The predictive skill of species distribution models for plankton in a changing climate

PHILIPP BRUN[1], THOMAS KIØRBOE[1], PRISCILLA LICANDRO[2] and MARK R. PAYNE[1]

[1]Centre for Ocean Life, National Institute of Aquatic Resources, Technical University of Denmark, DK-2920 Charlottenlund, Denmark, [2]Sir Alister Hardy Foundation for Ocean Science, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK

Abstract

Statistical species distribution models (SDMs) are increasingly used to project spatial relocations of marine taxa under future climate change scenarios. However, tests of their predictive skill in the real-world are rare. Here, we use data from the Continuous Plankton Recorder program, one of the longest running and most extensive marine biological monitoring programs, to investigate the reliability of predicted plankton distributions. We apply three commonly used SDMs to 20 representative plankton species, including copepods, diatoms, and dinoflagellates, all found in the North Atlantic and adjacent seas. We fit the models to decadal subsets of the full (1958–2012) dataset, and then use them to predict both forward and backward in time, comparing the model predictions against the corresponding observations. The probability of correctly predicting presence was low, peaking at 0.5 for copepods, and model skill typically did not outperform a null model assuming distributions to be constant in time. The predicted prevalence increasingly differed from the observed prevalence for predictions with more distance in time from their training dataset. More detailed investigations based on four focal species revealed that strong spatial variations in skill exist, with the least skill at the edges of the distributions, where prevalence is lowest. Furthermore, the scores of traditional single-value model performance metrics were contrasting and some implied overoptimistic conclusions about model skill. Plankton may be particularly challenging to model, due to its short life span and the dispersive effects of constant water movements on all spatial scales, however there are few other studies against which to compare these results. We conclude that rigorous model validation, including comparison against null models, is essential to assess the robustness of projections of marine planktonic species under climate change.

Keywords: climate change, comparative model performance, continuous plankton recorder, copepod, diatom, dinoflagellate, future prediction, positive predictive value, species distribution model, true skill statistic

Received 18 September 2015; revised version received 29 February 2016 and accepted 1 March 2016

Introduction

Global climate change has led to a sustained increase in ocean temperatures in the last decades (IPCC, 2013): sea surface temperature (SST) in the North Atlantic, for instance, has increased by about 0.5 °C during the past 50 years. Climate change profoundly alters the living conditions for marine organisms through changes in water temperature, water column stratification, acidity, and gas exchange (Barton *et al.*, 2013). Marine organisms react to this stress by changing their distribution and phenology, as has been observed for planktonic and benthic invertebrates, fish and mammals (Poloczanska *et al.*, 2013). To foresee emerging problems and mitigate potential damage, model-based projections of the impact of climate change on marine ecosystems are gaining increasing currency (Brander *et al.*, 2013). However, such projections are inherently uncertain and cannot be directly validated (Araujo *et al.*, 2005; Elith & Leathwick, 2009). A careful assessment of the expected accuracy and a cautious communication of the associated uncertainties, including the various components that contribute to the total uncertainty, are therefore crucial (Ladle *et al.*, 2004; Brander *et al.*, 2013; Payne *et al.*, 2016).

Projections of biogeographical change are often based on species distribution models (SDMs). SDMs are statistical tools to analyze and predict geographical ranges of species based on observed correlations with environmental variables (Guisan & Zimmermann, 2000). SDM extrapolations to future conditions are, however, associated with considerable uncertainty (Elith & Leathwick, 2009). SDMs are inherently empirical in nature and limited by the observations and datasets at hand: they typically cannot incorporate processes such as the response of an organism to novel or extreme environmental conditions, nor processes that are not parameterized, such as biotic interactions and dispersal limitations. On the other hand, it has been suggested that some of these limitations may be of less importance for planktonic organisms, which are at the mercy of their environment (Robinson *et al.*, 2011; Beaugrand

Correspondence: Philipp Brun, tel. +45 35 88 34 80, fax +45 35 88 33 33, e-mail: pgbr@aqua.dtu.dk

et al., 2013). However, the actual skill of planktonic distribution projections in the real world has not been thoroughly evaluated.

The validation of species distribution predictions has been the subject of controversy (Planque *et al.*, 2011), particularly with regard to three main problems. Firstly, validations typically summarize model skill by single-number metrics, which have been repeatedly criticized for their inability to capture the multifaceted nature of model performance (Fielding & Bell, 1997; Allouche *et al.*, 2006; Lobo *et al.*, 2008; Peterson *et al.*, 2008; Mouton *et al.*, 2010). More meaningful validations of presence/absence predictions may be possible by employing a multimetric approach (Liu *et al.*, 2011; Jiménez-Valverde, 2012). However, developing a species distribution model is only worthwhile if its predictions are more accurate and precise than an approximation that assumes constant distribution (i.e. 'no change'). Secondly, independent testing datasets, where they exist, are typically restricted in length and rarely exceed 20 years. Directly investigating the performance of predictions with lead times of 50–100 years, as they are used to project climate change impacts, is therefore usually not possible. Thirdly, the common validation approaches summarize skill over large, heterogeneous environments without considering the spatial distribution of model errors (Lobo *et al.*, 2008; Rocchini *et al.*, 2011).

The aim of this study is to perform a real-world test of the performance of SDM predictions of plankton, which thoroughly addresses all three problems mentioned above. We use 54 years of observational data in the North Atlantic from the Continuous Plankton Recorder, one of the world's longest running and most extensive marine biological monitoring programs (Richardson *et al.*, 2006), together with three common SDMs following current best-practices. Firstly, we report detailed results for four representative focus species of phyto- and zooplankton (two copepods, one dinoflagellate and one diatom) representing a range of typical North Atlantic plankton types that have undergone significant changes in abundance/distribution in recent decades. We follow the recommendations of Jiménez-Valverde (2012) and Liu *et al.* (2011) and report multiple measures of model performance. We investigate the predictive skill of SDMs for increasing temporal separation between the prediction and training datasets and compare it to the skill of 'no change' forecasts. Moreover, we assess the spatial distribution of the predictive skill. Secondly, we repeat the main analyses on an extended set of 20 species to both identify general overarching patterns and compare predictive skill between different plankton groups. Together, these results give us a novel insight into the true skill of SDM predictions for plankton, and therefore their usefulness in projecting the impacts of climate change in marine systems.

Materials and methods

CPR data

The Continuous Plankton Recorder (CPR) survey is a monitoring program that provides long-term data on plankton in the North Atlantic and adjacent seas (Fig. S1) (Batten *et al.*, 2003; Richardson *et al.*, 2006). The CPR is towed by ships of opportunity at a nominal depth of 7 m and, being equipped with a filtering mesh of 270 μm, is adapted to sample relatively large phytoplankton and mesozooplankton taxa. Each CPR sample corresponds to 10 nautical miles and approximately 3 m³ of seawater filtered. From 1958 to 2012, about 500 phytoplankton and zooplankton taxa have been identified in more than 200 000 CPR samples analyzed.

We used presence/absence observations of 20 representative plankton species from the CPR survey (Johns, 2015), including 10 copepods, 5 diatoms, and 5 dinoflagellates (Appendix S2). All species included have a clear taxonomic identification and were actively sampled during the entire time span (except *Coscinodiscus wailesii*, see below). Furthermore, they are relatively common in the North Atlantic with distributions centered in different parts of the investigated area. From this set, we chose four well-known 'focus' species with contrasting distributions for more detailed analyses: *Calanus finmarchicus* is a key zooplankton species in the food web and its spatial distribution has been investigated in numerous studies (e.g., Helaouët & Beaugrand, 2007; Reygondeau & Beaugrand, 2011; Chust *et al.*, 2013; Hinder *et al.*, 2014). *Calanus helgolandicus* is another common copepod, that is morphologically very similar to *C. finmarchicus* but with a more southern center of distribution. *Ceratium tripos* is a common dinoflagellate in the North Atlantic, and *Coscinodiscus wailesii* is a large disc-shaped diatom. *C. wailesii* is thought to be invasive in the northeastern North Atlantic, where it was first recorded in the English Channel in 1977, (Edwards *et al.*, 2001), although other hypotheses for its sudden appearance exist as well (Gómez & Souissi, 2010).

The focus area of this study was the North Atlantic and adjacent seas, extending from 80°W to 20°E and 25°N to 73°N. Raw CPR sample data from within this region was aggregated to the same temporal (monthly) and spatial (1° × 1°) resolution as the environmental data (see below) although with separation maintained between day and night: every grid cell with at least one presence observation was defined as a presence cell. The number of samples per grid cell was also recorded, and used to correct for variations in sampling effort in time and space.

Environmental variables

We used six environmental factors to describe the occurrence of the plankton species: sea surface temperature (SST), sea surface salinity (SSS), bathymetry, westerly winds, day length,

and time (day or night) of the sample. SST, SSS, and bathymetry have previously been shown to be powerful predictors of the distributions of copepods in the North Atlantic (Helaouët & Beaugrand, 2007; Beaugrand & Helaouët, 2008; Reygondeau & Beaugrand, 2011; Beaugrand *et al.*, 2013). Wind-induced mixing may be particularly relevant for phytoplankton but the occurrence of important copepods like *C. finmarchicus* has also been linked to westerly wind stress (Fromentin & Planque, 1996; Henson *et al.*, 2009). We extend the set by day length, a proxy for light availability which is crucial to phytoplankton photosynthesis, and time of sampling, a variable that accounts for the impact of diel vertical migration of copepods on the sampling device (Hays, 1994).

We used the HadISST1 data product (Rayner *et al.*, 2003) for monthly SST for each year between 1958 and 2012. Monthly SSS data were obtained from the World Ocean Atlas 2013 (Zweng *et al.*, 2013): six individual climatologies of SSS were available covering roughly one decade each (1955–1964, 1965–1974, 1975–1984, 1985–1994, 1995–2004 and 2005–2012). Bathymetry data stems from the ETOPO1 Global Relief Model (Amante & Eakins, 2009). Monthly westerly wind data, i.e. the west component of wind velocities, was merged from two re-analysis products from the European Centre for Medium-Range Weather Forecasts: for the period 1958–2010 the ERA-20CM model ensemble estimates were used (Hersbach *et al.*, 2015) and for the years 2011 and 2012 we used the ERA-Interim model ensemble estimates (Dee *et al.*, 2011). Both products were regridded to a $1° \times 1°$ resolution and normalized and standardized to correct for biases. The Pearson correlation coefficient of the two products during the overlapping period 1979–2010 was 0.99. Day length and day/night conditions were estimated with the 'MAPTOOLS' package (Bivand & Lewin-Koh, 2015) within R (R Core Team, 2013) for all samples based on the geographic location, date, and time of sampling.

Species distribution modeling techniques

We employ three common species distribution models with differing properties: MaxEnt (Phillips *et al.*, 2004, 2006), generalized additive models (Hastie & Tibshirani, 1990; Wood, 2006) and random forests (Breiman, 2001; Cutler *et al.*, 2007). We fitted MaxEnt models using the MAXENT software 3.3.3e (http://www.cs.princeton.edu/~schapire/maxent/). We disabled threshold features to avoid overly complex model fits which could reduce the predictive quality of the models. Since MaxEnt is a presence-only model, we discarded confirmed absence observations for model fitting (but not for validation) and instead supplied information about background conditions in the North Atlantic based on 20 000–100 000 randomly selected pixels, distributed uniformly in time and space. Consequently, we also removed the two predictors linked to sampling procedure: time of sampling (day or night) and number of samples taken per grid cell. MaxEnt models were therefore fitted based on a subset of the information supplied to the other SDMs.

Generalized additive models (GAMs) were fitted with the 'MGCV' package in R (Wood, 2006). We assumed our presence/absence data to follow a binomial error distribution and used the logit link function. We included the numerical variables (SST, SSS, depth, westerly wind, day length and number of samples) as smooth terms and time of sampling (day or night) as a factor.

Random forest is a statistical method based on classification trees (Breiman, 2001; Cutler *et al.*, 2007). Classification trees are built by recursive partitioning of a class variable (here presence/absence) into subgroups with binary decisions. The decisions are based on one of the predictor variables and optimized for a maximum homogeneity within the subgroups (Breiman *et al.* 1984). In random forest, an ensemble of classification trees is fitted, each on a subset of both the training observations and the predictor variables. Presence/absence probabilities are then predicted based on majority votes of the ensemble. We used the R package 'RANDOMFOREST' to fit the random forest models (Liaw & Wiener, 2002).

Spatial interpolations

We used spatial interpolations to produce overall distribution estimates and to generate decade-wise average-distributions used as 'no change' forecasts. For the decadal averages, we grouped presence/absence observations by months and for each decade before interpolating. We interpolated spatially to $1° \times 1°$ grids using inverse-squared-distance weighting, following the protocol of Beaugrand & Reid (2012). The search radius was restricted to 250 km and estimates were only made for pixels with a minimum of five samples within the search radius.

Model performance

Threshold-dependent metrics. Threshold-dependent model performance metrics were derived from the confusion matrix, a 2 by 2 matrix summarizing the four possibilities of the relationship between presence/absence model predictions and testing dataset (Table 1).

The prevalence error is the normalized difference between observed prevalence and predicted prevalence:

$$\text{Prevalence error} = \frac{\frac{a+c-(a+b)}{a+b+c+d}}{\frac{a+c}{a+b+c+d}} = \frac{c-b}{a+c} \qquad (1)$$

where $a–d$ are the elements of the confusion matrix from Table 1. The positive predictive value (PPV), the probability of

Table 1 Confusion matrix for the validation of presence/absence data; *a*: correctly predicted presences; *b*: erroneously predicted presences (commission errors); *c*: erroneously predicted absences (omission errors); *d*: correctly predicted absences

	Testing data	
	Presence	Absence
Model		
Presence	*a*	*b*
Absence	*c*	*d*

observing a species at locations where the model predicts presences, is:

$$PPV = \frac{a}{a+b} \qquad (2)$$

and the negative predictive value (NPV), the probability that the species is not detected where the model predicts absences, is given by:

$$NPV = \frac{d}{c+d} \qquad (3)$$

True skill statistic, an overall performance metric (TSS; Allouche *et al.*, 2006), is estimated as

$$TSS = \text{sensitivity} + \text{specificity} - 1, \qquad (4)$$

where sensitivity is the fraction of correctly predicted presences

$$\text{Sensitivity} = \frac{a}{a+c} \qquad (5)$$

and specificity is the fraction of correctly predicted absences

$$\text{Specificity} = \frac{d}{b+d} \qquad (6)$$

Model outputs and spatial interpolations were converted into binary presence/absence predictions by selecting a threshold probability: if the probability of occurrence exceeds this value, it is considered as a presence for the purpose of evaluating model performance. Several criteria to select thresholds exist, and the choice can strongly affect the resulting model performance (Freeman & Moisen, 2008). Here, we chose the threshold probability for which the predicted prevalence is identical to the observed prevalence, a criterion that has been shown to yield good presence/absence predictions performance (Freeman & Moisen, 2008).

Area under the curve. We supplemented our set of threshold-dependent model performance metrics with another overall performance metric, the area under the receiver operating characteristic curve (AUC), which is perhaps the most common model performance metric in species distribution modeling (Jiménez-Valverde, 2012). AUC is a threshold-independent method that indicates the average value of sensitivity over all possible values of specificity (Liu *et al.*, 2011).

Analyses

We investigated the predictive skill of our models in two phases (Table 2). In the first phase, we investigated the skill of SDM projections in detail for our focal species: we first fitted the models to the observations of the full time period to assess achievable skill with all available information. Then, we split the observations into decadal subsets, fitted the model to these datasets, and made predictions forward and backward in time to assess how predictive performance changes with temporal distance between the training and validation datasets. Finally, we used the decadal predictions to assess the spatial variations in model performance.

In the second phase, we repeated the main analyses on the full set of 20 plankton species. The aim here was to identify general patterns and to compare model predictions of copepods, diatoms, and dinoflagellates.

Phase one included four different procedures (Table 2). Firstly (phase 1a), model skill for the full time span was assessed by performing fivefold cross-validations. In total 111 351 presence/absence observations could be matched up with environmental data and were used to model the distributions of *C. finmarchicus*, *C. helgolandicus*, and *C. tripos*. For *C. wailesii*, we only considered the time span after its first occurrence in the CPR data in 1977 (62 769 observations). The years within the considered time spans were randomly split into five equally sized groups to generate the cross-validation subsets. We report means and 95% confidence intervals of TSS, AUC, PPV, and NPV. We used the full datasets (not cross-validation subsets) to create distribution estimates, visualizing the average distributions for the periods investigated.

We then (phase 1b) applied the analysis described above to the individual decades (training) and use the models to make projections into all the other decades (validation). We evaluate model skill within the same decade again with fivefold cross-validation. From this evaluation we additionally obtain the necessary thresholds to divide the continuous model predictions into binary presence/absence predictions. We thus determine the thresholds at the time where the model was trained and not at the time where we evaluate it (since this would be the unknown future in a practical application). Temporal projections are based on models tuned on the full data of the decade (not 80% as used for within-decade cross-validation). We evaluated all model projections and grouped them based on the temporal difference between the time when the model was trained and the time when it was evaluated. For each group we report means and standard deviations of TSS, PPV, and prevalence error. TSS scores for each group were then further compared to the scores of 'no change' forecasts (phase 1c) using two-sided, paired *t*-tests (6, 10, 8, 6, 4, and 2 data pairs for 0, 10, 20, 30, 40, and 50 years absolute time difference). Finally (phase 1d), we used all temporally extrapolated predictions from the best model and aggregated them into one by one degree cells to examine how performance varies in space. TSS scores are reported for cells for which at least five match-ups exist between model predictions and both presence and absence observations.

Phase 2 consisted of repeating phases 1a to 1c for the full set of 20 species (Table 2). We assessed TSS and PPV for the best models trained on the full time period. Based on these results, we also compared SDM skill between copepods, diatoms and dinoflagellates using Tukey honest significant difference (HSD) tests. Finally, for all species we produced decadal SDM projections and evaluated their absolute prevalence error and TSS as a function of temporal projection distance. Obtained TSS scores were compared again with 'no change' forecasts using *t*-tests to examine for which plankton groups, and for which models the most useful SDM predictions can be made.

Table 2 Overview over the analyses performed in this study. Items listed in square brackets are included in the supplementary material

Phase	Main focus	Training time span(s)	Considered species	SDM technique(s)	Validation metric(s)	Temporal projections
1a	Overall model performance	1958–2012	*Calanus finmarchicus* *Calanus helgolandicus* *Ceratium tripos* *Coscinodiscus wailesii*	MaxEnt GAM Random forest	TSS AUC PPV NPV	Interpolations only
1b	Absolute SDM skill under temporal extrapolation	Six decadal subsets*	*Calanus finmarchicus* *Calanus helgolandicus* *Ceratium tripos* *Coscinodiscus wailesii*	MaxEnt Random forest [GAM]	TSS prevalence error [PPV]	Interpolations and extrapolations
1c	SDM skill under temporal extrapolation relative to 'no change'	Six decadal subsets*	*Calanus finmarchicus* *Calanus helgolandicus* *Ceratium tripos*	Random forest	ΔTSS†	Interpolations and extrapolations
1d	Spatial distribution of predictive performance	Six decadal subsets*	*Calanus finmarchicus* [other focal species]	Random forest	Pixel-wise TSS	Extrapolations only
2	Generalization and comparison of plankton groups	1958–2012 and Six decadal subsets*	10 copepods 5 diatoms 5 dinoflagellates	Maxent GAM Random forest	TSS PPV prevalence error ΔTSS†	Interpolations and extrapolations

*Subsets are 1958–1964, 1965–1974, 1975–1984, 1985–1994, 1995–2004, 2005–2012.
†ΔTSS represents differences in TSS between SDM predictions and 'no change' forecasts as inferred by paired *t*-tests.

Results

Full models

Distributions. Based on its long-term average distribution *C. finmarchicus* was found to be the most widespread of the four focal species, mainly occurring north of 40°N and most frequently in the Labrador Sea, with probabilities of presence of up to 100%. *C. helgolandicus* was estimated to occur mainly in the eastern North Atlantic, in particular in the Celtic Sea, but also in the subtropical central North Atlantic. *C. tripos* was estimated to mainly grow in the western and northeastern North Atlantic, mostly close to the coast and seasonally restricted. Presence predictions for *C. wailesii* were found in parts of the presence areas of the former species, but this diatom was estimated to only occur during short time spans and not every year: few areas were found where the probability of presence of *C. wailesii* exceeded 5%. Mapped distribution estimates based on spatial interpolations and random forest predictions are shown in Fig. S3.

Model performance. We evaluated the skill of SDMs trained on the full time period with four model performance metrics (phase 1a). The model performance metrics suggested similar performance rankings of the models (Fig. 1). Random forest models typically performed best, followed by GAMs and MaxEnt. We also found clear differences in skill depending on the tested species: the metrics mostly identified highest skill for models of *C. finmarchicus* and *C. helgolandicus*, followed by *C. tripos*, while the models of *C. wailesii* achieved lowest performance.

However, model skill considerably varied according to the metric. The probability of correctly predicting presence (PPV) ranged from 15% for MaxEnt models of *C. wailesii* to 77% for random forest models of *C. finmarchicus* (Fig. 1c). These numbers were inversely related to the probability of correctly predicting absence (NPV) (Fig. 1d). Moreover, the two overall performance metrics AUC and TSS fundamentally disagreed about the skill of fits for *C. wailesii*: models for this diatom achieved highest AUC scores on average (mean AUC *C. wailesii* = mean AUC *C. helgolandicus* = 0.86), while TSS scores were by far the lowest (Fig. 1a, b). Overall, the results show that purpose and design of model performance metrics can strongly influence our perception of SDM skill.

Temporal extrapolation

Absolute SDM skill. No sustained loss in overall model performance (TSS) and in the precision of the presence predictions (PPV) was found for most focal species when predictions were made to increasingly distant
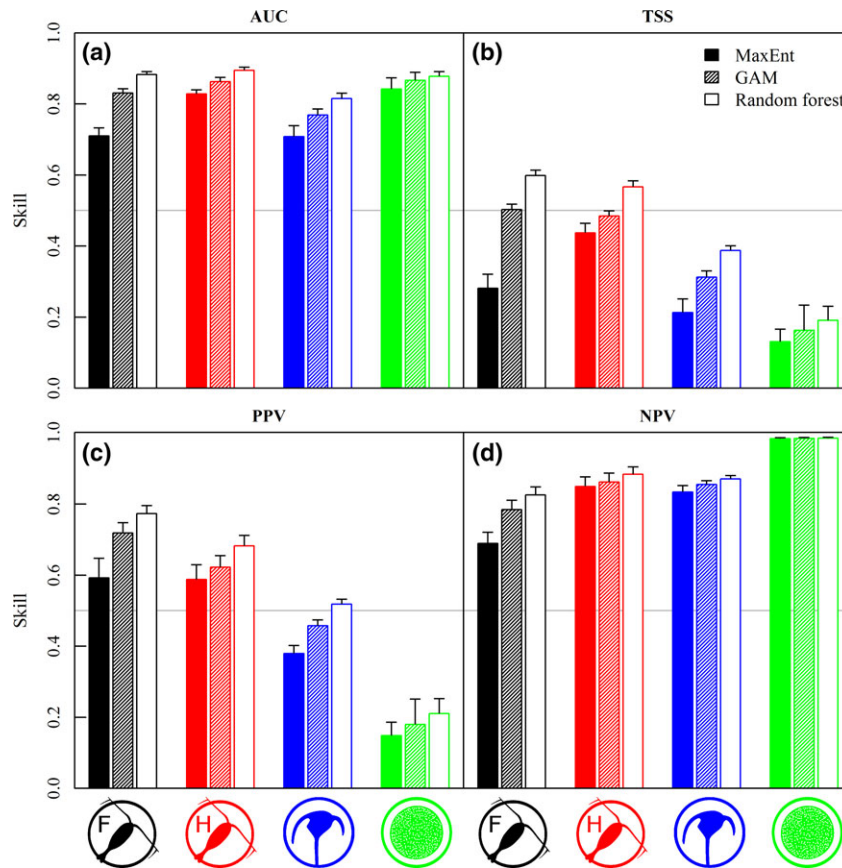
Fig. 1 Means (bars) and 95% confidence intervals (error bars) of model performance measures for all combinations of focal species and SDMs (phase 1a). Models are based on the entire dataset from 1958 to 2012 (1978–2012 for *Coscinodiscus wailesii*). Depicted measures are AUC (a), TSS (b), PPV (c), and NPV (d). Colors represent *Calanus finmarchicus* (black), *Calanus helgolandicus* (red), *Ceratium tripos* (blue), and *Coscinodiscus wailesii* (green).The different metrics are evaluated for MaxEnt (filled bars), GAMs (dashed bars), and random forest models (white bars).

times (phase 1b) (Figs 2a–c, S4). However, in some cases the curves showed distinct slopes; GAM predictions for *C. finmarchicus*, for instance, continuously increased in TSS the further into the future a projection was made (Fig. S4). The slopes of PPV scores were even steeper; predictions for *C. finmarchicus* tended to be more precise when they were made further into the past while the opposite was the case for *C. helgolandicus* (Fig. S4).

For *Coscinodiscus wailesii*, on the other hand, a rapid decrease in PPV and TSS consistently occurred for all SDMs when predictions were made into the past or the future, and the metrics approached 0 (no skill) for temporal distances of 20–30 years. Based on our set of observational data, the SDMs employed here were not able to make useful distribution predictions of *C. wailesii*, even though traditional performance metrics (AUC) suggested that this was a good model.

Errors in prevalence tended to increase for predictions into more distant times (Fig. 2d–f). Models

increasingly overestimated the prevalence of *C. finmarchicus* when projected into the future while the prevalence was more and more underestimated for predictions further into the past. The prevalence of *C. helgolandicus*, on the other hand, tended to be overestimated for predictions into the past but underestimated for future predictions. For *C. tripos* temporal trends were also apparent, but the magnitude was lower and the direction was not consistent among SDMs.

SDM skill relative to 'no change'. Predictions of the different SDMs showed variable skill relative to the performance of 'no change' forecasts (phase 1c). For *C. finmarchicus*, *C. helgolandicus,* and *C. tripos* MaxEnt predictions performed significantly worse than 'no change' forecasts for most extrapolation distances. GAM predictions did not perform significantly better than 'no change' forecasts for any time-lag tested. Only random forest predictions achieved higher skill for
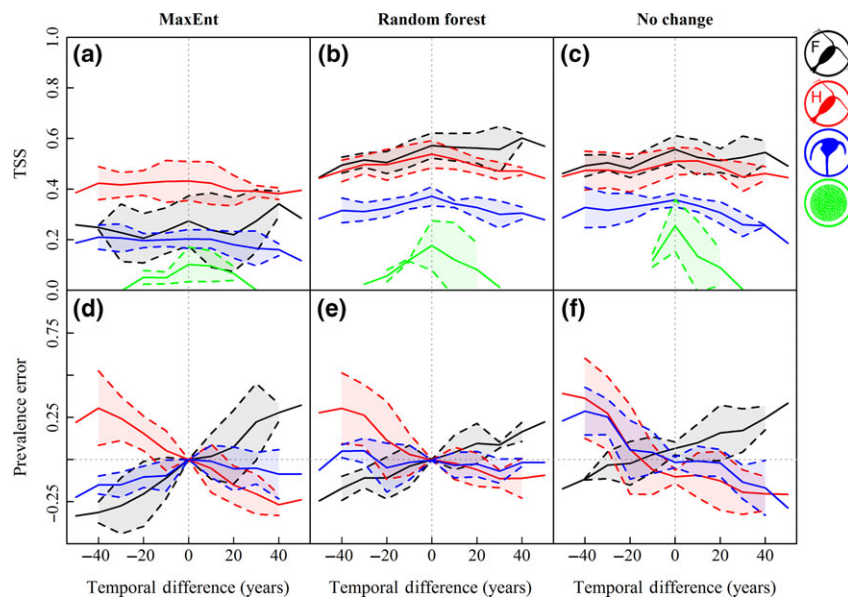
Fig. 2 Model performance metrics as a function of temporal difference (years) between training period and testing period for *Calanus finmarchicus* (black), *Calanus helgolandicus* (red), *Ceratium tripos* (blue), and *Coscinodiscus wailesii* (green). TSS is shown in the top row; prevalence error is shown in the bottom row. Columns represent the SDMs MaxEnt, and random forest, as well as 'no change' forecasts based on spatial interpolations. Solid lines depict the mean values of groups of predictions with equal temporal difference; dashed lines indicate means ± standard deviations. Prevalence error for *C. wailesii* is much larger than the scale and has been excluded for simplicity. Positive time differences correspond to future predictions.

temporal differences up to 30 years, in particular for *C. finmarchicus*. Detailed results of the comparisons are shown in Fig. S5.

Spatial distribution of predictive performance

Model performance showed substantial spatial variation for the focal species (phase 1d). The model predictions for *C. finmarchicus* (Fig. 3a) showed the highest TSS in the northern-central North Atlantic, while in the southern-central North Atlantic and in the North Sea the models had no skill in differentiating between presences and absences. Strong spatial variation was also seen for model predictions for *C. helgolandicus*, and *C. tripos* (Fig. S6). For *C. wailesii* only 90 cells located either in the southern North Sea or close to the New England coast could be evaluated (Fig. S6): spatial analysis confirmed that models for this species have little skill.

Local performance tended to be best in pixels where the species occurred in moderate frequencies. Pixels with poor local TSS predominantly occurred in areas where the species were expected to occur with a probability of <20% (see superimposed black lines in Figs 3a and S6). In the case of *C. finmarchicus* maximum TSS was found in pixels with moderate observed prevalence while TSS was poor in pixels where *C. finmarchicus* was either present or absent most of the time (Fig. 3b).

Generalization and comparison of plankton groups

True skill statistic and positive predictive value values of random forest models were on average relatively low and varied between plankton groups when considering the full timespan and all the 20 species (phase 2). Presences were predicted most precisely for copepods (average PPV = 0.50) and were significantly better ($P \leq 0.05$, Tukey HSD test) than for diatoms (PPV = 0.29) (Fig. 4a). The PPV for dinoflagellates (PPV = 0.48) was between the scores of the other two groups and not significantly different from any of them. For TSS we found two distinct groupings: Copepods and dinoflagellates both had significantly higher TSS scores than diatoms according to a Tukey HSD test, while no difference was found between their TSS scores (Fig. 4a).

We did not find a significant drop in TSS for increasing temporal differences between model training and model validation when averaged over all species, yet the absolute prevalence error was continuously increasing (Fig. 4b, c). For all SDM types TSS curves were flat but the level for the presence-only model (MaxEnt) was clearly below the levels for the presence/absence models. The absolute prevalence error, however, continuously increased with increasing time-lags in both directions of time. At absolute time lags of 50 years, the prevalence of species was over- or underestimated on average by about 50%, independent of SDM type.
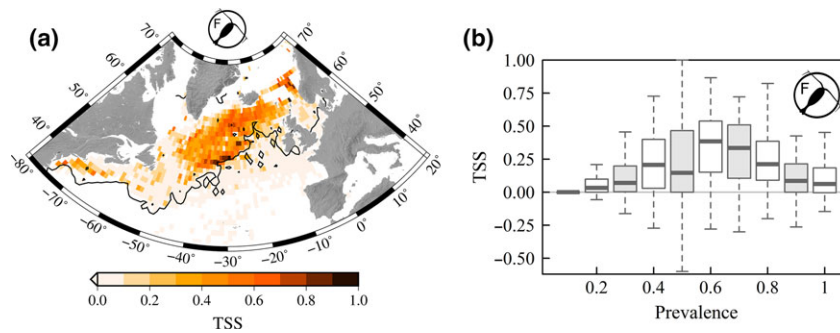
Fig. 3 Spatial patterns of TSS for random forest model predictions for *Calanus finmarchicus* (a). The isocline of 20% probability of presence is superimposed (black line). Only 1°×1° cells with at least five match-ups between model predictions and both presence and absence observations are shown. The box plots in panel (b) illustrate the relationship between pixel-wise TSS and observed prevalence. Thick lines on box plots illustrate median, boxes represent the inter quartile ranges and whiskers encompass the 95% confidence intervals. A TSS of zero corresponds to a random 'coin-toss'.
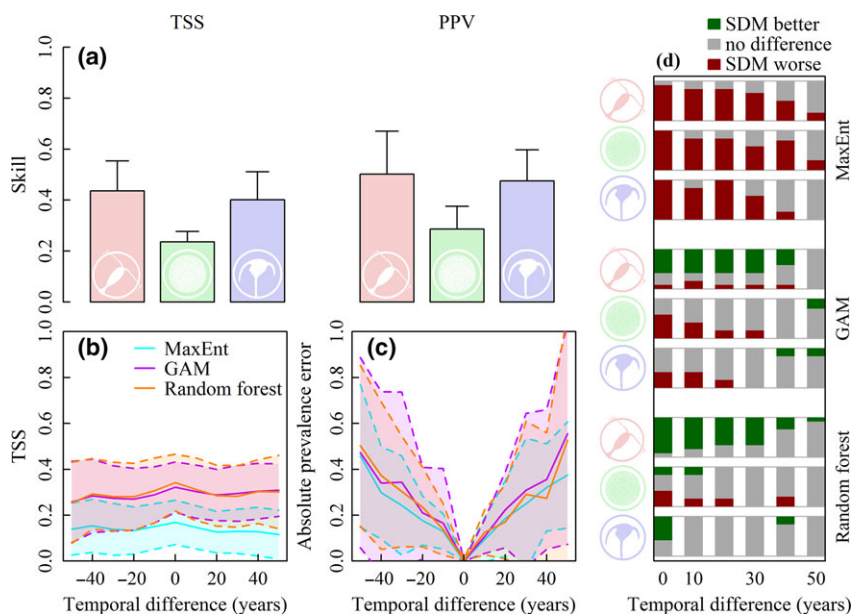

Fig. 4 Generalization of the analyses to a broader set of 20 species. In panel (a) TSS and PPV are shown for models trained on the full dataset for copepods (red), diatoms (green) and dinoflagellates (blue). Bars depict mean values; error bars illustrate standard deviations. Panels (b) and (c) show TSS and absolute prevalence error, respectively, grouped by SDM types and as a function of time lag between model training and validation. MaxEnt models are shown in cyan, GAMs in purple and random forest models in orange; solid lines depict means across all species, dashed lines show means ± standard deviations. Panel (d) shows comparisons of TSS between SDM predictions and 'no change' forecasts (inverse squared distance interpolation) for different temporal extrapolation distances (absolute time difference) for copepods (top), diatoms (middle), and dinoflagellates (bottom). Barplots depict fractions of species for which SDM predictions perform significantly better (green), equivalently (gray), or significantly worse (red) than no change predictions based on paired *t*-tests.

Species distribution model predictions significantly better than 'no change' forecasts were only found regularly for copepods with presence/absence models (Fig. 4d). For all plankton groups, MaxEnt models mostly performed worse than 'no change' forecasts. Random forest models for dinoflagellates produced predictions with TSS scores roughly equivalent to 'no change' forecasts. Otherwise, presence/absence SDM predictions (GAM and random forest) tended to be inferior to 'no change' forecasts for the two phytoplankton groups. Only in the case of copepods were presence/absence SDM predictions mostly significantly better than 'no change' forecasts for time lags of up to 30 years, particularly in the case of random forest models.

Discussion

Our analyses show that SDM predictions can provide valuable insight into how climate change will affect copepod biogeography, given that sound data and relevant environmental predictors are used. However, for the phytoplankton species as well as for the presence-only models tested here, this was not the case – predictions typically performed either as good as or worse than 'no change' forecasts (Fig. 4d). Furthermore, we identified a growing prevalence error for predictions with more temporal distance in time from their training dataset, suggesting that there is more to plankton prevalence than what these SDMs consider. Prevalence errors of climate-change projections may even exceed the percentages found here, given the growing nature of the error and the longer temporal horizon of such studies (often until 2100). Finally, we found strong variations of model skill in space: model performance was poor in areas where the species were less common, including the edges of the distribution ranges. This may have important implications on the accuracy of range shift estimates, which are a commonly reported outcome of SDM predictions in the literature (e.g., Weinmann *et al.*, 2013), as well as when predictions are intended to be used for ecosystem management purposes, with a local domain of interest.

These novel results deserve attention as they highlight the importance of a rigorous evaluation of the accuracy and precision of species distribution forecasts. In the following, we first discuss the relevance of our comparative approach, then we describe how the commonly used model performance metrics can promote overestimations of SDM skill and finally, we suggest three ecological explanations why our plankton distribution predictions did not perform better.

The two main conditions for meaningful comparative model validation have been fulfilled in this study: there were considerable changes in plankton distributions and the SDMs were fed with important environmental factors that changed. Comparative validations of model predictions relative to 'no change' predictions have been used in decadal climate forecasts (e.g., Matei *et al.*, 2012), but to our knowledge, this is the first time such an approach has been applied to set the skill of SDM predictions in context. Employing such a validation approach is only relevant if substantial distributional changes occur within study period, which was evident for the plankton species investigated (Fig. S7 and Poloczanska *et al.*, 2013). When comparing the distributions from 2005 to 2012 to those from 1958 to 1964 during the month of maximum prevalence, *C. finmarchicus* was lost from 21% of the cells where it was formerly present, and appeared new in 13%. For

C. helgolandicus and *C. tripos* the corresponding numbers were 23% or more. These changes arise from the combined effects of spatial relocations and possible shifts in phenology and should ideally be captured by model predictions.

However, to capture such change, SDMs need relevant environmental information. In our study, three gridded environmental variables were available with full temporal resolution for the entire investigated period, while a few more well-resolved alternatives would exist for more recent years, such as chlorophyll *a* concentration: other potentially important variables, such as nutrient fields, do not exist in the temporal and spatial resolutions required to enable their application in such modeling. Nevertheless, temperature and salinity are consistently among the most important environmental predictors in plankton SDMs (Helaouët & Beaugrand, 2007; Irwin *et al.*, 2012; Chust *et al.*, 2013). Furthermore, many environmental variables in the North Atlantic are highly correlated (Helaouët & Beaugrand, 2007) and intercorrelated sets of predictor variables may not greatly improve the performance of SDMs (Dormann *et al.*, 2012).

Model performance metrics did not always yield sensible conclusions – only by considering multiple metrics could a full picture of model skill be developed. For example, the overall performance metrics TSS and AUC assigned opposite relative skill to the models for *C. wailesii* and, in some cases, TSS and PPV oddly improved for predictions to increasingly distant times. These unexpected results arose from a weakness shared by all commonly used model performance metrics: they are either affected by the prevalence of a species or by erroneous prevalence estimates (Fielding & Bell, 1997; Allouche *et al.*, 2006; Lobo *et al.*, 2008; Mouton *et al.*, 2010). AUC, for instance, overestimates the skill of rare species (Lobo *et al.*, 2008), a common problem that also caused the inconsistent performance of *C. wailesii* models. A multimetric approach, with its capacity to unravel the impact of the different factors that affect the score of a single-value statistic, may more accurately identify the skill of a prediction and where its weaknesses are.

One of the weaknesses revealed by this study was the precision of presence predictions, which may be restricted by the characteristics of planktonic organisms and their marine habitat. When using the full dataset, the average precision for copepod predictions, the most suitable group tested, was only as good as coin a toss (50%) for the best SDM type. As PPV has rarely been estimated for SDM predictions (Liu *et al.*, 2009), we unfortunately cannot directly compare these numbers with those of other taxa. Nevertheless, PPV for plankton may be particularly limited: plankton, and in

particular phytoplankton, are short-lived organisms which undergo distinct phases of boom and bust (Mackas *et al.*, 2012; González Taboada & Anadón, 2014), creating local patchiness. Ocean currents and eddies lead to further local structuring of the pelagic environment and disperse plankton patches laterally beyond suitable areas (Barton *et al.*, 2010). The environmental predictors used here are too coarse to resolve many of these processes, a limitation that may be relaxed by using environmental predictors measured at the time of sampling. However, even co-sampled environmental information will not capture short-term phenomena like nutrient plumes which affect the prevalence of plankton with a time lag. Plankton occurrences may be constrained by the species' ecological niche but they are transitory, moving targets, which appear to be hard to accurately predict.

Modeling abiotic constraints alone may not be sufficient to understand how the prevalence of plankton changes. An obvious factor that is typically ignored by SDMs is biotic interactions (Elith & Leathwick, 2009), including both interspecific [e.g. alterations of food webs or trophic mismatch (Richardson & Schoeman, 2004; Kirby & Beaugrand, 2009)], and intraspecfic (e.g. competition and mating) forms. Biotic interactions may represent the stochastic part of the relationship between the spatial distribution of copepods and environmental conditions (Beaugrand *et al.*, 2013): however, they can also affect the abundance of the species and may thus be essential to understand how prevalence changes with time.

Relationships between the occurrence of species and environment are unique and may even vary for the same species within a large study area, undermining the potential of general SDM formulations across many species. For example, random forest models for *C. finmarchicus*, consistently made poor predictions for the North Sea, an area where the species regularly occurs. The environmental conditions in the North Sea are strongly affected by climate driven inflow events (Hjøllo *et al.*, 2009), which may be dominated by warm waters from the North Atlantic or by cold waters from the Norwegian Sea. Inflowing deep water from the Norwegian Sea, an important overwintering area of *C. finmarchicus*, directly transports the spring population of the species into the North Sea (Heath *et al.*, 1999). The necessary exclusion of such processes from our SDMs therefore clearly limits their skill in this region, even though much better skill is seen for the same models and species in other regions (Fig. 3).

Similarly, the history of *C. wailesii* poses a challenge to species distribution modeling. The species is invasive in the eastern North Atlantic and was expanding its range during the observed period with an initial phase of rather high abundance that eventually leveled off at lower numbers (Edwards *et al.*, 2001). Thus, for this species, a fundamental assumption of species distribution modeling was violated, i.e., the geographic distribution of *C. wailesii* was not in equilibrium with the environment (Phillips *et al.*, 2006). The broad tolerance range of *C. wailesii* in our key predictors temperature and salinity (Dürselen & Rick, 1999; Irwin *et al.*, 2012) further complicates the problem. SDMs for this cosmopolite may be improved through the inclusion of observations from its entire native range and more relevant environmental variables, such as river discharges in coastal areas (Gómez & Souissi, 2010; Jiménez-Valverde *et al.*, 2011). Yet, it may be doubted whether future predictions of a useful accuracy will be possible for this species at all. Powerful SDM algorithms do not guarantee reasonable predictions if a species' ecology is not thoroughly incorporated into them.

In summary, we have shown that the reliability and robustness of climate change projections of plankton biogeography based on species distribution modeling is not assured by powerful models and extensive datasets. The potential performance of a projection has to be thoroughly assessed, and critically communicated. Traditional single-value statistics are barely sufficient summaries of model performance in this context: their dependence on prevalence makes them unstable and difficult to compare. By validating with multiple uncommon approaches, we have shown that the understanding of model performance can be significantly improved. Reporting model performance with multiple metrics, including rarely used measures of predictive precision (i.e. PPV and NPV) provides a more exhaustive overall picture of model skill. Examining skill relative to 'no change' forecasts informs about the usefulness of the predictions, and resolving spatial aspects of model performance highlights local areas of poor skill. By incorporating such approaches routinely into the development and reporting of SDM climate-change forecasts, a more realistic and badly needed understanding of the accuracy of the predictions can be established.

References

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Amante C, Eakins BW (2009) *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, Boulder, Colorado.

Araujo MB, Pearson RG, Thuiller W *et al.* (2005) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Barton AD, Dutkiewicz S, Flierl G *et al.* (2010) Patterns of diversity in marine phytoplankton. *Science (New York, NY)*, **327**, 1509–1511.

Barton AD, Pershing AJ, Litchman E *et al.* (2013) The biogeography of marine plankton traits. *Ecology Letters*, **16**, 522–534.

Batten S, Clark R, Flinkman J *et al.* (2003) CPR sampling: the technical background, materials and methods, consistency and comparability. *Progress in Oceanography*, **58**, 193–215.

Beaugrand G, Helaouët P (2008) Simple procedures to assess and compare the ecological niche of species. *Marine Ecology Progress Series*, **363**, 29–37.

Beaugrand G, Reid PC (2012) Relationships between North Atlantic salmon, plankton, and hydroclimatic change in the Northeast Atlantic. *ICES Journal of Marine Science*, **69**, 1549–1562.

Beaugrand G, Mackas D, Goberville E (2013) Applying the concept of the ecological niche and a macroecological approach to understand how climate influences zooplankton: advantages, assumptions, limitations and requirements. *Progress in Oceanography*, **111**, 75–90.

Bivand R, Lewin-Koh N (2015) maptools: Tools for reading and handling spatial objects. R package version 0.8-37, Available at: https://CRAN.R-project.org/package=maptools (accessed 1 August 2015).

Brander K, Neuheimer A, Andersen KH *et al.* (2013) Overconfidence in model projections. *ICES Journal of Marine Science*, **70**, 1065–1068.

Breiman L (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, CA, USA.

Chust G, Castellani C, Licandro P *et al.* (2013) Are Calanus spp. shifting poleward in the North Atlantic? A habitat modelling approach. *ICES Journal of Marine Science*, **71**, 241–253.

Cutler DR, Edwards TC, Beard KH *et al.* (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Dee DP, Uppala SM, Simmons AJ *et al.* (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system (eds Levitus S, Mishonov A). *Quarterly Journal of the Royal Meteorological Society*, **137**, 553–597.

Dormann CF, Elith J, Bacher S *et al.* (2012) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Dürselen C-D, Rick H-J (1999) Spatial and temporal distribution of two new phytoplankton diatom species in the German Bight in the period 1988 and 1996. *Sarsia*, **84**, 367–377.

Edwards M, John AWG, Johns DG *et al.* (2001) Case history and persistence of the non-indigenous diatom *Coscinodiscus wailesii* in the north-east Atlantic. *Journal of the Marine Biological Association of the UK*, **81**, 207–211.

Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Freeman EA, Moisen GG (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 48–58.

Fromentin J-M, Planque B (1996) Calanus and environment in the eastern North Atlantic. 2. Role of the North Atlantic Oscillation on *Calanus finmarchicus* and *C. helgolandicus*. *Marine Ecology Progress Series*, **134**, 11–118.

Gómez F, Souissi S (2010) The diatoms odontella sinensis, coscinodiscus wailesii and thalassiosira punctigera in the european atlantic: recent introductions or overlooked in the past? *Fresenius Environmental Bulletin*, **19**, 1424–1433.

González Taboada F, Anadón R (2014) Seasonality of North Atlantic phytoplankton from space: impact of environmental forcing on a changing phenology (1998–2012). *Global Change Biology*, **20**, 698–712.

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hastie TJ, Tibshirani RJ (1990) *Generalized Additive Models*, vol. 43. Chapman and Hall, London.

Hays GC (1994) Zooplankton avoidance activity. *Nature*, **376**, 650–650.

Heath MR, Backhaus JO, Richardson K *et al.* (1999) Climate fluctuations and the spring invasion of the North Sea by *Calanus finmarchicus*. *Fisheries Oceanography*, **8**, 163–176.

Helaouët P, Beaugrand G (2007) Macroecology of *Calanus finmarchicus* and *C. helgolandicus* in the North Atlantic Ocean and adjacent seas. *Marine Ecology Progress Series*, **345**, 147–165.

Henson SA, Dunne JP, Sarmiento JL (2009) Decadal variability in North Atlantic phytoplankton blooms. *Journal of Geophysical Research*, **114**, C04013.

Hersbach H, Peubey C, Simmons A, Berrisford P, Poli P, Dee D (2015) ERA-20CM: a twentieth-century atmospheric model ensemble. *Quarterly Journal of the Royal Meteorological Society*, **141**, 2350–2375.

Hinder SL, Gravenor MB, Edwards M *et al.* (2014) Multi-decadal range changes vs. thermal adaptation for north east Atlantic oceanic copepods in the face of climate change. *Global Change Biology*, **20**, 140–146.

Hjøllo SS, Skogen MD, Svendsen E (2009) Exploring currents and heat within the North Sea using a numerical model. *Journal of Marine Systems*, **78**, 180–192.

IPCC (2013) *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, Cambridge.

Irwin AJ, Nelles AM, Finkel ZV (2012) Phytoplankton niches estimated from field data. *Limnology and Oceanography*, **57**, 787–797.

Jiménez-Valverde A (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, **21**, 498–507.

Jiménez-Valverde A, Peterson AT, Soberón J *et al.* (2011) Use of niche models in invasive species risk assessments. *Biological Invasions*, **13**, 2785–2797.

Johns DG (2015) Raw data for 20 plankton species (copepods, dinoflagellates and diatoms) in the North Atlantic (25–73N, 80W-20E) from 1958 to 2012 as recorderd by the Continuous Plankton Recorder. doi: 10.7487/2015.331.1.932.

Kirby RR, Beaugrand G (2009) Trophic amplification of climate warming. *Proceedings of the Royal Society of London B: Biological Sciences*, **276**, 4095–4103.

Ladle RJ, Jepson P, Araújo MB *et al.* (2004) Dangers of crying wolf over risk of extinctions. *Nature*, **428**, 799.

Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News*, **2**, 18–22.

Liu C, White M, Newell G (2009) Measuring the accuracy of species distribution models: a review. In: *Proceedings 18th World IMACs/MODSIM Congress* (eds Anderssen RS, Braddock RD, Newham LTH), pp. 4241–4247. Cairns, Australia.

Liu C, White M, Newell G (2011) Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, **34**, 232–243.

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Mackas DL, Greve W, Edwards M *et al.* (2012) Changing zooplankton seasonality in a changing ocean: comparing time series of zooplankton phenology. *Progress in Oceanography*, **97–100**, 31–62.

Matei D, Pohlmann H, Jungclaus J *et al.* (2012) Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model. *Journal of Climate*, **25**, 8502–8523.

Mouton AM, De Baets B, Goethals PLM (2010) Ecological relevance of performance criteria for species distribution models. *Ecological Modelling*, **221**, 1995–2002.

Payne MR, Barange M, Cheung WWL *et al.* (2016) Uncertainties in projecting climate-change impacts in marine ecosystems. *ICES Journal of Marine Science: Journal du Conseil*, fsv231.

Peterson AT, Papeş M, Soberón J (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, **213**, 63–72.

Phillips SJ, Avenue P, Park F *et al.* (2004) *A Maximum Entropy Approach to Species Distribution Modelling*. Proceedings of the 21st International Conference on Machine Learning, p. 83. ACM Press, Banff, Canada.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Planque B, Bellier E, Loots C (2011) Uncertainties in projecting spatial distributions of marine populations. *ICES Journal of Marine Science*, **68**, 1045–1050.

Poloczanska ES, Brown CJ, Sydeman WJ *et al.* (2013) Global imprint of climate change on marine life. *Nature Climate Change*, **3**, 919–925.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (accessed 15 December 2015).

Rayner NA, Parker DE, Horton EB *et al.* (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, **108**, 4407.

Reygondeau G, Beaugrand G (2011) Future climate-driven shifts in distribution of *Calanus finmarchicus*. *Global Change Biology*, **17**, 756–766.

Richardson AJ, Schoeman DS (2004) Climate impact on plankton ecosystems in the Northeast Atlantic. *Science (New York, NY)*, **305**, 1609–1612.

Richardson AJ, Walne AW, John AWG *et al.* (2006) Using continuous plankton recorder data. *Progress in Oceanography*, **68**, 27–74.

Robinson LM, Elith J, Hobday AJ *et al.* (2011) Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, **20**, 789–802.

Rocchini D, Hortal J, Lengyel S *et al.* (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Weinmann AE, Rödder D, Lötters S *et al.* (2013) Traveling through time: the past, present and future biogeographic range of the invasive foraminifera Amphistegina spp. in the Mediterranean Sea. *Marine Micropaleontology*, **105**, 30–39.

Wood S (2006) *Generalized Additive Models: An Introduction with R* (eds Carlin BP, Chatfiled C, Tanner M, Zidek J), CRC Press, Boca Raton, FL.

Zweng MM, Reagan JR, Antonov JI *et al.* (2013) *World Ocean Atlas* 2013, *Volume 2*: *Salinity* (eds Levitus S, Mishonov A), NOAA Atlas NESDIS 74, Silver Spring, MD.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Spatial distribution of CPR observations.
Appendix S2. List of species investigated.
Appendix S3. Comparison of spatially interpolated distributions and random forest predictions.
Appendix S4. Further measures for decade-wise performance.
Appendix S5. Comparative performance of the focal species.
Appendix S6. Spatial patterns of TSS for remaining focal species.
Appendix S7. Turnover of presence cells.