Received: 15 August 2024 | Accepted: 27 March 2025

DOI: 10.1111/2041-210X.70040

Methods in Ecology and Evolution

PRACTICAL TOOLS

CEPHALOPOD, a package to standardize marine habitat-modelling practices and enhance inter-comparability across biological observations

Alexandre Schickele[1] | Corentin Clerc[1] | Fabio Benedetti[1] | Daniele De Angelis[2,3] | Urs Hofmann Elizondo[1] | Matthias Münnich[1] | Jean-Olivier Irisson[4] | Meike Vogt[1]

[1]ETH Zürich, Institute of Biogeochemistry and Pollutant, Zurich, Switzerland

[2]Department of Biology and Biotechnology 'Charles Darwin', Sapienza University of Rome, Rome, Italy

[3]ISPRA, Institute for Environmental Protection and Research, Rome, Italy

[4]Laboratoire d'Océanographie de Villefranche, LOV, CNRS, Sorbonne Université, Villefranche-sur-Mer, France

Correspondence
Alexandre Schickele
Email: alex.schickele@gmail.com

Funding information
Horizon 2020 Research and Innovation Programme, Grant/Award Number: 101059915, 101094227 and 862923

Handling Editor: Fernando González Taboada

Abstract

1. As the volume of accessible marine pelagic observations increases exponentially, incorporating diverse data types such as metagenomics and quantitative imaging, the need for standardized modelling frameworks becomes critical to predict biogeographic patterns in space and time and across the diverse range of emergent sampling methods. In response, we introduce CEPHALOPOD (Comprehensive Ensemble Pipeline for Habitat modelling Across Large-scale Ocean Pelagic Observation Datasets), a standardized, highly automated and flexible framework designed to integrate and analyse heterogeneous marine data for multi-species habitat modelling following best practices in the field.
2. CEPHALOPOD is built on observational data from federating initiatives such as AtlantECO, OBIS, GBIF, associated with already existing statistical and machine learning methods that enable to extract and model information from heterogeneous, scarce and biased field observations. It is highly automated and follows explicit quality checks informing the user of the predictive accuracy and interpretability of the results.
3. Here, we document our statistical ensemble modelling approach and then assess its strengths and limitations with a virtual ecologist approach. We show how our framework performs in reproducing a range of distributions from biased field samples.
4. Our modelling framework serves as a foundation for the consistent generation of Essential Biodiversity and Ocean Variables (EBVs and EOVs) and carries the potential to significantly advance our comprehension of biodiversity and marine ecosystem functioning. Finally, it provides an unprecedented opportunity to foster collaborations in the field of marine science, sustainable ecological practices, and ultimately contribute to the preservation of global marine biodiversity.

KEYWORDS
biodiversity, data harmonization, marine species distribution modelling, standardized quality assessment, virtual species

1 | INTRODUCTION

Pelagic organisms including plankton and fish structure marine food webs and support ecosystem services such as carbon sequestration, oxygen production, fisheries and food security (Barbier, 2017). These organisms are incredibly diverse, and their presence or absence, taxonomic or functional composition, abundance and biomass are all accepted indicators of ocean ecosystem structure, health and resilience to climate change, known as Essential Biodiversity and Ocean Variables (EBVs and EOVs). However, marine biodiversity is still critically understudied, with thousands of species whose spatial distribution remains unknown (e.g., Carradec et al., 2018). Nowadays, biological data collection has rapidly increased due to the development of quantitative imaging (Lombard et al., 2019), genome-based approaches (Carradec et al., 2018) and online data repositories (e.g., Vogt et al., 2023). Nevertheless, their potential to fill knowledge gaps in marine biodiversity and ecosystem functioning remains hampered by their heterogeneity regarding spatio-temporal coverage and resolution, accessibility, metadata harmonization and specific ecological processes associated with each data type (e.g., Dallas & Hastings, 2018).

Niche theory, proposed by G.E. Hutchinson (1957), provided the foundation for the development of species distribution models (SDMs) which estimate the potential biogeographic distribution of a species based on the environmental conditions it is observed in (Peterson & Soberón, 2012). Initially developed for species occurrence data (Thuiller et al., 2009), SDMs are now also applied—although marginally used—to continuous biomass or abundance data (Knecht et al., 2023; Waldock et al., 2022) and proportions from metagenomic reads (Schickele et al., 2024). They benefit from methods and guidelines transferred from other fields of research, that enhance data integration, and lead to a more robust assessment of their predictive capabilities, explicit uncertainties and quality control guidelines (Descombes et al., 2022; Roberts et al., 2017; Waldock et al., 2022). However, associating each output with explicit, multimetric and community-approved quality control and metadata is still challenging. For instance, multiple taxonomic backbones are used or even developed, the associated metadata is highly heterogeneous and inherent data limitations remain, such as the management of heterogeneous sampling effort, biases in abundance measurements and the multivariate nature of omics data. Therefore, there is a need for a common comprehensive SDM framework that sources, combines and integrates this wealth of marine biological datasets across biodiversity metrics, data types, spatio-temporal scales and taxa (Miller et al., 2019), to produce high-quality, near real-time, standardized and comprehensive EOVs and EBVs, as acknowledged at the international level (IPBES, 2016; IUCN Standards and Petitions Subcommittee, 2017).

In this study, we introduce the Comprehensive Ensemble Pipeline for Habitat modelling Across large-scale Ocean Pelagic Observation Datasets (CEPHALOPOD), which unifies existing methods for presence only (Benedetti et al., 2021), continuous (Knecht et al., 2023) and proportion data (Schickele et al., 2024) within a common SDM framework tailored for the marine environment. Its novelty lies in the production of inter-comparable biogeographical estimates, EOVs and EBVs across multiple data types. To do so, it directly accesses—but is not limited to—a large and updated collection of fish and plankton datasets and a new collection of 47 standardized environmental predictors. CEPHALOPOD extrapolates information embedded in scarce and biased marine datasets, and it is associated with comprehensive, harmonized and explicit quality checks.

2 | MATERIALS AND METHODS

2.1 | Overview of the workflow

2.1.1 | Data preparation

Based on R (v4.2) and Python (v3.9) languages, CEPHALOPOD analyses relationships between presence only, continuous or proportional biological target variables and environmental conditions. To this end, we need to (i) access both in-situ biological observations and (ii) environmental climatologies, (iii) retrieve and harmonise the corresponding relevant metadata (i.e., including taxonomy) and (iv) ensure parsimonious inputs (i.e., discard biological outliers or non-informative environmental climatologies; see detailed methods in Section S1).

The pipeline first accesses in-situ observation within—but not limited to—the Ocean Biodiversity Information System (OBIS; http://obis.org), Global Biodiversity Information Facility (GBIF; http://gbif.org), AtlantECO (Vogt et al., 2023) and Marine Atlas of Tara Ocean Unigenes (MATOU; Carradec et al., 2018) datasets (Figure 1, steps 1–3). Metadata requirements correspond to a parsimonious subset of the Darwin Core format (Wieczorek et al., 2012), relevant for marine biogeographical data and harmonized against a reference taxonomy (i.e., World Register of Marine Species; WoRMS Editorial Board, 2024). After selecting target taxa, in-situ observations are associated with environmental variables known to influence plankton biogeography (Figure 1, step 4; Table S1). These characterize the physical (e.g., sea surface temperature), chemical (e.g., nutrient concentration, carbonate chemistry) and biological (e.g., Chlorophyll-a concentration, primary production) conditions of water masses as well as their circulation and turbulence patterns (e.g., Eddy Kinetic Energy; EKE). While this collection of environmental features is mainly targeted

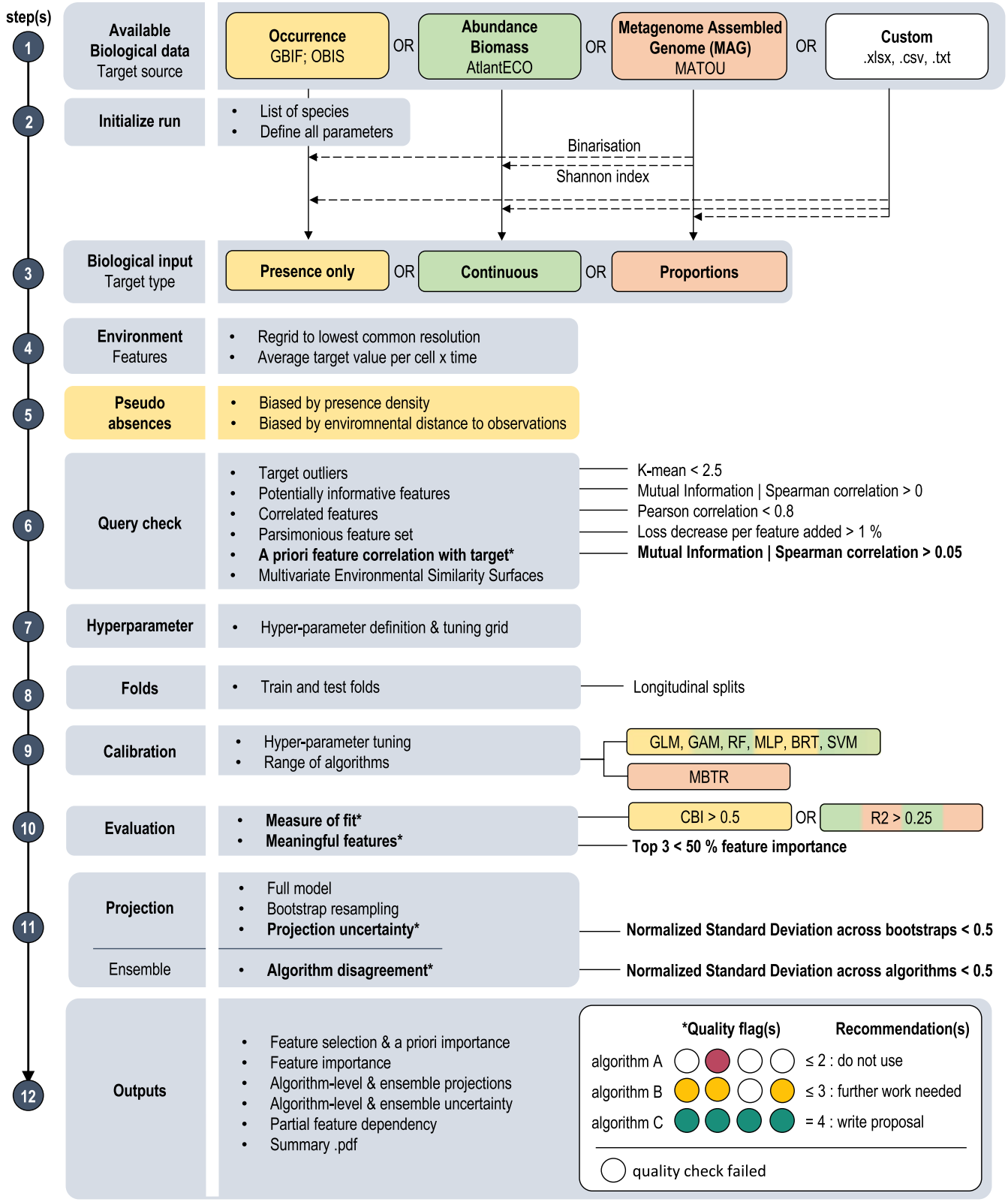

FIGURE 1 Synthetic diagram of the CEPHALOPOD indicating the essential list of modelling choices, quality checks and flags used in the workflow. The main steps are indicated on the left-hand side. The associated quality checks or modelling choices are displayed in the middle panels. The corresponding thresholds used for quality flags are indicated on the right-hand side. Those used in the final traffic light quality flag are indicated in bold. Grey, yellow, green or orange background corresponds to common, presence-only, continuous or proportions specific steps, respectively.

towards the epipelagic ocean, other depth layers are accepted by CEPHALOPOD (see Section S1). To compensate for the lack of true absences when using presence-only observations, CEPHALOPOD samples pseudo-absences following by default the nearby density of presences and in equal number and weight to the presences (additional options in Supporting Information), to minimize the prediction variance (Descombes et al., 2022; i.e., D-design theory; Montgomery, 2017).

Both biological (target) and environmental (features) data undergo a series of common pre-processing and quality checks before model fitting. To avoid overparameterization, CEPHALOPOD constructs a set of non-collinear and informative environmental features (Figure 1, step 6) by discarding environmental features that are unlikely to explain the target distribution, performing a pairwise correlation of the features at the target location, and a Recursive Feature Elimination procedure (RFE; Darst et al., 2018). Each pre-processing step is quality flagged (see Supplementary Information, section 1.4; Figure 1, step 6). Finally, to identify the regions where SDM extrapolates under novel conditions, CEPHALOPOD performs a Multivariate Environmental Similarity Surfaces (MESS; Elith et al., 2010). At this stage, CEPHALOPOD integrated, processed and quality-checked all input data and parameters necessary to calibrate an SDM algorithm.

2.1.2 | Predicting the distribution of the biological target

CEPHALOPOD employs an ensemble modelling approach, meaning it prioritizes spatial distribution estimates constructed from several algorithms (see detailed methods in Section S2). To account for spatial autocorrelation and optimize the algorithm training, CEPHALOPOD first selects the best hyperparameters using a Spatial Block Cross-Validation (SBCV; Figure 1, steps 7–9) (Roberts et al., 2017). SBCV offers a good compromise between spatial independence, balanced data partitioning and sufficient data for both training and evaluation (Roberts et al., 2017). Proposed alternatives such as leave-one-region-out or environmental block cross-validation require predefined regions or selected features to split the dataset, a criterion that can vary across taxa and spatial scales (Roberts et al., 2017). Therefore, SBCV offers a higher degree of automation and standardization across multiple applications of CEPHALOPOD. Each algorithm output is then evaluated on the same cross-validation splits to estimate its performance in reproducing observations. Furthermore, we compute the importance of each feature in explaining these observations (Figure 1, step 10). Both the model predictive performance and the interpretability of the feature importance are associated with a quality flag (Figure 1, step 10). To account for uncertainty in the spatial projections, CEPHALOPOD performs a bootstrap procedure for all combinations of algorithms and taxa passing the previous quality controls (Figure 1, step 11). We perform a final quality control on the projected uncertainty associated with each algorithm and across algorithms composing an ensemble projection (i.e., ensemble agreement; Figure 1, step 11).

CEPHALOPOD provides graphical outputs for each critical step and structured objects containing all inputs and outputs, enabling further analyses. The quality checks associated with each set of observations are represented as a traffic light with recommendations alongside monthly spatial projections of each taxa passing the above-mentioned quality checks (Figure 1, step 12). Finally, CEPHALOPOD can compute partial dependence plots for each feature included in the SDMs (Figure 1, step 12) and a range of Hill-based diversity indices to highlight biodiversity hotspots across the global ocean (Qiao et al., 2024).

2.2 | Quality assessment and flags

CEPHALOPOD features a multi-metric SDM evaluation integrating four standardized quality flags. First, the a-priori importance of the considered environmental features in explaining the biological target is tested by their average shared Spearman correlation and mutual information (Kinney & Atwal, 2014), which must exceed that with a null target by at least 0.05 (Figure 1, step 6) (i.e., adapted from Knecht et al., 2023). Second, CEPHALOPOD estimates the predictive performance of each algorithm by quantifying the quality of the corresponding evaluation split prediction (Figure 1, step 10). To do so, CEPHALOPOD computes (i) a Continuous Boyce Index (CBI; Hirzel et al., 2006), (ii) a univariate r-squared (R^2; Pearson) or (iii) a multivariate r-squared (Pearson) for presence only, continuous and proportional data, respectively. Metrics range from −1 to 1, with thresholds set at 0.5 for presence only data or 0.25 for continuous and proportional data to ensure sufficient predictive power (Figure 1, step 10). Third, CEPHALOPOD extracts the importance of each feature in explaining the observed patterns in the target (i.e., see function specification in the 'tidymodels' R library, version 1.1.1). To better interpret underlying ecological processes, informative features should explain a substantial part of the observed variability. Thus, CEPHALOPOD retains only algorithms with a cumulative variable importance above 50% for the top three features (Figure 1, step 10). Algorithms successfully passing the above-mentioned quality checks are retained to perform habitat suitability projections at the global scale. Finally, CEPHALOPOD computes the scale independent Normalized Standard Deviation (NSD), defined as the standard deviation normalized by the maximum average across bootstrap runs, for each algorithm, month and grid cell. To quantify the global projection uncertainty for each algorithm, the pipeline averages the NSD across the geographical projection domain and associates a quality check: CEPHALOPOD retains only algorithms with an NSD averaged across grid cells under 0.5 (Figure 1, step 11). The pipeline also computes the NSD between each algorithm considered in the ensemble projection as a proxy for disagreement among algorithms. CEPHALOPOD retains only ensembles with an NSD across algorithms under 0.5 (Figure 1, step 11).

Sequential quality checks are summarized with a traffic light system: red, yellow or green light indicates success across 2, 3 or all 4 quality checks, respectively (Figure 1, step 12, Figure S6), along with recommendations for interpretation.

3 | VALIDATION

3.1 | Virtual species design

We assessed CEPHALOPOD's skill in reproducing simulated species distributions across presence only, continuous and proportional data through a 'virtual ecologist approach' (Zurell et al., 2010). We simulated 20 virtual niches, (see detailed methods in Section S3), producing spatial distributions with prevalences (i.e., the proportion of occupied geographical cells) ranging from 0 to 1, to mimic both ubiquitous and rare species. These 20 simulated distributions served as ground truth and the diversity of niche shapes mimic various assemblages (e.g., ecosystems of equally abundant or co-existing species; one dominating species) at each geographical location. To generate discrete presence records from the continuous habitat suitability maps (where values are in [0,1]), we performed a binomial trial centred at 0.5 per sampled cell; if the value of the binomial trial was below the habitat suitability, the species was considered present in that cell.

We generated samples with three different sampling efforts ($N = 50$, 200, 1000) and three levels of geographical bias: (i) homogenous distribution across the global ocean, (ii) coastal bias and (iii) North Atlantic bias to replicate common spatial biases (Figure 2). The lower boundary corresponds to a conservative and commonly admitted minimum sample number in the field (Benedetti et al., 2021; Righetti et al., 2023). For continuous and proportional data, we added background noise to the sampled values (i.e., log-normal bias with standard deviation of 1 and 2; Figure 2), to simulate variability from local processes and measurement error. This virtual species design results in 180 datasets for presence-only and 540 datasets for continuous or proportional data.

3.2 | Predictive performance

The resulting projections that passed all four quality checks present a median Pearson correlation with the true distribution of 0.79, 0.82 and 0.58 for presence only, continuous and proportions data, respectively (Figure 3a–c). This means that CEPHALOPOD produces

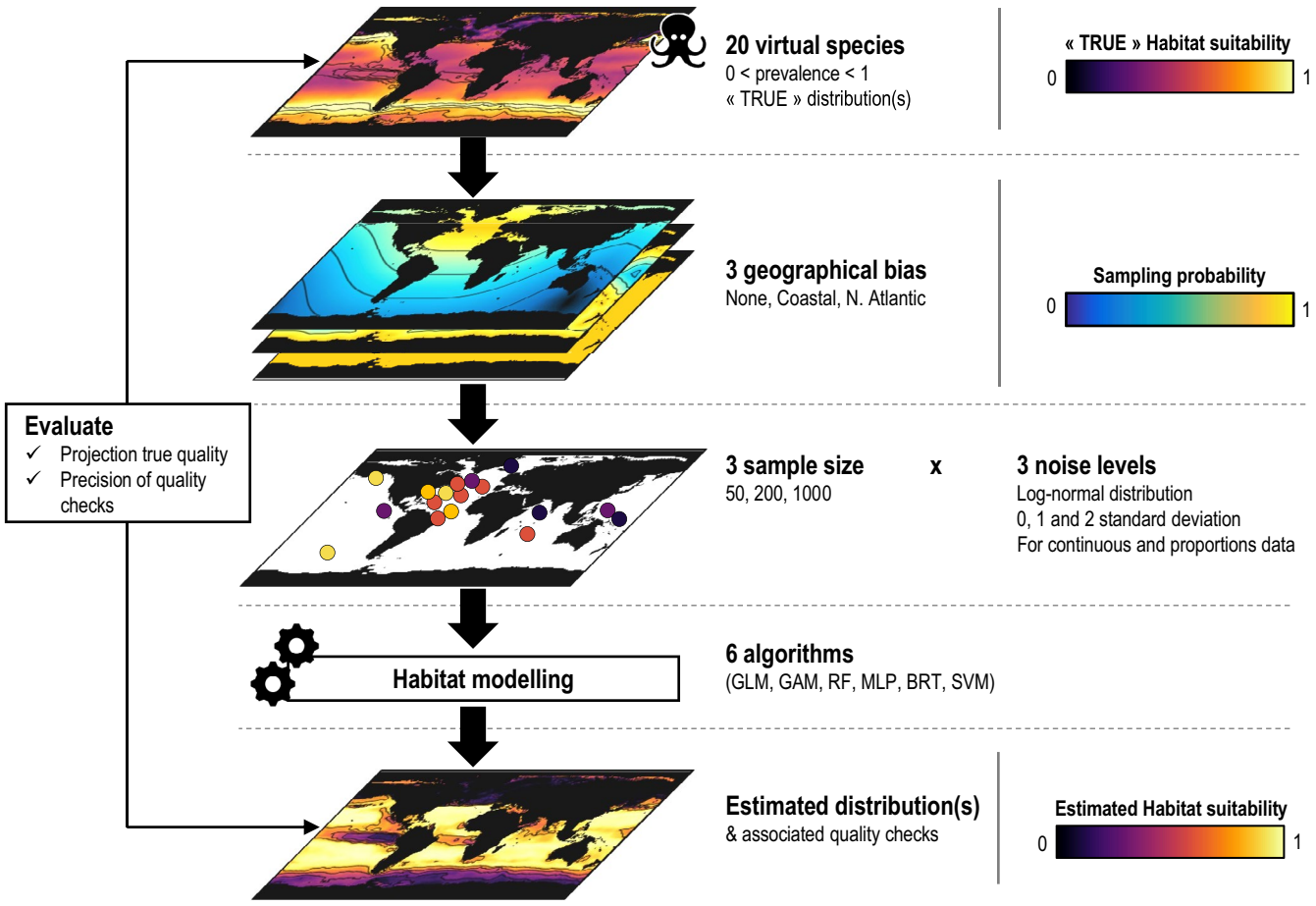

FIGURE 2 Synthetic diagram of the framework testing on virtual species. The virtual distributions, geographical biases, sample sizes and noise levels are indicated in lines.

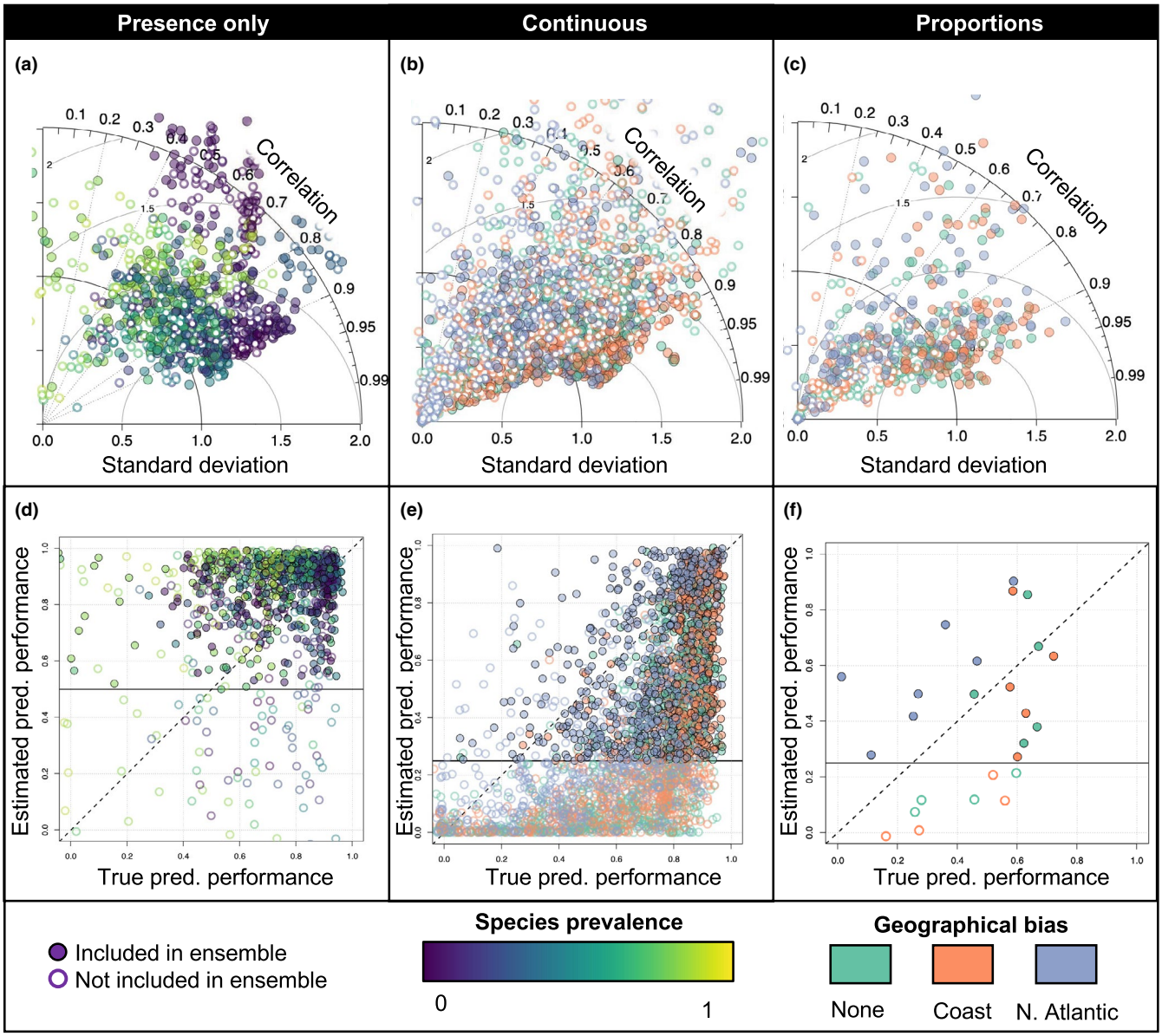

FIGURE 3 Habitat-modelling performance on 20 virtual species for presence only (a, d), continuous (b, e) and proportion data (c, f). Each circle corresponds to a combination of prevalence, sample size, background noise, geographical bias and algorithm. For each data type, the circles are coloured by the factor explaining the most variability. The top panels (a–c) compare the estimated and true distribution. The bottom panels (d–f) compare the estimated predictive performance (i.e., observed vs. predicted evaluation split) to the true performance (i.e., estimated vs. true distribution). The 1:1 dashed line indicates a perfect performance estimation. The horizontal black line indicates the threshold under which an algorithm is discarded. Note that the estimated predictive performance is computed simultaneously across all targets due to the multivariate nature of proportion data (*F*).

distribution estimates whose spatial patterns are consistent with the corresponding true distribution across all data types and sampling strategies. To assess CEPHALOPOD's skill further, we performed an analysis of variance (ANOVA) to identify which factors affect the model performance in predicting the true distribution. While most factors show a significant effect on true predictive performance, we highlight only a major contribution of prevalence (53%) when using presence only data and geographical bias (45% and 46%) when using continuous or proportional data (Figure S2). This suggests CEPHALOPOD may yield lower quality, though consistent, estimates for ubiquitous species with presence-only data (i.e., prevalences >0.8; Pearson correlation of 0.56; Figure 3a) or for regionally biased samples in continuous or proportional data (i.e., North Atlantic bias; Pearson correlation of 0.73 and 0.26, respectively; Figure 3b,c).

To estimate the accuracy of CEPHALOPOD's quality reporting, we compare the true predictive performance of each algorithm against the one estimated by the corresponding quality checks (i.e., CBI or R^2; see Supplementary Information, section 1.4; Figure 1, step 10). The deviation between the true predictive performance and the one estimated by CEPHALOPOD presents an RMSE of 0.27, 0.31

and 0.26 (Figure 3d–f). This means that the predictive performance estimated by CEPHALOPOD is consistent or under-estimated (i.e., for continuous data; Figure 3e), compared to the true predictive performance. To further investigate the robustness of our quality checks, we performed an ANOVA to identify which sampling strategy affects the false positive rate in the estimated predictive performance. While most factors show a significant effect, we highlight only a major contribution of prevalence (73%) when using presence only data and geographical bias (34% and 43%) when using continuous or proportional data (Figure S4). Thus, false positive quality flags may arise with presence only data for ubiquitous species (Figure 3d) and continuous or proportional data for regionally biased samples (i.e., north Atlantic bias; Figure 3e,f).

Overall, our virtual ecologist approach shows that each algorithm considered in CEPHALOPOD performs well across sampling factors and biases (median Pearson correlation of 0.72; Figure 3), with a limited risk of false positive quality flags on distribution estimates (Figure 3).

4 | DISCUSSION

4.1 | Advantage of CEPHALOPOD

The key strengths of CEPHALOPOD are (i) its ability to harmonize observed distributions, (ii) its standardized, multi-metric quality flags and (iii) its capacity to produce inter-comparable projections for various biological targets across multiple data types. The virtual ecologist approach highlights the overall good performance of CEPHALOPOD across multiple data types and sampling biases common to marine biogeographical data (Hughes et al., 2021) and its value in identifying cases of poor model performance, which limits the risk of misuse or misinterpretation of the estimated distributions. Moreover, the traffic light recommendations comprehensively report on the reliability of model outputs and thus match the requirements for minimum SDM standards (Flombaum & Martiny, 2021). However, its integration within international platforms such as EMODnet and the Digital Twin of the Ocean necessitates metadata templates designed for interoperability, distinct from traditional SDM formats (Zurell et al., 2020). Furthermore, CEPHALOPOD is largely automated and parallelized in its data pre-processing and modelling steps, therefore, greatly facilitating the estimation of species distribution and diversity across a large range of taxa. Indeed, while a minimum sample size of 50 observations is required by CEPHALOPOD, there is no predefined upper limit to the number of occurrences or species other than available computation resources. Therefore, CEPHALOPOD is well suited for large-scale (e.g., small pelagic fish, copepods; http://obis.org) or multi-species applications (e.g., species diversity estimates). Other SDM frameworks propose similar quality checks (e.g., predictive performance, uncertainty and extrapolation quantification), modelling options (e.g., ensemble modelling, hyper-parameter tuning, pseudo-absence selection) and outputs (e.g., projections, variable importance). However, they are often specifically designed for one type of data (e.g., occurrences; Thuiller et al., 2009), therefore, limiting the possibilities of integration and inter-comparison across studies or outputs from other data types such as omics or quantitative imaging.

4.2 | Caveats and limitations

The virtual species approach suggests CEPHALOPOD may overestimate the distribution quality when using presence-only to model ubiquitous species, or for geographically biased samples. The former is a known long-standing issue in species distribution modelling, as model evaluation is particularly difficult when the environmental conditions of a species' presence overlap with those of the background data (Tsoar et al., 2007). Geographically biased sampling effort is also known to bias distribution estimates when the environmental conditions of observations are not representative of the fundamental niche of a taxa, commonly known as niche truncation (Peterson & Soberón, 2012). While niche truncation and overestimation issues are inherent limitations of SDMs, CEPHALOPOD includes a range of automated quality checks and pre-processing strategies designed to alleviate their effect on projected distributions (e.g., biological outlier filtering, spatial discontinuity or extrapolation detection). However, CEPHALOPOD is designed to complement, not substitute, expert-driven ecological validation (Schickele et al., 2020). To further mitigate niche truncation, we recommend expert curation of occurrence records that can help identify whether observed distributions reflect true limits or sampling artefacts (Schickele et al., 2020). We also recommend avoiding basin-specific environmental variables such as salinity when regions of higher salinity such as the North Atlantic Ocean are disproportionally sampled (Benedetti et al., 2021). For overestimation in projections, we recommend the use of the expert target group approach (Righetti et al., 2023) and range restrictions (Waldock et al., 2022) further constrain extrapolation. Moreover, the collection of available environmental features is non-exhaustive and primarily optimized for epipelagic species to date. To address this, we allow expert input of environmental features, constructed outside the CEPHALOPOD framework, provided they follow the required format (e.g., bottom or demersal layers derived from the WOA). CEPHALOPOD offers the possibility to define fixed depth levels for each modelling run. However, depth layers are treated independently. Therefore, CEPHALOPOD should be considered flexible in matching environmental layers with the depth of the studied taxa, but not as a 3-dimensional depth-structured model (e.g., Owens & Rahbek, 2023). Although this version of CEPHALOPOD is not yet adapted to integrate depth-structured algorithms that are standardized across data types, this functionality remains a potential avenue for future development, alongside the integration of CMIP6 climate projections or diel vertical migration dynamics. Additionally, the quantitative observations retrieved from traditional biomass or omics-based data are not always comparable among sampling protocols, due to differing capturability or completeness of the sample across net types, mesh sizes, sequencing depth or divergent taxonomic nomenclatures (Lombard et al., 2019). To face these inherent

limitations, CEPHALOPOD incorporates best available practices and provides explicit insights into its performance and biases along with guidance towards informed decision-making in diverse ecological contexts.

5 | CONCLUSION

CEPHALOPOD advances traditional habitat-modelling tools by incorporating multidimensional and diverse data types to study marine ecosystems in an inter-operable and comparable format. As a highly automated algorithm, it is particularly suited for integration with international data repositories (e.g., OBIS; GBIF; Vogt et al., 2023) within the Digital Twin of the Ocean framework. The flexible and standardized formatting of in-situ observations datasets in CEPHALOPOD supports new observation types and has the potential to address evolving demands of multidimensional ecological research (Ashcroft et al., 2017). Its ability to generate inter-comparable EOVs and EBVs enhances our capacity to monitor, understand and predict marine ecosystem dynamics. As such, CEPHALOPOD provides a robust, scalable and operational tool of interest for a broad audience across ecological researchers, environmental managers, conservation practitioners and policymakers engaged in biodiversity assessment and ocean monitoring.

AUTHOR CONTRIBUTIONS

Meike Vogt conceived and supervised the study and provided funding. Alexandre Schickele, Corentin Clerc and Fabio Benedetti processed the input and predictor data and provided expert input. Alexandre Schickele wrote the first draft, conceived, designed the modelling pipeline, and performed the analysis. Jean-Olivier Irisson and Daniele De Angelis provided expertise on SDM modelling across data types. All authors substantially contributed to the successive versions of the modelling pipeline and of this manuscript.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 862923 (AtlantECO), no. 101094227 (Bluecloud2026) and no. 101059915 (BIOcean5D). This output reflects only the author's view, and the European Union cannot be held responsible for any use that may be made of the information contained therein. The authors want to thank all people involved in the Tara Oceans and AtlantECO projects for making data publicly available. M.V. would like to thank N. Knecht, L. Gregor for the conception of a first prototype during the first Bluecloud hackathon. M.V., F.B. and D.D.A. would like to acknowledge L. Maiorano and the AtlantECO WP2 team for data provision and model expertise. A.S., C.C., M.V. and U.H.E. would like to thank S. Pesant, R. Finn, L. Richardson, L. Meszaros, D. Obaton and the Bluecloud scientific and technical coordination team for fruitful discussion and advice on data structures and cloud-computing. A.S., C.C. and M.V. would like to acknowledge Cael B.B., G. Chust and the B5D modelling community for valuable input on modelling standards during a recent workshop. A.S., C.C., M.V. and U.H.E. would like to thank S. Lunven for extensive testing of the pipeline.

FUNDING INFORMATION

Bluecloud2026: Horizon 2020 Research and Innovation Programme, grant agreement no. 101094227. BIOcean5D: Horizon 2020 Research and Innovation Programme, grant agreement no. 101059915. AtlantECO: Horizon 2020 Research and Innovation Programme, grant agreement no. 862923.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70040.

DATA AVAILABILITY STATEMENT

Data and code available via zenodo https://zenodo.org/records/15168238 (Schickele & Hofmann Elizondo, 2025).

ORCID

Alexandre Schickele https://orcid.org/0000-0002-4842-6890
Corentin Clerc https://orcid.org/0000-0002-8436-4391

REFERENCES

Ashcroft, M. B., King, D. H., Raymond, B., Turnbull, J. D., Wasley, J., & Robinson, S. A. (2017). Moving beyond presence and absence when examining changes in species distributions. *Global Change Biology*, *23*(8), 2929–2940. https://doi.org/10.1111/gcb.13628

Barbier, E. B. (2017). Marine ecosystem services. *Current Biology*, *27*(11), R507–R510. https://doi.org/10.1016/j.cub.2017.03.020

Benedetti, F., Vogt, M., Elizondo, U. H., Righetti, D., Zimmermann, N. E., & Gruber, N. (2021). Major restructuring of marine plankton assemblages under global warming. *Nature Communications*, *12*(1), 5226. https://doi.org/10.1038/s41467-021-25385-x

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1), 373. https://doi.org/10.1038/s41467-017-02342-1

Dallas, T. A., & Hastings, A. (2018). Habitat suitability estimated by niche models is largely unrelated to species abundance. *Global Ecology and Biogeography*, *27*(12), 1448–1456. https://doi.org/10.1111/geb.12820

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*, *19*(1), 65. https://doi.org/10.1186/s12863-018-0633-8

Descombes, P., Chauvier, Y., Brun, P., Righetti, D., Wüest, R. O., Karger, D. N., Zurell, D., & Zimmermann, N. E. (2022). Strategies for sampling pseudo-absences for species distribution models in complex mountainous terrain. *bioRxiv*. https://doi.org/10.1101/2022.03.24.485693

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Flombaum, P., & Martiny, A. C. (2021). Diverse but uncertain responses of picophytoplankton lineages to future climate change. *Limnology and Oceanography*, *66*(12), 4171–4181. https://doi.org/10.1002/lno.11951

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*(2), 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., & Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, *44*(9), 1259–1269. https://doi.org/10.1111/ecog.05926

Hutchinson, G. E. (1957). *Concluding remarks*. Cold Spring Harb. Symp. Quant. Biol., 22, 145–159, Cold Spring Harbor Lab Press, Publications Dept.

IPBES. (2016). The methodological assessment report on scenarios and models of biodiversity and ecosystem services. *Zenodo*. https://doi.org/10.5281/zenodo.3235429

IUCN Standards and Petitions Subcommittee. (2017). *Guidelines for using the IUCN red list categories and criteria*.

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(9), 3354–3359. https://doi.org/10.1073/pnas.1309933111

Knecht, N. S., Benedetti, F., Hofmann Elizondo, U., Bednaršek, N., Chaabane, S., de Weerd, C., Peijnenburg, K. T. C. A., Schiebel, R., & Vogt, M. (2023). The impact of zooplankton Calcifiers on the marine carbon cycle. *Global Biogeochemical Cycles*, *37*(6), e2022GB007685. https://doi.org/10.1029/2022GB007685

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., Sosik, H. M., Schulz, J., Romagnan, J.-B., Picheral, M., Pearlman, J., Ohman, M. D., Niehoff, B., Möller, K. O., Miloslavich, P., Lara-Lpez, A., Kudela, R., Lopes, R. M., Kiko, R., … Appeltans, W. (2019). Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science*, *6*, 196. https://doi.org/10.3389/fmars.2019.00196

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/2041-210X.13110

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons.

Owens, H. L., & Rahbek, C. (2023). voluModel: Modelling species distributions in three-dimensional space. *Methods in Ecology and Evolution*, *14*(3), 841–847. https://doi.org/10.1111/2041-210X.14064

Peterson, A., & Soberón, J. (2012). Species distribution modeling and ecological niche modeling: Getting the concepts right. *Natureza & Conservação*, *10*(2), 102–107. https://doi.org/10.4322/NATCON.2012.019

Qiao, H., Orr, M. C., & Hughes, A. C. (2024). Measuring metrics: What diversity indicators are most appropriate for different forms of data bias? *Ecography*, *2024*(9), e07042. https://doi.org/10.1111/ecog.07042

Righetti, D., Vogt, M., Gruber, N., & Zimmermann, N. E. (2023). Mapping global marine biodiversity under sparse data conditions. *bioRxiv*. https://doi.org/10.1101/2023.02.28.530497

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-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

Schickele, A., Debeljak, P., Ayata, S.-D., Bittner, L., Pelletier, E., Guidi, L., & Irisson, J.-O. (2024). The genomic potential of photosynthesis in piconanoplankton is functionally redundant but taxonomically structured at a global scale. *Science Advances*, *10*(33), eadl0534. https://doi.org/10.1126/sciadv.adl0534

Schickele, A., & Hofmann Elizondo, U. (2025). *alexschickele/CEPHALOPOD*, *v1.1-publication_release*. https://doi.org/10.5281/zenodo.15168238

Schickele, A., Leroy, B., Beaugrand, G., Goberville, E., Hattab, T., Francour, P., & Raybaud, V. (2020). Modelling European small pelagic fish distribution: Methodological insights. *Ecological Modelling*, *416*, 108902.

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD—A platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, *13*(4), 397–405. https://doi.org/10.1111/j.1472-4642.2007.00346.x

Vogt, M., Sarmento, H., Benedetti, F., Huber, P., Arboleda-Baena, C., Bader, R. R., Eriksson, D., Knecht, N., Chénier, N., Jaillon, O., Frémont, P., Lombard, F., Guidi, L., Ricour, F., Sebille, E. V., Schmiz, S., Manral, D., Clerc, C., Santos, G., & Frölicher, T. (2023). *AtlantECO deliverable 2.1: AtlantECO-BASE1*. https://zenodo.org/records/7944433

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W. L., Edgar, G. J., Mouillot, D., Tjiputra, J., & Pellissier, L. (2022). A quantitative review of abundance-based species distribution models. *Ecography*, *2022*(1), e05694. https://doi.org/10.1111/ecog.05694

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

WoRMS Editorial Board. (2024). *World register of marine species* [dataset]. https://doi.org/10.14284/170

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., & Grimm, V. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, *119*(4), 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Pearson correlation coefficient and associated hierarchical clustering between environmental predictors (i.e., features) at the global scale.

Figure S2. Analysis of variance of the factors influencing the true predictive performance of the habitat-modelling framework for binary (A), continuous (B) and proportions (C) data.

Figure S3. Habitat-modelling performance on 20 virtual species for presence only (A, D), continuous (B, E, G) and proportion data (C, F, H).

BRITISH ECOLOGICAL SOCIETY Methods in Ecology and Evolution

Figure S4. Analysis of variance of the factors influencing the false positive rate in performance estimation by the habitat-modelling framework for binary (A), continuous (B) and proportions (C) data.

Figure S5. Habitat-modelling performance on 20 virtual species for presence only (A, D), continuous (B, E, G) and proportion data (F, H).

Figure S6. Combinations of quality flags and recommendations provided in the output of CEPHALOPOD as a traffic light system. Each circle represents one quality flag, that is coloured if successful.

Table S1. Collection of observation-based, monthly-resolved and global-scale environmental features.

How to cite this article: Schickele, A., Clerc, C., Benedetti, F., De Angelis, D., Hofmann Elizondo, U., Münnich, M., Irisson, J.-O., & Vogt, M. (2025). CEPHALOPOD, a package to standardize marine habitat-modelling practices and enhance inter-comparability across biological observations. *Methods in Ecology and Evolution*, *16*, 1126–1135. https://doi.org/10.1111/2041-210X.70040