Predicting the global distribution of reef fish abundance under climate change and human impacts

ReeFishBench v2

Description of the task:

In the context of ever-increasing human impacts and accelerating climate warming, we need to better understand and predict species occurrences and abundances over space and time. To this aim, Species Distribution Models (SDM) establish relationships between species occurrences and their habitat or social-environmental conditions enabling their persistence. By contrast, the prediction of species abundances over large scales is still challenging particularly for rare species or those showing aggregated distributions.

Local fish abundances are related to a myriad of contributions to Nature and People on temperate and tropical reefs worldwide such nutrient cycling or biomass production. So, the main Task of our DC Challenge is to predict the size class distribution of reef fish abundance using oceanographic (e.g. primary productivity), habitat (e.g. depth), and socioeconomic factors (e.g. marine protected areas). The goal is to cross-validate the models spatially and temporally to estimate reef fish abundances where data are sparse or lacking and in a near future according to climate and socioeconomic scenarios.

Datasets: Description of the proposed training, validation and test datasets (types, resolution, volume, etc.), of the required preprocessing steps

The dataset will come from the Reef Life Survey international program counting and monitoring more than 5,000 shallow reef fish species over 58 countries including France. More precisely this free to download dataset has gathered observations of 25,352,849 fish individuals sized and counted over 18,234 surveys (https://reeflifesurvey.com).

These surveys have been carried out with standardized protocols since 2006 and will constitute our training and validating datasets. Each survey corresponds to a site of 10,000 m^2 or 1ha with several transects. These surveys will continue during the challenge insuring perfect independent testing datasets through space (new locations) and time (future years 2025-2027). A large set of 24 oceanographic, habitat, and socioeconomic factors will be also provided by the consortium to feed the models at the training, validation and testing steps. Some additional factors could also come from other PPR Challenges providing data and predictions at the global scale to improve predictions.

Evaluation metrics: description of the proposed evaluation metrics, including the associated computational complexity and preprocessing steps if any.

We will use a series of complementary metrics to evaluate models' accuracy to fit with the challenge of predicting fish species abundances.

First, we will use the « L1 Log Loss »

$$(1)\ L1\ Log\ Loss = \frac{1}{n}\sum_{i=1}^{n}|log\ (y_i + 1) - log(\hat{y}_i + 1)|$$

where n = number of surveys, y_i = true abundance value of survey i, and $\hat{y}_i$ = predicted abundance value of survey i.

We will also use the D2 Absolute Log Error or D2log :

$$(2)\ D2log = 1 - \frac{\sum_{i=1}^{n}|log(y_i + 1) - log(\hat{y}_i + 1)|}{\sum_{i=1}^{n}|log(y_i + 1) - log(\tilde{y} + 1)|}$$

where n = number of species abundances, y_i = true abundance of a given species in a given survey, $\hat{y}_i$ = predicted abundance i, and $\tilde{y}$ = median of true abundances.

The R-squared on Log-transformed data or R2log will be computed as :

$$(3)\ R2log = 1 - \frac{\sum_{i=1}^{n}(log(y_i+1)-log(\hat{y}_i+1))^2}{\sum_{i=1}^{n}(log(y_i+1)-log(\bar{y}+1))^2}$$

where n = number of species abundances, y_i = true abundance i, $\hat{y}_i$ = predicted abundance i, and $\bar{y}$ = mean of true abundances.

Then the Spearman rank-order coefficient is adapted to regression problems with:

$$(4)\ \text{Spearman coefficient} = 1 - \frac{6\sum_{i=1}^{n}(u_i - \hat{u}_i)^2}{n(n^2 - 1)}$$

where n = number of species abundances, u_i = rank from smallest to largest of the i^{th} true specie abundance (in all true species abundances) and $\hat{u}_i$ = rank from smallest to largest of the i^{th} predicted specie abundance (in all predicted species abundances). Equal species abundances are assigned to the mean rank for their positions.

Baselines: description of the proposed baselines, both in terms of operational and learning-based baselines if any.

In the REEF-FUTURES project (https://www.biodiversa.eu/2022/10/31/reef-futures/), led by David Mouillot, we already predicted fish abundance using a set of Species Distribution Models (SDM) varying in complexity with respect to the relationship between species-specific fish biomass and 24 covariates (Appendix 1). This study entitled “A multi-model approach reveals the importance of environment, human and habitat for predicting species-specific fish abundance on shallow reefs worldwide” is currently under revision and will make the baseline for our DC proposal.

This baseline contains fish abundance predictions for GLMs, GAMs, random forests (RF) and gradient boosting machine (GBM). We also added two models that explicitly take into account the spatial structure in the data: a Spatial Mixed Model that includes geographical coordinates as a random effect, and a spatial RF that fits a local model at each RLS transect in addition to a global model.

All data, predictions, and codes will be free available in a Github repository related to the article even if not yet accepted.

Convolutional Neural Networks (CNN), Transformer or Large Language Models are expected to improve prediction accuracy by adopting a seascape approach, longer time series of oceanographic data and more relevant predictors, if any.

Related scientific references:

Caldwell, I.R., McClanahan, T.R., Oddenyo, R.M., Graham, N.A.J., Beger, M., Vigliola, L. *et al.* (2024). Protection efforts have resulted in ~10% of existing fish biomass on coral reefs. *Proceedings of the National Academy of Sciences*, 121, e2308605121.

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F. & Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLOS Computational Biology*, 17, e1008856.

Edgar, G.J., Stuart-Smith, R.D., Heather, F.J., Barrett, N.S., Turak, E., Sweatman, H. *et al.* (2023). Continent-wide declines in shallow reef life over a decade of ocean warming. *Nature*, 615, 858-865.

Estopinan, J., Servajean, M., Bonnet, P., Joly, A. & Munoz, F. (2024). Mapping global orchid assemblages with deep learning provides novel conservation insights. *Ecological Informatics*, 81, 102627.

Sanchez, L., Loiseau, N., Edgar, G.J., Hautecoeur, C., Leprieur, F., Manel, S. *et al.* (2024). Rarity mediates species-specific responses of tropical reef fishes to protection. *Ecology Letters*, 27, e14418.

Waldock, C., Stuart-Smith, R.D., Albouy, C., Cheung, W.W.L., Edgar, G.J., Mouillot, D. *et al.* (2022). A quantitative review of abundance-based species distribution models. *Ecography*, 2022.

Zamborain-Mason, J., Cinner, J.E., MacNeil, M.A., Graham, N.A.J., Hoey, A.S., Beger, M. *et al.* (2023). Sustainable reference points for multispecies coral reef fisheries. *Nature Communications*, 14, 5368.

Relevance of the proposal in terms of AI-native solutions for the DTO: (5-10 lines)

To provide a DTO on reef fish abundance which is of paramount importance to guide conservation strategies and better anticipate consequences of ongoing global change on marine resources, we will build-up an AI-native solution in the sense that predictive models will have AI at their core be they machine learning, natural language processing, or computer vision based. Such models will permit processing and analysing new data in real-time, which is crucial for recommendations in conservation like mitigating touristic frequentation or restricting fishing activities. They are also scalable and flexible models able to include more data, areas and years in the learning process. These AI-native solutions may also forecast future trends in fish abundance.

Moreover, our AI-native DTO for reef fish abundance can be interfaced with oceanographic DTOs from the PPR challenge predicting important factors shaping fish distribution patterns like sea surface temperature and primary productivity.

Relevance of the proposal in terms of topical demonstrations:

Our proposal will provide scientifically and socially motivated demonstrations through real-world datasets and benchmarks since the Reef Life Survey programme relies on citizen science, open data, and continuous data collection over space and time. So many free independent datasets will be generated for benchmarking and testing.

These demonstrations will contribute to the scientific priorities of the PPR “Océan & Climat” such as “Défi 3: Améliorer la protection et la résilience des milieux marins et le développement de nouvelles approches intégratives de gestion” with fish abundance assessments inside and outside marine protected areas and “Défi 4: Bénéficier durablement des ressources de l’océan en s’appuyant sur la science de la durabilité” by projecting reef fish abundance according to integrated climate and socioeconomic scenarios which is still rarely implemented but crucial to better understand spatiotemporal dynamics of marine resources.

Potential interactions with short-listed DC proposals

- DC1 SCOPE-Clim - A framework for benchmarking Single Column Ocean Physics Emulators for improving Climate predictions
- DC2 iGOR - emulating Global Ocean Reanalyses for ocean digital twins
- DC11 Probabilistic short-term forecasting of 3D ocean dynamics over the global ocean

We see a close interaction with these three challenges since regional and climatically relevant ocean geophysical dynamics are key drivers of benthic and fish assemblages. We could use these short-term probabilistic predictions to better forecast coastal fish abundance including uncertainty. It would reinforce our seascape approach.

- DC15 High-resolution tropical cyclones (TC) surface winds data challenge

Since our dataset is mainly tropical and that cyclones are notoriously known as major disturbances to coral reef habitats, we anticipate a close collaboration with this challenge. We plan to include the history of fine-scale wind intensity and direction as drivers of reef fish abundance and biodiversity.

- DC16 Combining AI and imaging to predict plankton and particulate matter at global scale in the ocean

Plankton is at the basis of most food webs in the ocean and acts on fish abundance. Yet current models in limited in their representation of this biological compartment both in terms of scale and composition. Advances proposed by this challenge could be integrated as forcing factors in our challenge. We also see some collaboration about implications of our findings on fisheries management and design of marine protection areas.

Even if our team has no “pure” specialist in ocean physics to identify relevant datasets and products we argue that:

- We have proven our ability to use available physico-chemical information in a relevant way to produce some breakthroughs in marine science:
    - https://www.science.org/doi/10.1126/sciadv.adn9660
    - https://www.mdpi.com/2072-4292/14/1/133
    - https://doi.org/10.1016/j.cub.2021.08.034

- We see opportunities to collaborate with other DC challenges having ocean physics as a main expertise.
- We have hired a research engineer in another project (Gaetan Morand: https://fr.linkedin.com/in/gaetanmorand) who will support our activities on data management (4 p.m.), oceanography, and IA.
- We will request 1 p.m. from the PPR to interconnect efficiently with other proposals and obtain the most relevant physico-chemical data.

Requested Engineer Resources

The proposed team will provide 4 p.m. (2 p.m. per versioning plan) to design and implement the proposed DC. This person will be Gaetan Morand (Research Engineer) from Ecole Centrale recruited during the next three years (2025-2027) on the ERC project led by David Mouillot.

Gaetan has expertise in data management and AI.
https://peercommunityjournal.org/articles/10.24072/pcjournal.471/
https://fr.linkedin.com/in/gaetanmorand
https://github.com/morand-g

He will also be a support for the hired postdoc.

We may need an additional 1 p.m. from the PPR to make our proposal interconnected with others, particularly in ocean physics.

DC Versioning Plan

In September 2025 we will provide the raw data used in the original baseline approach which is currently submitted under an international article entitled “A multi-model approach reveals the importance of environment, human and habitat for predicting species-specific fish abundance on shallow reefs worldwide”.

These data contain:
- Standardized fish abundance collected from 2006 to 2023 on 14,684 reef transects through the global Reef Life Survey (RLS) program (https://reeflifesurvey.com). Surveys consisted of underwater visual censuses by SCUBA divers along a 50 m transect line, where all fishes observed within 5 m of the transect line (5m wide and 5m high) were recorded and their abundance and size estimated. All data are freely available.
- 24 environmental, habitat and human covariates commonly used to explain fish abundance on shallow reefs (Table 1 in appendix).
- Free satellite images from Landsat and more recently Sentinel can be used as additional predictors towards a seascape approach using AI models based on imagery.

In September 2026 an update of this dataset will be provided since the Reef Life Survey (RLS) program is continuously collecting fish abundance worldwide. These new data will make ideal cross-validated forecasting tests. More precisely we will make fish abundance collected in 2024 and 2025 available in the same ready-to-use format. We also plan to include other relevant covariates (e.g. plankton, heat waves, cyclones) provided by other DC challenges.

All training and testing datasets will be available under the FAIR principles (Findability, Accessibility, Interoperability, and Reuse): https://www.nature.com/articles/sdata201618.

Scientific and technical staff involved: Description of the key personnel for the proposal. For each member, provide a short (<5 lines) CV and describe the key ocean and/or AI expertise for the proposal.

David Mouillot (University of Montpellier, MARBEC) has contributed to the study of social-ecological changes on coastal reef ecosystems worldwide to better understand the interlaced effects of climate change and socioeconomic factors, including marine protected areas, on fish biodiversity and ecosystem functioning. More recently, he developed new tools in artificial intelligence towards a better prediction of socio-ecosystem characteristics from fish biodiversity to human poverty. He has been recognized as Highly Cited Researcher in Ecology and Environment since 2016, published more than 350 articles, and is part of the MARBEC laboratory specialized in marine biodiversity and conservation with a focus on modelling approaches.
https://www.researchgate.net/profile/David-Mouillot/research

Alexis Joly (INRIA, LIRMM) has a background in analysis of high-dimensional data, representation learning, and efficient nearest neighbors methods. He is now applying these methods to biodiversity informatics and ecological modeling, in particular through the scientific direction of the Pl@ntNet research platform. He is experienced in AI Challenges through the organization of GeoLifeCLEF or through the design of new models such as dynamic species distribution models or hybrid AI models aimed at jointly modelling the sampling effort and the species distribution. He is part of the LIRMM laboratory having artificial intelligence and data science at the heart of activities for several decades.
https://www.researchgate.net/profile/Alexis-Joly

Rodolphe Devillers (IRD, ESPACE-DEV) works on sustainable development and coastal habitats IRD. He specializes in geospatial sciences applied to the marine environment, specifically in the areas of marine conservation. He coordinated the development of the Global Information System on Small-scale Fisheries 'ISSF' - the world's largest database on small-scale fisheries, available as open data. He is a part of the Espace-Dev research laboratory having a strong remote sensing component and focusing on various dimensions of global change that characterize the societal and environmental dynamics, as well as methodological research on data science and modeling.
https://www.researchgate.net/profile/Rodolphe-Devillers

Table 1: Set of covariates included in the DC by category (environmental, human, habitat) and scale. The root parts "_7days", "_1year", "_5year" represents a summary of a covariate (mean, min or max) across a 7 day, 1 year, 5 year period respectively, before a transect was conducted.

Covariate names	Scale	Category
mean_1year_Chlorophyll_a	local	environmental
mean_7days_Chlorophyll_a	local	environmental
max_5year_Degree Heating Weeks	local	environmental
min_5year_pH	local	environmental
mean_1year_Sea Surface Salinity	local	environmental
mean_7days_Sea Surface Temperature	local	environmental
min_1year_Sea Surface Temperature	local	environmental
max_1year_Sea Surface Temperature	local	environmental
Density of fishing vessels	local	human
Natural resource rent	country	human
Human Development Index (HDI)	country	human
Human Travel Time	local	human
Human Gravity	local	human
Growth Development Product	local	human
Marine Protected Areas	local	human
Number of Non-Governmental Organizations	country	human
Benthic Habitat	local	habitat

Algae	local	habitat
Sand_500mbuffer	local	habitat
Coral_ 500mbuffer	local	habitat
Rock_500mbuffer	local	habitat
Rubble_500mbuffer	local	habitat
Reef_extent_10kmbuffer	regional	habitat
Depth	local	habitat