Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Region-wide distribution models of aquatic species in the Baltic Sea

Edmond Sacre

edmond.sacre@slu.se

Swedish University of Agricultural Sciences

Antti Takolander

Finnish Environment Institute

Ants Kaasik

University of Tartu

Javier Lenzi

Jade University of Applied Sciences

Ulf Bergström

Swedish University of Agricultural Sciences

Elina Virtanen

Finnish Environment Institute

Jonne Kotta

University of Tartu

Mårten Erlandsson

Swedish University of Agricultural Sciences

Ronny Fredriksson

Swedish University of Agricultural Sciences

Maren Leiz

Jade University of Applied Sciences

Oscar Törnqvist

Geological Survey of Sweden

Jannica Haldin

Baltic Marine Environment Protection Commission

Charlotte Berkström

Swedish University of Agricultural Sciences

Katriina Juva

Finnish Environment Institute

Duncan Hume

Geological Survey of Sweden

Mikko Olin

Natural Resources Institute Finland

Jānis Gruduls

Institute of Food Safety, Animal Health and Environment “BIOR”

Laura Briekmane

Institute of Food Safety, Animal Health and Environment “BIOR”

Adam Lejk

National Marine Fisheries Research Institute

Christina Henseler

Johann Heinrich von Thünen-Institut

Justas Dainys

Nature Research Centre

Linas Ložys

Nature Research Centre

Roland Pesch

Jade University of Applied Sciences

---

data-descriptor

Keywords:

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10132050/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

As biodiversity across the world continues to decline at a concerning rate, regional strategies focused on conservation and spatial planning have intensified. To support these efforts, spatially continuous, full-coverage maps on species distributions are essential. The Baltic Sea is surrounded by nine countries, making effective and well-coordinated transboundary planning necessary to achieve positive conservation outcomes. However, existing spatial datasets in the Baltic Sea are limited in spatial coverage (e.g. national datasets), do not cover the full range of species, and are often difficult to compare due to the different modelling methods and input data used. Here we present a unified dataset of 630 species distribution models covering the whole Baltic Sea, including the Danish Straits and Kattegat. The dataset contains distribution models and maps for all fish, macrophytes, and invertebrates for which sufficient observation data within the region were available. The predicted distribution maps are provided in 250 m resolution and include several outputs intended to support spatial planning in the region, including spatial confidence estimates and expert evaluation scores for each species. Importantly, the dataset was produced using a harmonized methodology suitable across a wide range of species, allowing users to combine and compare different species and species groups for subsequent spatial analyses.

Background & Summary

The Baltic Sea is a semi-enclosed brackish water body in Northern Europe inhabited by a unique mixture of species adapted to freshwater and saltwater conditions. Its semi-enclosed nature, together with human habitation in surrounding lands, makes it vulnerable to a complex mixture of direct anthropogenic pressures, such as overfishing and habitat modification, and broad ranging environmental shifts, such as eutrophication and rising sea temperatures[1–5]. This has led to widespread declines in the abundance of key species and continuous degradation in the condition of habitats[6–10]. Despite these challenges, national and transboundary management efforts have been made to combat these declines, including fisheries regulation, habitat restoration, and spatial conservation planning[11–13]. However, one of the primary limitations to effective management across the region has hitherto been the lack of detailed spatial data on the distribution of species at ecologically relevant scales[14,15]. Such data are essential for effective spatial management practices, such as marine spatial planning, conservation, and spatial prioritization for marine protected areas.

In the coming years, substantial expansion of environmental management measures is expected in the Baltic Sea as part of regional and global policy objectives. Of particular importance among these is the Kunming-Montreal Global Biodiversity Framework, which aims to protect 30% of land and seas by the year 2030 (30x30)[16]. In the Baltic Sea, concrete policy targets for 30x30 have been established as part of the EU Biodiversity strategy[17]. Similarly, the Baltic Sea Action Plan has established regional objectives for returning the Baltic Sea to good environmental status[13]. Other relevant policy measures include the EU Marine Strategy Framework Directive (Directive 2008/56/EC, 2008), which requires members states

to achieve “good environmental status”, and the EU Nature Restoration Regulation (Regulation 2024/1991, 2024), which includes targets to implement restoration measures in 20% of sea (and land) areas in the EU by 2030[18]. Paramount to achieving these objectives is the availability of regional, full-coverage maps on species distributions to guide management decisions.

Species distribution models (SDMs) are a widely used and invaluable tool for generating full-coverage maps from sparse, irregular, or spatially biased observation data. SDMs are correlative models that use spatial layers of relevant variables (e.g., temperature, salinity, bathymetry) together with observation data (e.g., field surveys, citizen science, satellite imagery) to generate continuous predictions of species distributions. SDMs have a long history of application for spatial planning and can even provide useful insights into potential environmental or anthropogenic drivers of ecosystem change in the past or future[19–21]. For example, many conservation planning tools, such as Zonation, Marxan, and prioritizr, are dependent on SDMs as input data to identify priority areas for conservation action[22–24].

Production of SDMs is typically limited by the quality and breadth of observation data on which the models are trained, and the resolution and availability of relevant spatial predictor variables. In the Baltic Sea region, these challenges mean that, to date, region-wide distribution maps are limited to only a few species, typically low in spatial resolution, and produced using disparate methodologies[25]. Yet, such data are essential for regional transboundary planning. Here we present a dataset of species distribution models for 630 species (101 fish, 189 macrophytes, and 340 invertebrates) across the Baltic Sea HELCOM region. These models have been trained on a comprehensive dataset of species observations from across the Baltic Sea region and have been produced at a spatial resolution of 250 metres to support both local and regional planning scales.

Methods

Observation data for model training

Observation data for potential inclusion in the SDMs were gathered through a data call performed as part of the Protect Baltic EU Horizon project from November 2023 through March 2024. The data call targeted national agencies of Baltic Sea states responsible for environmental monitoring in aquatic areas. Additional datasets of species observations were obtained from public databases and through direct correspondence with national agencies. Observation data included surveys from the years 2000 to 2024. A full list of observation data sources is provided in the GitHub repository (linked below) in observations/data_sources.csv.

The observation data were separated into three categories based on the species groups that are typically targeted in different environmental monitoring and inventory surveys: fish, macrophytes, and invertebrates. Because the sampling methodology varied quite substantially among species groups, processing of the input observation data was handled separately for each species group. The full list of

species and their assigned modelling group is provided in the Zenodo repository (linked below) in WP3_species_list.csv. This pre-processing of observation data is described below.

Fish

Observation data for fish were restricted to scientific surveys conducted with fishing nets, including multi-mesh gillnets, fyke nets, benthic trawls, and pelagic trawls. Records from other survey types in which fish were not the primary target group (e.g. dive surveys) were excluded. These surveys also typically target adult or large juvenile fish, and as such the observation data and subsequent models are primarily a representation of these life stages. The rationale for including these survey types was to retain only broad-range sampling methodologies designed to record most species in the target group, so that a species not listed for a given survey could be assumed to have been absent rather than simply not targeted. In contrast, this assumption is likely false in species-targeted surveys, where non-detection of non-target species is uninformative about true absence.

Fish observations were then aggregated by date, latitude, longitude, and gear type to define unique observation events, with each event corresponding to a single deployment of a given gear at a given location and time. Any species present somewhere in the full retained dataset, but not recorded for a given event, was assigned as absent for that event.

Net-based fishing surveys also routinely capture and record several large and relatively mobile invertebrate species that are better sampled with fishing gear than with benthic samplers (e.g. large decapod crustaceans and squid). For these species, observation data processing and species distribution modelling followed the fish methodology described in this section, and the resulting model outputs were subsequently reassigned to the invertebrate species group. For the fish species group, all models were trained on a total of 65,882 observations.

Macrophytes

Macrophyte occurrence data were drawn from a mixture of survey types and treated accordingly. Complete dive surveys, in which all species present were recorded to species level, were used as presence–absence data; underwater video was treated equivalently for the few larger species reliably identifiable from video. Records from rakes, grab samplers, and species-targeted surveys (e.g. Danish eelgrass monitoring) were treated as presence-only. Background absences were drawn from deep soft-sediment benthos samples, capped at the number of presences for each species, except for very common species for which all 19,075 background points were retained. The data were aggregated either by latitude, longitude, and year (species with < 1000 raw occurrences) or to the 250 m modelling grid (species with ≥ 1000 raw occurrences), with any positive record within an aggregation unit classified as presence.

Differences in sampling intensity and protocol across countries occasionally produced biogeographic biases in model predictions, typically underprediction in parts of a species' range. Where these were judged substantial, three corrections were applied in sequence: (i) adding country as a sampling methodology covariate (see below); (ii) switching aggregation strategy (e.g. moving a species with < 1000 occurrences to grid-cell aggregation); and (iii) excluding Swedish observations from training. The last correction addressed several Kattegat macroalgae recorded in Denmark but not in nearby Swedish surveys despite similar environmental conditions, where retaining the Swedish data introduced large numbers of false absences and biased predictions toward Denmark.

To reduce spatial and environmental clustering, observations were partitioned into k-means clusters in environmental variable space, and 300 observations sampled per cluster, yielding datasets of approximately 30,000 (latitude/longitude/year aggregation) or 20,000 (grid cell aggregation) records. For rare species (prevalence < 0.02), where very low prevalence can degrade model performance[26], the clustering and sub-sampling were applied only to absences, with the number of clusters adjusted so that the final modelling prevalence was at least 0.02. For macrophytes, the total number of observations used to train the model varied by species, ranging from 1,770 to 29,979 after aggregating according to the aforementioned procedure.

Invertebrates

Observation data for invertebrates were restricted to those collected with grab or core samplers, the standard methods for soft-sediment benthic infauna. On hard bottoms, diver collected frame samples were used. Grab, core, and frame data are typically recorded as presence-only, meaning only the taxa identified in each sample are listed, and absences cannot be safely inferred across the full dataset because detection probability varies substantially with sampler type, sieve mesh size, and laboratory protocol. Processing was therefore carried out separately for each species, with absences constructed conservatively from sampling contexts in which the species was demonstrably detectable.

For each focal species, we first identified the combinations of country, year, and sampling method in which it had been recorded as present in at least one sample. Records of other invertebrate species made under these same (country, year, sampling method) combinations were then used to generate absences: where a sample within such a combination contained other invertebrates but did not list the focal species, it was assigned as an absence for the focal species. The union of the original presence records and the inferred absences was treated as the complete observation set for that species. This approach restricts absences to sampling contexts in which the focal species could plausibly have been detected, reducing the risk of treating non-detections under unsuitable methodology as true absences.

To reduce spatial and environmental clustering in intensively sampled sites that would otherwise dominate model training, the complete observation set was aggregated to the modelling grid. For each unique combination of year, sampling method, and grid cell with available data, a single observation was produced, classified as a presence if at least one presence occurred within the combination and as an

absence otherwise. As noted above, large and more mobile invertebrate species better captured by fishing gear were not processed through this pipeline, but followed the fish methodology, and their model outputs were subsequently reassigned to the invertebrate group. For the invertebrates species group, the total number of observations used to train the model varied by species, ranging from 472 to 29,872 after aggregating according to the aforementioned procedure.

Predictor variables

A set of 52 raster layers corresponding to predictor variables for potential inclusion in the species distribution models was sourced from the HELCOM Metadata catalogue[27]. These included various environmental variables, such as depth, salinity, temperature, and substrate type, among others. Next, predictor variable sets were further refined according to their relevant species subgroups, excluding variables with minimal ecological relevance for the species group. For example, salinity at the seafloor was used in the SDMs for species groups that are predominantly benthic, while salinity at the surface was excluded. Finally, for each predictor set, Variance Inflation Factor (VIF) tests were run with the goal of reducing VIF scores for all variables to below 10. Although a VIF score cutoff value of 10 is generally considered high, we deemed it as appropriate given the primary objective was to produce accurate species distribution models, rather than to identify the most important predictor variables correlated to species distributions. A list and description of the predictor variables used in the SDMs is provided in Table 1.

In addition to the determination of predictor variables described above, sampling methodology or sampling country was used as an additional covariate in a subset of the SDMs. For fish, potential inclusion of a fishing gear type covariate was considered. For macrophytes, potential inclusion of sampling country as a covariate was considered, because different countries typically used different sampling methodologies. For invertebrates, all observation data was collected using relatively similar methods, and so no sampling methodology covariate was used.

To determine if the inclusion of sampling method was necessary for a given species, we performed an initial run of all species with and without the sampling method covariate included. We then performed an internal review of the SDMs according to our expertise for each species group. For each species, a decision was then made to only include sampling method or country as a covariate if: (1) it substantially improved the model evaluation scores and (2) inclusion of the covariate eliminated extreme overestimation or underestimation of the species' distribution (e.g. the presence of freshwater species in areas of high salinity). Where improvement of the models with inclusion of the sampling method covariate was unclear, the covariate was not included in the final model in favour of choosing the simplest model. If sampling method was used in the SDM, when performing spatial predictions for each species, predictions were performed using the sampling method or country in which the species was most prevalent. For example, if a given fish species was most prevalent in pelagic trawl surveys, the

model was trained with the fishing gear type as a predictor, but final predictions were generated assuming a continuous layer corresponding to pelagic trawls as the sampling method.

Table 1
Predictor variables used in the species distribution models (SDMs).

Name	Selection in SDMs			
	Description	Fish	Macrophytes	Invertebrates
aspect	Aspect, i.e., direction in which the slope is facing out of 360 degrees (N, S, E, W)			✓
bpi_4km_20km	Bathymetric position index - ratio of average depth within 4 km radius to average depth within 20 km radius		✓	✓
bpi_500m_5km	Bathymetric position index - ratio of average depth within 500 m radius to average depth within 5 km radius	✓	✓	✓
chloro_a_mean	Mean annual mass concentration of chlorophyll-a in sea water from 2010 to 2023			✓
depth	Depth in metres below sea surface	✓	✓	✓
depth_attenuated_exposure	Depth attenuated exposure at the seafloor	✓	✓	✓
nitrate_surface_mean	Mean annual mole concentration of nitrate in sea water from 2010 to 2023		✓	
oxygen_bottom_mean	Mean annual mole concentration of dissolved molecular oxygen in sea water at the bottom from 2010 to 2023	✓		✓
ph_surface_mean	Mean annual sea water pH from 2010 to 2023		✓	
primary_prod_surface_mean	Mean annual net primary production of biomass expressed as carbon per unit volume in sea water from 2010 to 2023	✓		
rugosity	Rugosity within 3 cell radius			✓

Name		Selection in SDMs		
	Description	Fish	Macrophytes	Invertebrates
salinity_bottom_mean	Mean annual sea water salinity at the seafloor from 2010 to 2023	✓		✓
salinity_surface_mean	Mean annual sea water salinity at the surface from 2010 to 2023	✓	✓	✓
sea_ice_fraction_mean	Mean annual sea ice fraction from 2010 to 2023			✓
secchi_depth_mean	Mean annual secchi depth of sea water from 2010 to 2023		✓	
slope	Steepness of the seafloor slope in degrees	✓	✓	✓
substrate_coarse	Probability of coarse substrate		✓	
substrate_hard	Probability of hard substrate	✓	✓	
substrate_hard_binary	Binary classification of hard substrate			✓
substrate_mixed_binary	Binary classification of mixed substrate			✓
substrate_sand	Probability of sand substrate	✓	✓	
substrate_soft	Probability of soft substrate	✓	✓	
temperature_bottom_mean	Mean annual sea water potential temperature at seafloor from 2010 to 2023	✓		✓
temperature_surface_mean	Mean annual sea water potential temperature at surface from 2010 to 2023		✓	
wave_significant_height_mean	Mean annual sea surface wave significant height from 2010 to 2023	✓		✓

Species distribution modelling

Model training

All species distribution models were created using an ensemble approach with the 'biomod2' R package[28]. These ensemble models included three model algorithms: Generalized Additive Models (GAM), Random Forest (RF), and XGBoost (XGB). Five permutations of cross-validation were performed to evaluate the performance of each model and to weight model member in the final ensemble. Fo each permutation, all input observation data was separated into model training and testing data with a split of 80% and 20%, respectively. The training and testing data were allocated randomly within presence and absence observations, i.e. a random 80% of presence observations and a random 80% of absence observations. Thus, the prevalence (ratio of presences to absences) was maintained in both the training and testing dataset. This was done to avoid potentially, by chance, losing all presence observations in either the training or testing data for rare species with few presence observations. As such, for each species, each ensemble model contained 15 members (3 model algorithms, each with five cross-validation permutations).

No species-specific hyper-parameter tuning was performed, due to the large number of species and the computational time and resources that would be required to do so. Instead, suitable hyper-parameter values were chosen, based on our expertise, that are likely to perform well across a broad range of species and observation datasets. Hyper-parameter settings can be found in the source code within the GitHub repository for the species distribution models (described in the Data Records section). Model evaluation was performed using the True Skill Statistic (TSS), which is calculated as: sensitivity (true positive rate) + specificity (true negative rate) – 1. Evaluation was, therefore, based on binary predictions for the testing data, where the binary predictions are determined using the probability threshold that maximises TSS. TSS is a useful model evaluation measure that is suitable for species distribution models, particularly in cases where species are rare or have low prevalence in the observation data[29]. A TSS score of zero indicates a performance no better than random, while a score of one indicates perfect predictive accuracy. In the formation of the ensemble models, model members with TSS scores below 0.4 were excluded from the final ensemble model.

Model predictions

From the ensemble models, seven full-coverage map products were produced: (1) probability predictions, (2) binary predictions, (3) confidence estimates, (4) delta probabilities, (5) binary confidence maps, (6) strict binary maps, and (7) binary envelope maps. The probability predictions, binary predictions, and confidence estimates (1–3) are all direct predictions outputs from the model, while the other products (4–7) are derived from the direct outputs.

Probability predictions from the ensemble were produced using a weighted mean according to the following Eq. (1):

$$w = e^{5x}$$

where *w* is the weight of the member in the ensemble and *x* is the TSS score of each respective ensemble member. This provides a moderately steep exponential curve such that high-performing models have substantially greater weight in the final ensemble than poor-performing models. Weighted mean predictions are represented on a scale of 0-1000, where 1000 corresponds to a probability of 1. Binary prediction layers were generated using the presence/absence threshold that maximises TSS for the ensemble model and are represented as 0 (predicted absence) and 1 (predicted presence).

Confidence maps were produced using a "committee average" approach. Committee averaging was calculated as follows. First, predictions from each member of the final ensemble were converted to binary outputs, based on a threshold that maximises TSS. Then, the committee average was calculated as the mean of the binary predictions across members of the ensemble. As such, for a given cell, values of 1 indicate unanimous agreement on presence across all members, values of 0.5 indicate half of the members predicted presence and half predicted absence, and values of 0 indicate all members predicted absence. Thus, committee average values closer to 0.5 indicate disagreement between members of the ensemble. From the committee average outputs, we generated confidence maps according to the following Eq. (2):

$$c = 2\left|a - 0.5\right| \times \ 100$$

where *c* is the confidence and *a* is the committee average score for a given cell. Thus, a value of 100 indicates high confidence, while a value of 0 indicates low confidence.

The binary confidence maps categorise presence/absence predictions into four categories according to confidence scores. For these maps, confidence scores ≥ 50 were categorised as high confidence and scores < 50 as low confidence. The delta probability maps show the relative probability of areas predicted as presence. The delta probability maps were produced by multiplying the ensemble probability prediction maps by the binary maps and then rescaling the outputs from 0-1000. As such, in the delta probability maps, a value of 0 indicates absence, a value of 1 indicates the lowest probability of presence, and a value of 1000 indicates the highest probability of presence.

The strict binary maps categorise presence/absence predictions according to both confidence and probability predictions. The confidence categorisation follows the same method as described above. The probability categorisation separates high and low probabilities by the mid-point of the highest and lowest probabilities predicted above the presence threshold. The representation of these categories and scales is described further in the Data Record section below. Finally, the binary envelope maps were produced by applying a mask to the binary prediction map for species restricted to hard substrates. These species were identified by external experts throughout the expert evaluation process (described below). For these species, a binary map representing cells with > = 10% hard substrate[27] was multiplied by the binary prediction map.

After production of the SDMs, we applied technical exclusion criteria to determine which species models to include in the final dataset. To be included, the species model had to meet all of the following criteria:

1. More than one model algorithm type (GAM, RF, XGB) was included in the final ensemble (after the exclusion of members with TSS scores below 0.4).
2. The mean TSS score across all members of the final ensemble model was greater than or equal to 0.4.
3. The TSS of the final ensemble model was greater than or equal to 0.5.

Expert evaluation

In total, 630 species were successfully modelled and passed the technical criteria for inclusion in the final dataset (101 fish, 189 macrophytes, and 340 invertebrates). In addition to technical cross-validation for model evaluation (described above), the final set of SDMs were evaluated by both an internal panel of experts (within the project) and an external panel of experts.

For the internal evaluation, all SDM outputs were evaluated by modelling experts for each relevant species group. The internal expert panel consisted of four researchers from the Swedish University of Agricultural Sciences for fish, two researchers from the Finnish Environment Institute for macrophytes, and four researchers from the University of Tartu and Jade University of Applied Sciences for invertebrates. Species were then assigned a “flag” in one of the following categories: red, yellow, or green. Species were assigned red flags if the modelled distribution met any of the following criteria:

1. Substantial overprediction to locations where the species cannot feasibly exist (e.g., photosynthetic species in deep, aphotic areas or species predicted outside their known salinity tolerance limits)
2. Substantial underprediction in areas where the species is known to exist
3. Highly mobile or migratory species for which the sampling methodology and species distribution modelling approaches is likely to produce uncertain results, e.g. European eel (*Anguilla anguilla*), which migrates long distances across a wide range of environmental conditions.

Species were then assigned a yellow flag if application of the aforementioned criteria was unclear. Additionally, any species for which less than 150 presence observations were used in the training of the model (after the processing or aggregation of the observation data) were automatically assigned a yellow flag. Remaining species that did not meet any of these criteria were assigned a green flag.

For external evaluation, a panel of experts were nominated by national agencies of the HELCOM Contracting Parties. These experts were nominated based on their knowledge and expertise in the ecology and distribution of fish, macrophytes, or invertebrates. Experts were asked to assign a red, yellow, or green flag to each distribution map based on how well it reflects their knowledge of the geographic distribution of the species. Experts reviewed only the binary (confidence) outputs of the SDMs to simplify the review process. Mean expert evaluation scores for each species group are illustrated in Fig. 1. Based on both internal and external evaluations, each species was assigned overall evaluation scores ranging from 1 to 3, with 1 being the lowest score (red flag) and 3 being the highest (green flag). The external evaluation score represents the mean score across all expert scores for a given species.

Data record

All SDM data products are available on Zenodo: https://doi.org/10.5281/zenodo.20309311. All code and source data to reproduce the models are available in the following GitHub repository: https://github.com/helcomsecretariat/PROTECT-BALTIC-WP3-SDMs. All map products are provided in GeoTIFF format. All product maps and evaluation information can also be viewed on the following webpage: https://edsacre.github.io/WP3-SDMs/. In Table 2 we provide a summary of the different products available in the dataset.

Table 2
All data products provided in the dataset repository.

Product	Details
Probability predictions (probability_tif)	These maps are the probability prediction outputs of the SDMs. They are scaled between 0 and 1000, with higher values indicating higher probability of presence or higher suitability.
Binary predictions (binary_tif)	These maps are the binary prediction outputs of the SDMs. A value of 0 indicates absence and a value of 1 indicates presence.
Confidence estimates (confidence_tif)	These maps represent spatial confidence estimates, based on committee averaging across members of the ensemble SDM. Values range from 0 to 100, with higher values indicating higher confidence in predictions.
Delta probabilities (delta_tif)	These maps represent the relative probability of presence. They are calculated by multiplying the probability map by the binary map, then rescaling from 0 to 1000, such that a value of 0 represents absence, a value of 1 represents the lowest probability of presence, and a value of 1000 represent the highest probability of presence.
Binary confidence maps (binary_confidence_tif)	These maps are discrete classifications of habitat suitability for each species based on the binary predictions and confidence maps. Category values represent the following: 1 = Unsuitable (high confidence), 2 = Unsuitable (low confidence), 3 = Suitable (low confidence), 4 = Suitable (high confidence).
Strict binary maps (binary_strict_tif)	These maps are discrete classifications of habitat suitability for each species based on both confidence and predicted probability. Category values represent: 1 = Likely unsuitable, 2 = Possibly suitable, 3 = Likely suitable, 4 = Highly likely suitable.
Binary envelope maps (binary_envelope_tif)	These maps are binary outputs where areas with insufficient hard substrate cover have been masked out for species dependent on hard substrates. A value of 0 indicates absence and a value of 1 indicates presence.
WP3_species_list.csv	This spreadsheet contains a list of all the species and relevant information, such as model evaluation scores, sampling methodology covariates, species groups, etc.
WP3_species_review.csv	This spreadsheet contains the internal and external expert evaluation scores for each species.

Data overview

An overview of the seven map products is provided in Fig. 2. In Fig. 3, an example map is provided for a European perch (*Perca fluviatilis*) to illustrate the scale and resolution of the map products. In Fig. 4, mean spatial estimates of model confidence are displayed together with observation data density based on the number of observation points within each 10th -percentile across predictor variables.

Technical validation

Technical validation of the SDMs was done through cross-validation as described in the Methods section. Overall, the SDMs had good predictive accuracy as measured by TSS, although for some species, particularly rare species with few presence observations, performance was poor. The ensemble model outperformed ensemble members across all species (Fig. 5).

Declarations

Data availability

Map products for all 630 species are available on Zenodo: https://doi.org/10.5281/zenodo.20309311. All data for reproduction of the models, including observation data and predictor layers are available on GitHub: https://github.com/helcomsecretariat/PROTECT-BALTIC-WP3-SDMs.

Code availability

All code and data to reproduce the models are available on GitHub in the following repository: https://github.com/helcomsecretariat/PROTECT-BALTIC-WP3-SDMs.

Acknowledgements

This research was funded by the European Union through the Horizon Europe-funded project Protect Baltic (grant agreement 101112866). We would like to acknowledge and thank all of the institutions and people across the HELCOM region that have been involved in the collection, processing, and funding of projects for the collection of field data that have been used in this project. This work would not have been possible without these invaluable field monitoring and inventory programmes. We kindly thank all experts that reviewed the species distribution maps and provided invaluable feedback. We also thank the HELCOM Biodiversity Working Group for their feedback throughout the project.

Author contributions

The following authors contributed to drafting and writing of the manuscript: E.S., A.T., and A.K. All authors provided general feedback and comments on manuscript drafts. The following authors carried out the technical work and analysis: E.S., A.T., A.K., J.L., M.E., and R.F. The following authors contributed

to conceptualization and project management: E.S., A.T., A.K., J.L., U.B., E.V., J.K., M.E., R.F., M.L., O.T., J.H., C.B., K.J., D.H., and R.P. The following authors contributed to data collation and provision of unpublished data: M.O., J.G., L.B., A.L., C.H., J.D, and L.L.

References

1. Reckermann, M. *et al.* Human impacts and their interactions in the Baltic Sea region. *Earth Syst. Dyn.* 13, 1–80 (2022).
2. Reusch, T. B. H. *et al.* The Baltic Sea as a time machine for the future coastal ocean. *Sci. Adv.* 4, eaar8195 (2018).
3. HELCOM. *State of the Baltic Sea. Third HELCOM Holistic Assessment, 2016–2021*. (2023).
4. Andersen, J. H. *et al.* Long-term temporal and spatial trends in eutrophication status of the Baltic Sea. *Biol. Rev.* 92, 135–149 (2017).
5. Carstensen, J., Andersen, J. H., Gustafsson, B. G. & Conley, D. J. Deoxygenation of the Baltic Sea during the last century. *Proc. Natl. Acad. Sci.* 111, 5628–5633 (2014).
6. Casini, M. *et al.* Trophic cascades promote threshold-like shifts in pelagic marine ecosystems. *Proc. Natl. Acad. Sci.* 106, 197–202 (2009).
7. Conley, D. J. *et al.* Hypoxia is increasing in the coastal zone of the Baltic Sea. *Environ. Sci. Technol.* 45, 6777–6783 (2011).
8. Bergström, U. *et al.* Stickleback increase in the Baltic Sea – a thorny issue for coastal predatory fish. *Estuar. Coast. Shelf Sci.* 163, 134–142 (2015).
9. Eriksson, B. K. *et al.* Effects of altered offshore food webs on coastal ecosystems emphasize the need for cross-ecosystem management. *Ambio* 40, 786–797 (2011).
10. HELCOM. *HELCOM Red List II of the Baltic Sea Species in Danger of Becoming Extinct*. (2025).
11. Backer, H. *et al.* HELCOM Baltic Sea Action Plan – a regional programme of measures for the marine environment based on the Ecosystem Approach. *Mar. Pollut. Bull.* 60, 642–649 (2010).
12. European Parliament and Council. Directive 2008/56/EC establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive). (2008).
13. HELCOM. *HELCOM Baltic Sea Action Plan – 2021 Update*. (2021).
14. Sundblad, G., Bergström, U., Sandström, A. & Eklöv, P. Nursery habitat availability limits adult stock sizes of predatory coastal fish. *ICES J. Mar. Sci.* 71, 672–680 (2014).
15. Bučas, M. *et al.* Empirical modelling of benthic species distribution, abundance, and diversity in the Baltic Sea: evaluating the scope for predictive mapping using different modelling approaches. *ICES J. Mar. Sci.* 70, 1233–1243 (2013).
16. Convention on Biological Diversity. *Kunming-Montreal Global Biodiversity Framework*. (2022).
17. European Commission. EU Biodiversity Strategy for 2030: Bringing nature back into our lives. (2020).
18. European Parliament and Council. Regulation (EU) 2024/1991 on nature restoration. (2024).

19. Elith, J. & Leathwick, J. R. Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* 40, 677–697 (2009).
20. Guisan, A. *et al.* Predicting species distributions for conservation decisions. *Ecol. Lett.* 16, 1424–1435 (2013).
21. Reiss, H. *et al.* Benthos distribution modelling and its relevance for marine ecosystem management. *ICES J. Mar. Sci.* 72, 297–315 (2015).
22. Ball, I., Possingham, H. & Watts, M. Spatial Conservation Prioritisation: Quantitative Methods and Computational Tools. Chapter 14: Marxan and Relatives: Software for Spatial Conservation Prioritisation. in (Oxford, UK: Oxford Univ. Press, 2009).
23. Moilanen, A. *et al.* Novel methods for spatial prioritization with applications in conservation, land use planning and ecological impact avoidance. *Methods Ecol. Evol.* 13, 1062–1072 (2022).
24. Hanson, J. O. *et al.* Systematic conservation prioritization with the prioritizr R package. *Conserv. Biol.* 39, e14376 (2025).
25. HELCOM. HELCOM Map and data service. https://maps.helcom.fi/website/mapservice/index.html (2026).
26. Jiménez-Valverde, A., Lobo, J. M. & Hortal, J. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecol.* 10, 196–205 (2009).
27. HELCOM. Predictors used for spatial modelling of species and biotopes within the Protect Baltic project. HELCOM Metadata catalogue (2025).
28. Guéguen, M., Blancheteau, H., Lemaire-Patin, R. & Thuiller, W. *Biomod2: Ensemble Platform for Species Distribution Modeling*. (2026).
29. Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* 43, 1223–1232 (2006).

Figures

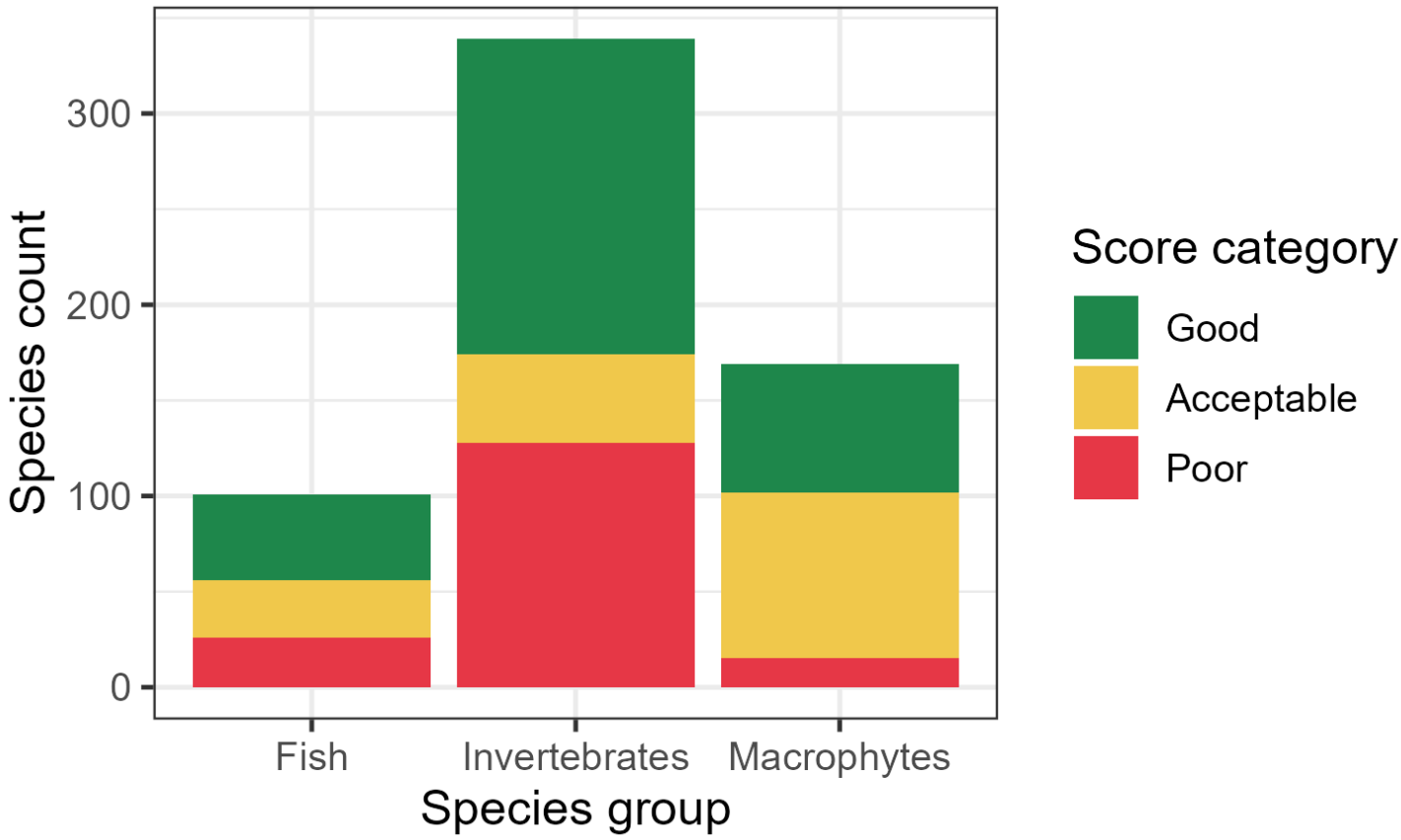


Figure 1

External expert evaluation score for the species distribution models across species groups. For the figure, mean expert evaluations score for each species, ranging from 1 to 3 were categorised as Poor (score between 1 and 1.67), Acceptable (score between 1.67 and 2.33), and Good (score between 2.33 and 3).

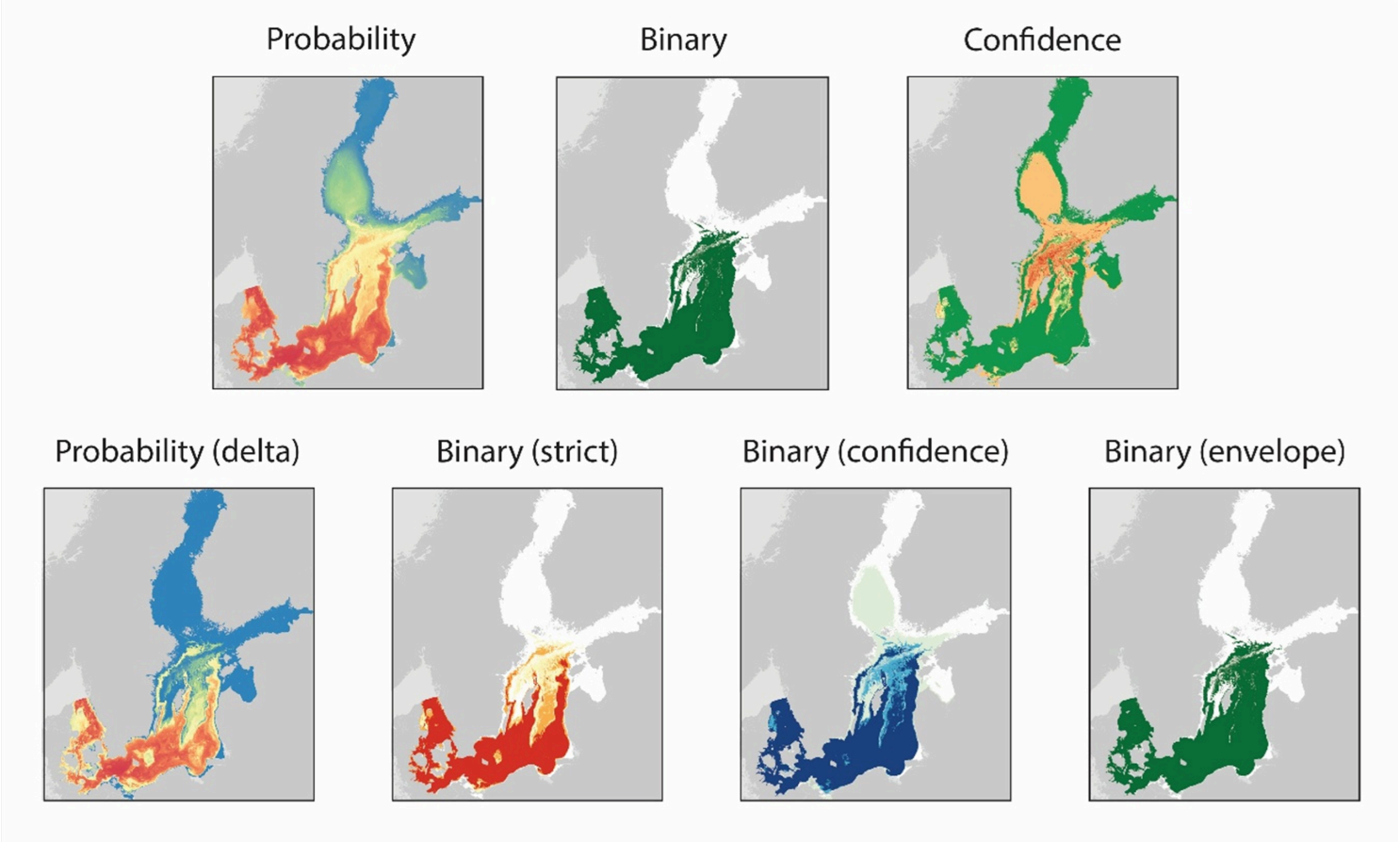


Figure 2

Map products available in the dataset. The map products in the figure are displayed using Atlantic cod (*Gadus morhua*) as an example species.

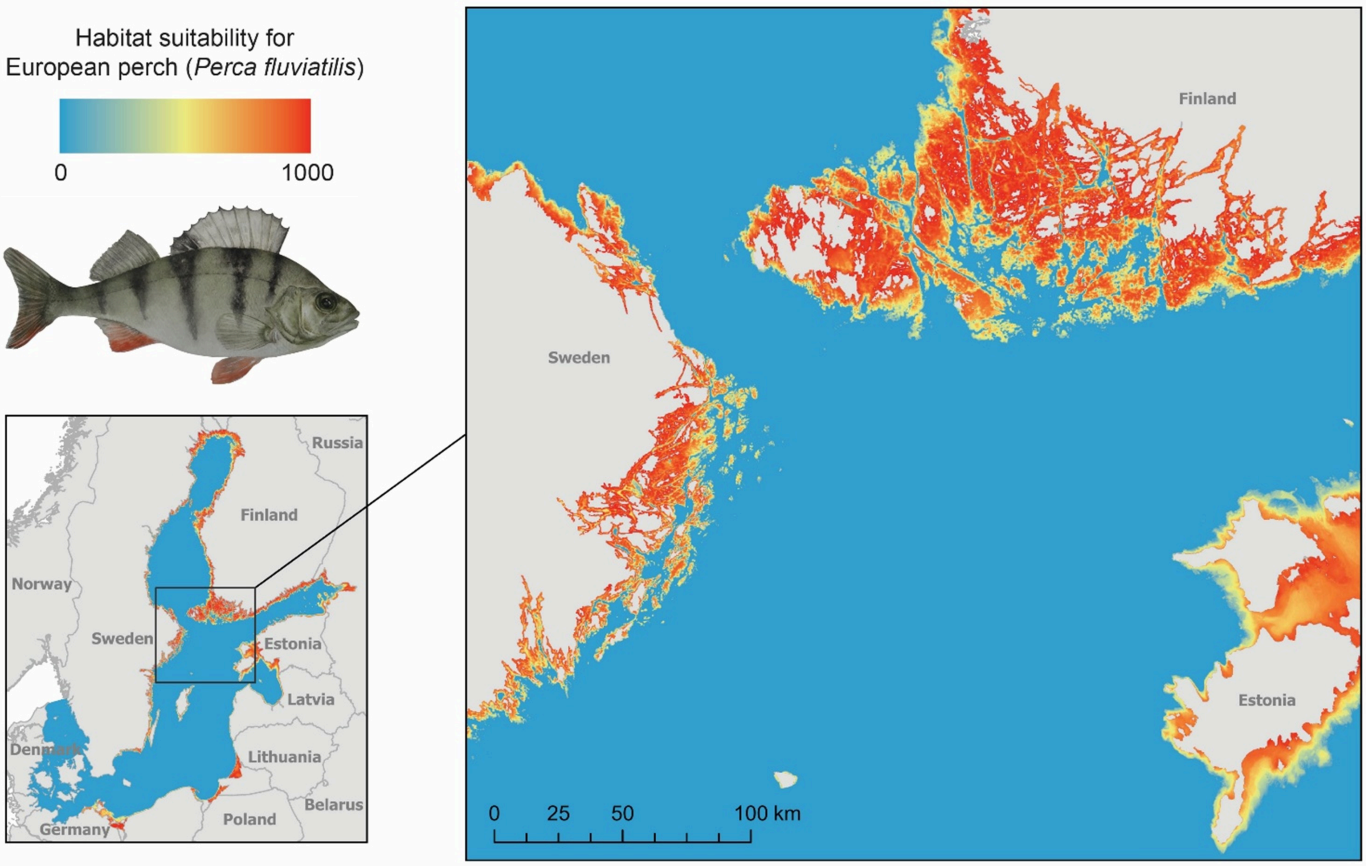


Figure 3

An example output product for European perch (*Perca fluviatilis*) showing the scale and resolution of the map products. The displayed example is the delta probability output, representing the relative probability of presence above the presence threshold (see Table 2). The perch illustration is an original painting done by Maria Eggertsen.

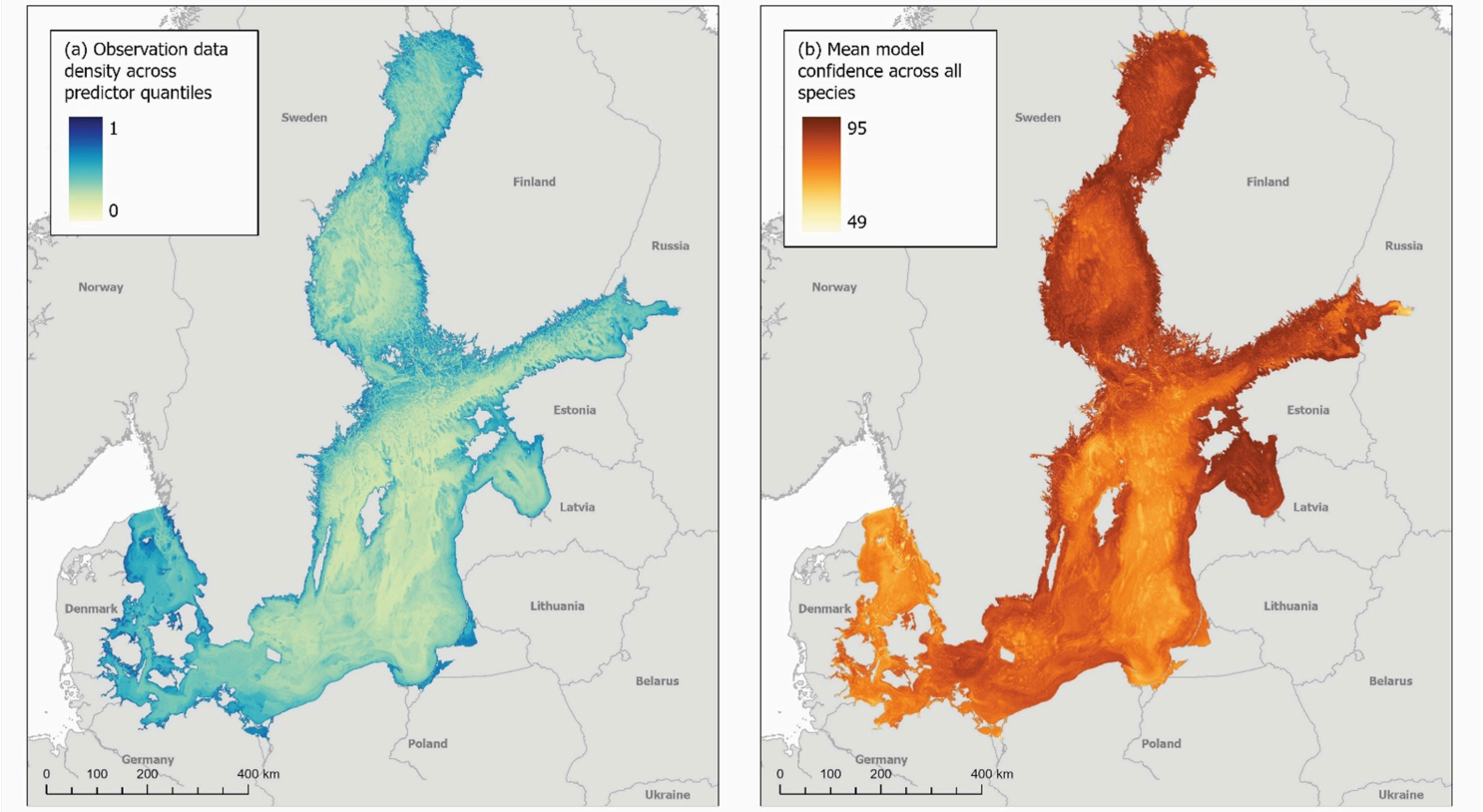


Figure 4

Spatial estimates of species distribution model confidence across the full dataset of 630 species. Panel (a) shows observation data density across all species, which was calculated by separating all values in all predictor layers into deciles and summing the number of observation points (presence or absence) within each decile. This measure was necessary because coordinates could not be shared for all observation data to produce a map of the number observation points per grid cell. Panel (b) shows the mean spatial confidence score across all species, as estimated by the level of agreement between members of the ensemble model.

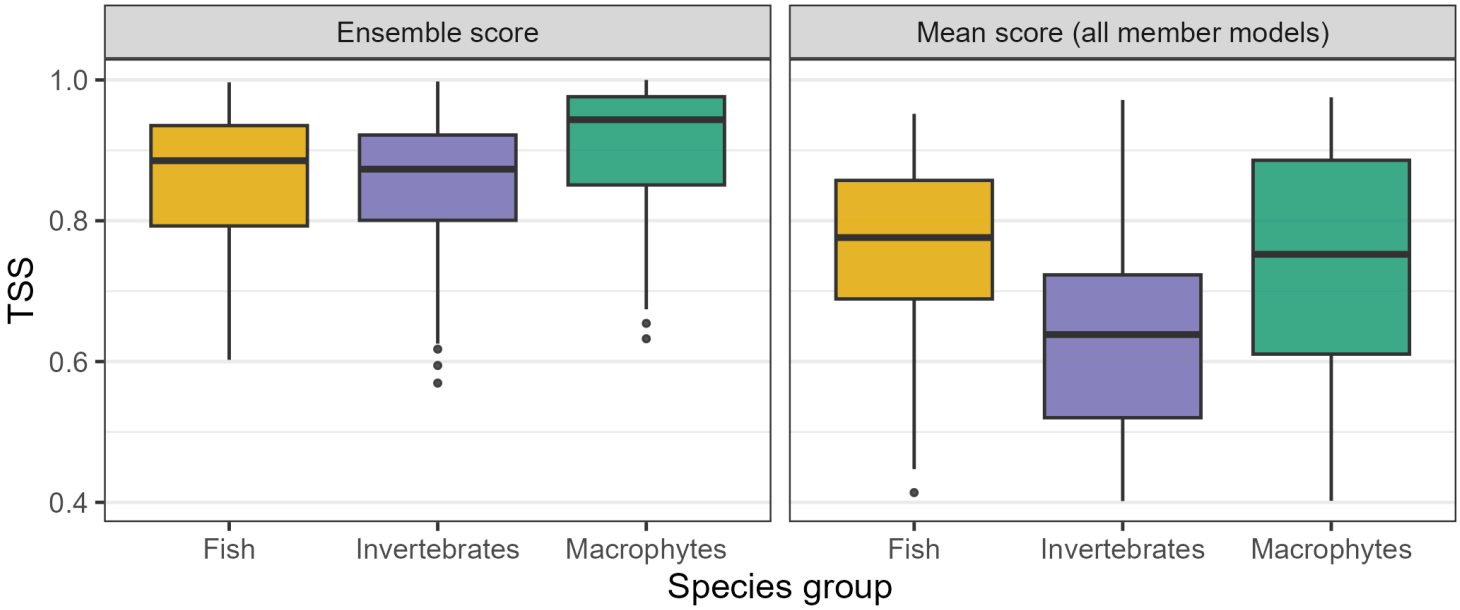


Figure 5

Technical evaluation scores of the species distribution models (SDMs) as measured using True Skill Statistic (TSS). The left panel shows TSS score across species groups for the final ensemble model used for the map products in the dataset. The panel on the right shows the mean TSS score across model members, showing that using the ensemble model improved performance overall.