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.

Building use-inspired species distribution models: using multiple data types to examine and improve model performance

Camrin D. Braun (cbraun@whoi.edu)
Woods Hole Oceanographic Institution
Martin C. Arostegui
Nima Farchadi
Michael Alexander
Pedro Afonso
Andrew Allyn
Steven J. Bograd
Stephanie Brodie
Daniel P. Crear
Emmett F. Culhane
Tobey H. Curtis
Elliott L. Hazen
Alex Kerney
Nerea Lezama-Ochoa
Katherine E. Mills
Dylan Pugh
Nuno Queiroz
James D. Scott
Gregory B. Skomal
David W. Sims
Simon R. Thorrold
Heather Welch
Riley Young-Morse
Rebecca Lewison

Research Article

Keywords: species distribution models, prediction, ecological forecasting, spatial ecology

Posted Date: April 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2802316/v1

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

Additional Declarations:
Competing interests: The authors declare no competing interests.

Building use-inspired species distribution models: using multiple data types to examine and improve model performance

Camrin D. Braun[1]*, Martin C. Arostegui[1], Nima Farchadi[2], Michael Alexander[3], Pedro Afonso[1,4], Andrew Allyn[5], Steven J. Bograd[6], Stephanie Brodie[6,7], Daniel P. Crear[8], Emmett F. Culhane[1, 9], Tobey H. Curtis[10], Elliott L. Hazen[6,7], Alex Kerney[5], Nerea Lezama-Ochoa[6,7], Katherine E. Mills[5], Dylan Pugh[5], Nuno Queiroz[11,12], James D. Scott[3,14], Gregory B. Skomal[15], David W. Sims[12,13], Simon R. Thorrold[1], Heather Welch[6,7], Riley Young-Morse[5] and Rebecca Lewison[2]

1. Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA USA
2. Institute for Ecological Monitoring and Management, San Diego State University, San Diego, CA USA
3. NOAA Earth System Research Laboratory, Boulder, CO USA
4. Okeanos and Institute of Marine Research, University of the Azores, 9901-862 Horta, Portugal
5. Gulf of Maine Research Institute, Portland, ME USA
6. Environmental Research Division, Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, Monterey, CA 93940, USA
7. Institute of Marine Sciences, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
8. ECS Federal, in Support of National Marine Fisheries Service, Atlantic Highly Migratory Species Management Division, Silver Spring, MD, USA
9. Massachusetts Institute of TechnologyWoods Hole Oceanographic Institution Joint Program in Oceanography-Applied Ocean Science and Engineering, Cambridge, MA 02139
10. National Marine Fisheries Service, Atlantic Highly Migratory Species Management Division, Gloucester, MA, USA
11. Research Network in Biodiversity and Evolutionary Biology, Universidade do Porto, Vairão, Portugal
12. Marine Biological Association of the United Kingdom, The Laboratory, Citadel Hill, Plymouth, UK
13. Ocean and Earth Science, National Oceanography Centre Southampton, University of Southampton, Southampton, UK
14. Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado
15. Massachusetts Division of Marine Fisheries, New Bedford, MA, USA

* Correspondence author. Email: cbraun@whoi.edu

Abstract

Species distribution models (SDMs) are becoming an important tool for marine conservation and management. Yet while there is an increasing diversity and volume of marine biodiversity data for training SDMs, little practical guidance is available on how to leverage distinct data types to build robust models. We explored the effect of different data types on the fit, performance and predictive ability of SDMs by comparing models trained with four data types for a heavily exploited pelagic fish, the blue shark (*Prionace glauca*), in the Northwest Atlantic: two fishery-dependent (conventional mark-recapture tags, fisheries observer records) and two fishery-independent (satellite-linked electronic tags, pop-up archival tags). We found that all four data types can result in robust models, but differences among spatial predictions highlighted the need to consider ecological realism in model selection and interpretation regardless of data type. Differences among models were primarily attributed to biases in how each data type, and the associated representation of absences, sampled the environment and summarized the resulting species distributions. Outputs from model ensembles and a model trained on all pooled data both proved effective for combining inferences across data types and provided more ecologically realistic predictions than individual models. Our results provide valuable guidance for practitioners developing SDMs. With increasing access to diverse data sources, future work should further develop truly integrative modeling approaches that can explicitly leverage strengths of individual data types while statistically accounting for limitations, such as sampling biases.

Key words: species distribution models, prediction, ecological forecasting, spatial ecology

Open research statement: The marker tag data used in this research is publicly available from the International Commission for the Conservation of Atlantic Tunas (ICCAT) Secretariat tag database which is archived at https://iccat.int/en/accesingdb.html, under "BSH" in the "Tagging" section. We certify that the electronic tag data will be publicly archived upon manuscript acceptance. A subset of the this data is currently published and publicly available on DataOne at https://search.dataone.org/#view/10.24431/rw1k329. The raw fishery-dependent observer dataset used in this study is considered confidential under the U.S. Magnuson-Stevens Act. Data can be requested by qualified researchers from the NOAA Pelagic Observer Program office by contacting popobserver@noaa.gov. We requested data representing all pelagic longline sets between the years 1993 and 2019. The code used in this analysis is publicly available on Github at https://github.com/camrinbraun/EcolApps_Data_Comparison.

1 Introduction

Species distribution models (SDMs) are an increasingly common tool used to understand species distributions and to predict species responses to changing environmental conditions (Elith et al., 2008; Guisan and Thuiller, 2005; Araújo et al., 2019). In the marine environment, SDMs have become an important tool to study biophysical drivers of habitat use that can be readily applied for conservation, spatial planning and fisheries management (Crear et al., 2021; Robinson et al., 2017; Araújo et al., 2019). While SDMs for marine species are often built using single data types (Grüss et al., 2019), there are a number of fishery-dependent and fishery-independent data sources that can be used to expand the scope and spatiotemporal scale of modeling efforts (Sequeira et al., 2013; Erauskin-Extramiana et al., 2019). Building robust SDMs is particularly important when faced with limited data, the need to understand how species will respond to a changing ocean, and to accurately assess exposure to various anthropogenic stressors including fisheries exploitation, habitat degradation, and energy development. Increasing human use of marine resources, climate variability and change, and limitations in data availability and scope require exploring best practices for leveraging multiple data types in marine conservation and management.

In addition to the typical fisheries datasets, such as vessel logbooks and fishery observers, a number of fishery-independent datasets have been developed that capture marine species occurrence, primarily as a product of targeted research or management efforts. Fishery-independent datasets include specific survey efforts, such as aerial or shipboard transect or trawl surveys (Di Sciara et al., 2015; Becker et al., 2019; Abrahms et al., 2019; Friedland et al., 2021), as well as electronic telemetry tags that track animal movement (*e.g.* Block et al. 2011, Queiroz et al. 2019). Electronic tags, in particular, represent species habitat use independent of fishing effort and are thus useful for representing the unbiased habitat use and environmental niche of tracked individuals. Despite the relatively high cost and low sample sizes, these datasets are growing and becoming increasingly available (Hussey et al., 2015), but guidance on best practices for building SDMs across disparate data types is lacking.

Here we develop a use-inspired comparison of SDMs built with four types of fishery-dependent and fishery-independent occurrence data using a heavily-exploited pelagic fish, the blue shark (*Prionace glauca*), as a model species to inform spatial management measures in a changing ocean. We use conventional marker tag, fishery observer, satellite-linked electronic tag, and pop-up archival tag data to fit data-specific SDMs in a comparative framework to inform important decisions in the model development process and identify tradeoffs associated with each data type. In addition to understanding differences among SDMs using a suite of validation and performance metrics, we tested the impact of data pooling and generating model ensembles for maximizing model utility and prioritizing model development in real-world applications.

2 Methods

2.1 Model species

Blue sharks occupy productive nearshore habitats in the North Atlantic Ocean during summer and fall (Carey and Scharold, 1990) and make extensive offshore migrations into the Gulf Stream and subtropical waters during winter (Campana et al., 2011; Vandeperre et al., 2014; Braun et al., 2019; Kohler and Turner, 2018; Queiroz et al., 2019). Blue sharks are typically caught as bycatch in longline fisheries that target swordfish and tunas, as well as recreational fisheries for large pelagic species (Aires-da Silva and Gallucci, 2007; Kohler and Turner, 2018). This species is also the target of a number of research efforts using electronic tags to study behavior and ecology across multiple ecosystems (*e.g.* Vandeperre et al. 2014; Braun et al. 2019). The relative abundance and widespread distribution of blue sharks results in a diverse set of occurrence data available for species distribution modeling (Druon et al., 2022), thus enabling evaluation of the data types and the associated model development process.

2.2 Fisheries-dependent datasets

2.2.1 Marker tag

We obtained marker tag data from the International Commission for the Conservation of Atlantic Tunas (ICCAT) Secretariat tag database (https://iccat.int/en/) for blue sharks in the Atlantic Ocean from 1959 to 2019. These marker (*e.g.* conventional or "spaghetti") tags are attached to a fish upon release and may be recorded again if the individual is later recaptured. This dataset consisted of 101,714 blue sharks tagged and released across a number of commercial and recreational fisheries. A total of 13,653 (~13%) tagged individuals were recaptured, yielding a total of 115,367 blue shark daily presence locations. The releases were dominated by three main gear types: 66% (n=67,085) were from rod and reel fisheries, 19% (18,826) from unclassified gear codes and 13% (13,022) from longline fisheries. Five gear types comprised the majority of marker tag recoveries: 34% (n=4,558) from longline, 21% (n=2,872) from rod and reel, 21% (n=2,806) from purse seine, 13% (n=1,728) from baitboat and 9% (n=1,197) from unclassified gear codes. These data were filtered to remove duplicate IDs and points on land, and only one tag event was retained for each day within a 0.01° grid to reduce autocorrelation structure in the data (Brodie et al., 2018a). The filtering steps retained 36,840 combined releases and recoveries in the North Atlantic during the oceanographic model time period (1993-2019) and were biased toward the NE U.S. shelf (Fig. 1a) during summer. Significant releases and recoveries occurred across the main footprint of the longline fleet in this region, spanning the area of impact of the Gulf Stream along the southeast U.S. and east of Cape Hatteras to the Azores and northern Europe.

2.2.2 Fisheries observer

The U.S. Atlantic pelagic longline fishery primarily targets swordfish (*Xiphias gladius*) and yellowfin tuna (*Thunnus albacares*). An at-sea observer program has been in place for this fishery since the early 1990s whereby independent observers catalog gear and catch information for every set made on ~10-15% of longline fishing trips (Beerkircher et al., 2002; Crear et al., 2021). These observer data were used to represent blue shark presence (catch) and absence through the spatial extent of the fishery concentrated in the northern Gulf of Mexico, along the east coast of the U.S. and along the southern and eastern edges of the Grand Banks (Fig. 1b). A total of 22,890 pelagic longline sets conducted between 1993-2019 were used in the analysis. A total of 8,057 and 14,833 sets recorded blue shark presence and absence, respectively.

2.3 Fisheries-independent datasets

2.3.1 Satellite-linked electronic tag

Satellite-linked tags (model SPOT, Wildlife Computers) were deployed on 70 individuals across a number of study sites in the North Atlantic, resulting in 6,430 unique individual tracking days over 12 years (2006-2018; Fig. 1c). Tags were attached to the dorsal fin of blue sharks in a manner similar to Braun et al. (2019). When at the surface, a wet-dry switch on the tag activated transmission to Argos satellites and a Doppler-based geoposition was calculated for the shark with associated location error (typically < 10km, Lopez et al. 2014). Resulting locations were then filtered using a speed filter (10 ms^{-1}) to remove unrealistic locations and regularized to daily location estimates by fitting a state-space model and predicting at daily time steps (R package `foieGras`, Jonsen et al. 2019, 2020).

2.3.2 Pop-up satellite archival transmitting tag

Pop-up satellite archival transmitting (aka "PSAT") tags (models PAT and miniPAT, Wildlife Computers) were deployed on 37 individuals in many of the same study locations, resulting in 5,136 unique individual tracking days over 8 years (2009-2017; Fig. 1d). Pop-up tags archive depth, temperature and light level data that are then used to estimate animal movements. However, tags that rely on light level for geolocation often exhibit large errors in daily position estimates (Nielsen and Sibert, 2007; Braun et al., 2015). We combined light and sea surface temperature measurements using a likelihood framework in a hidden Markov Model (Wildlife Computers "GPE3" geolocation software) which has been shown to provide realistic movement estimates to within <1° longitude and ~1-2° in latitude, particularly when datasets are high quality and target species are surface-oriented (Braun et al., 2018a). Fitted models provided daily location estimates and associated uncertainty for each tagged individual over the tag deployment period.

2.4 Environmental data

We included 10 environmental variables as potential predictor variables in the SDMs, which consisted of two static variables, seven dynamic surface variables and one dynamic subsurface variable to better represent the three-dimensional environment of this highly migratory species through time (Brodie et al., 2018b). The dynamic environmental data were sourced from the Global Ocean Physics Reanalysis (GLORYS, Copernicus Marine Environmental Monitoring Service; Lellouche et al. 2018). GLORYS is a global, data assimilating ocean model with daily outputs at 1/12° (~9km) horizontal resolution representing 50 vertical levels. The data assimilating nature of the model allows for regular data-driven updates to model predictions from *in situ* platforms and remote sensing observations that ensure realistic model outputs. The seven dynamic surface variables included: 1) sea surface temperature (SST; in °C) and 2) its spatial standard deviation (SST_sd; calculated over a 0.25° square), 3) sea surface height (SSH; in m) and 4) its spatial standard deviation (SSH_sd; calculated over a 0.25° square), 5) sea surface salinity (SSS; in PSU) and 6) its spatial standard deviation (SSS_sd; calculated over a 0.25° square) and 7) eddy kinetic energy (EKE; in m s^2). The dynamic subsurface variable, mixed layer depth (MLD; in m), was output from the model and used here as an index of water column structure. The two static variables included bathymetry (ETOPO1 obtained from https://www.ngdc. noaa.gov/mgg/global/global.html, coarsened to 1/12°; in m) and rugosity (calculated as the spatial standard deviation of bathymetry over a 0.25° square; in m). Each corresponding environmental value extracted from the presence/absence/pseudo-absence locations and times for each data type was included in the final dataframe. All environmental grids used the GLORYS native spatial (1/12°) and temporal (daily) resolution.

2.5 Species distribution models

The probability of species presence was modeled for each data type as a function of environmental variables using a boosted regression tree (BRT) framework (`dismo` R package, Elith et al. 2006). BRTs are non-parametric and use boosting (a numerical optimization technique) to determine optimal partitioning of variance. One of the advantages of using BRTs is their ability to handle correlation and collinearity effects of the environmental variables so *a priori* assessment of predictor variables is not needed (Elith et al., 2006). BRTs were fitted using a Bernoulli family appropriate to the binary nature of the response variable (presence / (pseudo)absence) and a fixed number of 2,000 trees with a learning rate of 0.005, a bag fraction of 0.75, and tree complexity of 5. Elith et al. (2008) present a thorough discussion of hyper-parameter tuning, therefore we fix these parameters here to isolate the effects of the different data types and our focal "treatments" (see below). The resulting models describe species-specific habitat suitability as continuous values ranging from 0 to 1.

2.6 Exploratory treatments: sample size, spatial extent, absences

In any SDM application, practitioners are faced with a number of decisions during model development that may impact the resulting model skill and applicability to the desired use case. We used the different data types to test the impact of three important aspects of our model framework: sample size, spatial extent, and representation of absences. To explore the effects of different sample sizes, models were trained with the maximum sample size available for each data type and then subsequently sub-sampled to 4,000 and 1,000 presences for subsequent model re-fitting.

We also explored how differing spatial extents affect model fit and performance. For our example use-case, we sought to build SDMs that could be predicted under climate change scenarios for the Northwest Atlantic Ocean. Therefore, our spatial extent of interest was the footprint of a down-scaled global climate model that spans from the Caribbean to the Grand Banks (Alexander et al., 2020), approximately equivalent to the extent of the fishery observer data and relatively restricted compared to the widespread coverage across the North Atlantic as represented by the other three data types. For spatial extent treatments, a model was trained for each data type with all available presence observations from the full spatial extent of each data type. Each data type was then subset to a common, limited spatial extent in the Northwest Atlantic within the spatial extent of the climate model as an example use-case. A second set of models for this treatment was then trained with the presence observations for each data type from this limited spatial extent. We subsequently compared predictions from the full extent and limited extent models within the spatial extent of the down-scaled climate model to understand the potential impacts of including training data from outside the study area.

A fundamental challenge of many data types for habitat modeling is that they are presence-only, and thus cannot provide information on animal absence. A number of techniques have been developed to simulate data representing where individuals were likely absent, often termed pseudo-absences (Barbet-Massin et al., 2012). These approaches include simple background sampling to more complex, biased sampling such as generating simulated animal movement trajectories using null animal movement models (Hazen et al., 2021; Pinti et al., 2022). For all datasets, we generated pseudo-absences using background sampling methods. Background sampling was performed by randomly drawing, without replacement, from the spatial extent of a given individual track from an electronic tag (background track sampling) or from the extent of the full dataset (background extent sampling). For electronic tags only, additional pseudo-absences were generated using correlated random walk simulations. To simulate realistic tracks and sample pseudo-absence locations, we conducted ten correlated random walk simulations per individual in each electronic tag dataset following Hazen et al. (2021). The fishery observer dataset does include observed fishing effort where blue sharks were not detected, but many of the fishing sets that recorded "absences" occurred in areas that were likely suitable blue shark habitat despite no blue sharks being captured,

presumably due to imperfect sampling as a function of gear-specific catchability. Thus, we also simulated pseudo-absences using the background method for the models fit with fishery observer data to compare to the "true" absences observed in these data. In all cases, dates were assigned to pseudo-absence locations by randomly drawing from the possible dates in the corresponding presence dataset. Simulated pseudo-absences were compared against all available presence data from all data types to avoid generating pseudo-absences for which a corresponding presence occurred in that month (regardless of year) and 0.1° grid cell (~10 km). Resulting pseudo-absence locations were randomly sub-sampled to generate a 1:1 presence/pseudo-absence dataset for each model training application.

Finally, we also explored two methods for combining data in SDMs. Pooling of data is common in species distribution modeling (Fletcher et al., 2019), especially when using opportunistic, presence-only data collated from multiple sources (Domisch et al., 2016). We created a pooled, all data model that was trained with all presences and associated pseudo-absences (from background sampling) combined across data types. Ensemble modeling techniques are also regularly applied to combine predictions across data types or model frameworks (Araújo and New, 2007). Thus, we also created an equal-weight, mean model ensemble that averaged across the predictions from each of the four data-specific models; in this case, each of the data-specific models relied on background pseudo-absence generation.

2.7 Comparing model performance

We evaluated model performance across three dimensions: explanatory power, predictive skill and ecological realism. Explanatory power indicates a models ability to explain the variability in a given dataset and was evaluated using percent explained deviance (R^2). Predictive skill indicates how well a model prediction can discern different actual outcomes (Norberg et al., 2019) and was evaluated with Area Under the Receiver Operating Characteristic Curve (AUC). These metrics were calculated using 10-fold cross-validation (Abrahms et al., 2019). We also calculated the sensitivity and specificity of each model (`caret` package for R, Kuhn 2015) that represent the proportion of true presences and true absences, respectively, correctly predicted by the model. Daily model predictions were generated for the full spatial extent of the data and predictions were classified as present when predicted suitability was greater than the 75% quantile of a given prediction surface and considered absent when less than the 25% quantile. We quantitatively assessed ecological realism for each model against its training data (*i.e.* in-sample) using median predicted habitat suitability at presences and pseudo-absences and qualitatively assessed realism using expert opinion of an example daily prediction for each model. The same quantitative approach was used for assessing each models predictive capacity (and thus ecological realism) against independent presence data (*i.e.* all true presences) from the three other data types (*e.g.* fisheries-observer SDM used to predict presences from the

three tagging datasets; repeated for all SDMs). Finally, we used pairwise correlation to quantify spatial variability among model predictions. We calculated Pearson's correlation coefficient in each grid cell by comparing monthly predictions (1993-2019; n=324) for each pair of data-specific models. For example, all monthly predictions from the marker tag model in a given grid cell were compared against all monthly predictions from the satellite tag model in the same grid cell by calculating the correlation between model predictions.

3 Results

After quality control and temporal filtering (1993-2019) to match available environmental data, we selected 56,240 presence observations for blue sharks in the North Atlantic from the 4 data types (Fig. 1). Our treatments identified a spectrum of model sensitivity to the different manipulations. The impact of successive reductions in sample sizes available for model training were minor based on metrics representing explanatory power, predictive skill and ecological realism (Table 1) and almost indiscernible among most example predictions (Fig. 2). In spatial extent manipulations, metrics for explanatory power, predictive skill, and ecological realism were relatively invariant for the three datasets that spanned the North Atlantic (marker, satellite and pop-up tags) and, in some cases, suggested minor improvements in model performance when spatial extent of the training data was limited to the NW Atlantic (Table 2, Fig. 3). In contrast, the performance of fishery observer models decreased across all metrics when comparing the full to limited spatial extent of training data.

Among the three treatments (sample size, spatial extent, representation of absences), manipulations in how absences were represented demonstrated the most significant impact on data-specific model performance. For both types of electronic tag data, pseudo-absences were either drawn from correlated random walk (CRW) simulations, randomly sampled from the extent of individual tracks (track extent) or randomly sampled from the extent of the full dataset pooled across individuals (background extent). In both cases, sampling pseudo-absences from the background extent resulted in the best performing model across all metrics compared to the track extent and CRW (Table 3). Among the two poorer performing pseudo-absence methods for electronic tag data (*i.e.* track extent and CRW), track extent pseudo-absence sampling consistently resulted in better predictive performance against all presence data across the four data types but within-sample metrics indicated slightly improved model performance using CRW-generated pseudo-absences (Table 3). The example predictions for the two electronic tag data types suggested the three pseudo-absence techniques resulted in significantly different predicted habitat suitability, with background extent sampling likely resulting in the most realistic predictions (Fig. 4). The background sampling of pseudo-absences also resulted in the most ecologically realistic predictions compared to models fit with "true" absence data in the observer dataset, despite the model performance metrics being largely invariant across absence and pseudo-absence based models for the observer data. For example, "true" absence models for

the fishery observer dataset predicted high habitat suitability in the subpolar North Atlantic and subtropical gyre for the example prediction day which contrasted with the almost complete absence of suitable habitat in these areas as predicted by the pseudo-absence based model (Fig. 4). The observed divergence across model predictions and, in some cases, between model validation metrics and ecological realism of model predictions (*e.g.* observer absence and pseudo-absence models, Table 3 & Fig. 4) highlights the utility in having experts assess the realism of model predictions in addition to commonly used model validation metrics.

Model performance also varied across data-specific models, with the marker tag model exhibiting the highest explanatory power and best predictive skill metrics (Table 4). Both fishery-dependent models indicated high performance metrics relative to fishery-independent models and resulted in spatially-constrained suitability in example predictions (Fig. 5, Table 4). In contrast, fishery-independent models predicted more widespread suitable habitat during the example July prediction; however, both satellite tag and pop-up tag-based models demonstrated better sensitivity when predicting to independent, out-of-sample presence data (Fig. 6). The marker tag model exhibited particularly high sensitivity predicting to both types of fishery-dependent presence observations, while the observer model indicated the lowest sensitivity of any model-data combination when predicting to the marker tag dataset. In contrast, the models trained with fishery-dependent data had higher specificity when predicting to true absences in the observer data.

Pairwise linear correlations among each model's prediction highlights where each pair of data-specific models tend to agree and disagree (Fig. 7). In general, there is large-scale agreement among models throughout the Slope Sea and along the U.S. East Coast and Gulf of Mexico. The most disagreement across models is apparent in the subpolar North Atlantic (Fig. 7a-c) and in subtropical waters east of the Mid-Atlantic Ridge. Overall, the model fit to all available presence data and the model ensemble (mean of each data-specific model prediction) provided similar example predictions (Fig. 5) and sensitivity when predicting to all available presence observations (Fig. 6). However, the data-pooled model and ensemble differed significantly in their in-sample predictive performance (Table 4), likely as a product of the ensemble predictions representing the mean suitability prediction across four data-specific models that were at times strongly divergent (Fig. 7).

4 Discussion

Species distribution models are an important tool to understand how species relate and respond to changing ocean conditions. Using data from a wide-ranging marine species, we found that inherent biases associated with both fishery-dependent and fishery-independent datasets, including spatial and temporal biases that arise from disproportionate sampling (*e.g.* fishing or tagging effort), must be considered when building models. Fishery-dependent datasets can be an effective and large-scale source for observations of marine species (*e.g.* Brodie et al. 2018a;

Arostegui et al. 2022). Despite the broad spatial extent and temporal coverage, models trained on these data are often influenced by non-random spatial and temporal distribution of fishing effort (*e.g.* Kroodsma et al. 2018). While both the marker tag and observer-based models were characterized by the highest model evaluation metrics, their performance when predicting to the fishery-independent datasets was generally poor, presumably as a result of heavily-biased sampling relative to environmental gradients (Baker et al., 2022). These results suggest that fishery-based models can reliably predict where blue sharks interact with specific fisheries (Stock et al., 2020; Crear et al., 2021). In contrast, the fishery-independent models exhibited generally lower evaluation metrics but were more broadly robust in their predictive performance and ecological realism, suggesting they may more accurately represent the realized environmental niche and geographic distribution of blue sharks beyond the footprint of the fishery. This distinction regarding the relative strengths of different data types may have even greater relevance for model projections to understand how species' distributions and their interactions with fisheries may shift under climate change (Karp et al., 2022).

In contrast to fisheries-dependent data, fisheries-independent electronic tags are critical for species that are rarely captured in fisheries or surveys and are otherwise data-limited with respect to their distribution. Archival, pop-up tags rely on *ad hoc* methods to estimate most probable movements of tagged animals (accuracy ≥1°, Nielsen and Sibert 2007; Wilson et al. 2007; Musyl et al. 2011; Braun et al. 2015, 2018b), whereas satellite-linked tags rely on communications to satellites at the surface, resulting in higher location accuracy (±5 km, Jonsen et al. 2020). This difference in accuracy between tag types suggests satellite-linked tags may provide superior occurrence data for SDMs; however, we found that the more error-prone observations from pop-up tags improved model performance. For both types of fishery-independent data, the environment was sampled for each presence location as the mean over the area encompassed by the estimated daily location ± the 95% confidence interval around that location. This approach explicitly accounts for location uncertainty and results in some averaging of environmental metrics over a broader area for the pop-up tags (due to higher uncertainty) compared to the specific environment sampled for the more accurate satellite tags. The improved model performance in our results is likely, in part, a product of smoothing the local environment to be more representative of regional scale environmental variability which has been shown to contribute disproportionately to SDM predictive performance (Brodie et al., 2021). The potential for environmentally-driven changes to drive the likelihood of surfacing behavior (*e.g.* Sepulveda et al. 2018), which is requisite for satellite-linked tag transmission, is likely another contributing factor to this data type exhibiting reduced model performance relative to pop-up tags. Models trained on satellite-linked tag data are biased to predict where the focal species engages in surfacing behavior (Pinti et al., 2022) akin to how fishery-based models are biased to predict where the focal species interacts with a fishery. Together, these results highlight important considerations for building SDMs with electronic tag data and suggest that relatively error-prone locations from archival tags may be suitable, or even superior in some applications, for model development.

4.1 Treatments: Sample size

With nearly an order of magnitude range in sample size across data types, we explored the impact of sample size on model validation metrics and ecological realism. Several efforts have demonstrated varying performance of different modeling approaches at very small sample sizes (<100; *e.g.* Hernandez et al. 2006; Wisz et al. 2008). However, such small sample sizes are becoming increasingly rare, particularly for marine species for which practitioners can leverage fishery interaction data and/or widespread tagging efforts (Hussey et al., 2015) that rapidly yield datasets in the hundreds to thousands. We demonstrate that the modeling framework used here was largely insensitive to changes in sample size in the thousands, even compared with full sample sizes with >36,000 occurrences. These results suggest that with the proper approach to model development, sample size should not inhibit habitat suitability models for most marine species, including rare or infrequently observed taxa (*e.g.* Lezama-Ochoa et al. 2020).

4.2 Treatments: Spatial extent

Information on species' occurrence over large scales is a fundamental need for basic and applied ecology studies. However, it is often time-consuming and expensive to develop survey-quality, large-scale species distribution datasets. Thus, practitioners often leverage opportunistic datasets that are available on smaller scales than the desired modeling application, when used with appropriate caution, to develop SDMs that can predict outside the original spatial extent (*e.g.* Stirling et al. 2016). While some work has shown that "scaling up" relatively small-scale, scientific survey data with opportunistic citizen science data can result in improved accuracy and spatial extent of SDMs (Robinson et al., 2020), our results suggest that survey-quality data may not be necessary when multiple, complementary, large-scale datasets exist, as is common for highly migratory marine species. Our results also corroborate previous findings that spatial mismatch between training data and the desired modeling application may not inhibit development of robust SDMs. For example, Abrahms et al. (2019) use electronic tag data from blue whales throughout >1,000,000 km^2 of the California Current to build SDMs that inform high collision risk areas and time periods in the ~6,000 km^2 Santa Barbara Channel located therein. While the authors did not explicitly test the impact of differing spatial extent between the blue whale occurrence data and desired modeling outcome, their model predictions proved consistent with independent sightings data and generally align with our results that differing spatial extent can be less important than other factors in training robust SDMs.

4.3 Treatments: Absences

The representation of absences proved the most important manipulation we tested during model development. Previous studies have indicated how critical pseudo-absence generation can be for modeling with presence-only

data (Barbet-Massin et al., 2012; Hazen et al., 2021; Pinti et al., 2022). Indeed, our findings align with suggestions by Hazen et al. (2021) that using background sampling to generate pseudo-absences results in the best model validation metrics and predictive skill. However, they also highlight that at least for their study species (blue whale) the expert opinion was that resulting model predictions were not biologically realistic compared to methods that leverage important characteristics of animal movement (*e.g.* autocorrelated step length and turn angles) such as the correlated random walk methods. In contrast, our blue shark models indicated that background sampling resulted in the best model metrics and most realistic models for this generalist species, highlighting the potential role of niche separation in presence versus pseudo-absence training data (O'Toole et al., 2021) and suggesting species-specific habitat specificity may be an important topic for future study.

The improved performance of fishery observer models trained with background pseudo-absences rather than "true" absences highlights the need to account for variable catchability of focal species when predicting their occurrence. Catchability is the efficiency of fishing gear in sampling a species' abundance and can change as a result of varying environmental conditions and fishing operational characteristics. Failing to account for catchability can obscure patterns in occurrence (Maunder and Punt, 2004). Most notably, the degree of vertical overlap between fishing gear and a species' habitat use modulates catchability. The diel change in depth distribution of many highly migratory marine species alters their susceptibility to being captured at a given depth (Ward and Myers, 2005), as does environmental variation in the water column that restricts species to near-surface waters or facilitates their increased occupation of deeper waters (*e.g.* Prince and Goodyear 2006; Arostegui et al. 2022). Similarly, modifications in fishery operations (*e.g.* changed hook and/or bait type) may also alter catchability (*e.g.* sea turtles and common mola – Arostegui et al. 2020) and can impact sympatric species in different ways (*e.g.* bigeye tuna versus porbeagle shark – Foster et al. 2012). Presence/absence data from fishery catches is, thus, more appropriately considered as detection/non-detection data due to the imperfect nature of such sampling (sensu MacKenzie et al. 2002). Models trained on fishery observer (or other catch) data must standardize for catchability when incorporating "true" absences or use pseudo-absences in their place. When catchability bias is unknown or variables contributing to catchability are unavailable, a background pseudo-absence approach (with filtering of pseudo-absences that conflict with known presences, as used here) may yield more realistic predictions.

4.4 Leveraging diverse data types

While previous studies have suggested that fishery-dependent and fishery-independent datasets can lead to consistent estimates of species' habitats (Pennino et al., 2016; Karp et al., 2022), our results suggest that models trained with heavily biased data may significantly diverge from less biased datasets, such as those collected with fishery-independent methods. Thus, we sought to leverage the diversity among data types to explore how to reconcile

the apparent differences among models. Combining multiple data sources is becoming increasingly common to model species distributions (Fletcher et al., 2019), often to supplement limited data (Fletcher et al., 2016) or to alleviate limitations of particular data types (Dorazio, 2014). While our pooled, all-data model demonstrated marginal performance from the perspective of traditional evaluation of model skill and ecological realism, the predictive performance to both fishery dependent and independent datasets was reasonable given disproportionate sample sizes among data types. Data pooling is the most common method of combining datasets (Fletcher et al., 2019), likely due to its simplicity, but does not account for the different assumptions and biases inherent in each data type. A number of studies have indicated empirical support for fitting independent models for distinct data types that are then combined through ensemble techniques (Araújo and New, 2007). Our approach to ensemble models assumed that the resulting model would better represent the spectrum of blue shark ecology from the fishery-independent datasets while still leveraging the significantly larger sample size from the fishery-dependent data. Indeed, our results suggest that even simple model ensembles may be an acceptable way to combine data for modeling species distribution as has been shown for other marine taxa (*e.g.* blue whale, Abrahms et al. 2019). Together, our results suggest that ensembles of independent models may be an appropriate compromise between: 1) data-rich fishery datasets that reliably predict a species fishery interaction probability but are not representative of the full extent of a species' distribution or habitat suitability; and 2) more ecologically-realistic predictions from fishery-independent models that tend to be more limited in spatial and temporal coverage.

Despite the relative success of model ensembles and data pooling shown here, a number of issues are apparent in this approach, including inability to explicitly account for uncertainty across datasets, leverage species-environment relationships across models, or incorporate spatial dependencies. Recent advances suggest that model-based data integration may be the most appropriate way to combine data (Fletcher et al., 2019) in order to retain the strengths of each dataset while explicitly accounting for data-specific biases (Isaac et al., 2020). Given the flexibility in these approaches, there are a number of opportunities for explicitly linking inference across datasets such that, for example, species-environment relationships can be derived using joint likelihood across diverse data types (Ahmad Suhaimi et al., 2021). Similarly, most SDMs – including those in this study – are spatially-implicit (and simple) in that they do not formally incorporate spatial dependencies in the data; although more complex in structure, spatially-explicit SDMs achieve greater predictive performance and are better suited to addressing management and conservation issues given their enhanced ability to represent local conditions (DeAngelis and Yurek, 2017; Domisch et al., 2019; Williamson et al., 2022). In applied science (such as spatial planning of marine protected areas), the ability to provide the most accurate species' occurrence predictions and their associated uncertainty (especially at local jurisdictional scales) is paramount; such information ultimately is used by managers in how they decide to balance the biological, economic, and social outcomes of fisheries that have real-world impact on fish and fishers (Anderson et al., 2019; Arostegui et al., 2021). As integrated and spatially-explicit SDMs

continue to gain traction in basic ecology and applied management (Zulian et al., 2021), practical guidance and best practices will make these approaches increasingly accessible to practitioners.

4.5 Conclusion

As SDMs become foundational in ecology, questions of how to use the ever-increasing volume of diverse data sets remain. While significant changes in sample size and spatial extent had relatively minor impacts on resulting models, our results demonstrate that how absences are represented in presence-absence models is a critical consideration in model development that can lead to varying model outcomes. Data-specific biases are inherent and in our results were clearly manifested in model predictions; these are integral considerations for modeling applications, particularly for models built with single data types. If multiple data types are available, our results suggest at minimum a comparison across models may illuminate important similarities and/or differences that can inform model utility for the desired application. We present an ensemble approach that leverages the desired strengths of the individual datasets while minimizing the inherent biases of each data type and provides the appropriate balance of predictive performance and ecological realism. In our use case, the divergence of the fishery observer model from the models trained with other data types, the variability among traditional model evaluation metrics, and the predictive performance of fishery-independent models together suggest an integrated approach to model development is needed to generate robust SDMs from diverse data types. While statistically reconciling, and even leveraging, diverse data types remains challenging for most practitioners, especially in a spatially-explicit model framework, increasing access to diverse data sources suggests explicit data integration is an important area for future work (Isaac et al., 2020) and will be instrumental in expanding and improving efforts to better understand the impacts of climate change on marine species.

5 Acknowledgements

We thank all those who supported tagging efforts, collection of observer program data, and those who contributed to the ICCAT marker tag program, including the NOAA Northeast Fisheries Science Center's Cooperative Shark Tagging Program. We thank the U.S. Atlantic pelagic longline fishery observers and data providers from the NOAA Southeast Fisheries Science Center including L. Beerkircher and S. Cushner. We are grateful to the numerous captains and crews who provided their expertise and ship time and thank J. Suca for helpful comments on an earlier version of this manuscript. This work was supported by a NASA Ecological Forecasting funded project (80NSSC19K0187) and NOAA's Integrated Ecosystem Assessment program. MCA was supported by the Postdoctoral Scholar Program at Woods Hole Oceanographic Institution with funding provided by the Dr. George D. Grice Postdoctoral Scholarship Fund.

6 References

Abrahms, B., et al., 2019: Dynamic ensemble models to predict distributions and anthropogenic risk exposure for highly mobile species. *Diversity and Distributions*, **25 (8)**, 1182–1193.

Ahmad Suhaimi, S. S., G. S. Blair, and S. G. Jarvis, 2021: Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Diversity and Distributions*, **27 (6)**, 1066–1075.

Aires-da Silva, a. M. and V. F. Gallucci, 2007: Demographic and risk analyses applied to management and conservation of the blue shark (Prionace glauca) in the North Atlantic Ocean. *Marine and Freshwater Research*, **58 (6)**, 570–580.

Alexander, M. A., S. I. Shin, J. D. Scott, E. Curchitser, and C. Stock, 2020: The response of the Northwest Atlantic Ocean to climate change. *Journal of Climate*, **33 (2)**, 405–428.

Anderson, C. M., et al., 2019: How commercial fishing effort is managed. *Fish and Fisheries*, **20 (2)**, 268–285.

Araújo, M. B. and M. New, 2007: Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22 (1)**, 42–47.

Araújo, M. B., et al., 2019: Standards for distribution models in biodiversity assessments. *Science Advances*, **5 (1)**, eaat4858.

Arostegui, M., C. Braun, P. Woodworth-Jefcoats, D. Kobayashi, and P. Gaube, 2020: Spatiotemporal segregation of ocean sunfish species (Molidae) in the eastern North Pacific. *Marine Ecology Progress Series*, **654**, 109–125.

Arostegui, M. C., C. M. Anderson, R. F. Benedict, C. Dailey, E. A. Fiorenza, and A. R. Jahn, 2021: Approaches to regulating recreational fisheries:ăbalancing biology with angler satisfaction. *Reviews in Fish Biology and Fisheries*, **31 (3)**, 573–598.

Arostegui, M. C., P. Gaube, P. A. Woodworth-Jefcoats, D. R. Kobayashi, and C. D. Braun, 2022: Anticyclonic eddies aggregate pelagic predators in a subtropical gyre. *Nature*, **609 (7927)**, 535–540.

Baker, D. J., I. M. Maclean, M. Goodall, and K. J. Gaston, 2022: Correlations between spatial sampling biases and environmental niches affect species distribution models. *Global Ecology and Biogeography*, **31 (6)**, 1038–1050.

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 (2)**, 327–338.

Becker, E. A., K. A. Forney, J. V. Redfern, J. Barlow, M. G. Jacox, J. J. Roberts, and D. M. Palacios, 2019: Predicting cetacean abundance and distribution in a changing climate. *Diversity and Distributions*, **25 (4)**, 626–643.

Beerkircher, L. R., E. Cortés, and M. Shivji, 2002: Characteristics of Shark Bycatch Observed on Pelagic Longlines off the Southeastern United States, 19922000. 1992–2000.

Block, B. A., et al., 2011: Tracking apex marine predator movements in a dynamic ocean. *Nature*, **475 (7354)**, 86–90.

Braun, C., M. Kaplan, A. Horodysky, and J. Llopiz, 2015: Satellite telemetry reveals physical processes driving billfish behavior. *Animal Biotelemetry*, **3 (1)**, 2.

Braun, C. D., B. Galuardi, and S. R. Thorrold, 2018a: HMMoce: An R package for improved geolocation of archival-tagged fishes using a hidden Markov method. *Methods in Ecology and Evolution*, **9**, 1212–1220, 0608246v3.

Braun, C. D., P. Gaube, T. H. Sinclair-Taylor, G. B. Skomal, and S. R. Thorrold, 2019: Mesoscale eddies release pelagic sharks from thermal constraints to foraging in the ocean twilight zone. *Proceedings of the National Academy of Sciences of the United States of America*, **116 (35)**, 17 187–17 192.

Braun, C. D., G. B. Skomal, and S. R. Thorrold, 2018b: Integrating archival tag data and a high-resolution oceanographic model to estimate basking shark (Cetorhinus maximus) movements in the western Atlantic. *Frontiers in Marine Science*, **5 (25)**.

Brodie, S., L. Litherland, J. Stewart, H. T. Schilling, J. G. Pepperell, and I. M. Suthers, 2018a: Citizen science records describe the distribution and migratory behaviour of a piscivorous predator, Pomatomus saltatrix. *ICES Journal of Marine Science*, **75 (5)**, 1573–1582.

Brodie, S., et al., 2018b: Integrating dynamic subsurface habitat metrics into species distribution models. *Frontiers in Marine Science*, **5 (June)**, 219.

Brodie, S., et al., 2021: Exploring timescales of predictability in species distributions. *Ecography*, **44 (6)**, 832–844.

Campana, S. E., A. Dorey, M. Fowler, W. Joyce, Z. Wang, D. Wright, and I. Yashayaev, 2011: Migration Pathways, Behavioural Thermoregulation and Overwintering Grounds of Blue Sharks in the Northwest Atlantic. *PLoS One*, **6 (2)**, e16 854.

Carey, F. G. and J. V. Scharold, 1990: Movements of blue sharks (Prionace glauca) in depth and course. *Marine Biology*, **106 (3)**, 329–342.

Crear, D. P., T. H. Curtis, S. J. Durkee, and J. K. Carlson, 2021: Highly migratory species predictive spatial modeling (PRiSM): an analytical framework for assessing the performance of spatial fisheries management. *Marine Biology*, **168 (10)**, 1–17.

DeAngelis, D. L. and S. Yurek, 2017: Spatially Explicit Modeling in Ecology: A Review. *Ecosystems*, **20 (2)**, 284–300.

Di Sciara, G. N., G. Lauriano, N. Pierantonio, A. Cañadas, G. Donovan, and S. Panigada, 2015: The devil we don't know: Investigating habitat and abundance of endangered giant devil rays in the North-Western Mediterranean Sea. *PLoS ONE*, **10 (11)**, 1–17.

Domisch, S., M. Friedrichs, T. Hein, F. Borgwardt, A. Wetzig, S. C. Jähnig, and S. D. Langhans, 2019: Spatially explicit species distribution models: A missed opportunity in conservation planning? *Diversity and Distributions*, **25 (5)**, 758–769.

Domisch, S., A. M. Wilson, and W. Jetz, 2016: Modelbased integration of observed and expertbased information for assessing the geographic and environmental distribution of freshwater species. *Ecography*, **39 (11)**, 1078–1088.

Dorazio, R. M., 2014: Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23 (12)**, 1472–1484.

Druon, J., et al., 2022: Global-scale environmental niche and habitat of blue shark (Prionace glauca) by size and sex: a pivotal step to improving stock management. *Frontiers in Marine Science*, **in press (April)**, 1–25.

Elith, J., J. R. Leathwick, and T. Hastie, 2008: A working guide to boosted regression trees. *Journal of Animal Ecology*, **77 (4)**, 802–813.

Elith, J., et al., 2006: Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29 (2)**, 129–151.

Erauskin-Extramiana, M., H. Arrizabalaga, A. J. Hobday, A. Cabré, L. Ibaibarriaga, I. Arregui, H. Murua, and G. Chust, 2019: Large-scale distribution of tuna species in a warming ocean. *Global Change Biology*, **25 (6)**, 2043–2060.

Fletcher, R. J., T. J. Hefley, E. P. Robertson, B. Zuckerberg, R. A. McCleery, and R. M. Dorazio, 2019: A practical guide for combining data to model species distributions. *Ecology*, **100 (6)**, 1–15.

Fletcher, R. J., R. A. McCleery, D. U. Greene, and C. A. Tye, 2016: Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, **31 (6)**, 1369–1382.

Foster, D. G., S. P. Epperly, A. K. Shah, and J. W. Watson, 2012: Evaluation of hook and bait type on the catch rates in the western North Atlantic Ocean pelagic longline fishery. *Bulletin of Marine Science*, **88 (3)**, 529–545.

Friedland, K. D., et al., 2021: Resource Occurrence and Productivity in Existing and Proposed Wind Energy Lease Areas on the Northeast US Shelf. *Frontiers in Marine Science*, **8 (April)**, 1–19.

Grüss, A., J. T. Thorson, and E. Jardim, 2019: Developing spatio-temporal models using multiple data types for evaluating population trends and habitat usage. *ICES Journal of Marine Science*, **76 (6)**, 1748–1761.

Guisan, A. and W. Thuiller, 2005: Predicting species distribution: offering more than simple habitat models. *Ecology letters*, **8 (9)**, 993–1009.

Hazen, E. L., B. Abrahms, S. Brodie, G. Carroll, H. Welch, and S. J. Bograd, 2021: Where did they not go? Considerations for generating pseudo-absences for telemetry-based habitat models. *Movement Ecology*, **9 (5)**, 1–13.

Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert, 2006: The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29 (5)**, 773–785.

Hussey, N. E., et al., 2015: Aquatic animal telemetry: A panoramic window into the underwater world. *Science (New York, N.Y.)*, **348 (6240)**, 1255 642–.

Isaac, N. J., et al., 2020: Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology and Evolution*, **35 (1)**, 56–67.

Jonsen, I. D., C. R. McMahon, T. A. Patterson, M. AugerMéthé, R. Harcourt, M. A. Hindell, and S. Bestley, 2019: Movement responses to environment: fast inference of variation among southern elephant seals with a mixed effects model. *Ecology*, **100 (1)**.

Jonsen, I. D., et al., 2020: A continuous-time state-space model for rapid quality control of argos locations from animal-borne tags. *Movement Ecology*, **8 (1)**, 1–13.

Karp, M. A., et al., 2022: Projecting species distributions using fishery-dependent data. *Fish and Fisheries*.

Kohler, N. E. and P. A. Turner, 2018: Distributions and movements of Atlantic shark species: a 52-year retrospective atlas of mark and recapture data.

Kroodsma, D. A., et al., 2018: Tracking the global footprint of fisheries. *Science*, **359 (6378)**, 904 LP – 908.

Kuhn, M., 2015: Caret: classification and regression training. *Astrophysics Source Code Library*, ascl–1505.

Lellouche, J.-M., et al., 2018: Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1 12° high-resolution system. *Ocean Science*, **14 (5)**, 1093–1126.

Lezama-Ochoa, N., M. G. Pennino, M. A. Hall, J. Lopez, and H. Murua, 2020: Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (Mobular mobular). *Scientific Reports*, **10 (1)**, 1–11.

Lopez, R., J.-P. Malarde, F. Royer, and P. Gaspar, 2014: Improving Argos doppler location using multiple-model Kalman filtering. *IEEE Transactions on Geoscience and Remote Sensing*, **52 (8)**, 4744–4755.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, A. A. Royle, and C. A. Langtimm, 2002: Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83 (8)**, 2248–2255.

Maunder, M. N. and A. E. Punt, 2004: Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, **70 (2)**, 141–159.

Musyl, M. K., R. W. Brill, D. S. Curran, N. M. Fragoso, L. M. McNaughton, A. Nielsen, B. S. Kikkawa, and C. D. Moyes, 2011: Postrelease survival, vertical and horizontal movements, and thermal habitats of five species of pelagic sharks in the central Pacific Ocean. *Fishery Bulletin*, **109 (4)**, 341–368.

Nielsen, A. and J. R. Sibert, 2007: Statespace model for light-based tracking of marine animals. *Canadian Journal of Fisheries and Aquatic Sciences*, **64 (8)**, 1055–1068.

Norberg, A., et al., 2019: A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, **89 (3)**, 1–24.

O'Toole, M., N. Queiroz, N. E. Humphries, D. W. Sims, and A. M. Sequeira, 2021: Quantifying effects of tracking data bias on species distribution models. *Methods in Ecology and Evolution*, **12 (1)**, 170–181.

Pennino, M. G., D. Conesa, A. Lopez-Quilez, F. Munoz, A. Fernández, and J. M. Bellido, 2016: Fishery-dependent and-independent data lead to consistent estimations of essential habitats. *ICES Journal of Marine Science*, **73 (9)**, 2302–2310.

Pinti, J., M. Shatley, A. Carlisle, B. A. Block, and M. J. Oliver, 2022: Using pseudo-absence models to test for environmental selection in marine movement ecology: the importance of sample size and selection strength. *Movement Ecology*, **10 (1)**, 1–17.

Prince, E. D. and C. P. Goodyear, 2006: Hypoxiabased habitat compression of tropical pelagic fishes. *Fisheries Oceanography*, **15 (6)**, 451–464.

Queiroz, N., et al., 2019: Global spatial risk assessment of sharks under the footprint of fisheries. *Nature*, **572 (7770)**, 461–466.

Robinson, N. M., W. A. Nelson, M. J. Costello, J. E. Sutherland, and C. J. Lundquist, 2017: A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. *Frontiers in Marine Science*, **4**, 421.

Robinson, O. J., V. Ruiz-Gutierrez, M. D. Reynolds, G. H. Golet, M. Strimas-Mackey, and D. Fink, 2020: Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions*, **26 (8)**, 976–986.

Sepulveda, C. A., S. A. Aalbers, C. Heberer, S. Kohin, and H. Dewar, 2018: Movements and behaviors of swordfish Xiphias gladius in the United States Pacific Leatherback Conservation Area. *Fisheries Oceanography*, **27 (4)**, 381–394.

Sequeira, A. M. M., C. Mellin, M. G. Meekan, D. W. Sims, and C. J. A. Bradshaw, 2013: Inferred global connectivity of whale shark Rhincodon typus populations. *J Fish Biol*, **82**, 367–389.

Stirling, D. A., P. Boulcott, B. E. Scott, and P. J. Wright, 2016: Using verified species distribution models to inform the conservation of a rare marine species. *Diversity and Distributions*, **22 (7)**, 808–822.

Stock, B. C., E. J. Ward, T. Eguchi, J. E. Jannot, J. T. Thorson, B. E. Feist, and B. X. Semmens, 2020: Comparing predictions of fisheries bycatch using multiple spatiotemporal species distribution model frameworks. *Canadian Journal of Fisheries and Aquatic Sciences*, **77 (1)**, 146–163.

Vandeperre, F., A. Aires-da Silva, J. Fontes, M. Santos, R. Serrão Santos, and P. Afonso, 2014: Movements of Blue Sharks (Prionace glauca) across Their Life History. *PloS one*, **9 (8)**, e103 538.

Ward, P. and R. A. Myers, 2005: Inferring the depth distribution of catchability for pelagic fishes and correcting for variations in the depth of longline fishing gear. *Canadian Journal of Fisheries and Aquatic Sciences*, **62 (5)**, 1130–1142.

Williamson, L. D., B. E. Scott, M. Laxton, J. B. Illian, V. L. Todd, P. I. Miller, and K. L. Brookes, 2022: Comparing distribution of harbour porpoise using generalized additive models and hierarchical Bayesian models with integrated nested laplace approximation. *Ecological Modelling*, **470 (April)**, 110 011.

Wilson, S. G., B. S. Stewart, J. J. Polovina, M. G. Meekan, J. D. Stevens, and B. Galuardi, 2007: Accuracy and precision of archival tag data: a multiple-tagging study conducted on a whale shark (Rhincodon typus) in the Indian Ocean. *Fisheries Oceanography*, **16 (6)**, 547–554.

Wisz, M. S., et al., 2008: Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14 (5)**, 763–773.

Zulian, V., D. A. Miller, and G. Ferraz, 2021: Integrating citizen-science and planned-survey data improves species distribution estimates. *Diversity and Distributions*, **27 (12)**, 2498–2509.

Table 1: Summary of model statistics for sample size manipulations. For each data type, a "full" model was built with all available presence observations (1st row of each data type) then randomly sub-sampled to smaller sample sizes. For all metrics except prediction at pseudo-absences, higher values indicate better model performance.

		Explanatory power	Predictive skill	Ecological realism			
Data type	**N**	**R^2**	**AUC**	**Median in-sample prediction at presences**	**Median in-sample prediction at pseudoabsences**	**Median prediction at all true presences**	**Figure panel**
Marker	36,840	0.71	0.97	0.98	0.06	0.93	2a
	4,000	0.73	0.97	0.98	0.06	0.93	2b
	1,000	0.79	0.96	0.97	0.06	0.93	2c
Observer	8,057	0.58	0.94	0.91	0.08	0.79	2j
	4,000	0.59	0.94	0.90	0.08	0.77	2k
	1,000	0.66	0.93	0.90	0.10	0.85	2l
Satellite	6,430	0.27	0.81	0.64	0.36	0.73	2d
	4,000	0.29	0.81	0.64	0.36	0.72	2e
	1,000	0.41	0.80	0.67	0.32	0.70	2f
Pop-up	4,913	0.50	0.93	0.79	0.18	0.52	2g
	4,000	0.49	0.92	0.78	0.19	0.58	2h
	1,000	0.58	0.92	0.80	0.18	0.70	2i

Table 2: Summary of model statistics for spatial extent manipulations. For each data type, a model was built with all available presence observations from the full spatial extent of each data type (1st row of each data type and see Fig. 1). Each data type was subset to a common, limited spatial extent in the Northwest Atlantic as an example study region of interest (2nd row for each data type), in this case representing the spatial extent of a downscaled global climate model. For all metrics except prediction at pseudo-absences, higher values indicate better model performance.

			Explanatory power	**Predictive skill**	**Ecological realism**			
Data type	**Spatial extent of data**	**N**	**R^2**	**AUC**	**Median in-sample prediction at presences**	**Median in-sample prediction at pseudoabsences**	**Median prediction at all true presences**	**Figure panel**
Marker	Full	36,840	0.71	0.97	0.98	0.06	0.98	3a
	Limited	8,950	0.79	0.98	0.97	0.02	0.96	3b
Observer	Full	8,057	0.58	0.94	0.91	0.08	0.81	3c
	Limited	2,572	0.39	0.85	0.76	0.23	0.59	3d
Satellite	Full	6,430	0.27	0.81	0.64	0.36	0.77	3e
	Limited	2,043	0.46	0.88	0.75	0.22	0.75	3f
Pop-up	Full	4,913	0.50	0.93	0.79	0.18	0.52	3g
	Limited	1,593	0.57	0.92	0.82	0.13	0.39	3h

Table 3: Summary of model statistics for "true" absence and pseudo-absence manipulations. Models based on observer data were fit with all absences (n=14,833; approx. 1:2 presence to absence ratio), sub-sampled true absences (to represent 1:1 presence to absence ratio) and pseudo-absences randomly sampled from the background extent of the dataset. The two types of electronic tag datasets (satellite and pop-up) were each treated with 3 different pseudo-absence generation techniques: correlated random walk, sampling from the extent of individual tracks and background sampling from the full spatial extent (see Methods). For all metrics except prediction at pseudo-absences, higher values indicate better model performance.

		Explanatory power	**Predictive skill**	**Ecological realism**			
Data type	**(Pseudo) absence method**	**R^2**	**AUC**	**Median in-sample prediction at presences**	**Median in-sample prediction at pseudoabsences**	**Median prediction at all true presences**	**Figure panel**
Observer	True (all)	0.57	0.94	0.85	0.05	0.70	4a
	True (1:1)	0.58	0.94	0.91	0.08	0.79	4b
	Bkgd extent	0.62	0.95	0.93	0.09	0.12	4c
Satellite	CRW	0.15	0.73	0.57	0.46	0.61	4d
	Track extent	0.13	0.70	0.53	0.45	0.71	4e
	Bkgd extent	0.24	0.81	0.64	0.35	0.73	4f
Pop-up	CRW	0.17	0.74	0.58	0.45	0.53	4g
	Track extent	0.14	0.70	0.54	0.47	0.65	4h
	Bkgd extent	0.49	0.92	0.79	0.18	0.66	4i

Table 4: Summary of model evaluation statistics for selected, final models for each data type and the all data model and model ensemble. *indicates values report the same metric. For all metrics except prediction at pseudo-absences, higher values indicate better model performance.

			Explanatory power	Predictive skill	Ecological realism			
Data type	**Pseudoabsence type**	**N**	**R^2**	**AUC**	**Median in-sample prediction at presences**	**Median in-sample prediction at pseudo absences**	**Median prediction at all true presences**	**Figure panel**
Marker tags	Background extent	36,840	0.71	0.97	0.98	0.06	0.93	5a
Fishery observer	Background extent	8,057	0.62	0.95	0.93	0.09	0.12	5b
Satellite tags	Background extent	6,430	0.27	0.81	0.64	0.36	0.73	5c
Pop-up tags	Background extent	4,913	0.50	0.93	0.79	0.18	0.68	5d
All data	Background extent	56,463	0.52	0.93	0.93*	0.14	0.93*	5e
Ensemble	Background extent	56,463	NA	0.92	0.67*	0.20	0.67*	5f

Figure 1. Presence locations for the marker tags (a), fishery observer data (b), and two types of electronic tags (c, satellite and d, pop-up). Marker tags and observer data are fishery dependent (a,b), and electronic tags are fishery independent (c,d). Observer data (b) also contains "true" absence locations (but see Discussion). Note that grid cells for the fishery observer locations that contained < 3 vessels were removed to protect confidentiality. Orange triangles in c and d indicate the locations where tags were deployed.

Figure 2. Predicted habitat suitability for an example day (2019-07-01) showing the impact of sample size manipulations for models trained with each data type. Yellow indicates highly suitable habitat and blue indicates low suitability.

Figure 3. Predicted habitat suitability for an example day (2019-07-01) showing the impact of spatial extent manipulations for each data type. The first column shows example predictions for data-specific models trained with the full spatial extent of each data type (see Fig. 1) and predicted to the extent of the downscaled climate model. The second column shows example predictions for models trained with occurrence data only from within the spatial extent shown.

Figure 4. Predicted habitat suitability for an example day (2019-07-01) showing the impact of absence and pseudo-absence manipulations for each data type. The observer data contain "true" absence locations that were all used for the first treatment (a; ~1:2 presence to absence ratio) and were sub-sampled to a 1:1 ratio for the second treatment (b). The third treatment (c) used pseudo-absences sampled from the background extent of the observer data. The electronic tag datasets (satellite and pop-up) are presence-only and thus require pseudo-absence generation. Three methods were tested: correlated random walk (d, g), sampling from the extent of individual tracks (e, h) and sampling from the background extent of the dataset (f, i).

Figure 5. Predicted habitat suitability for an example day (2019-07-01) using models fitted with each data type, the all data model (panel e) and the ensemble of panels a-d (panel f). Yellow indicates highly suitable habitat and blue indicates low suitability. The black grid cells indicate where presence data are available during any July in each dataset.

Figure 6. Proportion of presences (sensitivity, a) and "true" absences from the observer data (specificity, b) correctly predicted by each selected model (Table 4) and dataset combination. Model predictions were considered correct when predicted suitability was greater than the 75% quantile for presence observations and less than the 25% quantile for absences in the observer data. Model ensemble includes the selected model for each data type (Table 4), excluding the all data model (*i.e.* rows 1-4).

Figure 7. Pairwise linear correlation of monthly predictions during the GLORYS period (1993-2019) for each data-specific model. High positive correlation (red) indicates similarity in model predictions. High negative correlation (blue) indicates model predictions are in opposition.

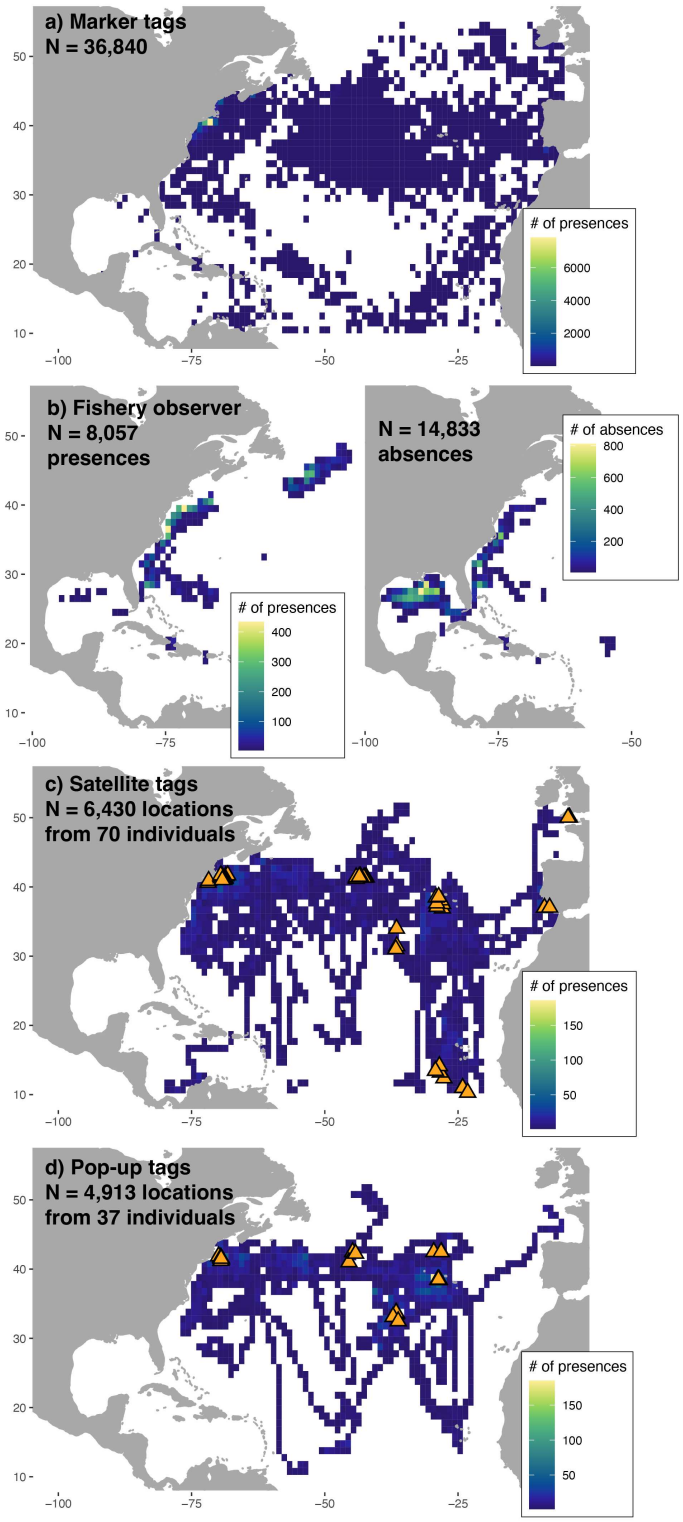
a) Marker tags
N = 36,840
# of presences
6000
4000
2000
b) Fishery observer
N = 8,057
presences
# of presences
400
300
200
100
N = 14,833
absences
# of absences
800
600
400
200
c) Satellite tags
N = 6,430 locations
from 70 individuals
# of presences
150
100
50
d) Pop-up tags
N = 4,913 locations
from 37 individuals
# of presences
150
100
50

Figure 1:

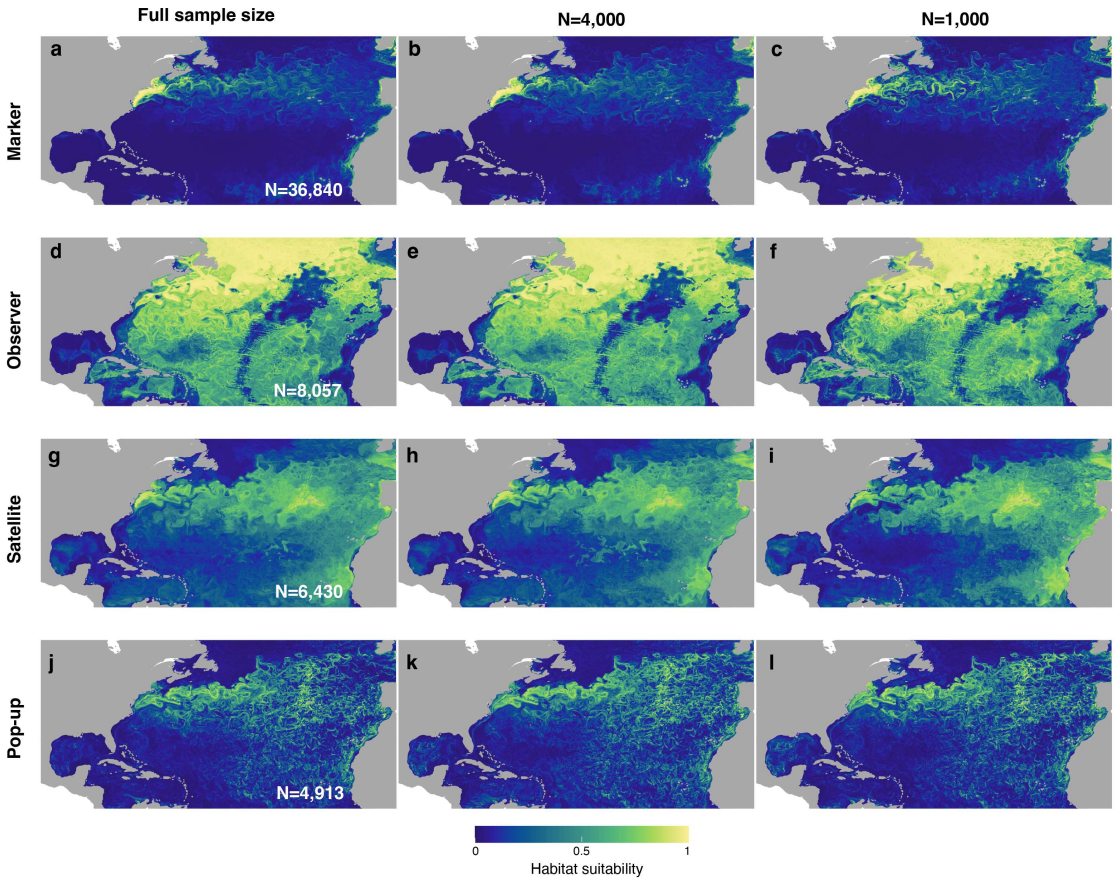
Full sample size
N=4,000
N=1,000
Marker
Observer
Satellite
Pop-up
a
b
c
d
e
f
g
h
i
j
k
l
N=36,840
N=8,057
N=6,430
N=4,913
0
0.5
1
Habitat suitability

Figure 2:

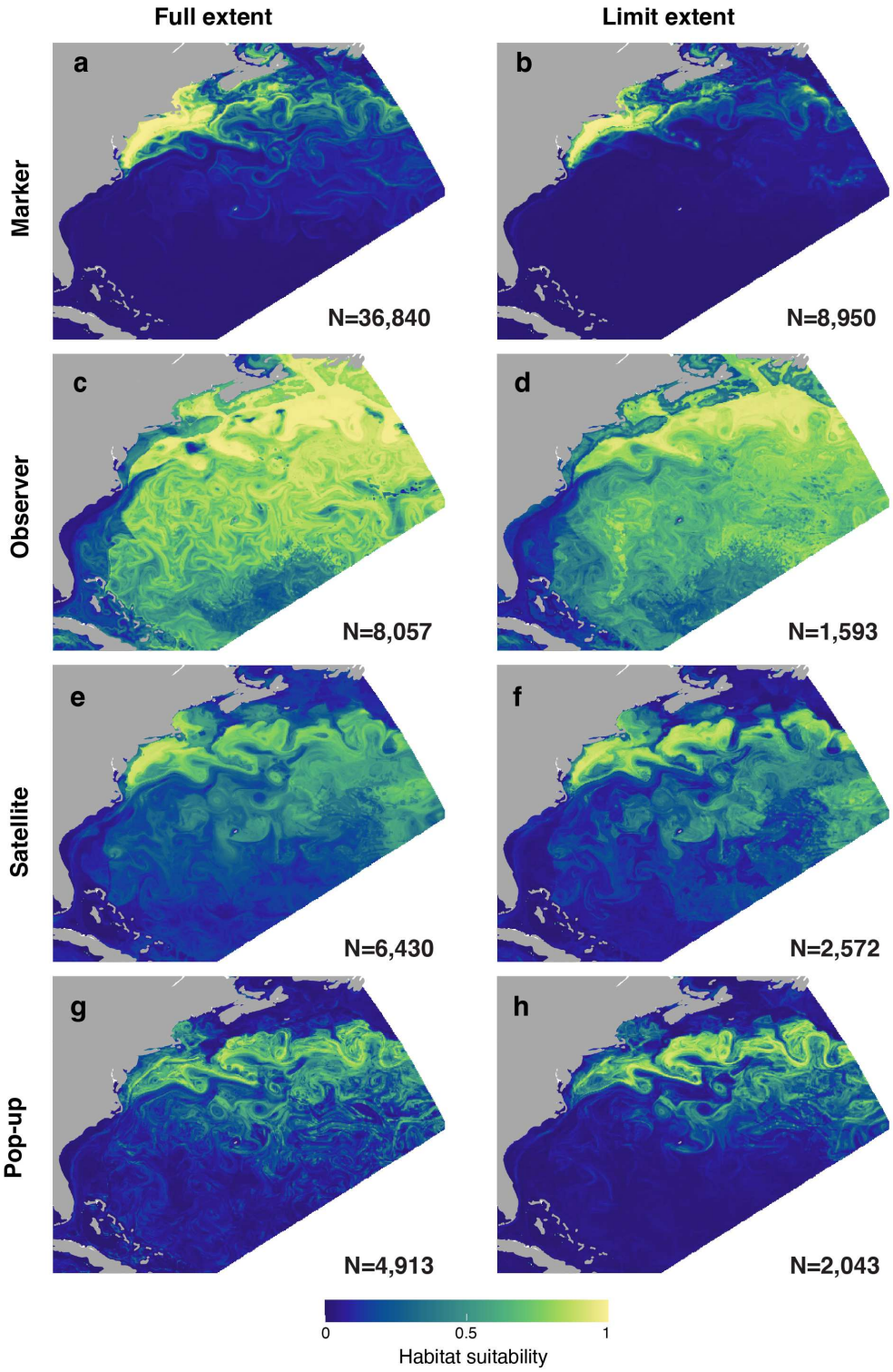
Full extent
Limit extent
Marker
Observer
Satellite
Pop-up
a
b
c
d
e
f
g
h
N=36,840
N=8,950
N=8,057
N=1,593
N=6,430
N=2,572
N=4,913
N=2,043
0
0.5
1
Habitat suitability

Figure 3:

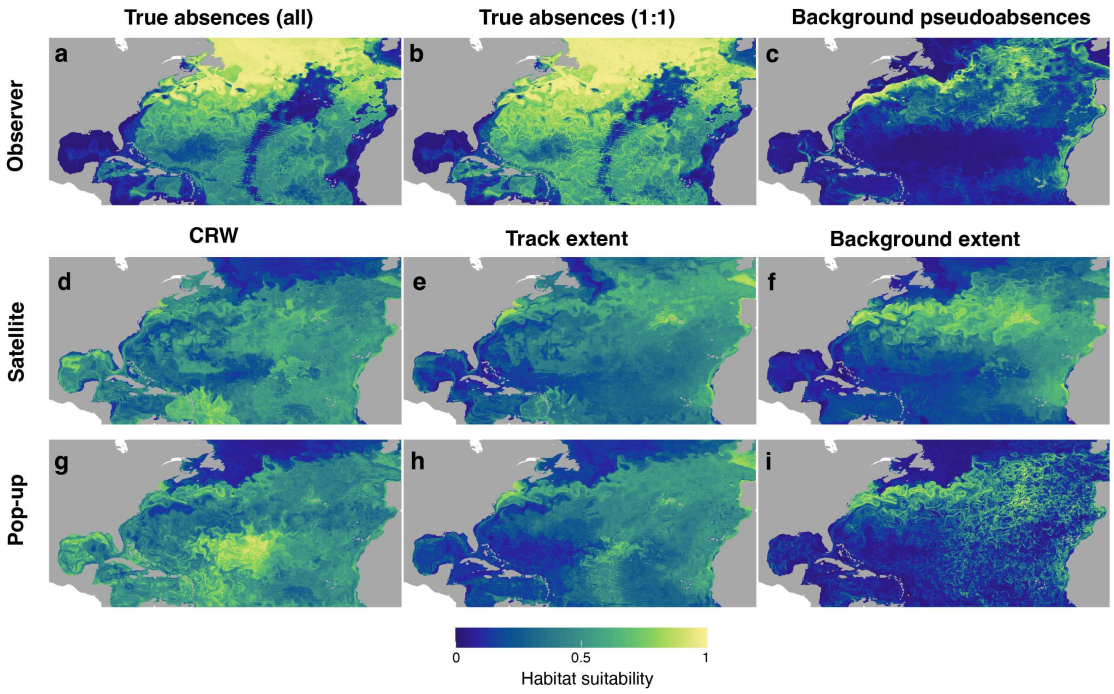
True absences (all)
True absences (1:1)
Background pseudoabsences
Observer
a
b
c
CRW
Track extent
Background extent
Satellite
d
e
f
Pop-up
g
h
i
0
0.5
1
Habitat suitability

Figure 4:

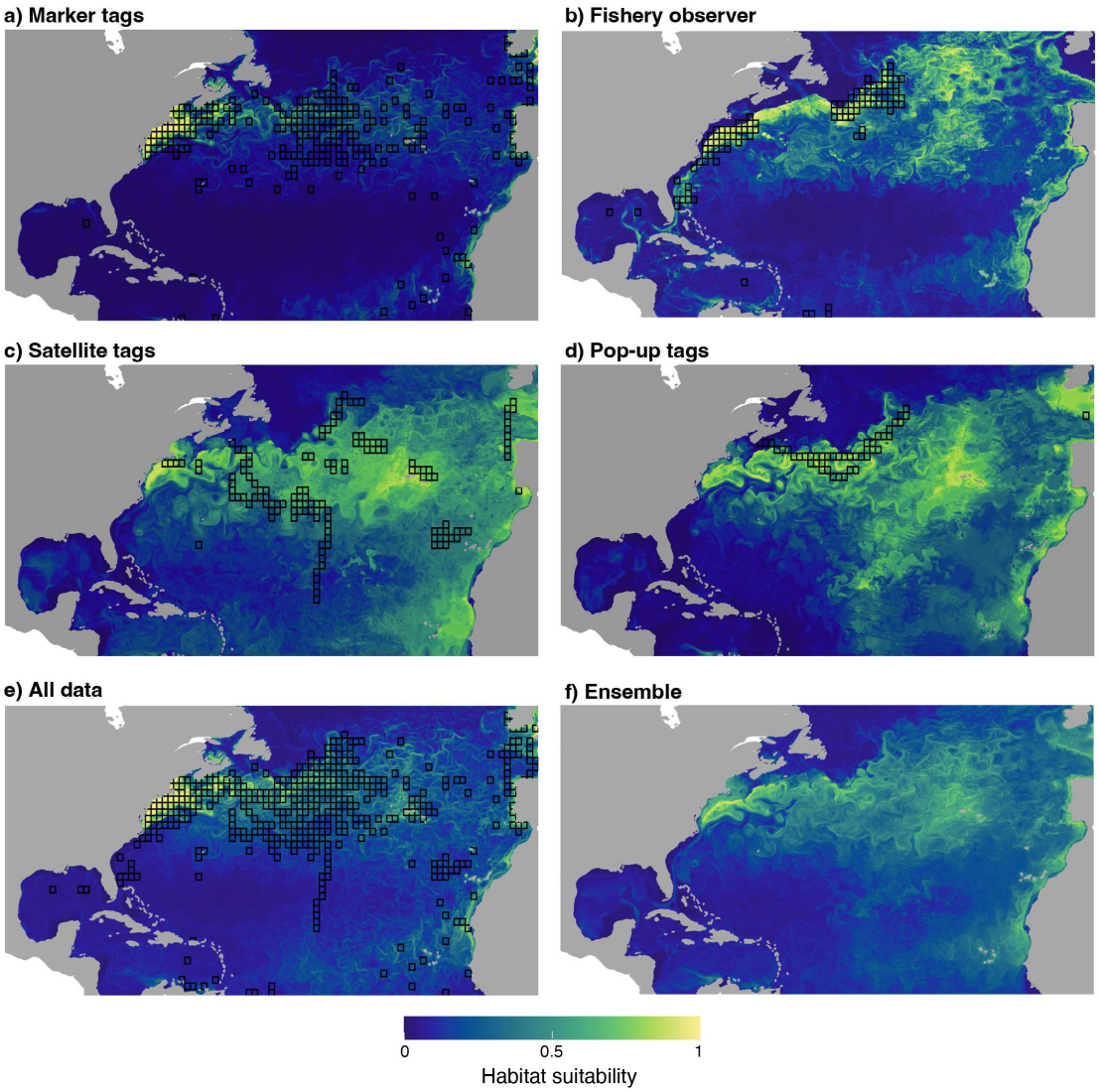
a) Marker tags
b) Fishery observer
c) Satellite tags
d) Pop-up tags
e) All data
f) Ensemble
0
0.5
1
Habitat suitability

Figure 5:

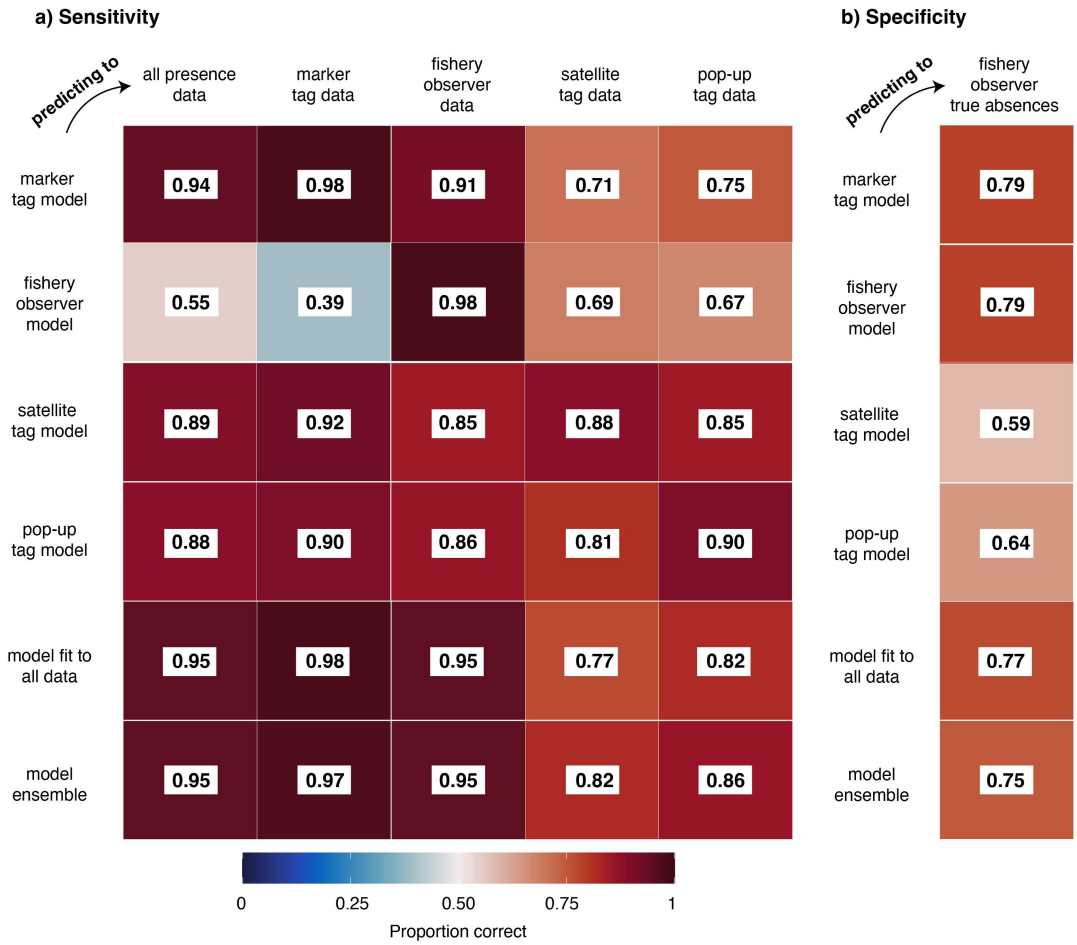
a) Sensitivity
predicting to
all presence data
marker tag data
fishery observer data
satellite tag data
pop-up tag data
marker tag model
fishery observer model
satellite tag model
pop-up tag model
model fit to all data
model ensemble
0.94
0.98
0.91
0.71
0.75
0.55
0.39
0.98
0.69
0.67
0.89
0.92
0.85
0.88
0.85
0.88
0.90
0.86
0.81
0.90
0.95
0.98
0.95
0.77
0.82
0.95
0.97
0.95
0.82
0.86
b) Specificity
predicting to
fishery observer true absences
marker tag model
fishery observer model
satellite tag model
pop-up tag model
model fit to all data
model ensemble
0.79
0.79
0.59
0.64
0.77
0.75
0
0.25
0.50
0.75
1
Proportion correct

Figure 6:

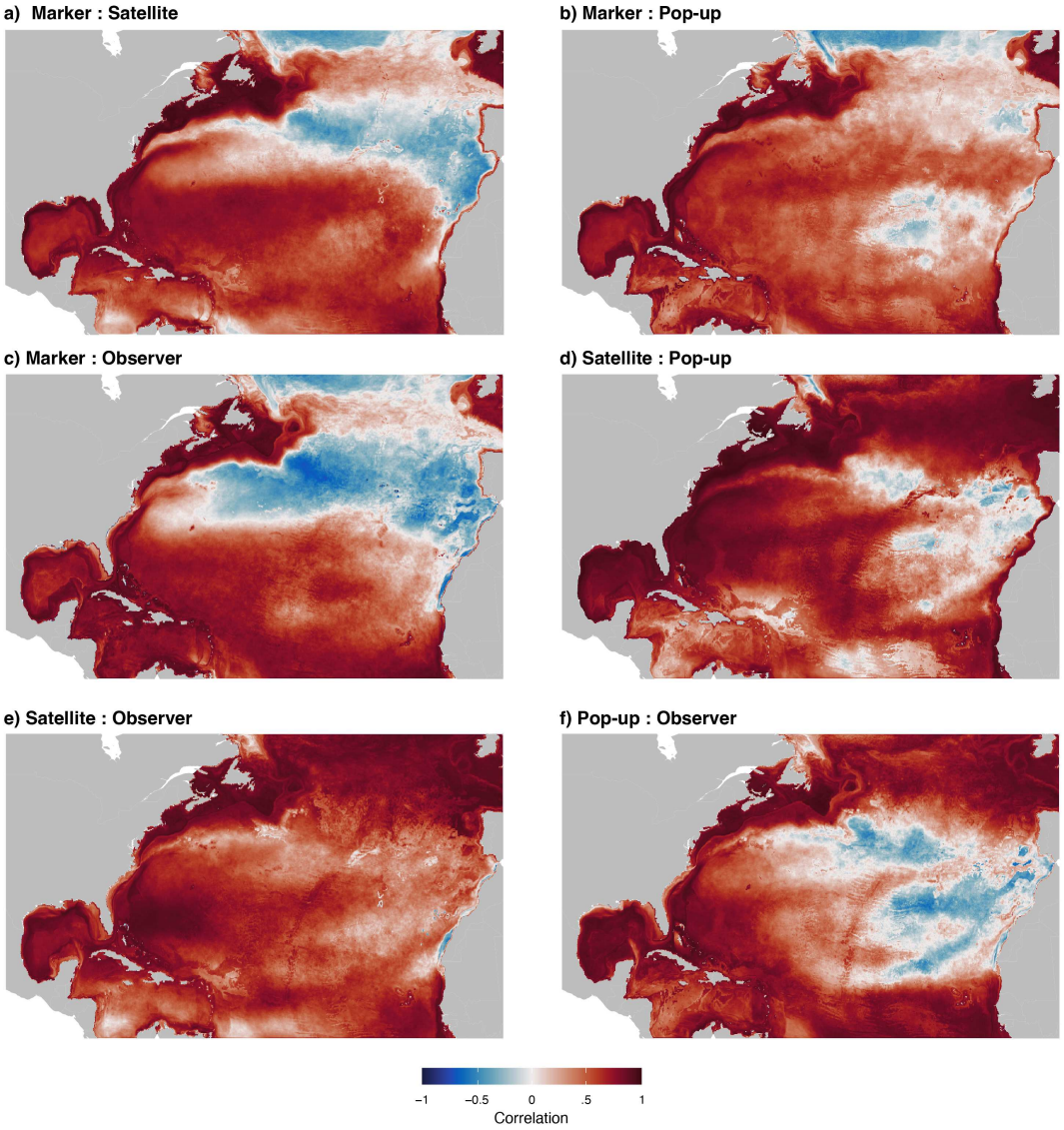
a) Marker : Satellite
b) Marker : Pop-up
c) Marker : Observer
d) Satellite : Pop-up
e) Satellite : Observer
f) Pop-up : Observer
−1
−0.5
0
.5
1
Correlation

Figure 7: