinira, Y. (2019). Simulation of potential distribution patterns of the invasive plant species *Xanthium spinosum* L. (Bathurst burr) in Xinjiang under climate change. *Shengtai Xuebao*, *39*(5), 1551–1559. Scopus. <https://doi.org/10.5846/stxb201802040299>

- Schwalm, C. R., Glendon, S., & Duffy, P. B. (2020). RCP8.5 tracks cumulative CO2 emissions. *Proceedings of the National Academy of Sciences*, 117(33), 19656–19657. <https://doi.org/10.1073/pnas.2007117117>
- Seebens, H., Gastner, M. T., & Blasius, B. (2013). The risk of marine bioinvasion caused by global shipping. *Ecology letters*, 16(6), 782–790.
- Simberloff, D. (2009). The role of propagule pressure in biological invasions. *Annual Review of Ecology, Evolution, and Systematics*, 40, 81–102.
- Simberloff, D., & Gibbons, L. (2004). Now you See them, Now you don't! – Population Crashes of Established Introduced Species. *Biological Invasions*, 6(2), 161–172. <https://doi.org/10.1023/B:BINV.0000022133.49752.46>
- Smith, A. B., Murphy, S. J., Henderson, D., & Erickson, K. D. (2023). Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Global Ecology and Biogeography*, 32(3), 342–355. <https://doi.org/10.1111/geb.13628>
- Soberon, J., & Peterson, A. T. (2005). Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. *Biodiversity Informatics*, 2. <https://doi.org/10.17161/bi.v2i0.4>
- Solidoro, C., Cossarini, G., Lazzari, P., Galli, G., Bolzon, G., Somot, S., & Salon, S. (2022). Modeling Carbon Budgets and Acidification in the Mediterranean Sea Ecosystem Under Contemporary and Future Climate. *Frontiers in Marine Science*, 8. <https://doi.org/10.3389/fmars.2021.781522>
- Solidoro, C., Libralato, S., Canu, D. M., Cossarini, G., & Giorgi, F. (2013). *Impacts of climate-driven changes on coastal marine ecosystem and related good and services*. <https://doi.org/10.17895/ices.pub.24753123.v1>
- Srèbalienė, G., Olenin, S., Minchin, D., & Narščius, A. (2019). A comparison of impact and risk assessment methods based on the IMO Guidelines and EU invasive alien species risk assessment frameworks. *PeerJ*, 7, e6965. <https://doi.org/10.7717/peerj.6965>
- Streftaris, N., & Zenetos, A. (2006). Alien marine species in the Mediterranean-the 100 'Worst Invasives' and their impact. *Mediterranean Marine Science*, 7(1), 87–118.
- Taherdoost, H., & Madanchian, M. (2023). Multi-Criteria Decision Making (MCDM) Methods and Concepts. *Encyclopedia*, 3(1), Articolo 1. <https://doi.org/10.3390/encyclopedia3010006>
- Tarkan, A. S., Yoğurtçuoğlu, B., Karachle, P. K., Kalogianni, E., Top Karakuş, N., & Tricarico, E. (2021). Editorial: Understanding the Impact and Invasion Success of Aquatic Non-native Species: How They Interact With Novel Environments and Native Biota. *Frontiers in Ecology and Evolution*, 9. <https://doi.org/10.3389/fevo.2021.790540>
- Tong, Y., Gao, X., Han, Z., Xu, Y., Xu, Y., & Giorgi, F. (2021). Bias correction of temperature and precipitation over China for RCM simulations using the QM and QDM methods. *Climate Dynamics*, 57(5), 1425–1443. <https://doi.org/10.1007/s00382-020-05447-4>
- Trombetti, M., Katsanevakis, S., Deriu, I., & Cardoso, A. C. (2013). EASIN-Lit: A geo-database of published alien species records. *Management of Biological Invasions*, 4(3), 261–264. <https://doi.org/10.3391/mbi.2013.4.3.08>
- Tsirintanis, K., Azzurro, E., Crocetta, F., Dimiza, M., Frogli, C., Gerovasileiou, V., Langeneck, J., Mancinelli, G., Rosso, A., & Stern, N. (2022). *Bioinvasion impacts on biodiversity, ecosystem services, and human health in the Mediterranean Sea*.

- Turicchia, E., Ponti, M., Rossi, G., & Cerrano, C. (2021). The Reef Check Med Dataset on Key Mediterranean Marine Species 2001–2020. *Frontiers in Marine Science*, 8. <https://www.frontiersin.org/articles/10.3389/fmars.2021.675574>
- Turicchia, E., Ponti, M., Rossi, G., Milanese, M., Di Camillo, C. G., & Cerrano, C. (2021). The Reef Check Mediterranean Underwater Coastal Environment Monitoring Protocol. *Frontiers in Marine Science*, 8. <https://www.frontiersin.org/articles/10.3389/fmars.2021.620368>
- Uyà, M., Bulleri, F., & Gribben, P. E. (2018). Propagules are not all equal: Traits of vegetative fragments and disturbance regulate invasion success. *Ecology*, 99(4), 957–965. <https://doi.org/10.1002/ecy.2168>
- Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Aroita, G. (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2), 225–232. <https://doi.org/10.1111/2041-210X.13107>
- Van Etten, J., & Hijmans, R. J. (2010). A Geospatial Modelling Approach Integrating Archaeobotany and Genetics to Trace the Origin and Dispersal of Domesticated Plants. *PLoS ONE*, 5(8), e12060. <https://doi.org/10.1371/journal.pone.0012060>
- Vantarová, K. H., Eliáš, P., Jiménez-Ruiz, J., Tokarska-Guzik, B., & Cires, E. (2023). Biological invasions in the twenty-first century: A global risk. *Biologia*, 78(5), 1211–1218. <https://doi.org/10.1007/s11756-023-01394-7>
- Vasquez, M., Allen, H., Manca, E., Castle, L., Lillis, H., Agnesi, S., Al Hamdani, Z., Annunziatellis, A., Askew, N., Bekkby, T., Bentes, L., Doncheva, V., Drakopoulou, V., Duncan, G., Gonçalves, J., Inghilesi, R., Laamanen, L., Loukaidi, V., Martin, S., ... Virtanen, E. (2021). *EUSeaMap 2021. A European broad-scale seabed habitat map*. <https://archimer.ifremer.fr/doc/00723/83528/>
- Verlaque, M., Durand, C., Huisman, J. M., Boudouresque, C.-F., & Le Parco, Y. (2003). On the identity and origin of the Mediterranean invasive *Caulerpa racemosa* (Caulerpales, Chlorophyta). *European Journal of Phycology*, 38(4), 325–339. <https://doi.org/10.1080/09670260310001612592>
- Wang, L., & Jackson, D. A. (2023). Effects of sample size, data quality, and species response in environmental space on modeling species distributions. *Landscape Ecology*, 38(12), 4009–4031. <https://doi.org/10.1007/s10980-023-01771-2>
- Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecological Applications: A Publication of the Ecological Society of America*, 21(2), 335–342. <https://doi.org/10.1890/10-1171.1>
- West, A. M., Kumar, S., Brown, C. S., Stohlgren, T. J., & Bromberg, J. (2016). Field validation of an invasive species Maxent model. *Ecological Informatics*, 36, 126–134. <https://doi.org/10.1016/j.ecoinf.2016.11.001>
- West, E. J., Barnes, P. B., Wright, J. T., & Davis, A. R. (2007). Anchors aweigh: Fragment generation of invasive *Caulerpa taxifolia* by boat anchors and its resistance to desiccation. *Aquatic Botany*, 87(3), 196–202. <https://doi.org/10.1016/j.aquabot.2007.06.005>
- Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE*, 7(1), e29715. <https://doi.org/10.1371/journal.pone.0029715>

- Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2013). Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, 4(3), 236–243. <https://doi.org/10.1111/2041-210x.12004>
- Young, D. S., Chen, X., Hewage, D. C., & Nilo-Poyanco, R. (2019). Finite mixture-of-gamma distributions: Estimation, inference, and model-based clustering. *Advances in Data Analysis and Classification*, 13(4), 1053–1082. <https://doi.org/10.1007/s11634-019-00361-y>
- Zenetos, A., & Galanidi, M. (2020). Mediterranean non indigenous species at the start of the 2020s: Recent changes. *Marine Biodiversity Records*, 13. <https://doi.org/10.1186/s41200-020-00191-4>
- Zenetos, A., Gofas, S., Morri, C., Rosso, A., Violanti, D., Raso, J. G., Çinar, M. E., Almogi-Labin, A., Ates, A. S., & AZZURRO, E. (2012). Alien species in the Mediterranean Sea by 2012. A contribution to the application of European Union's Marine Strategy Framework Directive (MSFD). Part 2. Introduction trends and pathways. *Mediterranean marine science*, 13(2), 328–352.
- Zhu, G. P., Liu, Q., & Gao, Y. B. (2014). Improving ecological niche model transferability to predict the potential distribution of invasive exotic species. *Biodiversity Science*, 22(2), 223–230. Scopus.
- Zunino, S., Canu, D. M., Zupo, V., & Solidoro, C. (2019). Direct and indirect impacts of marine acidification on the ecosystem services provided by coralligenous reefs and seagrass systems. *Global Ecology and Conservation*, 18, e00625. <https://doi.org/10.1016/j.gecco.2019.e00625>
- Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, 7(6), 636–645. <https://doi.org/10.1111/2041-210X.12577>