Journal of Biogeography

WILEY

Journal of Biogeography

METHOD OPEN ACCESS

Species Distribution Models for Mesopelagic Mesozooplankton Community

Yulia Egorova[1] | Gabriel Reygondeau[2] | William W. L. Cheung[2] | Evgeny A. Pakhomov[1,2]

[1]Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, British Columbia, Canada | [2]Institute for the Oceans and Fisheries, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence: Yulia Egorova (yulia.egorova@ubc.ca)

Received: 20 September 2023 | **Revised:** 10 August 2024 | **Accepted:** 1 September 2024

Keywords: deep ocean | habitat suitability | mesopelagic zone | mesozooplankton | species distribution models | twilight zone

ABSTRACT

Aim: We aimed to enhance our understanding of the distribution of mesopelagic mesozooplankton (MM) using species distribution models, assess the performance of various modelling techniques, identify key environmental predictors for MM distribution and compute their habitat suitability indices.

Location: Our study focused on the mesopelagic zone globally, with data analysed from different oceans.

Taxon: Our focus was primarily on mesopelagic mesozooplankton, gathering data on 861 different species from the Mesopelagic Mesozooplankton and Micronekton (MMM) Database.

Methods: We used an ensemble of species distribution models, applying 10 different modelling algorithms and three multi-model ensemble approaches. We explored two important factors that can affect model performance: subsampling and the choice of background points. We also estimated the relative importance of various environmental conditions such as mixed layer depth, temperature, salinity, net primary productivity, euphotic zone depth and dissolved nitrate concentration on the distribution of these species.

Results: Euphotic zone depth, salinity and dissolved nitrate concentration were identified as the most important variables for explaining the distribution of mesopelagic mesozooplankton. The ensemble modelling results were robust in areas with abundant observational records, but high uncertainty was observed in data-limited regions. We found a patchy habitat suitability map for zooplankton when modelled within their native range, largely due to uneven sampling. Unrestricted range models yielded smoother patterns but could inaccurately project species in areas where they do not occur.

Main Conclusions: Our study highlights the need for increased sampling effort in data-limited regions to improve the accuracy of mesopelagic species distribution models. Despite some inaccuracies, unrestricted range models, assuming ecological equivalence (where different species occupying a similar ecological niche in different geographical regions or different ecosystems exhibit similar adaptations and behaviours), provide a reasonable comparison for habitat suitability maps and model performance. It also confirms the significant impact of certain environmental conditions on mesozooplankton distribution.

1 | Introduction

Covering over two-thirds of the Earth's surface, the global ocean is a vast and complex ecosystem, with a multitude of physical, chemical and biological processes that influence its functioning and sustainability (Fleming et al. 2019). The mesopelagic zone encompasses 20% of the global ocean volume and extends on average 200–1000 m below the ocean's surface. Recently, it has been shown that depth range varies between ocean basins and latitudes (Reygondeau et al. 2018). This area is also known as

a twilight zone, layered between a sunlit epipelagic zone and a bathypelagic zone, which receives no sunlight. However, light reaching the mesopelagic region is insufficient for photosynthesis to occur (Boyd et al. 2019).

The mesopelagic food web is complex, connecting deep and surface oceans. Mesozooplankton (organisms in the size range of 0.2–20 mm) and micronekton (2–20 cm) communities contribute to CO_2 removal from the atmosphere (Guidi et al. 2016) through a process named the 'biological pump'. This pump transports organic carbon from the surface to the depths via sinking matter (passive transport) and vertical migration of zooplankton and micronekton (active transport) (Boyd et al. 2019; Koppelmann, Frost, and Mertens 2008; Tseitlin 1986; Vinogradov 1968). Although this carbon flux is widely acknowledged, the extent of the process is still poorly understood due to insufficient sampling efforts in the mesopelagic realm (Kwong and Pakhomov 2017; Robinson et al. 2010; Steinberg and Landry 2017).

Quantifying zooplankton vertical distributions, together with its biogeography, is fundamental in understanding pelagic food web structure and functioning (McIvor 2011; Robison 2009) and predicting vertical fluxes of organic matter to the deep ocean (Stemmann et al. 2008). Despite plankton/micronekton ecological significance and considerable biomass, remarkably little is known about global mesopelagic diversity, composition and distribution (Béhagle et al. 2015; St. John et al. 2016). Deep ocean sampling is logistically challenging, labour and cost-intensive (Berger 2009), with Ocean Biogeographic Information System records decreasing tenfold at 200 m and into the mesopelagic zone (Webb, Vanden Berghe, and O'Dor 2010) resulting in inconsistent global data and a lack of understanding the macroscale patterns (Proud, Cox, and Brierley 2017). The fundamental spatial correlation between community composition and its geographical location is required to fill in the missing gaps in knowledge of horizontal and vertical mesopelagic mesozooplankton (MM) distributions.

Species distribution models (SDMs) have been widely used to analyse species occurrence records and environmental data to project their distributions across space and time (Elith and Leathwick 2009). These models are used to predict how species distributions are influenced by the environment to employ them for ecological forecasting (Elith and Leathwick 2009). These approaches have been recently adapted for marine epipelagic species (Benedetti et al. 2021; Reygondeau and Beaugrand 2011; Righetti et al. 2019; Rombouts et al. 2009). However, limited efforts were made to investigate the global diversity of mesozooplankton in the entire mesopelagic ocean because the majority studies focused only on the upper mesopelagic zone (Buitenhuis et al. 2013; Drago et al. 2022).

This study represents a foundational step towards modelling the diversity and biomass of MM by developing and applying SDMs to determine habitat suitability indices. Through the construction and calibration of these models using the Mesopelagic Mesozooplankton and Micronekton (MMM) Database, we aim to establish a baseline understanding of the environmental conditions that influence MM distribution. This effort is pivotal in identifying the most relevant environmental predictors and calculating habitat suitability indices, thereby laying the groundwork for future research to accurately model the diversity and biomass of organisms within the mesopelagic zone.

2 | Materials and Methods

2.1 | Biotic Data

The MMM Database was compiled from a wide range of both published and unpublished sources to offer valuable insights into the distribution, abundance and quantitative biomass measurements of mesopelagic species (Egorova *et al.* in rev). We used records from the MMM Database to obtain a list of mesozooplankton and micronekton taxa occurring in the mesopelagic zone (list of species names and their Aphia ID can be found in Supporting Information). All statistical analyses, data manipulations and visualisations were performed using the R Statistical Software (R Core Team 2022).

For the analysis, we obtained a total of 1591 species from the MMM Database and reconciled the species names with their accepted taxonomic resolution using the World Register of Marine Species (WoRMS; WoRMS Editorial Board 2022). This reconciliation step was necessary as some species names in the MMM Database may not be consistent with WoRMS. We used the **worms** package to match the species names to their accepted WoRMS name (Holstein 2018). After the reconciliation process, we identified 1499 unique accepted species names (with AphiaID) and four species from the MMM Database that did not have a corresponding record in WoRMS: *Conchoecia belgicae*, *Conchoecia hettacra*, *Conchoecia acuticosta* and *Bylgides pelagica*. These four species were excluded from further analysis.

To increase the number of available observations and to account for the dynamic nature of the taxonomic classification, a list of synonyms was compiled for each species through the WoRMS database. However, to avoid having the same synonym for different species, only unique synonyms were kept. Additionally, various traits of species habitat were collected from FishBase (www.fishbase.org; Froese and Pauly 2022) and SeaLifeBase (www.sealifebase.org; Palomares and Pauly 2022) using the **rfishbase** (Boettiger, Lang, and Wainwright 2012) package. Since the available MMM Database did not contain a comprehensive representation of mesopelagic fishes (only 221 species were found in the database), we focused only on the mesozooplankton community (1278 species).

For most mesozooplankton species, SeaLifeBase did not contain reliable information on depth range (53% of zooplankton species lack information on their depth range) and their association to epipelagic or mesopelagic habitat (34% missing habitat classifications). More details on missing data can be found in Figure S1. In addition, many planktonic species caught in the mesopelagic zone were classified as benthic. For some organisms, it can be explained by different habitats occupied by different life stages (i.e., juveniles are planktonic while their adult form transforms into a benthic organism). Since SeaLifeBase predominantly provided information on adult species, we applied an empiric classification to determine the habitat of the zooplankton species. First, we updated the species' minimum and maximum depth range based on the depth

range found in the MMM Database. Second, we classified each species based on whether they could be found in the mesopelagic zone. In many studies, the mesopelagic zone is defined as an ocean area between 200 and 1000 m depths. However, Reygondeau et al. (2018) revealed that the vertical division of the zone is not constant over the global ocean but varies between ocean basins and latitude. Thus, we used the dynamic depth range of mesopelagic provinces from Reygondeau et al. (2018) for each location to update the information on the species' habitat. Species were classified into four groups: EPI, MESO, BOTH and NEITHER. Species were classified as epipelagic ('EPI'; $n = 96$) if they were found in the epipelagic depth range only. Species were classified as mesopelagic ('MESO'; $n = 314$) if species' occurrences were found only in the mesopelagic zone using variable mesopelagic depth from Reygondeau et al. (2018) and species were classified as both ('BOTH'; $n = 681$) if they occurred in both epi- and mesopelagic zones. Species whose occurrences were only found below the mesopelagic zone were classified as 'NEITHER' ($n = 187$). We focused on species that were part of the mesopelagic community, so species that were classified as NEITHER and EPI were excluded from the analysis.

2.2 | Occurrence Records

Species distribution records were collected from the Ocean Biogeographic Information System (https://iobis.org, accessed June 2022) using the **robis** (Provoost and Bosch 2021) package. Occurrence records for each accepted species name were combined with occurrence records from their synonym names. Additional occurrence data were added from the MMM Database. Occurrences records with a coordinate of '0°, 0°' generally indicated missing information. Also, records found on land were considered as error. These records were excluded from the analysis. To eliminate duplicated records from different data sources and limit the number of records from repeated sampling events in the same area, occurrence data were gridded on Map of Life (https://mol.org/) equal-area grid with a per cell area of 3091 km^2. Based on the occurrence data, the native range of each species was identified using the mesopelagic province developed by Sutton et al. (2017). If a province contained more than 1% of all the records, they were classified as part of the species native range.

2.3 | Environmental Data

Nine environmental variables that are available globally were used to model species bioclimatic envelopes: water temperature (Temp), salinity (Sal), Mixed Layer Depth (MLD), dissolved oxygen concentration (O_2), nitrate concentration (NO_3), phosphate concentration (PO_4), silicate concentration (SiO_2), net primary production (NPP) and euphotic zone depth (Zeu). Apart from NPP and Zeu, data were obtained from the World Ocean Atlas 2018. Zeu estimates were taken from the NASA MODIS database, and NPP was taken from the GDFL ES2M model (Table 1). All variables were interpolated to an equal-area grid for further analysis. Three sets of environmental conditions were calculated: EPI (mean environmental condition in each cell in the epipelagic zone following) Reygondeau et al. (2018), MESO (mean environmental condition in each cell in the mesopelagic zone) and BOTH (mean environmental condition in each cell for both epi-and mesopelagic zone). Since the goal of the study was to project the MM distribution, only environmental conditions for species found in MESO and BOTH were used for further analysis ($n = 861$). The distribution of environmental parameters is shown in Figures S2 and S3. NPP was removed from Non-Parametric Probabilistic Ecological Niche (NPPEN) model due to a problem of covariance in the matrix. To minimise the influence of coastal environmental dynamics on the open ocean, we excluded the data from shallow areas (< 200 m) by using the Sutton et al. (2017) mesopelagic boundary shapefile to mask areas that are outside of the shapefile boundary. We identified collinearity among the environmental parameters (Figure S4) using Pearson's correlation with $|r| > 0.7$ as a threshold level for collinearity (Dormann et al. 2013). For both MESO and BOTH sets of environmental variables, NO_3, SiO_2 and PO_4 were highly correlated ($r = 0.78–0.97$). PO_4 was removed due to its high Variance Inflation Factor. SiO_2 was also removed because it was highly correlated with both PO_4 and NO_3 and it did not provide meaningful information for modelling zooplankton distribution.

2.4 | Species Distribution Models

SDMs were implemented with the R **biomod2** package development version 4.1 (Thuiller et al. 2022, 2009). We fitted and compared eight of the statistical algorithms: generalised linear models (GLM; McCullagh 2019), generalised additive models (GAM; Hastie 1992), random forests (RF; Breiman 2001), artificial neural networks (ANN; Venables and Ripley 1999), flexible discriminant analysis (FDA; Marmion et al. 2009), classification tree analysis (CTA; Breiman 1984), surface range envelope (SRE or also known as BIOCLIM; Busby 1991) and maximum entropy modelling (MAXENT; Phillips, Anderson, and Schapire 2006). In addition, we added two presence-only models: Non-Parametric Probabilistic Ecological Niche (NPPEN; Beaugrand et al. 2011) and AquaMaps (www.aquamaps.org; Kaschner et al. 2010), to the analysis. NPPEN model was implemented using **nppen** package (github.com/jiho/nppen), and AquaMaps was modelled with the self-written package **aquamean** (github.com/yuliaUU/aquamean) and hence hereafter will be referred to as AQUAMEAN. We only used species that had more than nine records for the SDM process. Selection criteria were determined based on default settings recommended by Thuiller et al. (2009), Thuiller et al. (2022), Jones and Cheung (2015). Tunning parameters for models were chosen in accordance with Hattab et al. (2014) and can be found in Table S1. The overall modelling workflow for each species is depicted in Figure 1.

Several models integrated into **biomod2** required both presence and absence data (except MAXENT, which uses background points or SRE which is a presence-only model). Since absence of data for most marine species is unreliable due to the low catchability of sampling methods, we generated background points. The use of background points in presence–absence models can have a significant effect on the results of the model (Grimmett, Whitsed, and Horta 2020). The choice of background points and the method used to generate them can impact the accuracy and reliability of the model predictions (Jones et al. 2012). For

TABLE 1 | Available databases of the environmental parameters used in this study.

Environmental variables	Unit	Data source	Time period	Spatial resolution
Ocean temperature (Temp)	°C	World Ocean Atlas 2018 https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18.pl?parameter=t	Long-term annual mean (1981–2010)	1°
Salinity (Sal)	Unitless	World Ocean Atlas 2018 https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18.pl?parameter=s	Long-term annual mean (1981–2010)	1°
Mixed layer depth (MLD)	m	World Ocean Atlas 2018 https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18.pl?parameter=M	Long-term annual mean (1981–2010)	1°
Dissolved oxygen concentration (O_2)	$\mu mol \bullet kg^{-1}$	World Ocean Atlas 2018 https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18oxnu.pl?parameter=o	Annual mean	1°
Nitrate concentration (NO_3)	$\mu mol \bullet kg^{-1}$	World Ocean Atlas 2018 https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18oxnu.pl?parameter=n	Annual mean	1°
Silicate concentration (SiO_2)	$\mu mol \bullet kg^{-1}$	World Ocean Atlas 2018 https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18oxnu.pl?parameter=we	Annual mean	1°
Phosphate concentration (PO_4)	$\mu mol \bullet kg^{-1}$	World Ocean Atlas 2018 https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18oxnu.pl?parameter=p	Annual mean	1°
Net primary production (NPP)	$PgC \bullet m^{-3} \bullet y^{-1}$	GDFL ESM4	Long-term annual mean (1996–2015)	9 km
Euphotic depth (Z_{eu})	m	Aqua/MODIS Level-3 Mapped Euphotic Depth Data Version 2018 https://hermes.acri.fr/index.php?class=archive	Long-term annual mean (1997–2022)	4 km

example, if the background points are not appropriately placed to reflect the full range of environmental conditions that the species could potentially inhabit, the model will be biased towards the presence of the species in only the specific type of habitat where it is known to occur and will miss other potential habitats for the species. We used a random method to generate background points to avoid this issue.

To improve the models' predictive accuracy (Barbet-Massin et al. 2012), we set the prevalence (proportion of occupied locations relative to the number of pseudo-absence points) to 1:1, which generates the same number of background points as there are presence data sent to the training set. Background points were randomly generated in a predefined geographic area (native range and unrestricted range) except for the locations with presence data. For consistency, the same pseudo-absence points were selected as the background points in MAXENT (Thuiller et al. 2022), and the same pseudo-absence points were used to compute the ROC curve for NPPEN and AQUAMEAN models. We assumed that background points represented 'true' absence locations when we computed the AUC of the model predictions. Thus, predictions could be compared consistently across presence-only models using ROC curves and AUC scores.

2.5 | Evaluation Measures

Models were calibrated using the area under the curve (AUC) of the receiver operating characteristic (ROC) and the true skill statistic (TSS). TSS and AUC were used to evaluate model performance using the **pROC** package (Robin et al. 2011). TSS is defined as the sum of sensitivity (true positive rate) and

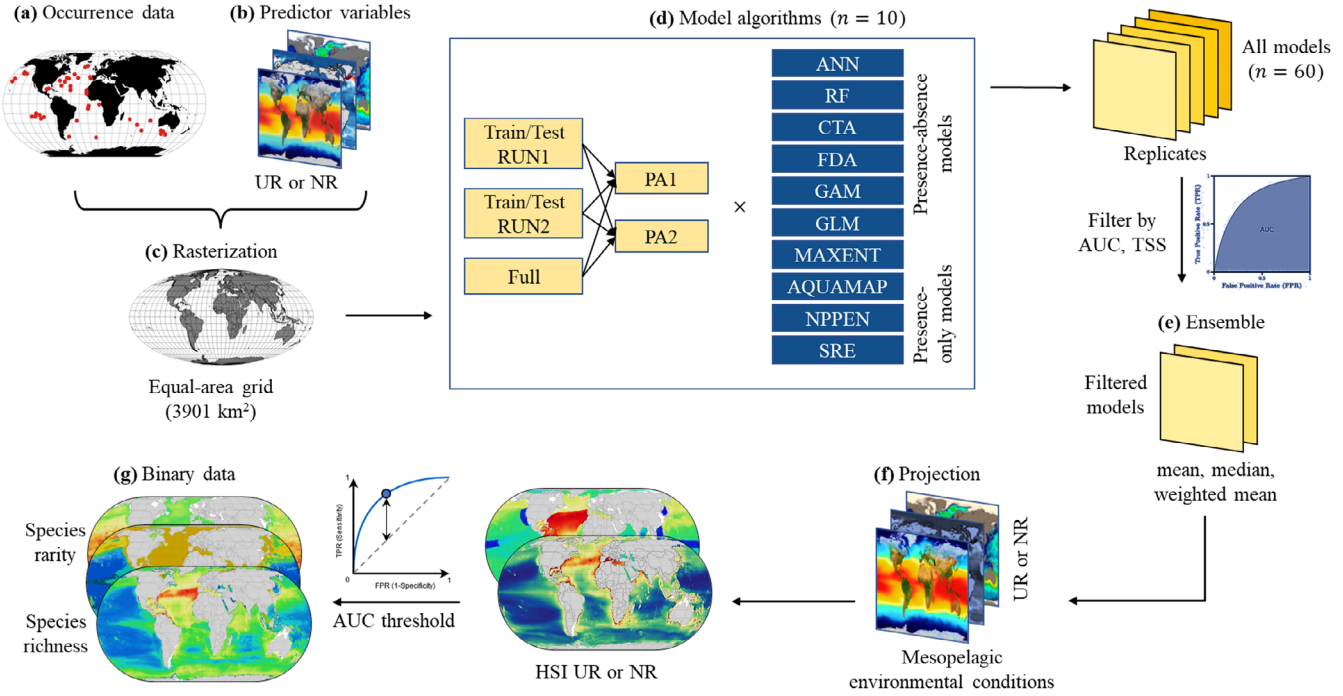

FIGURE 1 | The species distribution model workflow depicts the different analytical steps performed for each species. Depending on the sample size, the train/test split is different. AUC—area under the receiver operating curve; HSI—habitat suitability index; NR—species' native range; PA—sets of background points; TSS—true skill statistics; UR—species' unrestricted range.

specificity (true negative rate) minus one (Allouche, Tsoar, and Kadmon 2006). The AUC is a threshold-independent measure representing the relationship between sensitivity and the corresponding proportion of false positives (1-specificity). AUC ranges between 0 and 1, with values above 0.9 indicating excellent prediction and values below 0.5 indicating a projection no better than random AUC (Hajian-Tilaki 2013).

Two model projections were selected: unrestricted range and native range models. The former was run using all the species occurrences, and the model was projected into the entire mesopelagic area. The latter used only occurrences from the native range, and the projection was performed within the native range habitat.

For every species and algorithm, two replicate runs with different random subsamples (training and testing sets) were performed to evaluate algorithm accuracy. The data was split into training and validation sets based on the number of gridded occurrences: 95% training for datasets with ≤44 occurrences, 80% for datasets with 45–200 occurrences, and 70% for datasets with >200 occurrences. To explore the effect of a random choice of the pseudoabsences on algorithm projection and fit for each species, 2 pseudo-absence runs were run. For each species, 60 model runs were performed: 10 algorithms ×2 pseudo-absence runs ×3 runs (2 subsamples and 1 full run).

The difference in subsampling and choice of pseudoabsences was determined by computing the difference in AUC scores (ΔAUC) between 2 runs (same pseudo-absence) or between 2 pseudo-absence runs (same subsample), respectively. We consider an AUC of 0.1 to be an acceptable difference between the runs.

2.6 | Ensemble Modelling

We modelled the species' climatic envelope to improve the robustness of forecast abilities using an ensemble forecasting approach (Araújo and New 2007). Eight different modelling algorithms were combined into an ensemble model using three approaches: mean, weighted mean and median (note that AQUAMEAN and NPPEN algorithms were not included in the ensemble because they are designed for presence-only data). We only present here the weighted mean approach (probabilities from the selected algorithms are weighted according to their evaluation scores) because this method is significantly superior to other methods in predicting the accuracy of species distribution forecasts (Marmion et al. 2009). TSS and AUC evaluation statistics were used as a basis for weighting algorithms (Hao et al. 2019). Only robust replicate runs (individual algorithms with an AUC score above 0.6) were included in ensemble models.

The AUC score and 95% confidence interval (CI) were used to compare model performance to evaluate the best-performing model for each phylum or class. We report Variable Importance Plots (VIPs) for all the mesozooplankton groups and disaggregated per phylum. VIPs were constructed to explore environmental predictor variables' contribution to the ensemble weighted mean model output.

2.7 | Global Habitat Suitability Projections

Each model was run for every species and used to project the geographic location of the potential climatic niche. Then calibrated models were used to project species distributions using the set of climatic variables for the global mesopelagic zone

(MESO set). Here we only present the output of the ensemble weighted mean model. The model output is a raster layer consisting of equal-area grid cells, spanning the global mesopelagic zone (for unrestricted range models) or the native range mesopelagic provinces (for native range models).

The habitat suitability index (HSI) was computed for each species using weighted averages, means and median. HSI has a scale of 0 to 1000 (where 0 refers to low suitability and 1000 to a high suitability area) for all **biomod2** algorithms. Hence, the output of NPPEN and AQUAMEAN were converted to the same scale. Total HSI for the entire mesopelagic community was calculated as a sum of individual HSI for each species in each location. The total HSI score was normalised to a scale from 0 and 1 to be able to compare the unrestricted range and native range models. To compare the output of unrestricted range and native range models, we plotted the correlation of HSI between the models for each ocean basin.

3 | Results

In both unrestricted range and native range models, a total of 108,486 models were executed, covering 861 species, and utilizing 10 algorithms, 3 ensembles, 2 subsampling runs, 1 full run, and 2 pseudo-absence set runs. Further analysis and model evaluation were done only on the unrestricted range models since the native range model did not produce a realistic picture for HSI (see Figure 4b) due to potential sampling biases and lack of existing knowledge on mesopelagic zooplankton native range distribution (e.g., absence of range maps).

Out of all unrestricted range models, only 2.1% failed to run (2247 models), due to having an AUC score of less than or equal to 0.5, making it impossible to determine the cutoff point. Most of the failed algorithms were for species with a small number of occurrence points (interquartile range is 16–34 observations). The distribution of TSS and AUC scores is shown in Figure S5, and sensitivity and specificity are listed in Table S2. On average, a single species took 31 min to run, but the time varied from 6 min to over 2 h (Figure S6). The time it took algorithms to run grew logarithmically with the number of occurrences available.

3.1 | Evaluation of Models

Many individual algorithms showed good performance when applied to the test dataset (Table S2). Projections from FDA and RF were the most accurate, with AUC scores of 0.88 ± 0.16 and 0.94 ± 0.15, respectively. In contrast, predictions from AQUAMEAN and SRE had the lowest AUC scores of 0.59 ± 0.12 and 0.68 ± 0.13, respectively. Projections from CTA, GAM and GLM had very similar AUC scores, with values ranging from 0.85 to 0.86. Projections from MAXENT had an AUC score of 0.83 ± 0.16, while the presence-only NPPEN algorithm outperformed it with an AUC score of 0.84 ± 0.11. Finally, projections from ANN had an AUC score of 0.80 ± 0.18. AUC scores of the projections of the algorithm ensemble (0.94 ± 0.04) were higher than all individual algorithm projections except for RF in which their AUC scores were similar. The ranking and AUC scores were consistent when AUC scores of algorithm projections by taxonomic groups were compared (Figure 2).

Different methods of generating background points of species slightly affect the overall mean AUC scores of projections although the sensitivity varies between algorithms and species (Figure 3a,b). The mean difference in AUC between the two pseudo-absence runs was 0.07 ± 0.10. However, the effects of background points on the projections for 56 to 81 species were large with a difference in AUC score of ≥ 0.25 (Table S3). When excluding the presence-only algorithms (SRE, AQUAMEAN and NPPEN), the RF algorithm showed the smallest difference in performance, with a difference in AUC less than 0.1 for 88% of the species. FDA was the 2nd most accurate with 78% of species having delta AUC < 0.1. CTA and GAM algorithms show < 0.1 difference in AUC score in 74% of the species. MAXENT had a slightly lower performance, with 70% of the species showing an acceptable difference in pseudo-absence runs. An acceptable difference between pseudo-absence runs was observed only in 39% of species in ANN algorithms. Moreover, ANN had the largest number of species with at least one failed pseudo-absence run (29% of all species). FDA and GAM have 3–5% of species where at least one of the runs failed, and GLM only had less than 1% of failed runs. The remaining algorithms did not have any failed runs. The observed AUC difference between the pseudo-absence runs was correlated with the number of available presence records (Figure 3a). The difference between the pseudo-absence runs decreased exponentially with an increase in the available number of occurrence points (presence data).

Similar results were observed when comparing the difference between subsampling runs of the algorithms (Table S4). Interestingly, all runs in AQUAMEAN and NPPEN had acceptable AUC differences (ΔAUC < 0.1) between re-sampling runs. RF also showed consistent results for 83% of species. The difference between the resampling runs was < 0.1 only in 51% of the species in ANN algorithms. The larger difference in AUC values was also attributed to species with a low number of presences (Figure 3c). ANN, FDA, GAM and GLM were the only algorithms with failed runs (17.3%, 33.6%, 4.8% and 0.9%, respectively; Table S4).

3.2 | Environmental Variables of Importance

The distribution projections for the 861 species were primarily determined by salinity, dissolved nitrate, euphotic depth, and dissolved oxygen while NPP and mixed layer depth had the least effects on the projections (Figure S7). The relative importance of each environmental variable in determining species' distribution differed between phyla. For instance, for Annelida (10 species) and Ctenophora (5 species), dissolved nitrate, euphotic depth, dissolved oxygen and temperature were the most important variables while dissolved nitrate was the most important variable for Mollusca (33 species) and Chordata (15 species). Cnidaria (123 species) salinity and euphotic depth were the most important variables while it was temperature for Chaetognatha (31 species). NPP was consistently the least important predictor for all phyla.

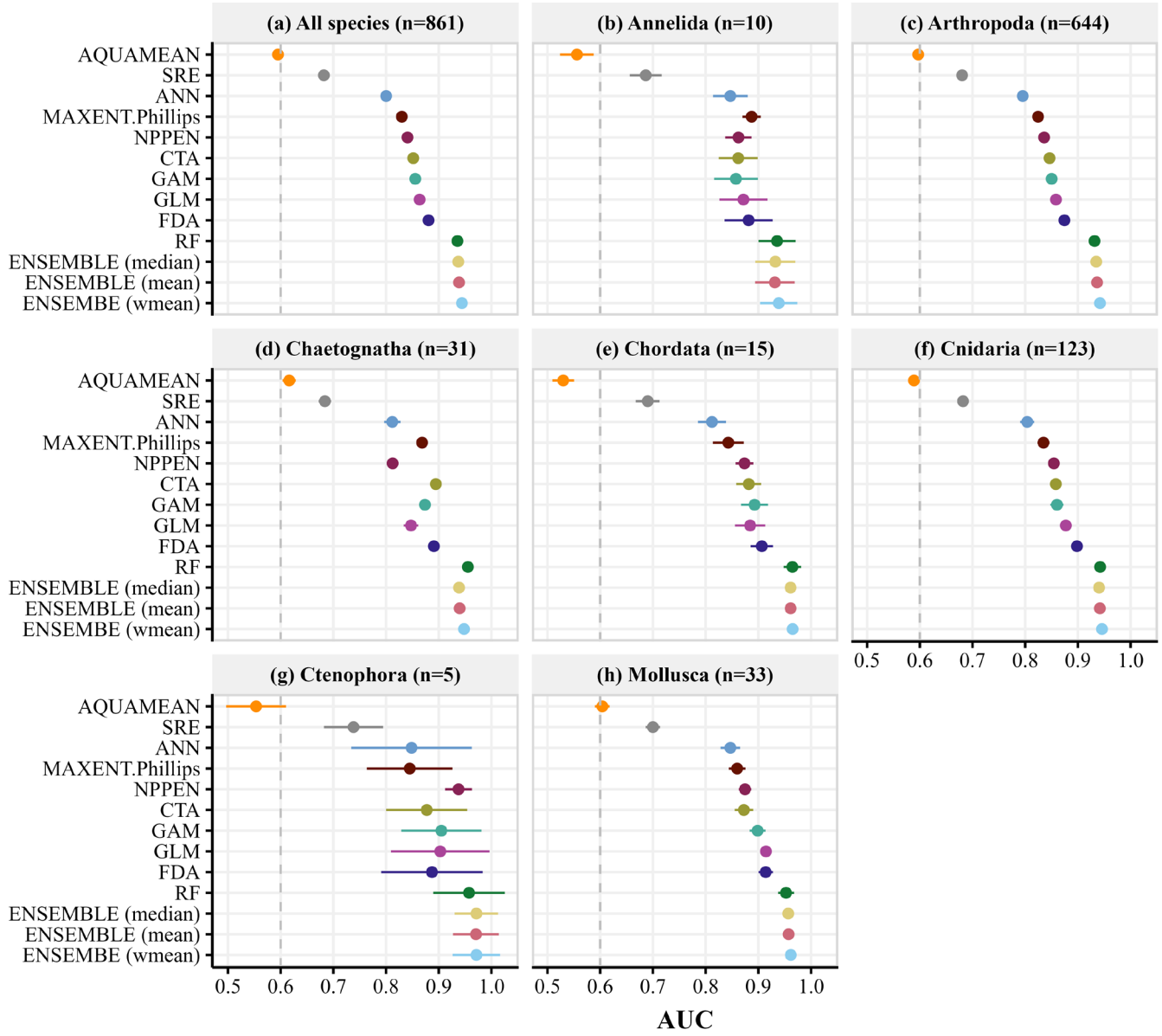

FIGURE 2 | Distributions of the area under the curve (AUC) scores of projections from individual species distribution algorithms and their ensemble. Means and 95% confidence intervals are presented for (a) all species combined ($n = 861$) and by phylum: (b) annelida ($n = 10$), (c) arthropoda ($n = 644$), (d) chaetognatha ($n = 31$), (e) chordata ($n = 15$), (f) cnidaria ($n = 123$), (g) ctenophora ($n = 5$), (h) mollusca ($n = 33$). The dotted vertical line represents the threshold (AUC score = 0.6) above which projections from the algorithms were included in the ensemble model.

3.3 | Comparing HSI in Unrestricted and Native Range Models

We utilised the habitat suitability index (HSI) values from the unrestricted range model to investigate species diversity patterns. The native range models do not perform well because the sampling within some mesopelagic provinces, which have been classified as native ranges for certain species, is insufficient (Figure 4b, Figure S8c).

For the unrestricted range model, areas with high HSI include the North Atlantic Ocean, coastal cells near the Mediterranean and Arabian Seas, and the Caribbean region (Figure 4a). Low HSI values were found in ocean basins, particularly in the Pacific Ocean. The Arctic Ocean generally has a medium HSI value, while the pattern of HSI in the Southern Ocean is more complex with lower values at the Circumglobal Subtropical Front and the sub-Antarctic waters.

For the native range model, only the North Atlantic basin was identified as an area with high HSI (Figure 4b). Such observation can be partially explained by high sampling bias with most of the observations found in the North Atlantic. Very low HSI values were observed in California Current, Equatorial Pacific, Guinea Basin and East Equatorial Atlantic, Peru Upwelling/Humboldt Current, Benguela Upwelling, Sea of Japan, South China Sea and Coral Sea mesopelagic provinces.

By comparing unrestricted and native range models, we can gain insights into the sufficiency of sampling. If the unrestricted

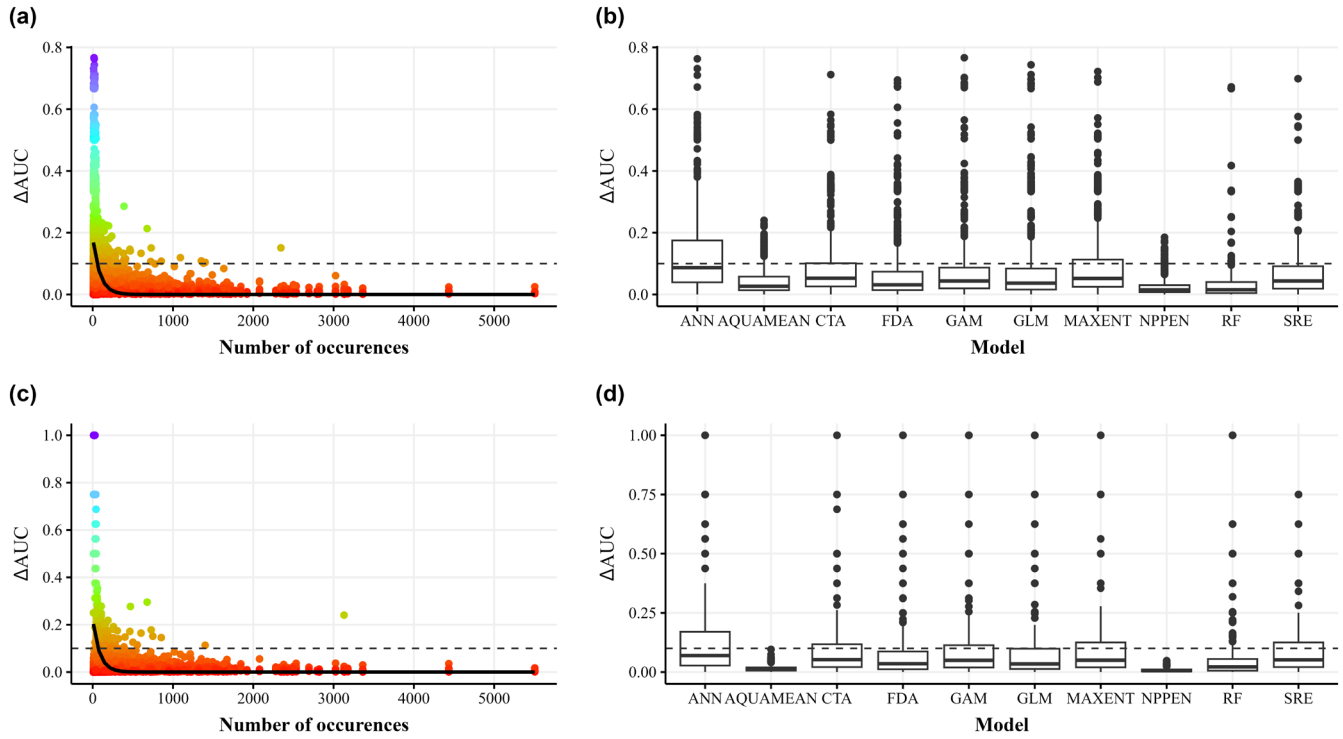

FIGURE 3 | Left column showing the difference in area under the curve scores (ΔAUC) as a function of number of occurrence points between two pseudoabsences runs (a) and between two subsampling runs (c). Right column showing the difference in ΔAUC disaggregated by the modelling algorithm for two pseudoabsences runs (b) and between two subsampling runs (d). Box plots showing the estimated median (horizontal black lines), 25th–75th percentiles (boxes) and the 1.5 interquartile range (whiskers) of AUC, points are outliers. Solid lines in panels (a) and (c) were fitted with exponential functions: (a) $\Delta\,\mathrm{AUC} = 0.19e^{-0.01N}$, $R^2 = 0.24$; (c) $\Delta\,\mathrm{AUC} = 0.23e^{-0.01N}$, $R^2 = 0.19$.

range model consistently outperforms the native range model in projecting species distributions, it suggests that the sampling may not be sufficient. This is because the unrestricted range model considers occurrences outside the native range, indicating that the species may have a broader distribution than initially believed. In contrast, if the native range model performs equally well or even better than the unrestricted range model, it implies that the sampling is likely sufficient, and the species' distribution is adequately captured within its native range. When comparing the HSI index between unrestricted range and native range models, it becomes apparent that only certain mesopelagic provinces have sufficient sampling (Figure S8). Thus, Central North Atlantic has more than sufficient sampling effort, like the Northern Central Pacific, and Eastern Tropical Pacific and part of the Antarctic/Southern Ocean mesopelagic zones. The rest of the mesopelagic provinces showed severe under-sampling in the mesopelagic realm.

4 | Discussion

4.1 | Model Evaluation

The ensemble approach is considered the optimal way to model mesozooplankton distribution patterns (Figure 2). However, the process is very time-consuming (Figure S6). To make the process faster, RF algorithms can be used as they have shown almost similar predictive performance (Figure 2) to the ensemble models, but computationally are substantially faster. In addition, RF algorithms show quite consistent results with different choices of background points or subsampling ratios (Figure 3b,d). However, there are several advantages to using the ensemble technique: ensemble methods offer a robust, flexible, and often more accurate alternative to relying solely on the best-performing single algorithm. They achieve this by combining the strengths of multiple learners, which can lead to improved performance (Geurts, Ernst, and Wehenkel 2006), better generalisation to new data, and enhanced robustness against overfitting (Polikar 2006) and noise (Sluban and Lavrač 2015).

We explored two important factors that can affect the model performance: subsampling and the choice of background points (Figure 3). Subsampling involves selecting a representative subset of data from a larger dataset for model training, which can significantly affect model performance. The design and strategy of subsampling are crucial for achieving reliable and robust model performance (Alhussein et al. 2019; Descombes et al. 2022; Phillips et al. 2009). Overfitting occurs when a model learns the noise in the training data to the extent that it negatively impacts the model's performance on new data. In the context of SDMs, using an excessive number of background points or poorly chosen background points might lead the model to overfit by learning patterns that do not genuinely represent the species' absence but rather the specific characteristics of the background points. Conversely, using too few background points or background points that are not representative may lead to underfitting, where the model fails to capture the underlying distribution patterns of the species. Phillips et al. (2009) highlight that the method of

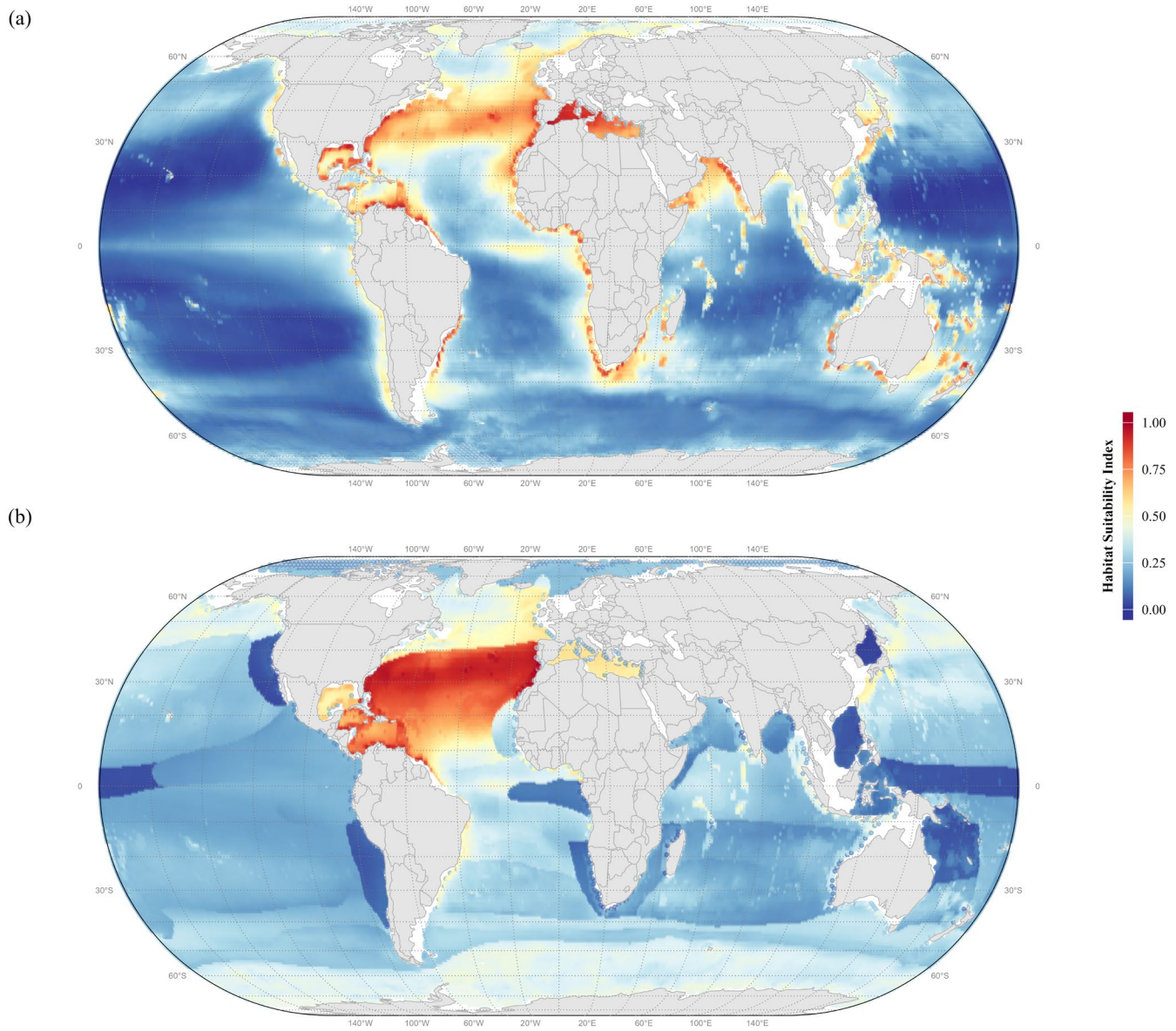

FIGURE 4 | Sum of habitat suitability indices for the mesopelagic mesozooplankton community ($n = 861$) modelled by the ensemble (weighted mean) approach using (a) unrestricted range and (b) native range. The sum of the HSI is normalised to a maximum of one. All maps in Eckert IV global equal area projection.

selecting background points can influence the model's ability to accurately project species distributions, indirectly suggesting that inappropriate selection (which could be related to the number or distribution of background points) might contribute to overfitting or underfitting. Chefaoui and Lobo (2008) emphasise that the strategy for selecting background points significantly affects model outcomes, including how well the model generalises beyond the training data. This again points to the importance of carefully considering the number and selection strategy of background points to avoid overfitting (Chefaoui and Lobo 2008).

Furthermore, characteristics of the species can typically guide the random sampling process and fine-tune the selection. In this study, we employed various data subsets to train the model and found that the difference in AUC scores was generally less than 0.2 across most of the algorithms (Figure 3d). However, when the number of occurrences was limited (Figure 3c), all algorithms displayed AUC value differences exceeding 0.2. Notably, we utilised different splitting ratios based on occurrence numbers, thus the mean AUC differences can vary for each split. Very similar results were observed when different locations of background points were chosen. For all models, most consistent results are obtained for species with higher occurrence records (> 500 observations; Figure 3a). In this case, the change in AUC is usually less than 0.1 for different background points and subsampling runs (Tables S3 and S4).

We would like to highlight the significance of carefully selecting the appropriate number of background points for SDM modelling. Although Barbet-Massin et al. (2012) recommended using 10,000 pseudo-absence for modelling with GLM and GAM algorithms, we find such recommendations not applicable to MM species. We did not find this approach feasible as the amount of

presence data for many plankton species is generally < 500, and in the absence of the expert range maps generally limited confidence in the location of background points such as in Barbet-Massin et al. (2012). Selecting too many pseudo-absence points in random locations appeared to introduce additional errors in estimating species occurrences, by placing, for example, species absence in a location optimal for its occurrence. Therefore, we used the same number of background points as presences, following a recommended ratio for CTA, BRT, RF algorithms suggested by Barbet-Massin et al. (2012), and applied them to all algorithms.

4.2 | Environmental Variables of Importance

According to the study results, the distribution of 861 species of marine organisms was primarily driven by salinity, dissolved nitrate, euphotic depth and dissolved oxygen, while NPP and mixed layer depth had the least impact (Figure S7). This pattern was particularly evident in the phylum Arthropoda, which made up 75% of the total species studied.

Salinity is important for mesopelagic zooplankton as it influences their distribution, physiology and ecological interactions. These marine organisms must maintain a stable internal environment to survive, regulating their internal ion concentrations and water content in response to changes in salinity (osmoregulation) (Aguilar, Diaz, and Buckle 1998; McNamara, Moreira, and Moreira 1983). Variations in salinity create physical or chemical barriers that affect zooplankton movement and distribution, with different species having specific salinity preferences (Ojeda et al. 2022). Indirectly, salinity influences mesopelagic zooplankton by affecting the distribution and productivity of phytoplankton, their primary prey, through changes in nutrient availability and water column stratification. Lastly, changes in salinity can alter the distribution and abundance of predators and competitors, affecting zooplankton survival and community dynamics (Yuan et al. 2020).

Changes in nitrate concentrations can indirectly impact zooplankton distribution, abundance and overall community structure. Upwelling delivers nutrient-rich waters to the surface, promoting phytoplankton blooms that enhance ecosystem productivity. However, these waters are usually low in dissolved oxygen, and the decay of phytoplankton by bacterial respiration of sinking detritus can further deplete oxygen concentrations at depth (Bakun et al. 2010), which can negatively affect mesopelagic organisms. In these zones, the concentration of dissolved oxygen is low, which can limit the survival and growth of many species, particularly those that require high levels of oxygen (Bakun 2022; Wishner et al. 2018).

Dissolved oxygen is essential for the respiration and metabolism of zooplankton, and variations in dissolved oxygen concentrations can impact their survival, growth and reproduction. High oxygen concentrations generally support a diverse and abundant zooplankton community, while low oxygen concentrations can create hypoxic or anoxic conditions, limiting the distribution of many zooplankton species and leading to reduced biodiversity (Soviadan et al. 2022). Some species have adapted to tolerate low-oxygen environments (Vaquer-Sunyer and Duarte 2008), but many are unable to survive under such conditions. Additionally, dissolved oxygen levels can influence the vertical distribution of zooplankton, as they often occupy specific depth ranges with suitable oxygen concentrations to meet their metabolic needs (Wishner et al. 2018). Furthermore, variations in oxygen concentrations can affect the distribution and behaviour of zooplankton predators and prey, ultimately shaping the structure and dynamics of mesopelagic food webs (Bakun 2022).

Euphotic depth, the depth at which sufficient light is available for photosynthesis, can indirectly influence zooplankton distribution by enhancing the search efficiency of visual predators in clearer waters, such as those found in the subtropical gyres. This improved visibility allows predators to locate and capture prey more effectively (Irigoien et al. 2014). Water temperature, although only found important for a couple of taxa (Figure S7) can affect the metabolic rates of zooplankton, with higher temperatures generally leading to increased metabolic activity, growth and reproduction rates (Rohde, 1992). As a result, it plays a significant role in determining the distribution, richness and abundance of different zooplankton species (Dipper 2022; Fenton et al. 2016), with each species occupying specific thermal niches in the water column. However, in the mesopelagic, the temperature gradient is smaller, and as a result temperature was not found to be an important factor in this study.

Across all phyla, NPP was not found to be the most important predictor, which is consistent with findings by Gagné et al. (2020). NPP was an important variable in many studies restricted generally to the epipelagic layer (Hernández-León et al. 2020; Rombouts et al. 2009; Vallina et al. 2014).

However, the importance of each environmental variable changed when the species were disaggregated by phyla. For instance, in Annelida and Ctenophora, salinity was one of the least important factors, along with mixed layer depth and NPP, while dissolved nitrate, euphotic depth, dissolved oxygen and temperature appeared as the most influential variables. Nevertheless, dissolved nitrate was the important predictor for Mollusca and Chordata, while for Cnidaria and Chaetognatha salinity/euphotic depth and temperature were prime factors, respectively. These results highlight the importance of considering the specific ecological requirements of different taxa when projecting their distributions. However, we presented an average importance of each environmental variable aggregated by phylum. It is important to note that results may differ when considering species-level distributions (Bosch et al. 2018), as different species within the same phylum may indeed have different ecological requirements due to specific evolutionary adaptations. Furthermore, while the results suggest that some environmental variables, that is, salinity, dissolved nitrate, euphotic zone depth and dissolved oxygen, have a major impact on the distribution of marine species, it is plausible that other factors, such as biotic interactions, dispersal ability and historical events, may also play a role in determining species distributions. Therefore, it is crucial to consider the results of this study as the first attempt to model species distributions and acknowledge the potential influence of additional factors on projections of the individual species distributions.

4.3 | Comparing HSI in Unrestricted and Native Range Models

In various SDM models, it is a widespread practice to use species range maps to help establish their geographical occurrence boundaries and avoid projecting their appearances in areas where species cannot be found. It also helps with the selection of background points, as they can be taken outside the species range. However, range maps are not widely available for zooplankton groups, except for a few species of plankton for whom the printed range maps are published (Siccha and Kucera 2017; van der Spoel and Heyman 1983) but hardly numerically adapted. In this work, we have demonstrated that if we only model zooplankton into the native range area (which is determined by available species occurrence points), the produced HSI map is very patchy (Figure 4b), with apparent areas of the mesopelagic realm where almost no zooplankton species are projected (since few to no sampling available in the area). Hence, the patchiness we see in the native range HSI can be explained by an uneven sampling effort (Figure S9). The Pacific Ocean, South Atlantic and basin of Indian Oceans are severely undersampled, while the polar regions are characterised by comprehensive coverage and moderate number of observed species (Figure S9). As a result, many provinces lack consistent sampling or only have a few species sampled within them, and the environmental envelope that represents those areas is not adequately represented in SDM.

However, if mesozooplankton species can occur in any part of the ocean as long as environmental conditions are optimal, the HSI map produces more consistent, smooth patterns of their distribution (Figure 4a). The discrepancy between the two approaches comes from the severe undersampling in many mesopelagic provinces (Figure S8). For example, mesopelagic provinces in Arctic, South Atlantic and Pacific Ocean lack consistent sampling, especially their central ocean basins. One of the disadvantages of unrestricted range approach is that species can be projected in the areas where they do not occur. It was observed that 637 of the modelled species have occurrence records in both hemispheres at latitudes greater than 40°, thus the majority of species (74%) are already projected in both hemispheres. However, 26% of modelled species are projected into the hemisphere where they were not observed. For instance, species that can only be found in the Pacific Ocean, can be projected in Atlantic Ocean where they may or may not been recorded, or species that only found in the Northern hemisphere can be projected into the Southern Hemisphere. Thus, projections of unrestricted range models are less accurate than native range models since the majority of species are not found across the global ocean. However, it is plausible to assume that similar, ecologically equal, species (e.g., *Calanus finmarchicus* and *Calanus pacificus*) that have comparable niches could potentially be found in every basin with similar environmental settings. Therefore, the use of unrestricted range models to compare habitat suitability maps and explore model performance can be reasonably justified.

We acknowledge that the produced habitat suitability map is a simplified version of reality assuming that every species can be everywhere all at once. However, we rest on the plankton paradox (Hutchinson 1961) to provide a close to reality picture of what the mesozooplankton diversity gradient should look like when relying on knowledge of the relationship between available occurrence and mesopelagic environment.

4.4 | Limitations

This work provides the first attempt at modelling the species distribution of the mesopelagic mesozooplankton. More work is indeed required to determine optimal tuning parameters for each species and broader taxonomic groups. Additional research should be done to explore the effect of subsampling ratio for species with low occurrences and a choice of the pseudo-absence numbers as well as effects of different evaluation metrics (AUC vs. TSS). The SDMs are machine learning algorithms that can project certain species into the area where they are not occurring even though the area has an optimal environmental niche. To avoid it, expert knowledge has to be applied in the post-processing step to create zooplankton species range maps (similar to the ones available for many fish species).

Another big limitation of this study is the lack of biological information on the habitat and depth range of many mesopelagic species. We acknowledge that our own classification of the organism into meso- or epipelagic groups is quite coarse and is limited by the amount of available data in the MMM Database. For instance, if sampling were only done in the upper layer of the water column, the species would be classified as epipelagic even though it may be present at mesopelagic depths too. In addition, the boundary of the mesopelagic zone is dynamic and can change on seasonal and yearly timescales (Reygondeau et al. 2018). Moreover, the spatial extent of each mesopelagic zone can vary with time. This study used a static position of the mesopelagic provinces that can both affect the classification of the organism into meso/epipelagic species but also affect the distribution of the organisms, especially when modelling the native range distributions.

Author Contributions

Y.E., G.R. and E.A.P. designed the methods; Y.E. and G.R. performed data collection and analysis; Y.E. analyzed data and created figures and images for the paper; Y.E., G.R. and E.A.P. wrote the paper; Y.E., G.R., W.W.L.C. and E.A.P. edited the paper.

Acknowledgements

We thank the reviewers for their comments which help to improve the manuscript. We also would like to thank Sutton et al. (2017) for sharing the data on mesopelagic boundaries. No permits were needed to work on this project.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Species Distribution Models for Mesopelagic Mesozooplankton at https://doi.org/10.6084/m9.figshare.26499265.v1.

Code Availability

Any computer code used to generate the results of the study are freely available upon request to the authors and all R codes are currently

stored on the GitHub account of Y.E. (https://github.com/yuliaUU/SDM-MesoZoop).

References

Aguilar, M., F. Diaz, and L. F. Buckle. 1998. "The Effect of Salinity on Oxygen Consumption and Osmoregulation of Macrobrachium Tenellum." *Marine and Freshwater Behaviour and Physiology* 31, no. 2: 105–113. https://doi.org/10.1080/10236249809387066.

Alhussein, O., P. D. Yoo, S. Muhaidat, and J. Liang. 2019. "Semiparametric Subsampling and Data Condensation for Large-Scale Data Analytics." In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, 1–6. Edmonton, AB, Canada: IEEE. https://doi.org/10.1109/CCECE.2019.8861966.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS)." *Journal of Applied Ecology* 43, no. 6: 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

Araújo, M. B., and M. New. 2007. "Ensemble Forecasting of Species Distributions." *Trends in Ecology & Evolution* 22, no. 1: 42–47. https://doi.org/10.1016/j.tree.2006.09.010.

Bakun, A. 2022. "Adjusting Intuitions as to the Role of Oxygen Constraints in Shaping the Ecology and Dynamics of Ocean Predator–Prey Systems." *Environmental Biology of Fishes* 105, no. 10: 1287–1299. https://doi.org/10.1007/s10641-022-01317-6.

Bakun, A., D. B. Field, A. N. Redondo-rodriguez, and S. J. Weeks. 2010. "Greenhouse Gas, Upwelling-Favorable Winds, and the Future of Coastal Ocean Upwelling Ecosystems." *Global Change Biology* 16, no. 4: 1213–1228. https://doi.org/10.1111/j.1365-2486.2009.02094.x.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. "Selecting Pseudo-Absences for Species Distribution Models: How, Where and How Many?" *Methods in Ecology and Evolution* 3, no. 2: 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Beaugrand, G., S. Lenoir, F. Ibañez, and C. Manté. 2011. "A New Model to Assess the Probability of Occurrence of a Species, Based on Presence-Only Data." *Marine Ecology Progress Series* 424: 175–190. https://doi.org/10.3354/meps08939.

Béhagle, N., G. Roudaut, E. Josse, et al., eds. 2015. "Acoustics Characterization of Micronekton Spatial Distribution in Indian Ocean Using Scientific Surveys and Integrated Marine Observing System Database: Acoustics Characterization of Micronekton." In *IEEE/OES Acoustics in Underwater Geosciences Symposium (RIO Acoustics)*. Piscataway: IEEE, 5 p, RIO Acoustics 2015, Rio de Janeiro (BRA), July 29–31, 2015. https://doi.org/10.1109/RIOACOUSTICS.2015.7473616.

Benedetti, F., M. Vogt, U. H. Elizondo, et al. 2021. "Author Correction: Major Restructuring of Marine Plankton Assemblages Under Global Warming." *Nature Communications* 12: 6256. https://doi.org/10.1038/s41467-021-26564-6.

Berger, W. H. 2009. *Ocean: Reflections on a Century of Exploration*. Berkeley: University of California Press.

Boettiger, C., D. T. Lang, and P. Wainwright. 2012. "Rfishbase: Exploring, Manipulating and Visualizing FishBase Data From R." *Journal of Fish Biology* 81: 2030–2039. https://doi.org/10.1111/j.1095-8649.2012.03464.x.

Bosch, S., L. Tyberghein, K. Deneudt, F. Hernandez, and O. De Clerck. 2018. "In Search of Relevant Predictors for Marine Species Distribution Modelling Using the MarineSPEED Benchmark Dataset." *Diversity and Distributions* 24, no. 2: 144–157. https://doi.org/10.1111/ddi.12668.

Boyd, P. W., H. Claustre, M. Levy, D. A. Siegel, and T. Weber. 2019. "Multi-Faceted Particle Pumps Drive Carbon Sequestration in the Ocean." *Nature* 568, no. 7752: 327–335. https://doi.org/10.1038/s41586-019-1098-2.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. 1st ed. Chapman and Hall/CRC: Routledge. https://doi.org/10.1201/9781315139470.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45, no. 1: 5–32. https://doi.org/10.1023/A:1010933404324.

Buitenhuis, E. T., M. Vogt, R. Moriarty, et al. 2013. "MAREDAT: Towards a World Atlas of MARine Ecosystem DATa." *Earth System Science Data* 5, no. 2: 227–239. https://doi.org/10.5194/essd-5-227-2013.

Busby, J. R. 1991. "BIOCLIM—A Bioclimate Analysis and Prediction System." *Plant Protection Quarterly* 6, no. 1: 8–9.

Chefaoui, R. M., and J. M. Lobo. 2008. "Assessing the Effects of Pseudo-Absences on Predictive Distribution Model Performance." *Ecological Modelling* 210, no. 4: 478–486. https://doi.org/10.1016/J.ECOLMODEL.2007.08.010.

Descombes, P., Y. Chauvier, P. Brun, et al. 2022. "Strategies for Sampling Pseudo-Absences for Species Distribution Models in Complex Mountainous Terrain." https://doi.org/10.1101/2022.03.24.485693.

Dipper, F. 2022. "Chapter 2—The Seawater Environment and Ecological Adaptations." In *Elements of Marine Ecology*. 5th ed., 37–151. Butterworth-Heinemann: Elsevier. https://doi.org/10.1016/B978-0-08-102826-1.00002-8.

Dormann, C. F., J. Elith, S. Bacher, et al. 2013. "Collinearity: A Review of Methods to Deal With It and a Simulation Study Evaluating Their Performance." *Ecography* 36, no. 1: 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Drago, L., T. Panaïotis, J.-O. Irisson, et al. 2022. "Global Distribution of Zooplankton Biomass Estimated by In Situ Imaging and Machine Learning." *Frontiers in Marine Science* 9: 894372. https://doi.org/10.3389/fmars.2022.894372.

Elith, J., and J. R. Leathwick. 2009. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time." *Annual Review of Ecology, Evolution, and Systematics* 40, no. 1: 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Fenton, I. S., P. N. Pearson, T. Dunkley Jones, and A. Purvis. 2016. "Environmental Predictors of Diversity in Recent Planktonic Foraminifera as Recorded in Marine Sediments." *PLoS One* 11, no. 11: e0165522. https://doi.org/10.1371/journal.pone.0165522.

Fleming, L. E., B. Maycock, M. P. White, and M. H. Depledge. 2019. "Fostering Human Health Through Ocean Sustainability in the 21st Century." *People and Nature* 1, no. 3: 276–283. https://doi.org/10.1002/pan3.10038.

Froese, R., and D. Pauly. 2022. "*FishBase*." Accessed June 2022 www.fishbase.org.

Gagné, T. O., G. Reygondeau, C. N. Jenkins, et al. 2020. "Towards a Global Understanding of the Drivers of Marine and Terrestrial Biodiversity." *PLoS One* 15, no. 2: e0228065. https://doi.org/10.1371/journal.pone.0228065.

Geurts, P., D. Ernst, and L. Wehenkel. 2006. "Extremely Randomized Trees." *Machine Learning* 63, no. 1: 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Grimmett, L., R. Whitsed, and A. Horta. 2020. "Presence-Only Species Distribution Models Are Sensitive to Sample Prevalence: Evaluating Models Using Spatial Prediction Stability and Accuracy Metrics." *Ecological Modelling* 431: 109194. https://doi.org/10.1016/j.ecolmodel.2020.109194.

Guidi, L., S. Chaffron, L. Bittner, et al. 2016. "Plankton Networks Driving Carbon Export in the Oligotrophic Ocean." *Nature* 532, no. 7600: 465. https://doi.org/10.1038/nature16942.

Hajian-Tilaki, K. 2013. "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation." *Caspian Journal of Internal Medicine* 4, no. 2: 627–635.

Hao, T., J. Elith, G. Guillera-Arroita, and J. J. Lahoz-Monfort. 2019. "A Review of Evidence About Use and Performance of Species Distribution Modelling Ensembles Like BIOMOD." *Diversity and Distributions* 25, no. 5: 839–852. https://doi.org/10.1111/ddi.12892.

Hastie, T. J., ed. 1992. *Statistical Models in S*. 1st ed., 624. New York: Routledge. https://doi.org/10.1201/9780203738535.

Hattab, T., C. Albouy, F. B. R. Lasram, S. Somot, F. Le Loc'h, and F. Leprieur. 2014. "Towards a Better Understanding of Potential Impacts of Climate Change on Marine Species Distribution: A Multiscale Modelling Approach." *Global Ecology and Biogeography* 23, no. 12: 1417–1429. https://doi.org/10.1111/geb.12217.

Hernández-León, S., R. Koppelmann, E. Fraile-Nuez, et al. 2020. "Large Deep-Sea Zooplankton Biomass Mirrors Primary Production in the Global Ocean." *Nature Communications* 11, no. 1: 6048. https://doi.org/10.1038/s41467-020-19875-7.

Holstein, J. 2018. Worms: Retriving Aphia Information From World Register of Marine Species. https://CRAN.R-project.org/package=worms.

Hutchinson, G. E. 1961. "The Paradox of the Plankton." *American Naturalist* 95: 137–145.

Irigoien, X., T. A. Klevjer, A. Røstad, et al. 2014. "Large Mesopelagic Fishes Biomass and Trophic Efficiency in the Open Ocean." *Nature Communications* 5: 3271. https://doi.org/10.1038/ncomms4271.

Jones, M. C., and W. W. L. Cheung. 2015. "Multi-Model Ensemble Projections of Climate Change Effects on Global Marine Biodiversity." *ICES Journal of Marine Science* 72, no. 3: 741–752. https://doi.org/10.1093/icesjms/fsu172.

Jones, M. C., S. R. Dye, J. K. Pinnegar, R. Warren, and W. W. Cheung. 2012. "Modelling Commercial Fish Distributions: Prediction and Assessment Using Different Approaches." *Ecological Modelling* 225: 133–145. https://doi.org/10.1016/j.ecolmodel.2011.11.003.

Kaschner, K., J. Rius-Barile, K. Kesner-Reyes, et al. 2010. *AquaMaps: Predicted Range Maps for Aquatic Species*. World Wide Web Electronic Publication. Version 08/2010. www.aquamaps.org.

Koppelmann, R., J. Frost, and L. P. Mertens, eds. 2008. *Biological Oceanography Research Trends*. NY: Nova Science Publishers, Inc., 67–130.

Kwong, L. E., and E. A. Pakhomov. 2017. "Assessment of Active Vertical Carbon Transport: New Methodology. *Uchenye Zapiski Kazanskogo Universiteta*." *Seriya Estestvennye Nauki* 159, no. 3: 492–509.

Marmion, M., M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. "The Performance of State-Of-The-Art Modelling Techniques Depends on Geographical Distribution of Species." *Ecological Modelling* 220, no. 24: 3512–3520. https://doi.org/10.1016/j.ecolmodel.2008.10.019.

Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. "Evaluation of Consensus Methods in Predictive Species Distribution Modelling." *Diversity and Distributions* 15, no. 1: 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x.

McCullagh, P. 2019. *Generalized Linear Models*. 2nd ed. NY: Routledge. https://doi.org/10.1201/9780203753736.

McIvor, I. 2011. *A Comparison of the Geographical and Vertical Distribution of Mesozooplankton Communities From 0 to 1000 m at a Global Scale [MSc]*. Vancouver, Canada: University of British Columbia.

McNamara, J. C., G. S. Moreira, and P. S. Moreira. 1983. "The Effect of Salinity on Respiratory Metabolism, Survival and Moulting in the First Zoea of *Macrobrachium amazonicum* (Heller) (Crustacea, Palaemonidae)." *Hydrobiologia* 101, no. 3: 239–242. https://doi.org/10.1007/BF00009880.

Ojeda, V., B. Serra, C. Lagares, et al. 2022. "Interannual Fluctuations in Connectivity Among Crab Populations (*Liocarcinus depurator*) Along the Atlantic-Mediterranean Transition." *Scientific Reports* 12, no. 1: 9797. https://doi.org/10.1038/s41598-022-13941-4.

Palomares, M., and D. Pauly. 2022. "SeaLifeBase." www.sealifebase.org Version (04/2022).

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190, no. 3–4: 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Phillips, S. J., M. Dudík, J. Elith, et al. 2009. "Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data." *Ecological Applications: A Publication of the Ecological Society of America* 19, no. 1: 181–197. https://doi.org/10.1890/07-2153.1.

Polikar, R. 2006. "Ensemble Based Systems in Decision Making." *IEEE Circuits and Systems Magazine* 6, no. 3: 21–45. https://doi.org/10.1109/MCAS.2006.1688199.

Proud, R., M. J. Cox, and A. S. Brierley. 2017. "Biogeography of the Global Ocean's Mesopelagic Zone." *Current Biology: CB* 27, no. 1: 113–119. https://doi.org/10.1016/j.cub.2016.11.003.

Provoost, P., and Bosch, S. 2021. *Robis: Ocean Biodiversity Information System (OBIS) Client*. https://CRAN.R-project.org/package=robis

R Core Team. 2022. *R: [Computer Software]*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reygondeau, G., and G. Beaugrand. 2011. "Future Climate-Driven Shifts in Distribution of *Calanus finmarchicus*." *Global Change Biology* 17, no. 2: 756–766. https://doi.org/10.1111/j.1365-2486.2010.02310.x.

Reygondeau, G., L. Guidi, G. Beaugrand, et al. 2018. "Global Biogeochemical Provinces of the Mesopelagic Zone." *Journal of Biogeography* 45: 1–15. https://doi.org/10.1111/jbi.13149.

Righetti, D., M. Vogt, N. Gruber, A. Psomas, and N. E. Zimmermann. 2019. "Global Pattern of Phytoplankton Diversity Driven by Temperature and Environmental Variability." *Science Advances* 5, no. 5: eaau6253. https://doi.org/10.1126/sciadv.aau6253.

Robin, X., N. Turck, A. Hainard, et al. 2011. "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12: 77.

Robinson, C., D. K. Steinberg, T. R. Anderson, et al. 2010. "Mesopelagic Zone Ecology and Biogeochemistry—A Synthesis." *Deep Sea Research Part II: Topical Studies in Oceanography* 57, no. 16: 1504–1518. https://doi.org/10.1016/j.dsr2.2010.02.018.

Robison, B. H. 2009. "Conservation of Deep Pelagic Biodiversity." *Conservation Biology: The Journal of the Society for Conservation Biology* 23, no. 4: 847–858. https://doi.org/10.1111/j.1523-1739.2009.01219.x.

Rombouts, I., G. Beaugrand, F. Ibanez, S. Gasparini, S. Chiba, and L. Legendre. 2009. "Global Latitudinal Variations in Marine Copepod Diversity and Environmental Factors." *Proceedings. Biological Sciences* 276, no. 1670: 3053–3062. https://doi.org/10.1098/rspb.2009.0742.

Siccha, M., and M. Kucera. 2017. "ForCenS, A Curated Database of Planktonic Foraminifera Census Counts in Marine Surface Sediment Samples." *Scientific Data* 4, no. 1: 1–12.

Sluban, B., and N. Lavrač. 2015. "Relating Ensemble Diversity and Performance: A Study in Class Noise Detection." *Neurocomputing* 160: 120–131. https://doi.org/10.1016/j.neucom.2014.10.086.

Soviadan, Y. D., F. Benedetti, M. C. Brandão, et al. 2022. "Patterns of Mesozooplankton Community Composition and Vertical Fluxes in the Global Ocean." *Progress in Oceanography* 200: 102717. https://doi.org/10.1016/j.pocean.2021.102717.

St. John, M. A., A. Borja, G. Chust, et al. 2016. "A Dark Hole in Our Understanding of Marine Ecosystems and Their Services: Perspectives

From the Mesopelagic Community." *Frontiers in Marine Science* 3: 1–6. https://doi.org/10.3389/fmars.2016.00031.

Steinberg, D. K., and M. R. Landry. 2017. "Zooplankton and the Ocean Carbon Cycle." *Annual Review of Marine Science* 9: 413–444. https://doi.org/10.1146/annurev-marine-010814-015924.

Stemmann, L., A. Hosia, M. J. Youngbluth, H. Søiland, M. Picheral, and G. Gorsky. 2008. "Vertical Distribution (0–1000m) of Macrozooplankton, Estimated Using the Underwater Video Profiler, in Different Hydrographic Regimes Along the Northern Portion of the mid-Atlantic Ridge." *Deep Sea Research Part II: Topical Studies in Oceanography* 55, no. 1–2: 94–105. https://doi.org/10.1016/j.dsr2.2007.09.019.

Sutton, T. T., M. R. Clark, D. C. Dunn, et al. 2017. "A Global Biogeographic Classification of the Mesopelagic Zone." *Deep Sea Research Part I: Oceanographic Research Papers* 126: 85–102. https://doi.org/10.1016/j.dsr.2017.05.006.

Thuiller, W., D. Georges, M. Gueguen, E. Robin, F. Breiner, and L. Bruno. 2022. "*biomod2: Ensemble Platform for Species Distribution Modeling*."

Thuiller, W., B. Lafourcade, R. Engler, and M. B. Araújo. 2009. "BIOMOD - A Platform for Ensemble Forecasting of Species Distributions." *Ecography* 32, no. 3: 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x.

Tseitlin, V. 1986. *Energetics of Deep-Water Communities*. Moscow: Nauka.

Vallina, S. M., M. J. Follows, S. Dutkiewicz, J. M. Montoya, P. Cermeno, and M. Loreau. 2014. "Global Relationship Between Phytoplankton Diversity and Productivity in the Ocean." *Nature Communications* 5: 4299. https://doi.org/10.1038/ncomms5299.

van der Spoel, S., and R. P. Heyman. 1983. *A Comparative Atlas of Zooplankton*. Heidelberg: Springer Berlin. https://doi.org/10.1007/978-3-662-02366-2.

Vaquer-Sunyer, R., and C. M. Duarte. 2008. "Thresholds of Hypoxia for Marine Biodiversity." *Proceedings of the National Academy of Sciences of the United States of America* 105, no. 40: 15452–15457. https://doi.org/10.1073/pnas.0803833105.

Venables, W. N., and B. D. Ripley. 1999. "Modern Applied Statistics With S-PLUS." In *Statistics and Computing*. 1st ed. New York: Springer.

Vinogradov, M. E. 1968. *Vertical Distribution of the Oceanic Zooplankton*. Moscow: NAUKA.

Webb, T. J., E. Vanden Berghe, and R. O'Dor. 2010. "Biodiversity's big wet Secret: The Global Distribution of Marine Biological Records Reveals Chronic Under-Exploration of the Deep Pelagic Ocean: The Global Distribution of Marine Biological Records Reveals Chronic Under-Exploration of the Deep Pelagic Ocean." *PLoS One* 5, no. 8: e10223. https://doi.org/10.1371/journal.pone.0010223.

Wishner, K. F., B. A. Seibel, C. Roman, et al. 2018. "Ocean Deoxygenation and Zooplankton: Very Small Oxygen Differences Matter." *Science Advances* 4, no. 12: eaau5180. https://doi.org/10.1126/sciadv.aau5180.

WoRMS Editorial Board. 2022. "*World Register of Marine Species*." https://www.Marinespecies.Org at VLIZ. Accessed June 7, 2023. https://doi.org/10.14284/170.

Yuan, D., L. Chen, L. Luan, Q. Wang, and Y. Yang. 2020. "Effect of Salinity on the Zooplankton Community in the Pearl River Estuary." *Journal of Ocean University of China* 19, no. 6: 1389–1398. https://doi.org/10.1007/s11802-020-4449-6.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.