Received: 6 November 2020 | Revised: 13 March 2021 | Accepted: 30 March 2021

DOI: 10.1111/faf.12563

ORIGINAL ARTICLE

FISH and FISHERIES WILEY

Effect of data spatial scale on the performance of fish habitat models

Ismael Núñez-Riboni | Anna Akimova | Anne F. Sell

Thünen Institute of Sea Fisheries, Bremerhaven, Germany

Correspondence
Ismael Núñez-Riboni, Thünen Institute of Sea Fisheries, Herwigstraße 31, 27572 Bremerhaven, Germany.
Email: ismael.nunez-riboni@thuenen.de

Abstract

Habitat models are widely used to explore past and predict future shifts in fish distribution. Our literature review reveals a widespread practice of using in situ data or data with the highest possible resolution to train fish habitat models. Using examples of six fish species at two life stages in the North Sea, we demonstrate that the choice of the data resolution is crucial for a model's performance. We matched fish abundance data from a 51-year long scientific survey at three spatial scales with environmental parameters at seven spatial scales, obtaining a total of 240 data sets. We varied the resolution used for model training and for model predictions and evaluated model performance with various metrics on training and cross-validating data. Contrary to the common notion, training the model with low-resolution data generally improved the performance metrics when compared to models built upon in situ or high-resolution data. The optimal resolution for fish and environmental data was roughly twice the average distance between observations. Training the model with data of higher resolutions often yielded unrealistic fish multidecadal distributional shifts. In turn, best model predictions were achieved with data of higher resolution than the training data. We explain these results with scale-dependent ecological responses, subscale noise in the raw data, failure of interpolation to create information and failure to comply with the Nyquist–Shannon sampling theorem. This study shows that the choice of an appropriate spatial scale is crucial to correctly predict shifts in fish distribution under climate change.

KEYWORDS

climate change, commercially exploited fish species, fish habitat model, Nyquist, Shannon theorem, spatial scale

1 | INTRODUCTION

One of the core goals in ecology is to understand the spatial distribution of organisms in their physical environment, as well as to link changing habitat conditions and the thriving of populations. This understanding is crucial to explain currently observed and to predict future shifts in species distributions in terrestrial and marine ecosystems under a changing climate (e.g. Cheung et al., 2009; Parmesan & Yohe, 2003; Pecl et al., 2017; Pinsky et al., 2020; Root et al., 2003). Habitat or species distribution models are a widely used approach to study distributional shifts of organisms through analysing statistical relations between

species data and environmental parameters describing their habitat.

Species and environmental data used in a habitat model are often not taken simultaneously at the same geographical positions, and their native resolutions may considerably differ both in space and time. Matching such data, that is their transformation to a common grid, is an unavoidable step prior to the habitat model fitting (Guisan et al., 2017). Two matching methods can be used: upsampling (e.g. attempt of increasing the resolution of one or both data sets beyond their native resolution) and downsampling (e.g. reduction in the resolution of the fine data set to the resolution of the coarse data set, or of both data sets to an even coarser resolution).

The role of the matching scale for the performance of statistical habitat models, as well as the choice of the best scale, has been topics of debate over the last 20 years. Some authors have argued that using coarse data in habitat modelling might obscure the correct climate–species relationship (Guisan et al., 2017), reduce model performance or accuracy (Dyer et al., 2013; Ferrier & Watson, 1997) and poorly predict the species' distributional area (Seo et al., 2009). On the contrary, other modellers have claimed that coarse data produce better test statistics (Guisan et al., 2017; Luoto et al., 2007; Rahbek & Graves, 2001; Tobalske, 2002), whereas high-resolution data do not necessarily improve the model fit (Becker et al., 2010; Guisan et al., 2007; Johnson et al., 2002; Mitchell et al., 2001; Núñez-Riboni et al., 2019; Redfern et al., 2008; Thuiller et al., 2005). A third group of authors did not favour one resolution over another but emphasized that the scale of the data aggregation in a habitat model should reflect the characteristic scale of the environmental and biological processes of interest (e.g. Bellier et al., 2010; Kärcher et al., 2019; de Knegt et al., 2010; Pearson et al., 2004; Redfern et al., 2006).

Yet, as to date, there have been only few attempts to quantify the effect of the data matching scale in a systematic way, although such studies could advise in the choice of the best options when setting up a habitat model. Furthermore, previously published habitat models have largely ignored (with few exceptions like Thuiller et al., 2005) the possibility that the resolution of the data used to train or fit a habitat model (further on "training data set") does not necessarily have to be the same as the resolution of the data used for habitat predictions (further on "predicting data set").

To our best knowledge, the role of the scale in statistical habitat modelling has been mainly discussed in the terrestrial ecology (e.g. Elith & Leathwick, 2009), with fewer examples in the marine environment (e.g. Becker et al., 2010; França & Cabral, 2016; Redfern et al., 2008). However, there are several reasons to specifically study the effect of scale in marine ecosystems. Statistical habitat models are widely used in the marine environment (e.g. a review paper of Melo-Merino et al., 2020) to study a broad range of processes including climate-driven changes in species distribution. The field observations in marine realm are typically scarce with native data resolutions being up to two orders of magnitude coarser (Guisan et al., 2017) and data uncertainties being considerably larger (Elith & Leathwick, 2009) in comparison with the terrestrial environment. Furthermore, it has been claimed that species' local extinction in marine realm due to climate change happens with at least the same (Webb & Mindel, 2015) or even higher rates than in terrestrial ecosystems (Pinsky et al., 2019).

1 INTRODUCTION	955
2 DATA AND METHODS	957
2.1 Fish abundance and environmental data	957
2.2 Construction of data sets at various levels of resolution	959
2.3 Habitat model	960
2.4 Assessment of performance and realism of habitat model	960
3 RESULTS	961
3.1 Model performance	961
3.2 Realism of model results	962
4 DISCUSSION	962
4.1 Effect of scale on the performance of the fish habitat model	962
4.2. Reasons behind the improved performance with downsampled data	963
4.2.1 Scale-dependent ecological response	965
4.2.2 Noise	966
4.2.3 Failure of interpolation to create information	966
4.2.4 Sampling resolution and Nyquist–Shannon theorem	967
4.3. Spatial autocorrelation and pseudoreplication	968
4.4 Temperature response function	968
4.5 Precautionary remark for scientific survey design	970
5 CONCLUSIONS	970
ACKNOWLEDGEMENTS	970
DATA AVAILABILITY STATEMENT	970
REFERENCES	971
SUPPORTING INFORMATION	973

The importance of scale in the marine environment is widely acknowledged and is intrinsically anchored on the geophysical nature of the ocean, where different mechanisms drive variability at different spatial and temporal scales. For instance, tides and eddies exert their effects at scales of days and few tens of kilometres, while interannual modes of variability like the North Atlantic Oscillation act at scales of several years and hundreds of kilometres (Stommel, 1963). Climate change affect oceanic ecosystems at even larger scales of centuries and thousands of kilometres (Dickey, 2003). The response of individual organisms and their entire populations to their environment is scale-dependent as well, and involves biological mechanisms (e.g. physiological response, behaviour, acclimatization and colonization), which vary across a wide range of spatial and temporal scales (Wiens, 1989 and, e.g. figure 2 in Pinsky et al., 2020).

In the particular case of spatial distribution of marine fishes, some mechanisms can be important at relatively small spatial and temporal scales, like schooling, avoidance of fishing gear, direct prey–predator interactions, eddies and tides. Other processes drive fish distribution at larger and longer scales, like distribution of water masses with physiologically optimal properties, oceanic currents and geographical distribution of prey and predators (as opposed to their

likelihood of local encounter). This influence of scale could completely change the outcome of a fish habitat model depending on the resolution of the training data.

According to our review of 43 recently published fish habitat models, in situ environmental and in situ fish data are a common choice for training data sets (42% of the reviewed studies, Table 1). When in situ data are lacking, gridded environmental data (e.g. meteorological or oceanographic data products, hydrodynamic models and satellite data) are usually spatially interpolated over the positions of fish hauls or both fish and environmental data sets are upsampled to a resolution of the finer data set or even finer (44% of all studies). Only few reviewed studies (19%) apply downsampling. In further 14% of the studies, the data matching method is not clearly described and rather treated as an unimportant detail of the analysis.

In total, in situ data and upsampling were used in 86% of the studies, while addressing processes at a wide range of scale from seasonal variability to climate change (see "main focus" in Table 1). Therefore, our literature review points out a general practice of using the highest possible resolution in fish habitat modelling, independently of the scale of the investigated processes. Seemingly, many fishery scientists rely on the habitat model to disentangle the interrelation between scale and ecological response, independently of the scale of the input data. We argue here that this notion is unfounded, since to our knowledge no previous study has systematically examined the role of the matching scale in fish habitat modelling or has demonstrated that in situ data outperforms gridded data in fish habitat models, particularly with a focus on large-scale distributional shifts due to climate change.

Our study aimed at demonstrating that matching of environmental and fish data is not an unimportant step in the model design but, on the contrary, choice of the correct matching scale is fundamental to unveil the correct relations between fish and environment, whereas the wrong scale can mask them. We followed an approach put forward by Guisan et al. (2007) and examined the effects of scale of both training and predicting data sets on the model performance. Our study builds on a few similar analyses (like those quoted above) and is, to our knowledge, the most comprehensive study of the effects of scale in habitat modelling in fishery science. In particular, we show how in situ and high-resolution data are, in comparison with coarse data, counterproductive as training data sets for studies about the influence of climate change on shifts of fish distribution: At this "climate scale" (i.e. time periods of several decades and spatial scales of several hundreds to thousands of kilometres), the detail provided by high-resolution data is not only unnecessary but even impairs model performance. We discuss our findings to the light of scale-dependent ecological response and signal processing theory.

TABLE 1 Number and proportion (%) of 43 reviewed habitat modelling studies categorized by four criteria: (i) method used to match environmental and fish data; (ii) main focus of the study (targeted scales); (iii) environmental covariates included in the habitat model; and iv) type of model

Criteria/Categories		
Method of matching environmental and biotic data	**Main focus**	**Number of studies (%)**
In situ data	Climate	7 (16)
	Interannual	2 (5)
	Seasonal	4 (9)
	Other	5 (12)
	Total	18 (42)
Downsampling	Climate	5 (12)
	Interannual	1 (2)
	Seasonal	1 (2)
	Other	1 (2)
	Total	8 (19)
Upsampling	Climate	8 (19)
	Interannual	2 (5)
	Seasonal	2 (5)
	Other	7 (16)
	Total	19 (44)
Not clearly described	Climate	5 (12)
	Interannual	0 (0)
	Seasonal	0 (0)
	Other	1 (2)
	Total	6 (14)
Environmental covariates included in the analyses		
Temperature		42 (98)
Depth		23 (53)
Salinity		11 (23)
Sediment type		5 (12)
Statistical method		
Generalized additive models (GAM)		30 (70)
Generalized linear models (GLM)		13 (30)
Maximal Entropy (MaxEnt)		5 (12)
Random Forest (RF)		5 (12)

Note: The method categories "in situ data," "upsampling" and "downsampling" are described in the text. The focus categories include "climate" (records of more than 20 years and/or climate projections), "interannual," "seasonal" and "other" (studies aiming at understanding other aspects than the temporal variability like spatial patterns or model performance). Criteria and categories in the table are not mutually exclusive. See Section S1 in Supplement for further details.

2 | DATA AND METHODS

2.1 | Fish abundance and environmental data

Data of fish abundance (catch per unit effort; CPUE) were collected during the North Sea International Bottom Trawl Survey (NS-IBTS) in the first quarter (Q1) of each year over five decades (from 1967 to 2017). The study area and an example of the data distribution are shown in Figure 1 (upper left insert). The data were obtained from the Database of Trawl Surveys (DATRAS, 2020) of the International

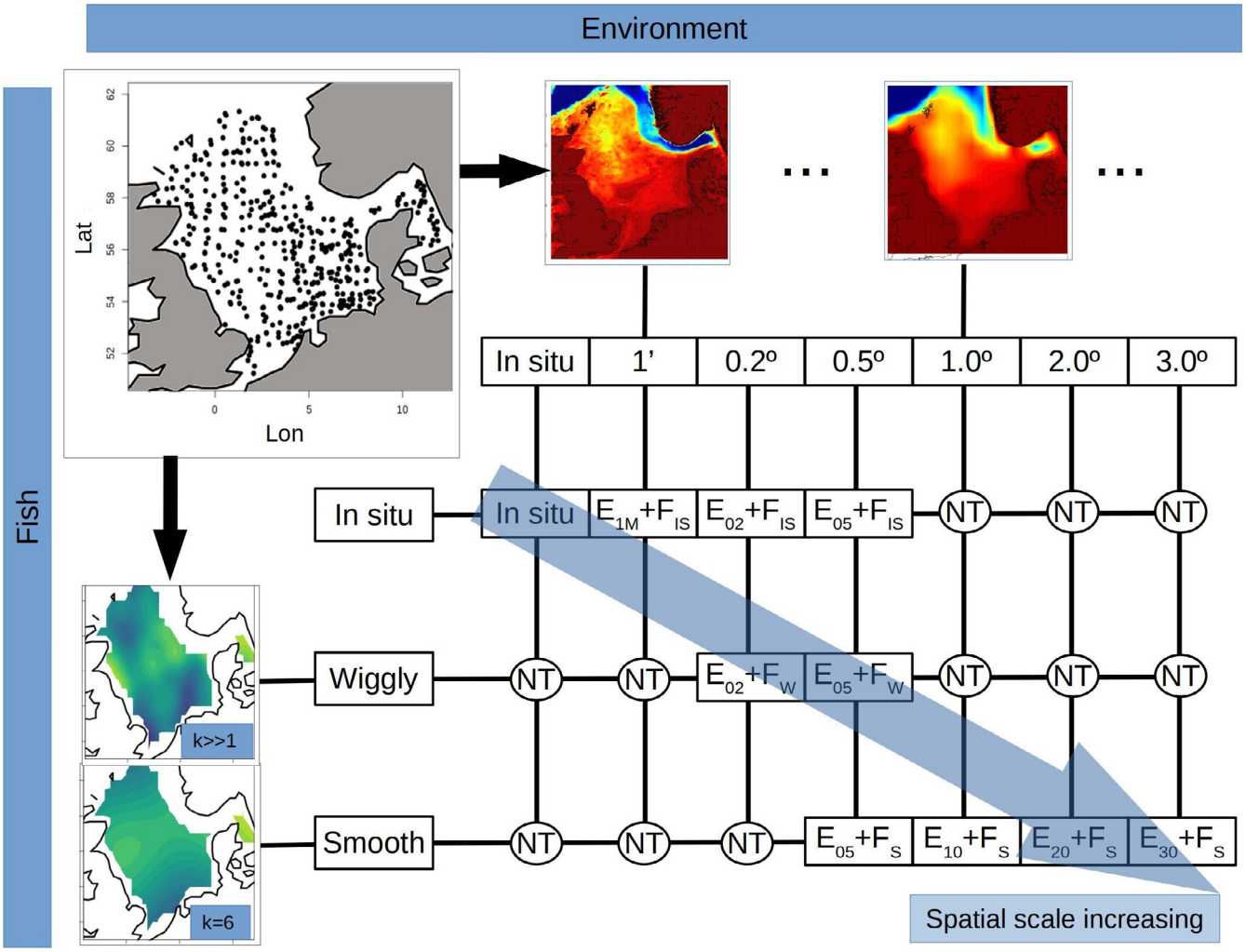

FIGURE 1 Diagram summarizing the matching of fish and environmental data at different spatial scales and the resulting data sets used in the present study for Q1 temperature. The case of Q3 temperatures is identical, but excluding the in situ environmental data, because these data were not recorded simultaneously with the fish abundance from Q1 (see text). Examples of distribution of in situ data (upper left insert) as well as of maps of bathymetry (top inserts) and fish abundance (left inserts) at two scales are shown. "NT" refers to combinations of spatial scales "not tested" in this study

Council for the Exploration of the Sea (ICES). All roundfish areas, ships and gears were included. While we are aware of methodological changes of the survey, which have occurred within this time frame (Annex 2 in ICES, 2015), we accounted for abundance variations over time due to natural or artificial reasons with a year-effect term in the habitat model (see Section 2.3 below). Q1 was chosen for being the fisheries survey with the longest uninterrupted time series in the North Sea and, thus, suitable also for climate studies. In contrast, the summer NS-IBTS started only in 1991 and data records shorter than four decades are inappropriate to detect the full climate signal in the North Sea (see for instance Henson et al., 2017, their Figure 1e). The NS-IBTS fisheries data were sampled following a grid of ICES rectangles with a resolution of 0.5° latitude by 1° longitude, equalling roughly 30 x 30 nautical miles. In each of these grid cells, typically two fisheries hauls per survey season are performed at random positions. We chose the six most abundant demersal and benthopelagic fish species (Table 2) and modelled their abundance separately for two life stages, "juvenile" (immature fish) and "adults" (mature fish) based on their length (FishMap, 2006; Heessen et al., 2015). This yielded a total of 12 "Species at Life Stages" (SLS, further on).

We used temperature and depth as the only environmental covariates, as they have been previously identified as the most important environmental factors influencing fish habitats (98% and 53% of studies in our literature review, correspondingly; Table 1). In situ temperature were measured simultaneously to the NS-IBTS fishery hauls by a Conductivity–Temperature–Depth (CTD) profiler. For the analysed fish species, bottom temperature was chosen since this appears to be the most logical choice for demersal and benthopelagic species and has been widely used before (e.g. Perry et al., 2005; Punzón et al., 2021). In situ bathymetry observations were recorded with the on-board echo sounder. Gridded bathymetry was taken from the 1-min ETOPO1 (Amante & Eakins, 2009). Gridded bottom temperature for 1967–2017 was obtained from a recent run of the Adjusted Hydrography Optimal Interpolation (AHOI; Núñez-Riboni

TABLE 2 Fish species and life stages ("species at life stages" or "SLS") used in the present study

Fish species	Common name	Habitat Type	Length	
			Juvenile	Adults
Eutrigla gurnardus	Grey gurnard	Demersal	<20 cm	≥20 cm
Gadus morhua	Cod	Benthopelagic	<40 cm	≥40 cm
Limanda limanda	Comon dab	Demersal	<20 cm	≥20 cm
Melanogrammus aeglefinus	Haddock	Benthopelagic	<30 cm	≥ 30 cm
Merlangius merlangus	Whiting	Benthopelagic	<20 cm	≥20 cm
Trisopterus esmarkii	Norway pout	Benthopelagic	<15 cm	≥15 cm

& Akimova, 2015). AHOI temperature has a native resolution of 0.2° and is mainly based on field observations, including those taken during the NS-IBTS. We used winter (Q1, mean from February to March) and summer (Q3, mean from July to September) temperatures in the analysis. The later was included to test for possible lagged responses of fish species to the conditions from the previous summer, similar to Pinsky et al., (2019).

2.2 | Construction of data sets at various levels of resolution

Training data sets for the habitat model were constructed with fish abundance and environmental data matched at different spatial scales. Please refer to Figure 1 for a summarizing diagram of this procedure. The most obvious choice was to combine in situ fish abundance with the simultaneously taken in situ environmental data, where no change of scale is necessary. As we mentioned above, this is one of the most popular approaches in habitat modelling (42% of the reviewed papers; Table 1). This can be regarded as data matching at the smallest possible scale (or highest resolution) because environmental and fishery data are taken nearly simultaneously and on scales of only few kilometres (the trawling distance).

To match in situ fish abundance data with gridded environmental data at different scales, the AHOI temperatures in Q1 and Q3 and ETOPO bathymetry were downsampled to regular grids of resolution L following these two steps: 1) Low pass filter: Every gridded data point Γ has been replaced by the weighted average of surrounding data inside radii L = 0.2°, 0.5°, 1.0°, 2.0° and 3.0°. The averaging weights were constructed with a Gaussian function depending only on the distance to Γ. 2) Decimation: Data within radii L were removed to leave only the data point Γ. This measure prevents potential data pseudoreplication as discussed in detail in Section 4.3 below.

These downsampled environmental data were then spatially interpolated over the fish haul positions and sampling times. Such interpolations are a popular upsampling approach to match environmental and fishery data in habitat models. The respective training data sets will be herewith denoted as $E_L + F_{IS}$ where E and F refer to "environment" and "fish abundance," respectively. The subscript "IS" represents in situ data, and L indicates the resolution of the downsampled environmental data (e.g. $E_{05} + F_{IS}$ labels the combination of environmental data downsampled to 0.5° and matched to in situ fish abundance, see Figure 1). Both temperature and bathymetry data represent the same scale in all data sets except $E_{1\ M} + F_{IS}$, where bathymetry and temperature were interpolated from data with their native resolutions of 1' and 0.2°, respectively.

Contrary to an in situ observation, a gridded datum reflects the oceanic conditions not at a point but within a cell area (in case from AHOI, for instance, ca. 200 km^2). Therefore, we further constructed data sets where both fish abundance and environmental data were gridded on the same regular grid of the various spatial resolutions L mentioned above. Fish abundance data in individual years were gridded using a generalized additive model (GAM; Hastie & Tibshirani, 1986) with a Tweedie distribution (Augustin et al., 2013; Tweedie, 1984) and log link function:

$$\log(\hat{y}) = s_M(lon, lat), \quad (1)$$

where $\hat{y}$ is the gridded abundance, *lon* is longitude and *lat* is latitude and $s_M(lon,lat)$ is a thin plate spline smoother (Wood, 2017), which is an optimal two-dimensional smooth representation of the observed fish abundance.

The spline smoother s_M is the sum of k basis splines, and its complexity (i.e. in the number of knots) increases with k. The larger the basis dimension k, the more complex or wiggly s_M is and *vice versa*. We varied k to control the smoothness of the maps of fish abundance, obtaining coarser data sets analogous to the downsampled environmental data. We chose two smoothness levels: k as low and as large as practically possible. The lowest possible k was $k = 6$. The largest k was constrained by the number of fish abundance data M in each year and computing time. We chose $k = M/5$ as a good trade-off between computation time and smoother complexity. These two smoothness or scale levels of the obtained maps of fish abundance will be denoted here as F_w ("wiggly," k»1) and F_s ("smooth," k = 6; Figure 1).

Gridded fish abundance and environmental data sets were matched only in a geographical region with sufficient abundance

estimates defined by a convex hull calculated with a Delaunay triangulation (Swan & Sandilands, 1995). For simplicity, we matched fish and environmental data only at roughly similar spatial scales (Figure 1). Spectral analysis of the gridded fish abundance data sets (no figure shown) indicated that the smooth (coarser) maps F_S contain large amounts of variability on scales of 1 to 2°. Thus, we excluded some of the possible combinations of *E* and *F* from our analysis and matched smooth abundance data sets F_S only to environmental data sets with large filter scale *L* (1.0° to 3.0°), and wiggly data sets F_W to the *E* data sets with smaller filter scales *L* (0.2° and 0.5°). In total, we analysed 240 data sets (10 matching scales ×12 SLS ×12 annual quarters for temperature).

2.3 | Habitat model

We modelled fish habitat with a generalized regression model, a statistical method widely used for habitat modelling (e.g. Guisan et al., 2017; Table 1). Specifically, a GAM with Tweedie distribution and logarithmic link was fitted to each data set described above (Figure 1):

$$\log(\hat{y}) = \alpha_0 + \alpha_1 \cdot T + \alpha_2 \cdot T^2 + \beta_1 \cdot B + s_R(lon, lat) + s_t(year), \quad (2)$$

where α_0, α_1, α_2 and β_1 are model parameter to be determined, $\hat{y}$ is modelled fish abundance, *T* is bottom temperature (either from Q1 or Q3 but not both simultaneously since they are strongly co-linear) and *B* is bathymetry. The temperature response function was intentionally modelled as a second-order polynomial to obtain (in combination with the logarithm) a unimodal response curve (a Gaussian bell). A justification for this choice over the more popular penalized smoothers is given in Section 4.4 below.

The smoother s_t deals with year-to-year variations of the total abundance of a SLS. This smoother is unpenalized and has a basis dimension *k* equal to the length of the time series to reproduce the stock's annual variations with maximum flexibility and the smallest temporal autocorrelation. The smoother s_R is a random effect for the geographical position called Gaussian process smooth (Kammann & Wand, 2003; Wood, 2017), playing a double role in our model. On the one hand, it accounts for "geographical attachment" (Planque et al., 2011), that is for the relationship between fish abundance and factors other than those explicitly modelled, like salinity, the distribution of prey and predators, etc. On the other hand, s_R is used to account for the spatial autocorrelation in fish distribution due to intrinsic factors like aggregation and dispersal (Beale et al., 2010). A good value for the spatial autocorrelation of s_R was found by trial and error to be 2°. This means that variations of abundance at distances smaller than 2°, which are not fully explained by variations of the environmental variables, were explained by s_R. The basis dimension k was equal to the number of data points in each training data set if it was less than 150 and limited to $k = 150$ otherwise.

Some SLS could possibly be modelled better with slightly different models and/or covariates. However, the same model configuration was intentionally applied to all SLS, that is the same structure of response functions and covariates. This choice is similar to Hawkins et al., 2007 and is further motivated by de Knegt et al., (2010), who show with synthetic data that if a habitat model addresses the wrong scale, it behaves similar to a model in which an important covariate is omitted (fitting the data poorly). Therefore, if using different amount (or combination) of covariates or different degrees of freedom with models trained at different scales can potentially mask model misspecification due to scale, keeping the model unchanged over all scales seems the appropriate way of isolating the effect of scale on model performance. Nevertheless, please note that model parameters were fitted for each SLS separately.

2.4 | Assessment of performance and realism of habitat model

The model performance at different matching scales was assessed by a threefold cross-validation. The $F_{IS} + E_{IS}$ data sets were split by individual hauls into three subsets, each containing randomly selected 1/3 of the original data per SLS. Environmental and abundance data of two subsets were used to train the model (Equation 2). Values of environmental data in the third subset were used to predict fish abundances. The training $F_L + E_{IS}$ data sets were downsampled and matched to gridded environmental data at scale *L* to construct training $F_W + E_L$ and $F_S + E_L$ data sets as described in Section 2.2. All combinations of training and predicting scales *L* were tested. Modelled estimates of fish abundance were compared to the corresponding in situ observations of the third data set using the total residual deviance *D* for Tweedie distribution (appendix 2 of Candy, 2004; equation 7 of Shono, 2008):

$$D = 2 \cdot \sum_{i=1}^{N} d_i = 2 \cdot \sum_{i=1}^{N} \frac{\left(y_i^{(2-p)} - (2-p) \cdot y_i \cdot \hat{y}_i^{(1-p)} + (1-p) \cdot \hat{y}_i^{(2-p)}\right)}{(1-p) \cdot (2-p)}, \quad (3)$$

where y_i is the abundance observations, $\hat{y}_i$ is their model estimates, *N* is the size of the validating subset and *p* is the "power parameter" of the Tweedie distribution ($1 < p < 2$). Note that all validating sets have the same amount of data *N* throughout all scales and SLS (=1/3 of the total in situ data). *D* is a reasonable metric to evaluate model performance because it is a generalization of the residual sum of squares for generalized linear and additive models (McCullagh & Nelder, 1989), that is it correctly deals with heteroscedasticity and is nearly normal in spite of the non-normality of the data. The procedure was repeated with all three subsets, and *D* was averaged over the three realizations.

Although the cross-validation is the most objective method to assess model performance because it uses independent data, we have additionally calculated metrics on the data sets used to train the habitat model. These metrics are described in Section S2 and provide additional information about the effect of the data downsampling on some model characteristics like signal-to-noise ratio and parameter errors.

Because not always the most realistic model yields the best metrics (Burnham & Anderson, 1998), we further evaluated the realism of the habitat model based on two criteria. The first criterion was the behaviour of the modelled temperature response function, following notions from Elith and Leathwick (2009). A realistic habitat model should reflect reasonable response of fish abundance to temperature variations. The temperature response curve was considered realistic only if not inverted (growing to infinity), but instead with a local maximum within the known temperature range for the modelled species (see Section S3 for details). Similar arguments have been used by Burnham and Anderson (1998) to select more realistic models over the models with optimal values of statistical metrics like the Akaike information criterion (AIC). In analogy to Núñez-Riboni et al., (2019), we calculated the partial effect of temperature by averaging all model terms of Equation 2, excluding temperature:

$$\log(\hat{y}_T) = A_0 + \alpha_1 \cdot T + \alpha_2 \cdot T^2, \tag{4}$$

where

$$A_0 = \alpha_0 + \beta_1 \cdot -\hat{B} + -\hat{s}_R(lon, lat) + -\hat{s}_t(year)$$

and the overbars denote averaging over all observed values. $\hat{y}_T$ in Equation 4 is called herewith the "temperature curve" because it represents changes of abundance (or habitat) as a function of temperature alone.

The second criterion to evaluate model realism was its ability to reproduce the observed changes in fish habitat at climate scale. Habitat suitability is understood here as the occupancy resulting from all factors which influence the local abundance of fish, including the two environmental parameters used in the model. We estimated changes in the observed habitat suitability H_O by first calculating the median fish abundances within distances of 500 km centred on grid points of the AHOI grid (0.2° × 0.2°). Such large radius was needed to isolate large-scale climate-related fish preference while ignoring small-scale ecological and environmental processes. The obtained abundance maps were then scaled with (1) the median annual abundance to deal with interannual changes of the fish biomass and (2) with the overall historical maximum (98th percentile) to scale the habitat suitability H_O between 0 and 100.

Changes in the modelled fish habitat H_M, were estimated with a method similar to Núñez-Riboni et al., (2019): relocation of fish $\hat{y}_T(lon, lat, T)$ due to temperature alone was isolated by averaging the population size effect $s_t(year)$ in Equation 2. This modelled abundance was transformed into an estimate of habitat suitability H_M by scaling $\hat{y}_T(lon, lat, T)$ with the historical maximum (also 98th percentile):

$$H_M(lon, lat, T) = \frac{100 \cdot \hat{y}_T(lon, lat, T)}{max(\hat{y})}. \tag{5}$$

We acknowledge that marine fishes change their spatial distribution due to factors other than temperature. However, in agreement with a solid body of literature on climate change and marine ecosystems (IPCC, 2014), we assume that the distributional shifts at climate scale are fundamentally related to temperature. Therefore, we expect to evaluate the model ability to reproduce climate-induced shifts by comparing the differences between H_M and H_o. We estimated these differences for each data set using two metrics:

(1) The percentage Ω of all grid cells where local differences ΔH_M and ΔH_O matched between the periods 1970–1980 and 2007–2017:

$$\Omega\left(lon_i, lat_i\right) = \frac{100}{Z} \sum_{i=1}^{i=z} \delta\left(lon_i, lat_i\right), \tag{6}$$

where

$\delta(lon_i, lat_i) = sign(H_M(lon_i, lat_i)) \cdot sign(H_O(lon_i, lat_i))$, Z the number of grid points and only $\delta(lon_i, lat_i) > 0$ considered in the sum.

(2) The median absolute deviation (MAD) between the observed and modelled habitats:

$$MAD = median(|H_M(lon_i, lat_i) - H_O(lon_i, lat_i)|), \tag{7}$$

which is a metric independent of data distribution (Pham-Gia & Hung, 2001).

Finally, model output was considered realistic only when $\Omega > 65\%$ and MAD < MAD, where MAD was the average over all SLS.

A summary of our complete analysis is shown as a flow chart with cross-references to this section in Figure S1 of the Supplement.

3 | RESULTS

3.1 | Model performance

In the case of Q1 temperature data, the minimum residual deviance D between the observed and modelled fish abundances was obtained with the training data set $E_{10} + F_S$ for most of the SLS (8 out of 12; Figure 2a, left column), suggesting that environmental data downsampled to 1.0° and smoothed fish abundance data are the best training data sets for Q1 temperature. In situ data yielded minimum D only in 2 of the 12 SLS test cases. The results obtained with Q3 temperature were similar, with 7 of 12 SLS showing minimum D with $E_{10} + F_S$ (Figure 2b, left column). For both Q1 and Q3, D decreased strongly from the scale of 0.5° to 1.0°: data sets with resolutions of 0.5° or smaller resulted in an average D of roughly e^{40}, while data sets with resolutions of 1.0° and coarser resulted in an average D of roughly e^{9}.

In contrast to the training data sets, the best model performance was achieved with the predicting data sets at the smallest possible scale: E_{IS} for Q1 and $E_{1\,M}$ for Q3 yielded minimum D in 11 of 12 SLS (Figure 2a,b, right columns). Overall minimum D was obtained when training the model with $E_{10} + F_S$ and predicting with E_{IS} for Q1 ($e^{5.32}$; Table 3) and when training with $E_{20} + F_S$ and predicting with $E_{1\,M}$ for Q3 ($e^{5.19}$; Table 4). The choice of resolution in the training data set had a much stronger effect on model performance than the resolution of the predicting data set: Note that the colour gradients in Tables 3 and 4 are stronger between rows than between columns.

3.2 | Realism of model results

The realism of the habitat model is exemplified here with two SLS, juvenile Atlantic cod (Gadus morhua, Gadidae) and adult grey gurnard (Eutrigla gurnardus, Triglidae), which showed distinctly different results depending on the scale of the training data sets. In the northern North Sea, the observed habitat suitability H_O (occupancy) has increased for both SLS during the past five decades, whereas in the southern North Sea, suitability has decreased for cod (Figure 3a) and remained almost unchanged for grey gurnard (Figure 3b). These changes correspond to a northward displacement of biomass distribution for cod and an increase of biomass in the northern North Sea for grey gurnard (possibly also a northward displacement).

Figure 4a,b shows the respective modelled changes of the habitat suitability H_M for these two SLS when using in situ (for juvenile cod) and small-scale data (for adult grey gurnard). The habitat model trained with these data sets suggests an unchanged or increased habitat suitability for cod everywhere in the North Sea. For grey gurnard, modelled suitability decreased everywhere in the North Sea, except the region at the flanks of the Dogger Bank, where suitability increased. These results differ considerably from the observed changes in the abundance of both fish species (Figure 3). Figure 4c,d shows the underlying temperature curves corresponding to the model output from Figure 4a,b. Training the model with in situ or small-scale data yielded unrealistic temperature curves for both SLS, with suitability growing unbounded to infinity as temperature approaches the edges of the observed temperature range.

In contrast to the model fitted with the small-scale training data, training the model with large-scale data (E_{10} + F_S with Q3 temperature for both SLS) yielded a modelled habitat suitability similar to the observed one (Figure 5): in the northern North Sea suitability increases for both SLS and in the southern North Sea suitability decreases for cod (Figure 5a) and remains nearly unchanged for grey gurnard (Figure 5b). These changes indicate an increase of biomass in the northern North Sea for both SLS, similar to Figure 3. This visual match is supported by a larger percentage of overlapping changes Ω (Equation 6) and smaller differences (MAD; Equation 7) between the observed and the modelled changes in the habitat suitability. Ω increased from 56.7% to 88.6% for cod and from 13.5% to 65.5% for grey gurnard when downsampled training data sets were used instead of the small-scale ones (Table S5). MAD decreased from 4.4 to 3.3 for cod and from 28.0 to 20.8 for grey gurnard (Table S6). Furthermore, reasonable, bounded temperature curves (Figure 5c,d) were obtained with the downsampled training data sets. The maximal abundance was predicted at temperatures within the observed range: roughly 11.4°C for cod and 13.1°C for grey gurnard (Table S4).

Although not all tested SLS presented such strong effect of scale of the training data on the model realism, the habitat models trained with large-scale data generally yielded more realistic changes in the modelled habitat suitability at climate scale. This is indicated by the number of model outputs complying our three realism criteria (Table 5): E_{10} + F_S (see Figure 1) yielded realistic model output for nine SLS (seven with Q3 and 2 with Q1), whereas small-scale data sets only for four SLS (two with Q1 E_{IS} + F_{IS}, two with Q3 $E_{1\,M}$ + F_{IS}).

4 | DISCUSSION

4.1 | Effect of scale on the performance of the fish habitat model

In situ data have often been regarded as the first choice for fitting a habitat model, even for climate studies (e.g. Kirkman et al., 2013; Pinsky et al., 2013), whereas aggregated or gridded data appear to be regarded as deficient. In the absence of in situ data, the majority of modellers favour the use of high-resolution (or small-scale) gridded environmental data and often upsample (interpolate) them over the positions of fish

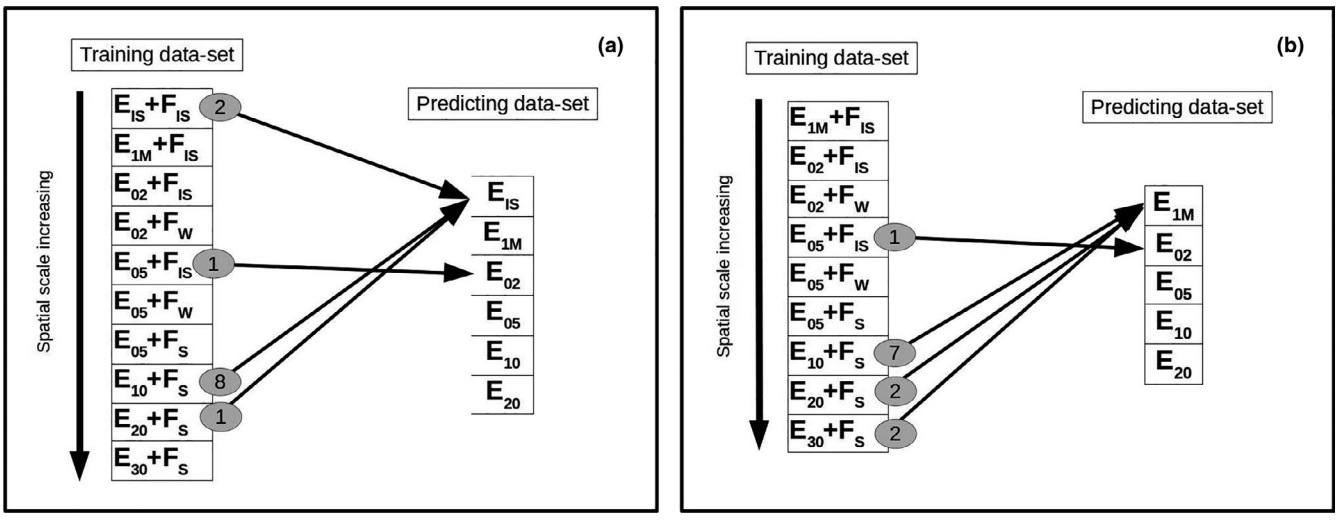

FIGURE 2 Results of the threefold cross-validation for the case of Q1 temperature (panel a; including in situ temperature) and Q3 temperature (panel b). Numbers in the ellipses indicate the number of SLS showing minimum residual deviance *D* (Equation 3) when training and predicting the model with the data sets connected by the arrows. For a description of the data set names, see Section 2.2

TABLE 3 Logarithm of total residual deviance D (Equation 3) averaged over all 12 SLS for habitat models trained with Q1 temperature

log(average D)		Predicting data set					
		E_{IS}	E_{1M}	E_{02}	E_{05}	E_{10}	E_{20}
Training data set	$E_{IS}+F_{IS}$	15.10	16.20	16.01	15.95	16.01	15.82
	$E_{1M}+F_{IS}$	50.87	34.41	69.47	65.12	37.10	96.58
	$E_{02}+F_{IS}$	51.60	30.96	37.98	51.51	15.89	16.03
	$E_{02}+F_{W}$	20.96	20.73	20.76	20.71	20.71	37.41
	$E_{05}+F_{IS}$	39.60	39.62	15.91	15.94	16.01	16.03
	$E_{05}+F_{W}$	27.94	20.48	20.35	20.36	20.35	19.99
	$E_{05}+F_{S}$	87.65	124.69	85.25	124.92	87.02	85.13
	$E_{10}+F_{S}$	**5.32**	9.38	9.13	10.00	9.45	9.53
	$E_{20}+F_{S}$	5.33	10.63	11.19	10.36	11.15	11.34
	$E_{30}+F_{S}$	5.80	9.75	9.74	10.23	9.73	10.01

Note: Infinity values of D have been ignored in the averages. The cell colour coding is proportional to the numerical values (green, small; yellow, intermediate; and red, large). The overall minimum is printed in bold font.

TABLE 4 Like Table 3 but for Q3 data sets

log(average D)		Predicting data set					
		E_{1M}	E_{02}	E_{05}	E_{10}	E_{20}	E_{30}
Training data set	$E_{1M}+F_{IS}$	69.90	59.35	69.81	114.19	79.05	15.40
	$E_{02}+F_{IS}$	30.96	73.38	97.03	15.50	15.87	15.30
	$E_{02}+F_{W}$	23.57	20.59	20.64	22.81	25.69	20.12
	$E_{05}+F_{IS}$	44.89	16.26	42.76	16.32	15.96	15.56
	$E_{05}+F_{W}$	20.52	20.30	20.46	25.56	20.53	31.20
	$E_{05}+F_{S}$	31.51	57.90	55.78	34.70	59.11	21.68
	$E_{10}+F_{S}$	5.33	11.12	12.27	11.58	10.69	10.78
	$E_{20}+F_{S}$	**5.19**	10.45	10.33	10.16	10.00	9.77
	$E_{30}+F_{S}$	5.54	14.71	14.22	13.19	13.62	13.79

abundance observations (Table 1), probably following the intuitive idea of working with highest possible resolution and detail.

However, most of the evidence in our study pointed to the opposite direction, that is neither in situ nor the most finely resolved gridded data used to train the model yielded the best performance (Figure 2; Tables 3 and 4). For most of our SLS, best predictions were made when both environmental and abundance training data were downsampled to roughly 1.0° (E_{10} + F_S). Five out of 12 SLS suggest that downsampling to even coarser resolutions might improve model performance: One SLS for Q1 (Figure 2a) and four SLS for Q3 (Figure 2b) yielded minimum D for coarser resolutions. The overall minimum D for Q3 was obtained with E_{20}, while E_{10} had the second smallest D.

Predictions of the large-scale shifts of thermal suitability of fish habitats obtained in our study were considerably different depending on the scale of the training data set and led to contradictory pictures (Figures 4 and 5). The suitability changes predicted with the low-resolution model clearly indicated northward displacements of fish biomass, which is in agreement with the observed distributional shifts reported here (Figure 3) and in previous studies on North Sea cod (Engelhard et al., 2014; Hedger et al., 2004; Núñez-Riboni et al., 2019) and grey gurnard (Perry et al., 2005). In contrast to cases with downsampled training data sets, the habitat model trained with high-resolution data failed to reproduce the observed changes in the habitat suitability (occupancy) for both species and produced unrealistic results (Table 5) and poor statistics (Tables S1 and S2).

One reason for the large differences in the habitat model predictions obtained with the training data sets matched at different spatial scales is the shape of the temperature curve. In the particular cases of cod and grey gurnard, the temperature curves obtained with in situ or high-resolution data, respectively, were unbounded (Figure 4c and d) and, thus, unrealistic in comparison with the curves obtained with E_{10} + F_S data sets (Figure 5c and d). We deepen the discussion on the temperature response function in Section 4.4. In view of these findings, the adequate choice of the matching scale seems fundamental for the predictive skill of fish habitat models, particularly at climate scale.

4.2 | Reasons behind the improved performance with downsampled data

Although the importance of scale is widely acknowledged by marine ecologists (e.g. Hale et al., 2019; Pinsky et al., 2020; Redfern

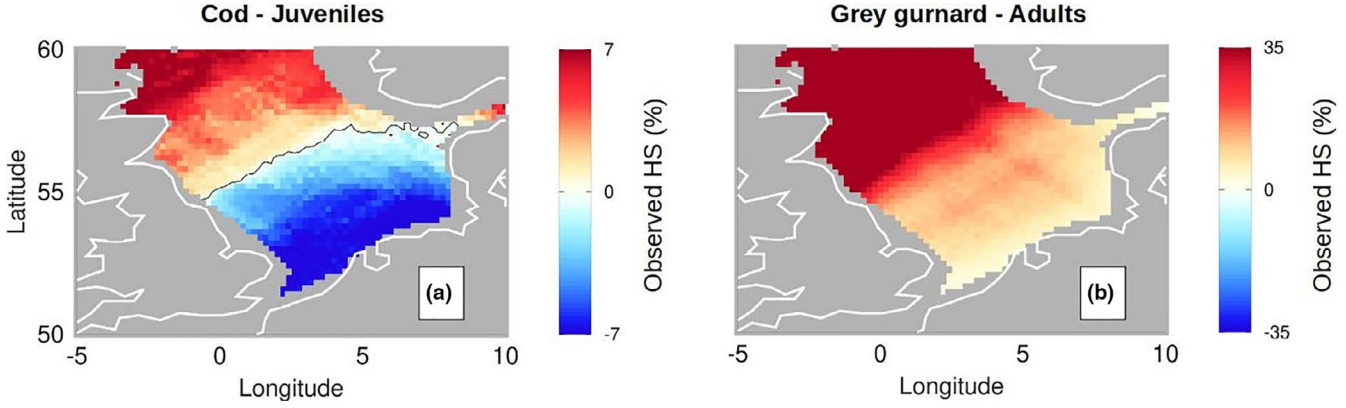

FIGURE 3 Differences of the observed fish abundance H_O between decades 1970–1980 and 2007–2017 for (a) juvenile cod and (b) adult grey gurnard. The black curve represents no change

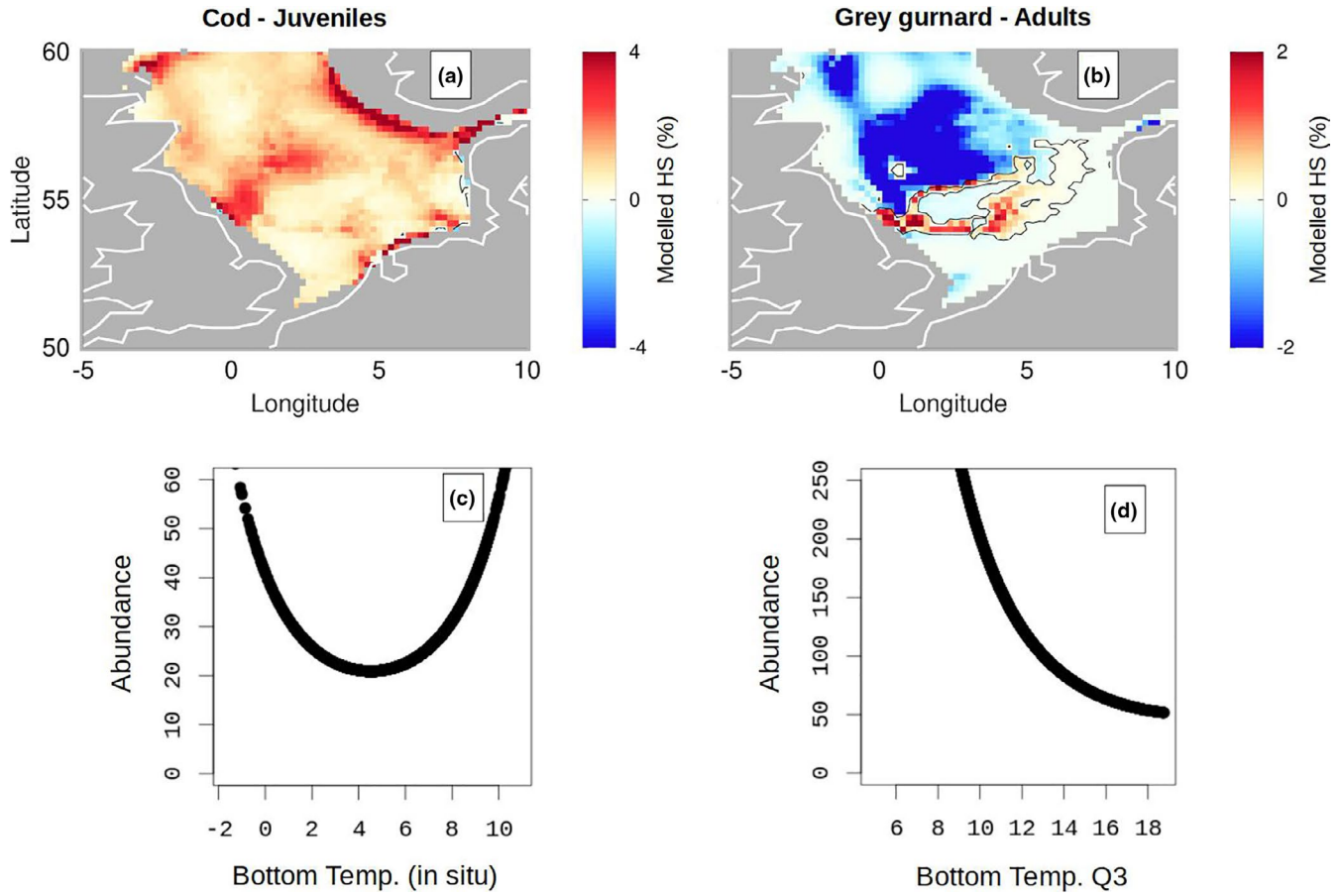

FIGURE 4 Upper panels (a, b): As in Figure 3, but for modelled habitat suitability H_M (Equation 5) trained with small-scale data: data set $E_{IS} + F_{IS}$ (see Figure 1) for cod (Q1 temperature) and data set $E_{1\,M} + F_{IS}$ (Q3 temperature) for grey gurnard. Lower panels (c, d): Corresponding temperature curves (Equation 4)

et al., 2006), studies specifically aiming at unveiling the relation between scale and performance of fish habitat models are scarce. Within comparable studies, we were only aware of França and Cabral (2016), who observed, contrary to us, a decrease of the model performance at their large scale. However, a direct comparison regarding the disagreement is difficult since these authors changed their model by using different predictors for each tested scale, whereas we intentionally kept the model unchanged (see Section 2.3). Hale et al., (2019) find strongest relation between environment and coral reef fish at an intermediate scale from four scales tested, in perfect agreement with our findings. Nonetheless, this is not a modelling but rather an observational study based on similarity matrices. While our study is specific to marine fish, the scarcity of similar modelling studies in fisheries science leads us to involve examples from other ecological disciplines in the discussion as well.

Some of the previous studies have also claimed that coarse data reduces model accuracy (Dyer et al., 2013; Ferrier & Watson, 1997; Guisan et al., 2007; Ross et al., 2015; Seo et al., 2009). However, many other studies have found similar results to ours, with use of high-resolution data not improving (Becker et al., 2010; Lowen et al., 2016; Mitchell et al., 2001; Redfern et al., 2008; Thomas et al., 2002) or even decreasing predictive power of their habitat models (Guisan et al., 2007; Johnson et al., 2002; Luoto et al., 2007; Rahbek & Graves, 2001; Tobalske, 2002).

In an attempt to elucidate why the scale issue seems so evasive, we propose herewith the following conceptual model: Let us assume that changes of environment E_T and fish abundance A_T can be decomposed in large and small-scale variations as follows:

$$E_T = E_L + E_S \tag{8}$$

and

$$A_T = A_L + A_S + A_{\text{noise}}. \tag{9}$$

The sub-indices T, L and S stand for "total," "large" and "small," respectively. E_S can represent short-scale, high-frequency oceanic variations like eddies or current meanders, while A_S is the small-scale deterministic response of fish to those variations. A_{noise} is the noise in fish abundance data that is unrelated to the environmental conditions (more details in Section 4.2.2 below). Environmental noise (variations with no impact on the fish abundance field, like observational error) are also possible but omitted from this discussion for simplicity. Based on this conceptual model, we attempt to explain our findings in terms of four (closely related) concepts in the following subsections: scale-dependent ecological response (Section 4.2.1),

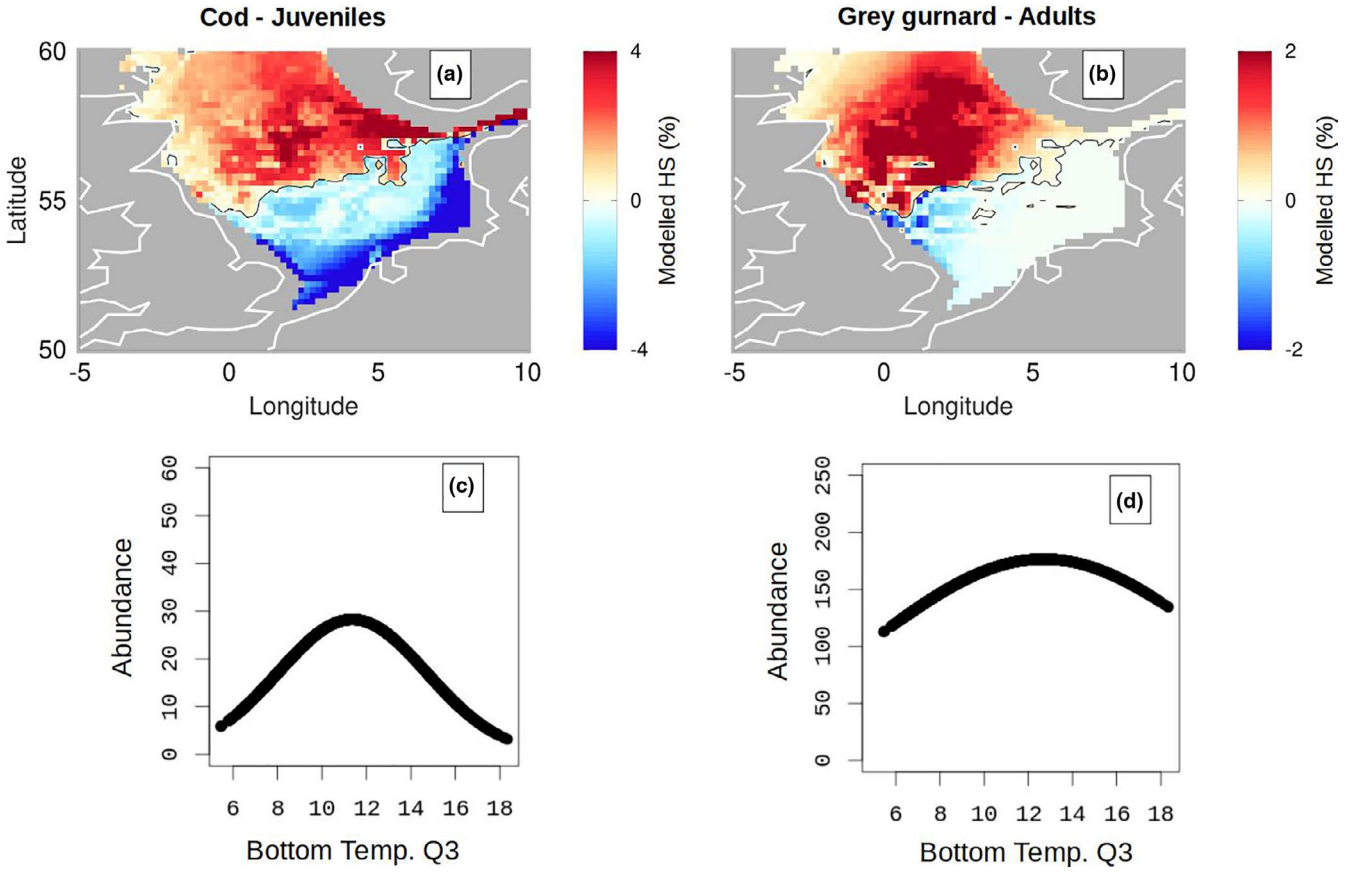

FIGURE 5 As Figure 4 but for model trained with large-scale data (E_{10} + F_S with Q3 temperatures from previous year for both SLS)

noise (Section 4.2.2), effects of spatial interpolation (Section 4.2.3) and Nyquist–Shanon sampling theorem (Section 4.2.4).

4.2.1 | Scale-dependent ecological response

Several studies on habitat modelling favoured the use of neither fine nor coarse data, but pointed out profound effects of scale of training data on model performance and suggested that the choice of the proper data scale should match the research question addressed by a study (e.g. Austin & Van Niel, 2011; Bellier et al., 2010; García-Callejas & Araújo, 2016; Pearson et al., 2004). This agrees with the concept of a scale-dependent response of organisms (including fish) to their changing environment that is broadly recognized as an important issue in ecology (see Wiens, 1989; Mitchell et al., 2001; Luoto et al., 2007; Hale et al., 2019). Organisms respond differently to the environmental changes depending on their spatial and temporal scales via physiological adjustment, behaviour, acclimatization and colonization (de Knegt et al., 2010; Nyström Sandman et al., 2013; Levin, 1992; Pinsky et al., 2020).

The notion of a scale-dependent ecological response implies that A_T in Equation 8 should depend on E_T in Equation 9 through two (probably non-linear) response functions G_L and G_S:

$$A_T = G_L\left(E_L\right) + G_S\left(E_S\right) + A_{\text{noise}}. \quad (10)$$

TABLE 5 Model realism assessed by three criteria based on the metrics described in Section 2.4. "RTC" stands for "realistic temperature curve"

Training data set	Realism Criteria	Norway Pout		Haddock		Cod		Grey gurnard		Whiting		Common dab	
		juv	ad	juv	ad	juv	ad	juv	ad	juv	ad	juv	ad
Q1 E_{IS}+F_{IS}	MAD< $\overline{MAD}$		NA										
	Ω<65%												
	RTC												
Q1 E_{10}+F_S	MAD< $\overline{MAD}$		NA										
	Ω<65%												
	RTC												
Q3 E_{1M}+F_{IS}	MAD< $\overline{MAD}$		NA										
	Ω<65%												
	RTC												
Q3 E_{10}+F_S	MAD< $\overline{MAD}$		NA										
	Ω<65%												
	RTC												

Note: The habitat models for each SLS ("juv": juveniles and "ad": adults) were trained with two small-scale (Q1 E_{IS} + F_{IS} and Q3 $E_{1\,M}$ + F_{IS}) and two large-scale data sets (Q1 E_{10} + F_S and Q3 E_{10} + F_S). Colours indicate whether the model fulfils the criteria (green: yes and red: no). Thick contours highlight realistic models complying all three criteria. For Norway pout (Trisopterus esmarkii, Gadidae, marked with NA, i.e. "not available"), the number of observations at the beginning of the record was too small for this analysis.

Our results indicate that, for most of the adult and juvenile fishes tested here, G_L is more important than G_S or, alternatively, G_S is not properly resolved in the data (see Sections 4.2.2. and 4.2.4 below). However, G_S might be more important that G_L for some fish species,

explaining why in situ data yields best results for two SLS in our study (Figure 2). These notions agree with Becker et al., (2010), who found improvement of habitat model performance for the majority (but not all) of their species of marine mammals with coarse resolution. On the other hand, our improved model performance with high-resolution predicting data indicates that both response functions G_L and G_S are probably similar. For instance, both response curves should reach maximum at the same optimum value but could differ in their amplitude or width, indicating a different tolerance to environmental changes at different scales.

Another aspect of the scale-dependent ecological response is that the importance of environmental variables can vary with scale. While one variable can be important at a particular scale, another one could be the important one at another scale. Our results relating the Q1 and Q3 temperatures agree with this notion: While the concurrently occurring Q1 temperature seems more important at the small scale, the Q3 temperature seems more important at large scales for most SLS used in this study (Table 5). This issue has been discussed previously by Pinsky et al., (2019), who argued that fish living at the southern edge of their suitable habitat are mainly driven by summer temperatures at climate scale. Note, however, that while the choice of Q3 temperature may improve the model performance over Q1 temperature for several SLS, the choice of the adequate scale (downsampling) improves it considerably more (for both quarters). This can be seen in the range of D within Tables 3 and 4 (average of e^{80}) in comparison with the differences of D between the two tables (average of e^{13}). Moreover, grey gurnard, the only investigated species yielding realistic results with Q1 in situ (Table 5), shows a total residual deviance D considerably smaller with the downsampled data set Q1 $E_{10} + F_S$ (e^5) than with Q1 in situ (e^{10}; no table shown).

4.2.2 | Noise

Tobalske (2002) and Redfern et al., (2008) explicitly suggested that data aggregation could eliminate data "noise" obscuring patterns between environment and species variables. Tobalske (2002) used this explanation to justify a better predictive accuracy of her habitat model (for birds) with the coarse resolution data in comparison with the fine resolution. In our study, both fish abundance and characteristics of the marine environment are indeed subject to high-frequency and small-scale variations. For temperature, these variations arise from tides, internal waves, eddies, atmospheric low- and high-pressure regimes (Meyers et al., 1991) and might (or might not) drive similar short-scale variations in fish abundance. Short-scale variations in fish abundance, probably unrelated to the environmental variations, could arise from schooling, non-deterministic changes of swimming direction and avoidance of the fishing gear, as argued by Wood (2017, his Section 7.5). Therefore, while data matched at scales of, say, 10 km might be good to study the effect of tides and eddies on the spatial distribution of fish, the same data set could have too much superfluous information, that is noise, making it difficult for a statistical habitat model to correctly isolate signal at the interannual or climate scales from the higher-frequency variations.

One important result of our study is that the best model predictions of fish habitat were mainly obtained when the model was trained with low-resolution and the prediction was performed with high-resolution data (Figure 2, Tables 3 and 4). Thuiller et al., (2005) obtained similar results with a model of plant habitats. The competing effects of data detail and noise might be a one possible explanation. We have explored this notion with frequency diagrams of in situ and gridded AHOI bottom temperature (Figure 6). The in situ data (dashed curve) are more scattered than the gridded data (continuous curve), showing more frequent extreme values, like temperatures cooler than 3°C and warmer than 9°C. Additionally, in situ data have a higher peak indicating a larger amount of the most frequent observations near 7°C. These small-scale temperature variations in the in situ data (denoted as E_S in Equation 8) occur at scales smaller than AHOI's resolution and are, thus, absent in the gridded data (more about this "subscale" noise in Sections 4.2.3 and 4.2.4 below). Therefore, the diagram of the gridded data is a "blurred" version of the in situ diagram, with summit and extreme values reduced in frequency, while intermediate temperatures (near 5°C and 8°C) being more frequent. The sharper summit of the in situ diagram indicates a higher level of detail but more frequent extreme temperatures indicate more subscale noise as well. According to our results, the higher level of noise seems to be adverse for the model training. But at the same time, more details have a positive effect when using the model for a prediction. Therefore, we advocate here model training with low-resolution data but model prediction with fine-resolution data.

A bias of 0.3°C in the gridded temperature data has been intentionally removed in Figure 6 by aligning the diagram of gridded data with the diagram of in situ data. This systematic bias in AHOI has been thoroughly discussed in Núñez-Riboni and Akimova (2015). Such a bias equally affects all temperature values in all modelled years and, therefore, has no influence on the model training. However, this bias would manifest when predicting fish distribution using in situ temperatures. This and other potential negative impacts of the gridded data seem to be overcome by the increase of model performance from training the model with coarse data, as reflected by the total residual deviance *D* from the cross-validation: variations of *D* with the training scale are several orders of magnitude larger than variations with the predicting scale (roughly e^{30} against e^8; Tables 3 and 4). Therefore, the key factor improving model performance in our study is the scale of the training and not of the predicting data.

4.2.3 | Failure of interpolation to create information

Statistical methods to grid environmental data are specifically designed to reduce high-frequency, short-range subscale noise, underscoring the resolvable scales of available measurements (Clancy, 1983; Hiller & Kaese, 1983; Meyers et al., 1991; Núñez-Riboni & Akimova, 2015). Therefore, while in situ fishery abundance represents conditions on scales of only few kilometres (the trawl distance), the gridded environmental data often represent conditions on scales at least one order of magnitude larger (in our particular

case ca. 20 km × 20 km for AHOI temperature). This implies that the small-scale variations E_S are normally absent in environmental gridded data, and, thus, the small-scale variations A_S present in fishery in situ data have no environmental counterpart. If small-scale changes in fish and environment do not correspond, training a habitat model with such data would "mislead" the model, reducing its performance.

While this scale mismatch might be intuitive and known, many modellers try to overcome it by interpolating the gridded environmental data on the positions of fish abundance data (12 out of 19 studies in the "upsampling" category in Table 1). However, we must bear in mind that interpolation cannot restore the missing information E_S: Non-recorded (or removed) eddies or fine-scale bathymetric features cannot be generated by a simple spatial interpolation of temperature or water depth. Therefore, the problem of matching in situ fish abundance and gridded environmental data can only be correctly solved by removing the high-frequency, short-scale variations A_S and A_{noise} from the fishery data as well.

Similarly to the noise reduction in environmental data mentioned above, a smoother such as the one used in Equation 1 would eliminate the small-scale variations in the fish data, allowing to focus on the resolvable scales. This measure reduces Equation 10 to a simpler equation $A_L = G_L(E_L)$ and helps the statistical model to isolate this signal from the data. Our study underpins this idea by showing how the smoother s_M (Equation 1) reduces noise in fish abundance considerably, as seen in Tables S1 and S2 (compare column $E_{02} + F_{IS}$ with $E_{02} + F_W$ or column $E_{05} + F_{IS}$ with $E_{05} + F_W$). In agreement with this notion, Redfern et al., (2008) point out how different data resolutions changed the signal-to-noise ratio of their dolphin sights due to a reduction in zero observations.

4.2.4 | Sampling resolution and Nyquist–Shannon theorem

Once it has been accepted that fishery data should be gridded to correctly match them to the gridded environmental data, the corresponding gridding scale must be chosen. The native resolution of the available data (in our case a scientific bottom trawl survey) constrains the choice of the gridding scale due to a fundamental principle of data sampling, that is the Nyquist–Shannon sampling theorem (Nyquist, 1928; Shannon, 1948). This theorem postulates that a signal of length *B* can only be completely resolved if recorded with sampling interval *B/2* or smaller (i.e. both, crests and valleys of a signal wave, must be sampled to resolve it completely). When this does not happen, "aliases" of signals at scales smaller than B/2 are recorded (Oppenheim et al., 1999). These aliases are large-scale phantom signals arising from sampling short-scale signals with a long sampling distance.

As a consequence of the Nyquist–Shannon theorem, data sets with two different sampling resolutions should not be matched at the fine resolution because the coarsely sampled data does not completely resolve fine-scale signals contained in the fine data. Thus, the average distance between observations sets a lower limit to the gridding resolution of fishery in situ data. The fish abundance data used in this study has an average sampling distance between neighbouring observations of 0.4° or approximately 35 km (as given by the average length of the Delaunay triangles used to define the mapping region; Section 2.2). Therefore, the fish abundance data can only correctly resolve signals at scales equal to or longer than 2 × 0.4° (i.e. 0.8°) or 70 km, namely the "Nyquist distance" or effective scale of the NS-IBTS survey. This agrees well with our results showing the best resolution of the training data set of 1.0° for most of the SLS. Variations in abundance at smaller scales than the Nyquist distance (like those potentially related to eddies or fish schooling) would be perceived as a (long) stochastic signal (i.e. aliased noise).

Downsampling data to resolutions coarser than the Nyquist distance seem to reduce the aliased noise. In agreement with this notion, Tables 3 and 4 show a strong improvement of model performance as soon as the filter scale exceeds the Nyquist distance (i.e. from 0.5° to 1.0°). It is, however, important to stress that data resolution should not be reduced indefinitely. Our results indicate

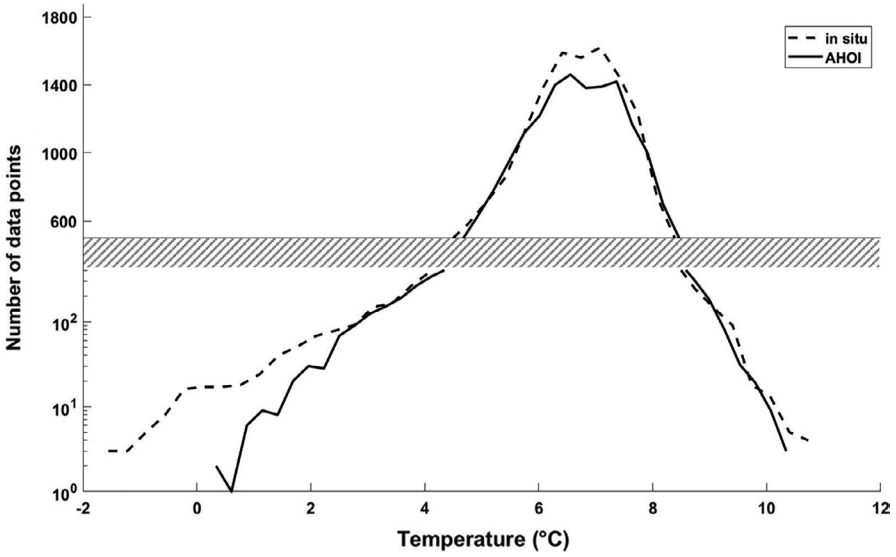

FIGURE 6 Frequency diagrams of in situ and gridded (AHOI) bottom temperatures. Note the different scales in the vertical, with the bottom subplot being logarithmic. A positive bias of roughly 0.3°C in the AHOI temperatures (discussed in Núñez-Riboni & Akimova, 2015) was intentionally removed to align both diagrams and compare them better

that habitat model performance does not increase unlimited when reducing data resolution. Both a too coarse and too high resolutions of training data impair the model performance (Tables 3, 4, S1 and S2). When the resolution is too coarse, the environment-species relation $A_L = G(E_L)$ can be essentially "eroded," that is extreme or seldom observed combinations of observations potentially playing an important ecological role are eliminated. Therefore, the model cannot be trained correctly and will poorly represent the relation between environment and fish abundance. Thus, more than data downsampling improving the fit of a habitat model, matching data at too fine scales (i.e. scales smaller than the Nyquist distance) worsens it.

A quantitative, graphical example encompassing all concepts discussed in this section is given in Section S5 of the Supplement. There, we exemplify how downsampled data can improve model performance over in situ and high-resolution data. While all these concepts seem to explain our results well, a complete understanding of our results demands further research about the scales and mechanisms of the unresolved small-scale variations in environmental and fish data.

4.3 | Spatial autocorrelation and pseudoreplication

In addition to the Nyquist–Shannon theorem, pseudoreplication is another important reason to avoid gridding in situ environmental and fish abundance data at small spatial scales. Pseudoreplicated data are data which partially depend on each other, either because they were sampled with a rate higher than the natural autocorrelation scale or because the data user has intentionally tried to increase the sampling rate or data resolution (e.g. by interpolating observations; see for instance Millar & Anderson, 2004). Such pseudoreplication inflates the significance of the model terms, leading to invalid statistical measures and wrong decisions when designing the model based on such significance (Beale et al., 2010; Lennon, 2000). As described in Section 2.3, we did not design the habitat model in this study, but choose one de facto. However, for clarity, it is important to discuss the role that autocorrelation and pseudoreplication play in our study.

Because a data filter used to downsample the data (Section 2.2) removes short-scale variations, it also changes the autocorrelation function, making it smoother, flatter and increasing the decorrelation scale Δ (e.g. the first zero-crossing of the autocorrelation function). Therefore, filtering environmental and abundance data increases the number n_Δ of the auto-correlated data within Δ, potentially pseudoreplicating the data. To deal with this issue, data downsampling includes not only filtering but also decimation, that is removing data points as to leave only one datum inside the filter scale L (Section 2.2). To verify this notion, we calculated n_Δ using the example of ETOPO1 bathymetry, showing that such downsampling did not increase n_Δ (Figure 7).

An additional and more important measure to deal with pseudoreplication was the explicit modelling of the spatial and temporal autocorrelation with the random effect s_R and the temporal smoother s_t, respectively (Equation 2). Dealing with pseudoreplication with random effects is one of the approaches suggested by Millar and Anderson (2004). The model term errors reaching a minimum for E_{05} (Tables S1 and S2) instead of continuously decreasing with the scale L also indicates that our measures to avoid pseudoreplication were effective.

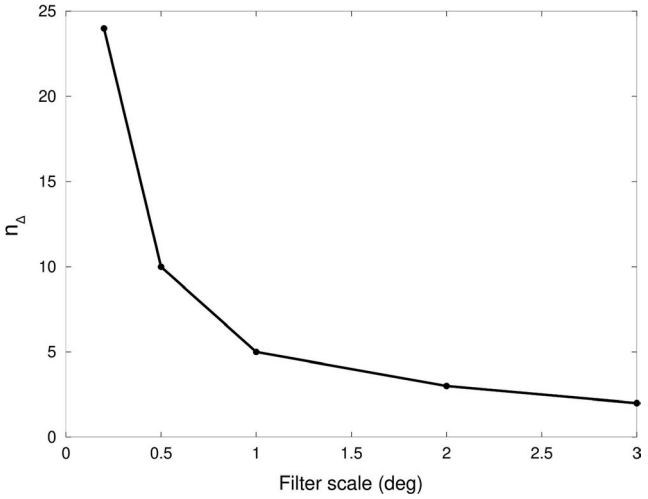

FIGURE 7 Average number of data points n_Δ of ETOPO1 bathymetry within the decorrelation length Δ against filter scale L (as example of how data downsampling does not pseudoreplicate our data). Δ was calculated with the first zero-crossing of the isotropic autocorrelation function

4.4 | Temperature response function

In this study, the temperature response function was intentionally modelled as a second-order polynomial to obtain, in combination with the logarithmic link, a curve with a single maximum (Equation 4), that is a Gaussian bell (GB further on). Such unimodal response functions are motivated by the concept of ecological niche (Hutchinson, 1957), where there is a range of suitable temperature values, including a single optimum, beyond which the habitat suitability decreases to zero. In various previous studies of fish habitat (e.g. Beare & Reid, 2002; Borchers et al., 1997; Bruge et al., 2016; Brunel et al., 2017; Lindegren et al., 2013; Rutterford et al., 2015) and marine mammals (Becker et al., 2010; Redfern et al., 2008), unimodal response curves have been achieved with penalized spline smoothers with a low basis dimension k (between 3 and 7). Modelling the temperature response with a GB is rather unconventional in fish habitat modelling (with only Núñez-Riboni et al., 2019 known to us). Therefore, it is valid to ask whether the results of this study would be different if a penalized smoother is used instead of the GB.

To answer this question and justify our choice for the GB, we have repeated our analysis using penalized smoothers for temperature. We tested two values for the basis dimension k: one with low degrees of freedom ($k = 4$) and one with large ($k = 50$) to inspect the behaviour of smooth and wiggly responses. In what follows, these smoothers will be

called s_W and s_S, respectively. For both smoothers, the relative values of total deviance *D* from the cross-validation and of the metrics on the training data sets were similar to those obtained with the GB, supporting downsampling of data to $E_{10} + F_S$ for Q1 and $E_{20} + F_S$ for Q3 (no tables shown for brevity). This indicates that our major conclusions are independent from the choice of temperature response function.

Whether a smoother would generally perform better in comparison with the GB is not easy to answer: Some of the metrics favour the use of s_W, while some other favour the use of the GB. This topic is beyond the scope of this study, but we consider important to advocate the use of the GB, which seems rather unknown in the modelling of fish habitat. We show here three examples from our SLS where the GB clearly resulted in more ecologically meaningful response than the smoothers (Figure 8). Because the maximum of the response curve should represent a preferred temperature, the unimodal GB (Figure 8, panels c, f, i) seems a more realistic representation of habitat than s_W, which often has two or more local maxima (panels a, d, g).

With its unimodal response, s_S is a better representation of habitat niche than s_W. However, even s_S can yield unrealistic estimates when predicting with values beyond the range of the training data

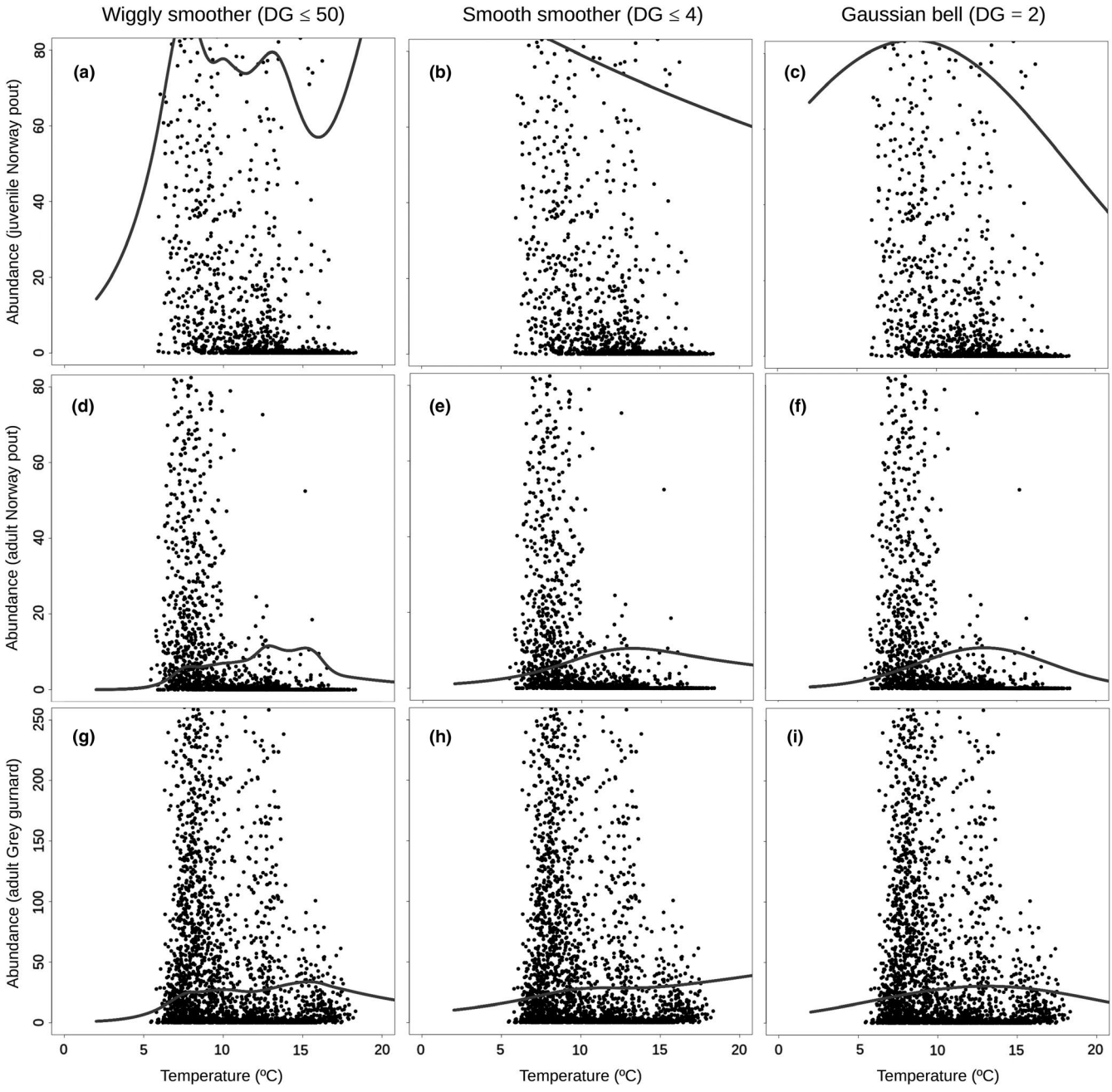

FIGURE 8 Examples of three temperature response curves (grey curves) for juvenile (top panels) and adult (middle panels) Norway pout, as well as for adult grey gurnard (bottom panels). The left panels show the wiggly smoother s_W, the middle panels the smooth smoother s_S and the right panels the Gaussian bell. DG stands for "degrees of freedom". The model has been trained with the data set $E_{10} + F_S$, which is shown with black dots

(Figure 8, panels b and h) or decrease extremely slow to zero (panel e). Projections of future fish distribution under climate change are highly sensitive to such behaviour of the response function (Thuiller et al., 2004). To counteract this behaviour, some authors have inserted an arbitrary amount of artificial zeroes beyond the data range (for instance, Beare & Reid, 2002). In our habitat model, the GB decreased to zero beyond the data range (Figure 8, panels c,f,i) without this questionable insertion of zeroes. Therefore, we see advantages of using the GB response in studies projecting future habitat under climate change, since the model must be able to predict fish distribution at temperatures higher than those observed in the past and used to train the model (Núñez-Riboni et al., 2019; Thuiller et al., 2004).

4.5 | Precautionary remark for scientific survey design

Some words are needed to prevent misinterpretation of our findings relating to the required sampling intensity of the fishery-independent scientific surveys. Although we claim here that downsampling, that is resolution reduction, of training data sets improves the performance of the fish habitat model, this should not be interpreted as an argument to reduce the sampling intensity within scientific fish surveys. A technical reason is that downsampling as applied in this study relies on all measured values, and integrates them over larger grid cells, where a greater number of samples per cell improves the quality of the estimate for the respective grid cell. A second, ecological reason is that surveys like the North Sea International Bottom Trawl Survey, in particular, have multiple objectives and are important sources of information for fisheries management and for numerous studies about ecological processes taking place at different spatial scales, for example predator–prey interactions, species productivity and biodiversity. Hence, any consideration of options to reduce the survey effort would need to take technical, statistical and ecological aspects into account and weight them against the priorities of the respective survey.

5 | CONCLUSIONS

In situ, high-resolution and interpolated data are often regarded as the best input for fish habitat modelling, while aggregated data appear to be regarded as deficient (Table 1). Contrary to these notions, we demonstrated here that training a fish habitat model with environmental and abundance data downsampled to a resolution of 1.0° considerably increased model performance. The best predictions of fish habitat were in our study achieved with environmental data of the highest available resolution (Figure 2). Still, the key factor affecting model performance was the scale at which the training data were matched, and not the scale of the predicting data (Tables 3 and 4).

These results appear to arise from four (intrinsically related) reasons: (1) Scale-dependent ecological response of fishes to changes in their environment is dominated by large-scale processes; (2) small-scale variations like fish schooling, eddies, tides and frontal meanders are either mechanistically unrelated (i.e. are noise) or only partially resolved at the scale of the sampling length (i.e. are subscale noise); (3) the interpolation of data sampled at a coarser resolution cannot create the information missing at finer scales; (4) the Nyquist–Shannon theorem sets a lower limit to the scale at which fishery data should be matched to gridded environmental data. Because in our case the average distance between IBTS hauls is 0.4°, the data can only resolve signals with lengths of 0.8° or longer. Downsampling of both fish and environmental data prior to model fitting deals with all four issues: it eliminates noise, focuses only on the large-scale ecological response, avoids interpolation and complies with the Nyquist–Shannon theorem.

Our study underpins the importance of bearing in mind the characteristic temporal and spatial scales of ecological and environmental processes in focus, as well as the native resolution of the available observations. Only by considering these issues, it seems possible to correctly model fish habitat. Instead of analysing spatial observations neglecting their complex interplay with scale, fishery scientists should consciously target a particular scale of interest in all modelling efforts. If the topic of interest is the effect of small-scale environmental changes, both in situ fishery and in situ environmental data should be used, if available. For research questions regarding interannual to climate scales, or if non-observed or lagged variables are used (e.g. previous season or year), gridded environmental data are needed. In such cases, both environmental and fish abundance data should be downsampled to the relevant effective scale under the consideration of the Nyquist–Shannon sampling theorem. Failure to do so can have serious consequences on model predictions, as demonstrated with the ability of our model to reproduce displacements of fish distribution at climate scale (Figures 3–5).

ACKNOWLEDGEMENTS

We appreciate discussions with our colleagues of the modelling group of the Thünen Institute of Sea Fisheries, particularly Miriam Püts, Marc Taylor and Alexander Kempf. Elucidating explanations about signal analysis from Professor Dudley Chelton (Oregon State University; USA) are specially acknowledged. We thank all NS-IBTS survey teams for making their data sets freely available through the databases at ICES. Comments of two anonymous reviewers considerably improved the manuscript.

DATA AVAILABILITY STATEMENT

Most of the data used in this study are openly available: Fish abundance data are available from the DATRAS portal (DATRAS, 2020), in situ temperature and bathymetry from the ICES portal of oceanographic data at https://www.ices.dk/data/data-portals/Pages/ocean.aspx and from the World Ocean Database at https://www.ncei.noaa.gov/access/world-ocean-database-select/dbsearch.html, while gridded ETOPO1 bathymetry are available from the US National Oceanic and Atmospheric Administration at https://www.ngdc.noaa.gov/mgg/global/. Sources of raw temperature data used in AHOI are also openly available: please see Núñez-Riboni &

Akimova, 2015 for a list of all data providers and links to their data portals. The final (gridded) AHOI data used in this study (until 2017) are available from the corresponding author upon request, while a shorter run (until 2014) is openly available from the Thünen Institute of Sea Fisheries at https://www.thuenen.de/de/sf/projekte/ein-physikalisch-statistisches-hydrographie-modell-fuer-fischerei-und-oekologiestudien-ahoi/.

ORCID

Ismael Núñez-Riboni https://orcid.org/0000-0002-7059-9050

REFERENCES

Amante, C., & Eakins, B. W. (2009). Etopo1 1 arc-minute global relief model: procedures, data sources and analysis. NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center. Marine Geology and Geophysics Division, Boulder, Colorado.

Augustin, N. H., Trenkel, V. M., Wood, S. N., & Lorance, P. (2013). Space-time modelling of blue ling for fisheries stock management. *Environmetrics*, *24*, 109–119. https://doi.org/10.1002/env.2196

Austin, M. P., & Van Niel, K. P. (2011). Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography*, *38*, 1–8. https://doi.org/10.1111/j.1365-2699.2010.02416.x

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. *Ecology Letters*, *2*, 246–264. https://doi.org/10.1111/j.1461-0248.2009.01422.x

Beare, D. J., & Reid, D. G. (2002). Investigating spatio-temporal change in spawning activity by Atlantic mackerel between 1977 and 1998 using generalized additive models. *ICES Journal of Marine Science*, *59*, 711–724. https://doi.org/10.1006/jmsc.2002.1207

Becker, E. A., Forney, K. A., Ferguson, M. C., Foley, D. G., Smith, R. C., Barlow, J., & Redfern, J. V. (2010). Comparing California Current cetacean–habitat models developed using in situ and remotely sensed sea surface temperature data. *Marine Ecology Progress Series*, *413*, 163–183. https://doi.org/10.3354/meps08696

Bellier, E., Certain, G., Planque, B., Monestiez, P., & Bretagnolle, V. (2010). Modelling habitat selection at multiple scales with multivariate geostatistics: An application to seabirds in open sea. *Oikos*, *119*, 988–999. https://doi.org/10.1111/j.1600-0706.2009.17808.x

Borchers, D. L., Buckland, S. T., Priede, I. G., & Ahmadi, S. (1997). Improving the precision of the daily egg production method using generalized additive models. *Canadian Journal of Fisheries and Aquatic Sciences*, *12*, 2727–2742. https://doi.org/10.1139/f97-134

Bruge, A., Alvarez, P., Fontan, A., Cotano, U., & Chust, G. (2016). Thermal niche tracking and future distribution of Atlantic mackerel spawning in response to ocean warming. *Frontiers in Marine Science*, *3*, 86. https://doi.org/10.3389/fmars.2016.00086

Brunel, T., van Damme, C. J. G., Samson, M., & Dickey-Collas, M. (2017). Quantifying the influence of geography and environment on the northeast Atlantic mackerel spawning distribution. *Fisheries Oceanography*, *27*, 159–173. https://doi.org/10.1111/fog.12242

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference*. Springer. https://doi.org/10.1007/b97636

Candy, S. G. (2004). Modelling catch and effort data using generalised linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science*, *11*, 59–80.

Cheung, W. W., Lam, V. W., Sarmiento, J. L., Kearney, K., Watson, R., & Pauly, D. (2009). Projecting global marine biodiversity impacts under climate change scenarios. *Fish and Fisheries*, *10*, 235–251. https://doi.org/10.1111/j.1467-2979.2008.00315.x

Clancy, R. M. (1983). The effect of observational error correlations on objective analysis of ocean thermal structure. *Deep Sea Research Part A. Oceanographic Research Papers*, *9*, 985–1002. https://doi.org/10.1016/0198-0149(83)90053-5

DATRAS (2020). *Database of Trawl Surveys (DATRAS) of the International Council for the Exploration of the Sea (ICES)*. https://datras.ices.dk/Data_products/Download/Download_Data_public.aspx

de Knegt, H. J., van Langevelde, F., Coughenour, M. B., Skidmore, A. K., de Boer, W. F., Heitkönig, I. M. A., Knox, N. M., Slotow, R., van der Waal, C., & Prins, H. H. T. (2010). Spatial autocorrelation and the scaling of species–environment relationships. *Ecology*, *91*, 2455–2465. https://doi.org/10.1890/09-1359.1

Dickey, T. D. (2003). Emerging ocean observations for interdisciplinary data assimilation systems. *Journal of Marine Systems*, *7963*, 5–48. https://doi.org/10.1016/S0924-7963(03)00011-3

Dyer, J. J., Brewer, S. K., Worthington, T. A., & Bergey, E. A. (2013). The influence of coarse-scale environmental features on current and predicted future distributions of narrow-range endemic crayfish populations. *Freshwater Biology*, *6*, 1071–1088. https://doi.org/10.1111/fwb.12109

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Engelhard, G. H., Righton, D. A., & Pinnegar, J. K. (2014). Climate change and fishing: A century of shifting distribution in North Sea cod. *Global Change Biology*, *8*, 2473–2483. https://doi.org/10.1111/gcb.12513

Ferrier, S., & Watson, G. (1997). *An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity*. NSW National Parks and Wildlife Service. Department of Environment, Sport and Territories.

FishMap (2006). North Sea fish species fact sheets. Available at ICES website http://www.ices.dk/marine-data/maps/Pages/ICES-FishMap.aspx

França, S., & Cabral, H. (2016). Predicting fish species distribution in estuaries: Influence of species' ecology in model accuracy. *Estuarine, Coastal and Shelf Science*, *180*, 11–20. https://doi.org/10.1016/j.ecss.2016.06.010

García-Callejas, D., & Araújo, M. B. (2016). The effects of model and data complexity on predictions from species distributions models. *Ecological Modelling*, *326*, 4–12. https://doi.org/10.1016/j.ecolmodel.2015.06.002

Guisan, A., Graham, C. H., Elith, J., & Huettmann, F. & the NCEAS Species Distribution Modelling Group (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, *3*, 332–340. https://doi.org/10.1111/j.1472-4642.2007.00342.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). Habitat suitability and distribution models. With applications in R. Ecology, biodiversity and conservation. https://doi.org/10.1017/9781139028271

Hale, R., Colton, M. A., Peng, P., & Swearer, S. E. (2019). Do spatial scale and life history affect fish–habitat relationships? *Journal of Animal Ecology*, *88*, 439–449. https://doi.org/10.1111/1365-2656.12924

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *3*, 297–310. https://doi.org/10.1214/ss/1177013604

Hawkins, B. A., Diniz-Filho, J. A. F., Mauricio Bini, L., De Marco, P., & Blackburn, T. M. (2007). Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. *Ecography*, *30*, 375–384.

Hedger, R., McKenzie, E., Heath, M., Wright, P., Scott, B., Gallego, A., & Andrews, J. (2004). Analysis of the spatial distributions of mature cod (Gadus *morhua) a*nd haddock (Mel*anogrammus aeglefinus) a*bundance in the North Sea (1980–1999) using generalised additive models. *Fisheries Research*, *1*, 17–25. https://doi.org/10.1016/j.fishres.2004.07.002

Heessen, H. J., Daan, N., & Ellis, J. R. (Eds.) (2015). *Fish atlas of the Celtic Sea, North Sea, and Baltic Sea: Based on international research-vessel surveys*. Wageningen Academic Publishers.

Henson, S., Beaulieu, C., Ilyina, T., John, J. G., Long, M., Séférian, R., Tjiputra, J., & Sarmiento, J. L. (2017). Rapid emergence of climate change in environmental drivers of marine ecosystems. *Nature Communications*, *8*, 14682. https://doi.org/10.1038/ncomms14682

Hiller, W., & Kaese, R. H. (1983). Objective analysis of hydrographic data sets from mesoscale surveys. *Berichte Aus Dem Institut Für Meereskunde an Der Christian-Albrechts-Universitaet Kiel*, *116*, 78.

Hutchinson, G. E. (1957). Concluding remarks. *Cold Spring Harbor Symposium on Quantitative Biology*, *22*, 415–427. https://doi.org/10.1101/SQB.1957.022.01.039

ICES (2015). Manual for the international bottom trawl surveys. *Series of ICES Survey Protocols SISP*, *10*, IBTS IX.

IPCC. (2014). Summary for policymakers. In C. B. Field, V. R. Barros, D. J. Dokken, K. J. Mach, M. D. Mastrandrea, T. E. Bilir, & M. Chatterjee et al (Eds). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1–32) Cambridge University Press.

Johnson, C. M., Johnson, L. B., Richards, C., & Beasley, V. (2002). Predicting the occurrence of amphibians: An assessment of multiple-scale models. In J. M. Scott, P. J. Heglund, F. Samson, J. Haufler, M. Morrison, M. Raphael, & B. Wall (Eds.), *Predicting species occurrences: Issues of accuracy and scale* (pp. 157–169). Island Press.

Kammann, E. E., & Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society. Applied Statistics - Series C*, *52*, 1–18. https://doi.org/10.1111/1467-9876.00385

Kärcher, O., Frank, K., Walz, A., & Markovic, D. (2019). Scale effects on the performance of niche-based models of freshwater fish distributions. *Ecological Modelling*, *405*, 33–42. https://doi.org/10.1016/j.ecolmodel.2019.05.006

Kirkman, S. P., Yemane, D., Kathena, J., Mafwila, S. K., Nsiangango, S. E., Samaai, T., Axelsen, B., & Singh, L. (2013). Identifying and characterizing demersal fish biodiversity hotspots in the Benguela Current Large Marine Ecosystem: Relevance in the light of global changes. *ICES Journal of Marine Science*, *70*, 943–954. https://doi.org/10.1093/icesjms/fst040

Lennon, J. J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, *23*, 101–113. https://doi.org/10.1111/j.1600-0587.2000.tb00265.x

Levin, A. (1992). The problem of pattern and scale in ecology: The Robert H. MacArthur Award Lecture. *Ecology*, *73*(6), 1943–1967. https://doi.org/10.2307/1941447

Lindegren, M., Checkley, D. M. Jr, Rouyer, T., MacCall, A. D., & Stenset, N. C. (2013). Climate, fishing, and fluctuations of sardine and anchovy in the California Current. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 13672–13677. https://doi.org/10.1073/pnas.1305733110

Lowen, J. B., McKindsey, C. W., Therriault, T. W., & DiBacco, C. (2016). Effects of spatial resolution on predicting the distribution of aquatic invasive species in nearshore marine environments. *Marine Ecology Progress Series*, *556*, 17–30. https://doi.org/10.3354/meps11765

Luoto, M., Virkkala, R., & Heikkinen, R. K. (2007). The role of land cover in bioclimatic models depends on spatial resolution. *Global Ecology and Biogeography*, *1*, 34–42. https://doi.org/10.1111/j.1466-8238.2006.00262.x

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall. 511. https://doi.org/10.1002/bimj.4710290217

Melo-Merino, M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, *415*, 108837. https://doi.org/10.1016/j.ecolmodel.2019.108837

Meyers, G., Phillips, H., Smith, N., & Sprintall, J. (1991). Space and time scales for optimal interpolation of temperature - Tropical Pacific Ocean. *Progress in Oceanography*, *3*, 189–218. https://doi.org/10.1016/0079-6611(91)90008-A

Millar, R. B., & Anderson, M. J. (2004). Remedies for pseudoreplication. *Fisheries Research*, *70*, 397–407. https://doi.org/10.1016/j.fishres.2004.08.016

Mitchell, M. S., Lancia, R. A., & Gerwin, J. A. (2001). Using landscape-level data to predict the distribution of birds on a managed forest: Effects of scale. *Ecological Applications*, *6*, 1692–1708. https://doi.org/10.2307/3061089

Núñez-Riboni, I., & Akimova, A. (2015). Monthly maps of optimally interpolated in situ hydrography in the North Sea from 1948 to 2013. *Journal of Marine Systems*, *151*, 15–34. https://doi.org/10.1016/j.jmarsys.2015.06.003

Núñez-Riboni, I., Taylor, M. H., Kempf, A., Püts, M., & Mathis, M. (2019). Spatially resolved past and projected changes of the suitable thermal habitat of North Sea cod (Gadus morhua) under climate change. *ICES Journal of Marine Science*, *76*, 2389–2403. https://doi.org/10.1093/icesjms/fsz132

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *AIEE Transactions*, *47*, 617–644. https://doi.org/10.1109/T-AIEE.1928.5055024

Nyström Sandman, A., Wikström, S. A., Blomqvist, M., Kautsky, H., & Isaeus, M. (2013). Scale-dependent influence of environmental variables on species distribution: a case study on five coastal benthic species in the Baltic Sea. *Ecography*, *36*(3), 354–363. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0587.2012.07053.x

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-time signal processing*. Prentice Hall.

Parmesan, C., & Yohe, G. (2003). Globally coherent fingerprint of climate change impacts across natural systems. *Nature*, *421*, 37–42. https://doi.org/10.1038/nature01286

Pearson, R. G., Dawson, T. P., & Liu, C. (2004). Modelling species distributions in Britain: A hierarchical integration of climate and land-cover data. *Ecography*, *27*, 285–298. https://doi.org/10.1111/j.0906-7590.2004.03740.x

Pecl, G. T., Araújo, M. B., Bell, J. D., Blanchard, J., Bonebrake, T. C., Chen, I.-C., Clark, T. D., Colwell, R. K., Danielsen, F., Evengård, B., Falconi, L., Ferrier, S., Frusher, S., Garcia, R. A., Griffis, R. B., Hobday, A. J., Janion-Scheepers, C., Jarzyna, M. A., Jennings, S., … Williams, S. E. (2017). Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science*, *355*, 6332. https://doi.org/10.1126/science.aai9214

Perry, A. L., Low, P. J., Ellis, J. R., & Reynolds, J. D. (2005). Climate change and distribution shifts in marine fishes. *Science*, *5730*, 1912. https://doi.org/10.1126/science.1111322

Pham-Gia, T., & Hung, T. L. (2001). The mean and median absolute deviations. *Mathematical and Computer Modelling*, *7*, 921–936. https://doi.org/10.1016/S0895-7177(01)00109-1

Pinsky, M. L., Eikeset, A. M., McCauley, D. J., Payne, J. L., & Sunday, J. M. (2019). Greater vulnerability to warming of marine versus terrestrial ectotherms. *Nature*, *569*, 108–111. https://doi.org/10.1038/s41586-019-1132-4

Pinsky, M. L., Selden, R. L., & Kitchel, Z. J. (2020). Climate-driven shifts in marine species ranges: Scaling from organisms to communities. *Annual Review of Marine Science*, *12*(1), 153–179. https://doi.org/10.1146/annurev-marine-010419-010916

Pinsky, M. L., Worm, B., Fogarty, M. J., Sarmiento, J. L., & Levin, S. A. (2013). Marine taxa track local climate velocities. *Science*, *341*(6151), 1239–1242. https://doi.org/10.1126/science.1239352

Planque, B., Loots, C., Petitgas, P., Lindstrøm, U., & Vaz, S. (2011). Understanding what controls the spatial distribution of fish populations using a multi-model approach. *Fisheries Oceanography*, *20*, 1–17. https://doi.org/10.1111/j.1365-2419.2010.00546.x

Punzón, A., López-López, L., González-Irusta, J. M., Preciado, I., Hidalgo, M., Serrano, A., Tel, E., Somavilla, R., Polo, J., Blanco, M.,

Ruiz-Pico, S., Fernández-Zapico, O., Velasco, F., & Massuti, E. (2021). Tracking the effect of temperature in marine demersal fish communities. *Ecological Indicators*, *121*, 107142. https://doi.org/10.1111/j.1365-2419.2010.00546.x

Rahbek, C., & Graves, G. R. (2001). Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences*, *8*, 4534–4539. https://doi.org/10.1073/pnas.071034898

Redfern, J. V., Barlow, J., Ballance, L. T., Gerrodette, T., & Becker, E. A. (2008). Absence of scale dependence in dolphin–habitat models for the eastern tropical Pacific Ocean. *Marine Ecology Progress Series*, *363*, 1–14. https://doi.org/10.3354/meps07495

Redfern, J. V., Ferguson, M. C., Becker, E. A., Hyrenbach, K. D., Good, C., Barlow, J., Kaschner, K., Baumgartner, M. F., Forney, K. A., Ballance, L. T., Fauchald, P., Halpin, P., Hamazaki, T., Pershing, A. J., Qian, S. S., Read, A., Reilly, S. B., Torres, L., & Werner, F. (2006). Techniques for cetacean-habitat modeling. *Marine Ecology Progress Series*, *310*, 271–295. https://doi.org/10.3354/meps310271

Root, T., Price, J., Hall, K., Schneider, S. H., Rosenzweig, C., & Pounds, J. A. (2003). Fingerprints of global warming on wild animals and plants. *Nature*, *421*, 57–60. https://doi.org/10.1038/nature01333

Ross, L. K., Ross, R. E., Stewart, H. A., & Howell, K. L. (2015). The influence of data resolution on predicted distribution and estimates of extent of current protection of three "listed" deep-sea habitats. *PLoS One*, *10*(10), e0140061. https://doi.org/10.1371/journal.pone.0140061

Rutterford, L. A., Simpson, S. D., Jennings, S., Johnson, M. P., Blanchard, J. L., Schön, P.-J., Sims, D. W., Tinker, J., & Genner, M. J. (2015). Future fish distributions constrained by depth in warming seas. *Nature Climate Change*, *5*, 569–574. https://doi.org/10.1038/nclimate2607

Seo, C., Thorne, J. H., Hannah, L., & Thuiller, W. (2009). Scale effects in species distribution models: Implications for conservation planning under climate change. *Biology Letters*, *1*, 39–43. https://doi.org/10.1098/rsbl.2008.0476

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, *1*, 154–162. https://doi.org/10.1016/j.fishres.2008.03.006

Stommel, H. (1963). Varieties of oceanographic experience. *Science*, *139*(3555), 572–576. https://doi.org/10.1126/science.139.3555.572

Swan, A. R. H., & Sandilands, M. (1995). *Introduction to geological data analysis*. Blackwell.

Thomas, K. A., Keeler-Wolf, T., & Franklin, J. (2002). A comparison of fine- and coarse-resolution environmental variables toward predicting vegetation distribution in the Mojave dessert. In J. M. Scott, P. J. Heglund, F. Samson, J. Haufler, M. Morrison, M. Raphael, & B. Wall (Eds.), *Predicting species occurrences: Issues of accuracy and scale* (pp. 133–139). Island Press.

Thuiller, W., Brotons, L., Araújo, M. B., & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, *27*, 165–172. https://doi.org/10.1111/j.0906-7590.2004.03673.x

Thuiller, W., Lavorel, A., Araújo, M. B., Sykes, M. T., & Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Science USA*, *23*, 8245. https://doi.org/10.1073/pnas.0409902102

Tobalske, C. (2002). Effects of spatial scale on the predictive ability of habitat models for the Green Woodpecker in Switzerland. In J. M. Scott, P. J. Heglund, F. Samson, J. Haufler, M. Morrison, M. Raphael, & B. Wall (Eds.), *Predicting species occurrences: Issues of accuracy and scale* (pp. 197–205). Island Press.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In J. K. Ghosh, & J. Roy (Eds.). *Statistics: Applications and new directions* (pp. 579–604). Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Indian Statistical Institute.

Webb, T. J., & Mindel, B. L. (2015). Global patterns of extinction risk in marine and non-marine systems. *Current Biology*, *25*(4), 506–511. https://doi.org/10.1016/j.cub.2014.12.023

Wiens, J. A. (1989). Spatial scaling in ecology. *Functional Ecology*, *3*, 385–397. https://doi.org/10.2307/2389612

Wood, S. (2017). *Generalized additive models: An Introduction with R* (2nd ed). Chapman & Hall/CRC Texts in Statistical Science. https://doi.org/10.1201/9781315370279

SUPPORTING INFORMATION

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

How to cite this article: Núñez-Riboni I, Akimova A, Sell AF. Effect of data spatial scale on the performance of fish habitat models. *Fish Fish*. 2021;22:955–973. https://doi.org/10.1111/faf.12563