

Appendix 6.A. Statistical Methods

6.A.1. Model Overview

We adopt a two-stage approach that separates a model of the presence probability of a species from a model of its relative abundance when it is present. This approach has been successfully used to model highly zero-inflated marine distribution data (e.g., Stefánsson, 1996; Ver Hoef and Jansen, 2007; Winter et al., 2011). This technique is also referred to in the statistical literature as a hurdle model (Cragg, 1971; Potts and Elith, 2006; Ver Hoef and Jansen 2007). In our case we refer to the two parts of the model as Stage I and Stage II. Stage I models the probability, $p_i(x,y)$, that species or group i is observed in a survey at location (x,y) in a given season (models were repeated for each season, but seasonal subscripts are omitted for simplicity):

$$p_i(x,y) \equiv \text{Prob}(i \text{ observed at } \langle x,y \rangle \text{ in a single 15-minute survey}) \quad \text{Eq. 1 (Stage I)}$$

Here, $p_i(x,y)$ is treated as a spatial random variable whose value is a probability; the details of how it is modeled are discussed below and in Sections 6.A.4., 6.A.5., and 6.A.6. We do not distinguish between observation and presence; the probability $p_i(x,y)$ is assumed to be equal to the probability that the species was actually present during a single 15-minute survey conducted over the 9-year study period. In other words, probability of detection when the species is present is assumed to be 1; consequences of this assumption are discussed in Section 6.A.14.

Stage II models $E\{Z_i(x,y) \mid P_i(x,y)=1\}$, the long-term mean of the observed relative abundance (SPUE), $Z_i(x,y)$, of species or group i at location (x,y) when the species or group is present:

$$E\{Z_i(x,y) \mid P_i(x,y)=1\} \quad \text{Eq. 2 (Stage II)}$$

Here $Z_i(x,y)$ is a continuous random variable representing relative abundance (number of individuals sighted per 15-minute survey per km² of survey area), and $P_i(x,y)$ is a Bernoulli random variable whose probability of success in a single trial is given by $p_i(x,y)$. Note that $E\{A|B\}$ represents the conditional expectation operator, which returns the expected value (arithmetic average over many trials) of the random variable A, given the value of the random variable B. This expected value can be thought of as the average SPUE that would have been recorded if the same location had been visited many times, instead of only once, during the 9-year survey period, and only non-zero values were included in the average. In this model, the observed value of SPUE at each location is our single observation of the random variable $Z_i(x,y)$, conditional on the outcome of $P_i(x,y)$ at that location (0 if species i is absent, 1 if present). Over a 9-year period, assuming 6 hours of potential survey per day, approximately 20,000 temporally non-overlapping surveys *could have been* conducted at each location in each season. If hypothetical repeat surveys were conducted and averaged (excluding zero observations), then the value of that average would approach that of equation 2 as the number of repeat surveys increased, if the relevant assumptions outlined in Section 6.A.14. are also met.

The MBO seabird data, processed as described in Section 6.8.3., are conceptually modeled as a set of outcomes of the purely spatial (non-temporal) random variables $P_i(x,y)$ (Stage I) and $Z_i(x,y)$ conditional on $P_i(x,y)=1$ (Stage II). This relies on the basic assumption that the parameters that define these random variables (described in more detail below) do not vary over time within a season or among survey years. Implications of this assumption are discussed in Section 6.A.14. The use of spatial random variables without an explicit temporal component is termed a spatial climatological approach and has been used elsewhere to map “hotspots” and “coldspots” in long-term average patterns of species distribution (e.g., Santora and Reiss, 2011). The word climatology in this context means long-term average.

Both Stage I and Stage II of the model are themselves comprised of two sub-models: a trend model and a residual model, described in more detail below. The trend models are implemented as generalized linear models (GLMs), and predict large-scale variation in a species’ distribution from environmental variables. The residual models are implemented as geostatistical models (kriging) to account for spatial autocorrelation in the residuals from the trend (Cressie, 1993; Pebesma, 1998).

The GLM trend component was necessary because exploratory data analysis showed that both probability of presence (Stage I) and abundance when a species is present (Stage II) showed large-scale trends that were related to environmental variables. Notably, presence/absence often showed different large-scale spatial patterns than abundance when the species was present, motivating the two-stage approach. Other types of trend models are possible, and could be explored in future work (e.g., generalized additive models, classification and regression trees).

The geostatistical component was necessary because the data are clustered and unevenly distributed in space, and preliminary analysis after removal of large-scale trends with GLM revealed autocorrelation in the spatial pattern of residuals. When this is the case, spatial dependence must be explicitly modeled to obtain unbiased estimates of GLM coefficients, as well as to properly model uncertainty at unsampled locations (Cressie, 1993; Chiles and Delfiner, 1999). A major advantage of the hybrid GLM-geostatistical approach is that predictions are accompanied by spatially explicit estimates of uncertainty, because spatial dependence in error fields is explicitly modeled (Pebesma, 1998).

The final seasonal model prediction of SPUE is the product of Stage I and Stage II maps, which gives the unconditional expected value of $Z_i(x,y)$:

$$E\{Z_i(x,y)\} = p_i(x,y) * E\{Z_i(x,y) | P_i(x,y)=1\} \quad \text{Eq. 3 (Stage I x II)}$$

This result follows directly from application of basic laws of probability and conditional expectation for random variables (Cragg, 1971; Ross, 2007). The final predicted value represents the average number of birds that would be seen if a site was surveyed repeatedly (using the same standardized 15-minute surveys), *including times when the species was not seen as values of 0*.

The seasonal modeling process can be summarized as follows. For each species and group, for each season that can be modeled, the following steps are performed:

1. Transform potential predictor variables for linearity. See Section 6.A.2. below.
2. Divide data into training and validation (“holdout”) subsets for cross-validation purposes. See Section 6.A.3. below.
3. Stage I trend model: Use a GLM (binomial distribution, logit link) to generate a predictive map of the mean probability of species occurrence. See Section 6.A.4. below.
4. Stage I residual model: Use ordinary indicator kriging (OIK) to predict the “residual” probability map, where “residual” is defined as the probability that the regression model leads to an incorrect classification of the presence state ($P_i(x,y)$) of a given location. See Section 6.A.5. below.
5. Final Stage I model: Adjust the trend-predicted probability map using the kriged residual probability map from step 4. The trend from step 3 and residual from step 4 are combined using probability laws. See Section 6.A.6. below.
6. Stage II trend model: Use a GLM (normal distribution, identity link) to generate a predictive map of the mean abundance of a species when it is present. Data were transformed for normality for this part of the analysis using a Box-Cox type transformation (Box and Cox 1964), described further below, and back-transformed for final maps. See Section 6.A.7. below.
7. Stage II residual model: Use Simple Kriging (SK) to predict residual map of the regression model of abundance. See Section 6.A.8. below.
8. Final Stage II model: Add the trend map from step 6 and the residual map from step 7. See Section 6.A.9. below.
9. Final Stage I x II model prediction: Multiply the predicted probability of occurrence at each location by the predicted abundance if present to produce the final prediction of the expected value (long-term average) of abundance at each location. See Section 6.A.10. below.

10. Relative uncertainty calculation: scaled relative uncertainty values were calculated for the trend, residual, and final models for Stage I and Stage II, and for the final Stage IxII prediction. See Section 6.8.11. below.
11. Model evaluation, cross-validation, and relative uncertainty calibration. See Section 6.A.12. below.

The sections below describe each of these steps in detail.

6.A.2. Transformation of potential predictor variables for linearity

Transforming independent variables in a multiple linear regression context for normality, centrality, and homogeneity of variance is often desirable for stabilizing estimates of regression parameters, and can also help to linearize relationships between predictors and response (Sokal and Rohlf, 1995). The family of power-law transformations studied by Box and Cox (1964) is particularly useful for improving both normality and linearity. A Box-Cox transformation is defined as follows, where X denotes the original variable and X^* the transformed variable:

$$X^* = \begin{cases} X^\lambda, & \text{if } \lambda \neq 0 \\ \ln(X), & \text{if } \lambda = 0 \end{cases} \quad \text{Eq. 4}$$

A maximum-likelihood procedure (Box and Cox, 1964; Dror, 2006) was used to estimate the Box-Cox transformation parameter λ for each potential predictor variable, and guide the final choice of stabilizing transformation for each predictor. *A priori* knowledge about the types of transformations likely to be justified for different types of variables was also considered (Sokal and Rohlf, 1995). Predictor transformations expressions are shown in Table 6.A.1. Note that the transformation of some of these variables changes the sign of the linear relationship between variable and response; care must therefore be taken in interpreting the signs of regression coefficients for transformed predictors. Details of transformation choices and pre- and post-transformation distributions are given in Appendix 6.B. and Online Supplement 6.1.

Table 6.A.1. Predictor variable transformations.

PREDICTOR VARIABLE	TRANSFORMATION EXPRESSION	NOTES
BATH	$X^* = (1-x)^{-0.4}$	For all $X \leq 0$
SLOPE	$X^* = X^{-0.4}$	
DIST	$X^* = X^{0.6}$	
SSDIST	$X^* = X$	Not transformed
SST	$X^* = 11605/(X+273.15)$	Arrhenius transform (Laidler, 1997)
STRT	$X^* = X$	Not transformed
TUR	$X^* = 1/X$	
CHL	$X^* = 1/(X+1)$	
ZOO	$X^* = X$	Not transformed
SLPSLP	$X^* = X^{-0.3}$	
PHIM	$X^* = 1/(X+3)$	

Transformed predictor variables were centered and standardized prior to each GLM fit, using the set of values of each predictor variable at the data locations under consideration (centering and standardization was performed each time just prior to running the GLM, because different patterns of missing predictor data could cause different data points to be used, requiring re-centering and re-standardization).

6.A.3. Selection of training and validation subsets for cross-validation

50% of the observation locations were selected at random to be used in subsequent model-fitting (henceforth referred to as the training set), with the remaining 50% withheld for cross-validation (henceforth referred to as the validation or holdout set). All model selection and model fitting (Sections 6.A.4. to 6.A.10.) was carried out using only the training set. Cross-validation statistics were calculated by comparing model predictions at the holdout locations to the true data values at the holdout locations. Final predictive maps, however, used all available data by applying the models selected and fit based on training data to the entire original dataset. Cross-validation error estimates are thus conservative in the sense that they were derived from a model fit to a dataset one half the size of the final dataset.

6.A.4. Stage I trend model

The trend component of the Stage I model, $\mu_i^t(x,y)$, was estimated as follows.

Observed data $Z_i(x,y)$ were first transformed to a binary indicator variable $P_i(x,y)$, whose value was 1 if $Z_i(x,y) > 0$ and 0 otherwise. The initial set of 11 potential predictor variables was then pre-screened to remove any predictors whose pattern of missing values would too greatly influence the data points that could be used to estimate the GLM. Pre-screening criteria are given in Table 6.A.2.

Table 6.A.2. Criteria for inclusion of a predictor variable in the set of potential predictors evaluated for a given seasonal Stage I or Stage II GLM model ("pre-screening criteria"). The set of points for which both data and predictor values were available had to meet all of these criteria for a predictor variable to be considered.

CRITERION	CONDITION
Fraction of all data eliminated	$\leq 30\%$
Fraction of presences eliminated	$\leq 20\%$
Fraction of absences eliminated	$\leq 50\%$
Number of presences remaining	≥ 15

Predictor variables not excluded in the pre-screening process were centered, standardized, and the R package 'glmulti' (Calcagno and Mazancourt, 2010; Calcagno, 2011) was used to search for the model with lowest AICc from the set of possible generalized linear models, allowing two-way interaction effects to be included, but requiring that both corresponding main effects be in the model if an interaction term were to be included (marginality requirement). GLM model used a binomial distribution with a logit link function (Fox, 2008).

The search method used depended on the size of the possible model space, which was restricted by the elimination of some potential predictors in the pre-screening stage (above) and by an upper bound on the number of terms determined by the number of observations. The number of terms in a model (not including the intercept) was restricted to be no greater than the number of observations divided by 10 (Sokal and Rohlf, 1995; Fox, 2008). If the number of predictors and/or maximum number of terms was sufficiently small, then the model space was searched exhaustively for the model with the lowest corrected Akaike's Information Criterion (AICc; Sokal and Rohlf, 1995). If the number of predictors and/or maximum number of terms was intermediate, then a genetic algorithm with the default parameters and stopping criteria of $\Delta AICc = 0.5$, $\text{conseq} = 5$ was used (Calcagno and Mazancourt, 2010; Calcagno, 2011). If the number of predictors and/or maximum number of terms was too large for the genetic algorithm to enumerate the model space, then an exhaustive search was performed of all possible models with 5 or fewer main effects (allowing for two-way interactions within each subset).

The selected model structure was then fit to the data using Matlab Statistics Toolbox function 'glmfit', which implements standard Generalized Linear Model fitting by iteratively re-weighted least-squares (Bjorck, 1996; Fox, 2008). As before, a binomial distribution and logit link function were used. Use of binomial distributions and logit link functions involves assumptions that are discussed in Section 6.A.14. Parametric ± 1 standard error confidence bounds on GLM estimates were calculated using Matlab function 'glmval' (following equations in Fox, 2008).

A standard array of GLM diagnostics was produced, including effect tests, deviance goodness-of-fit tests, several 'pseudo- R^2 ' measures designed for logistic regression, residual leverage and influence plots, and a variety of other diagnostic measures (for details see diagnostic tables in main text and Online Supplement 6.2). An ROC curve analysis was also performed to assess accuracy of the Stage I trend prediction (see Online Supplement 6.2).

6.A.5. Stage I residual model

The residual component of the Stage I model, $\varepsilon_i^t(x,y)$, was estimated as follows.

First, ROC curve analysis was used to determine the optimal cutoff value of the trend probability, $\mu_i^t(x,y)$, to use for classifying the presence/absence data (Cardillo, 2008). ROC curve analysis identifies the cutoff probability for classification that optimizes the tradeoff between sensitivity and specificity, given a training dataset. This cutoff was then applied to transform the trend prediction map $\mu_i^t(x,y)$ into a binary classification

map (0=predicted absence, 1=predicted presence). Use of this ROC curve method to classify the trend can result in global bias of the classification toward the less-common class (usually presences), and the implications of this are discussed in Section 6.A.14.

A binary indicator variable (the “misclassification indicator”) was then created that took the value 1 if the binary classification map based on the trend was correct at a data location, and 0 if not. Indicator variograms were estimated and modeled from this misclassification indicator, and Ordinary Indicator Kriging (OIK) was used to produce a map of predicted misclassification probabilities. Kriging predictions >1 or <0 were set to 1 or 0, respectively, to satisfy order relations for probabilities (Deutsch and Journel, 1998; Pebesma, 1998), and the resulting map was the residual component of Stage I, $\varepsilon_i^I(x,y)$. Because misclassification of 0’s as 1’s and 1’s as 0’s were considered equivalent, the OIK geostatistical model makes the assumption that the spatial patterns of misclassification of 1’s and 0’s are equivalent (symmetry). Implications of this symmetry assumption are discussed in Section 6.A.14.

Variogram models were fit automatically by a non-linear weighted least-squares minimization algorithm (Pebesma, 1998; Pardo-Igúzquiza, 1999), using weights proportional to N/h^2 (the number of pairs of observations used to estimate each observation divided by the square of the lag distance), as described by Pebesma (1998). Following standard geostatistical practice, the functional form of the variogram and an initial-guess parameter set was specified prior to the least-squares minimization by inspection of the empirical variogram (Issaks and Srivistava, 1989; Cressie, 1993; Deutsch and Journel, 1998; Chiles and Delfiner, 1999).

OIK produces parametric estimates of uncertainty (kriging standard error) for each location in the residual prediction map (Pebesma, 1998; Deutsch and Journel, 1998). An ROC curve analysis was also performed to assess accuracy of the Stage I residual prediction (see Online Supplement 6.2).

6.A.6. Final Stage I model

Because the trend and residual components of the Stage I model are probabilities, they can be combined using the laws of conditional probability to arrive at the full Stage I model as follows (Ross, 2007):

$$p_i(x,y) = \text{Prob}([\text{trend model predicts } i \text{ is present AND trend model is not wrong}] \text{ OR } [\text{trend model predicts } i \text{ is not present AND trend model is wrong}]) \quad \text{Eq. 5}$$

which can be translated to,

$$p_i(x,y) = \mu_i^I(x,y) \cdot (1 - \varepsilon_i^I(x,y)) + (1 - \mu_i^I(x,y)) \cdot \varepsilon_i^I(x,y) \quad \text{Eq. 6}$$

which simplifies to the final Stage I model:

$$p_i(x,y) = \mu_i^I(x,y) + \varepsilon_i^I(x,y) - 2 \cdot \mu_i^I(x,y) \cdot \varepsilon_i^I(x,y) \quad \text{Eq. 7}$$

Parametric ± 1 SE confidence intervals for the final Stage I model, $p_i(x,y)$, were derived by applying Equation 7 to the parametric confidence intervals for $\mu_i^I(x,y)$ and $\varepsilon_i^I(x,y)$ calculated using the GLM model and the geostatistical (OIK) model, respectively.

6.A.7. Stage II trend model

The trend component of the Stage II model, $\mu_i^{II}(x,y)$, was estimated as follows.

Data at non-zero locations were first transformed for normality using a Box-Cox power transform (see Section 6.A.2.) whose parameter λ was chosen by a maximum likelihood procedure (Figure 6.A.1) (Box and Cox, 1964; Dror, 2006). Power-law family models have recently been found to outperform other often-used statistical models (e.g., Poisson) for describing distributions of seabird group sizes in our study region (Beauchamp, 2011), lending further motivation to the use of the Box-Cox family of transformations for this purpose.

The initial set of 11 potential predictor variables was then pre-screened to remove any predictors whose pattern of missing values would too greatly influence the data points that could be used to estimate the GLM. Pre-screening criteria are given in Table 6.A.2.

The predictor variables were centered, standardized, and the R package 'glmulti' (Calcagno and Mazancourt, 2010; Calcagno, 2011) was used to search for the model with lowest AICc in the same way described for Stage I (Section 6.A.4.), except that in this case the GLM model used a normal distribution with an identity link function (Fox, 2008).

The selected model structure was then fit to the data using Matlab Statistics Toolbox function 'glmfit', which implements standard Generalized Linear Model fitting by iteratively re-weighted least-squares (Bjorck, 1996; Fox, 2008). A normal distribution and identity link function were used. Use of the normal distribution here involves assumptions that are discussed in Section 6.A.14. Parametric ± 1 standard error uncertainty bounds on GLM estimates were calculated using Matlab function 'glmval' (following equations in Fox, 2008).

Because spatial autocorrelation biases the estimation of GLM parameters, we followed an iterative procedure to fit the final GLM in gstat (Pebesma, 1998; Chiles and Delfiner, 1999).

1. Calculate residuals and estimate residual variogram (see Section 6.8.).
2. Re-calculate fit with gstat, using residual variogram
3. Re-calculate residuals and repeat fitting with gstat (steps 2 and 3) until residual variogram has converged (determined by inspection).

A standard array of GLM diagnostics was produced, including effect tests, goodness-of-fit F tests, R^2 and several 'pseudo- R^2 ' measures to allow comparison with the Stage I logistic regression, residual leverage and influence plots, and a variety of other diagnostic measures (for details, see diagnostic tables in main text and Online Supplement 6.2).

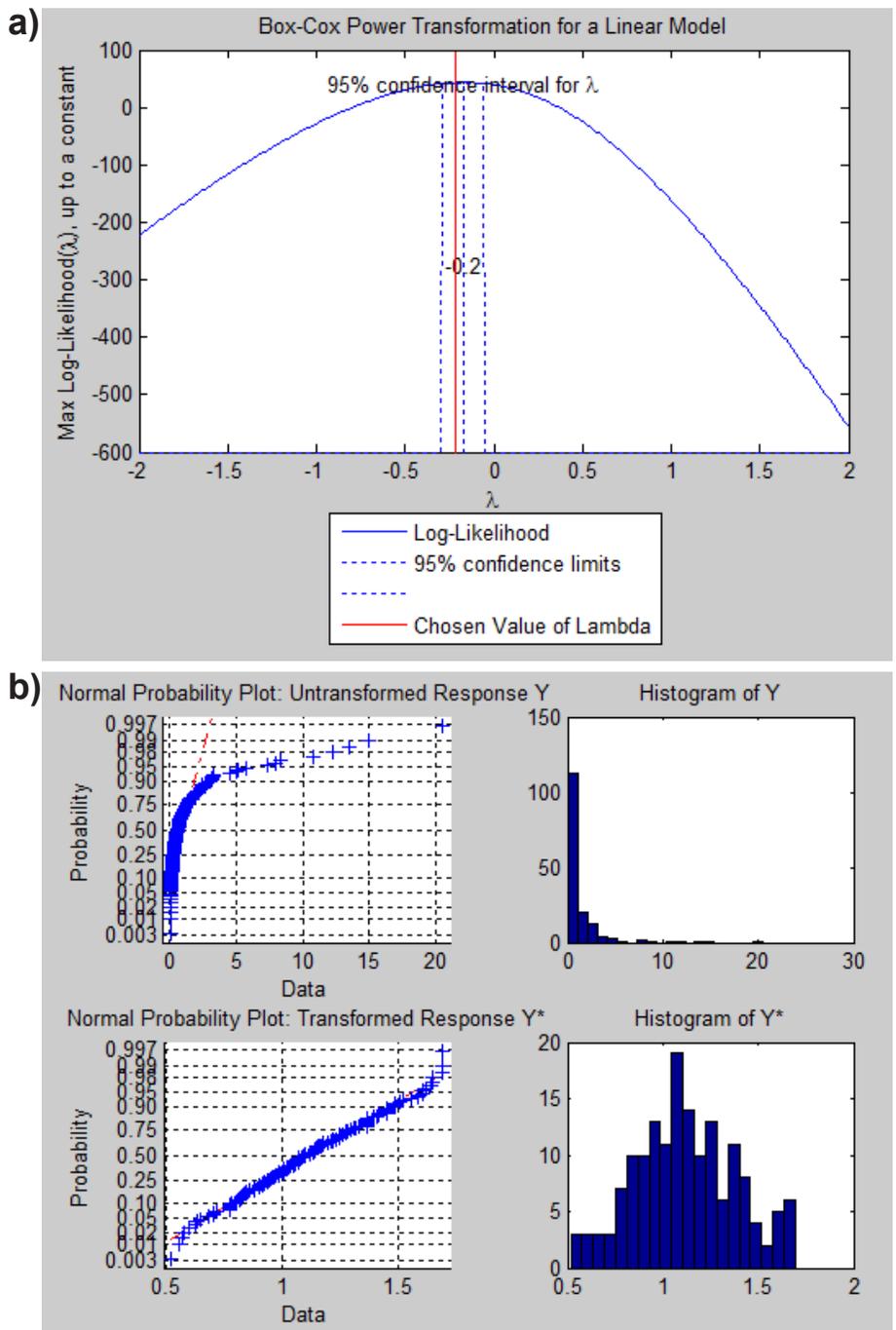


Figure 6.A.1 Box-Cox transformation of non-zero relative abundance (SPUE) data. Example for Dovekie in Winter. (a) selection by maximum likelihood procedure. (b) normal probability plots and histograms before and after transformation.

6.A.8. Stage II residual model

The residual component of the Stage II model, $\varepsilon_i''(x,y)$, was estimated as follows.

First, residuals from the trend model fit were calculated by subtracting the observed values from predicted values. Residuals were calculated in Box-Cox transformed space to satisfy normality assumptions of geostatistical methods. Residual variograms were then estimated and modeled using gstat, and Simple Kriging (SK) was used to produce a map of predicted residuals. The resulting map was the residual component of Stage II, $\varepsilon_i''(x,y)$.

Variogram models were fit automatically by a non-linear weighted least-squares minimization algorithm (Pebesma, 1998; Pardo-Igúzquiza, 1999), using weights proportional to N/h^2 (the number of pairs of observations used to estimate each observation divided by the square of the lag distance), as described by Pebesma (1998). Following standard geostatistical practice, the functional form of the variogram and an initial-guess parameter set was specified prior to the least-squares minimization by inspection of the empirical variogram (Issaks and Srivistava, 1989; Cressie, 1993; Deutsch and Journel, 1998; Chiles and Delfiner, 1999).

SK was also used to produce parametric estimates of uncertainty (kriging standard error) for each location in the residual prediction map (Pebesma, 1998; Deutsch and Journel, 1998).

6.A.9. Final Stage II model

In Box-Cox transformed space, the final Stage II model is simply the sum of trend and residual components:

$$E\{Z_i^{Transformed}(x,y) | P_i(x,y)=1\} = \mu_i''(x,y) + \varepsilon_i''(x,y) \quad \text{Eq. 8}$$

The result can be back-transformed to yield a prediction in the original units of SPUE:

$$E\{Z_i(x,y) | P_i(x,y)=1\} = \begin{cases} (E\{Z_i^{Transformed}(x,y) | P_i(x,y)=1\})^{1/\lambda}, & \text{if } \lambda \neq 0 \\ \exp(E\{Z_i^{Transformed}(x,y) | P_i(x,y)=1\}), & \text{if } \lambda = 0 \end{cases} \quad \text{Eq. 9}$$

Back-transforms were constrained to lie between 0 and 110% of the observed data maximum.

Parametric ± 1 SE confidence intervals for the final back-transformed Stage II model, $E\{Z_i(x,y) | P_i(x,y)=1\}$, were derived by applying Equations 8 and 9 to the parametric confidence intervals for $\mu_i''(x,y)$ and $\varepsilon_i''(x,y)$ calculated using the GLM model and the geostatistical (SK) model, respectively.

6.A.10. Final Stage I x II model

Stage I and Stage II models were combined as described in Section 6.A.1. (Equation 3) to produce each seasonal predictive map of the unconditional expected value of SPUE, which we will refer to as the "Stage I x II" prediction map or $E\{Z_i(x,y)\}$. Specifically, $E\{Z_i(x,y)\}$ is equal to the product of Equation 9 (the final back-transformed Stage II prediction) and Equation 7 (the final Stage I model prediction). Note that the Stage I x II predictions are in back-transformed units (SPUE).

Parametric uncertainty bounds (± 1 SE) for the final Stage I x II maps were obtained by plugging the confidence intervals for $\mu_i^I(x,y)$, $\varepsilon_i^I(x,y)$, $\mu_i''(x,y)$, and $\varepsilon_i''(x,y)$ described above into equations 7 and 9 and multiplying equation 7 by equation 9 for each set of uncertainty bounds.

6.A.11. Relative uncertainty calculations

In order to simplify comparison of uncertainties among different model components, uncertainties were converted to relative values that fall between 0 and 1, with 0 representing low uncertainty (high certainty) and 1 representing high uncertainty (low certainty). To further aid in interpretation, relative certainty classes were defined as follows: high certainty class (relative uncertainty ≤ 0.5), medium certainty class ($0.5 < \text{relative uncertainty} \leq 0.65$), and low certainty class (relative uncertainty > 0.65). The implications of a particular relative

uncertainty value or certainty class for model performance can be determined by examining the diagnostic tables in the main text, which give cross-validation error statistics for each certainty class, and the cross-validation relative uncertainty calibration plots in Appendix 6.C. (described in Section 6.A.12. below).

6.A.11.1. Stage I

The relative uncertainty of Stage I model predictions is expressed as the scaled negative log (odds ratio), *SNLOR*. The negative log odds ratio, *NLOR*, is the negative natural logarithm of the ratio of the odds of correct binary classification (absence= 0, presence= 1) using the Stage I model to the odds of correct binary classification under a null model:

$$NLOR = -\ln\left(\frac{Odds_{model}}{Odds_{null}}\right) \quad \text{Eq. 10}$$

To calculate the odds of correct classification under the Stage I model, $Odds_{model}$, we first consider uncertainty of the Stage I model prediction relative to the cutoff probability c used for binary classification (in this case, the optimal cutoff probability determined by ROC curve analysis). The uncertainty around the Stage I model prediction p can be modeled by a normal curve on the logit scale, with mean equal to the Stage I prediction and standard deviation equal to the larger of the upper and lower 1SE confidence intervals:

$$z_p \sim N(\text{logit}[p], \max(\text{logit}[p^{+1SE}] - \text{logit}[p], \text{logit}[p] - \text{logit}[p^{-1SE}])). \quad \text{Eq. 11}$$

Then the probability of the true predicted value lying above the cutoff probability c is given by

$$p_{above} = \text{Prob}(z_p > \text{logit}(c)), \quad \text{Eq. 12}$$

and the probability of the true predicted value falling below the cutoff probability is

$$p_{below} = \text{Prob}(z_p < \text{logit}(c)). \quad \text{Eq. 13}$$

The classifier itself is subject to error, which we estimate by its performance in cross-validation: the true positive (\hat{p}_{TP}), true negative (\hat{p}_{TN}), false positive (\hat{p}_{FP}), and false negative (\hat{p}_{FN}), rates of the classifier from the cross-validation confusion matrix at cutoff value c . The odds of correct classification using the Stage I model can then be calculated as:

$$SNLOR = \frac{\ln\left(\frac{Odds_{model}^{max}}{Odds_{null}}\right) - \ln\left(\frac{Odds_{model}}{Odds_{null}}\right)}{\ln\left(\frac{Odds_{model}^{max}}{Odds_{null}}\right) - \ln\left(\frac{Odds_{model}^{min}}{Odds_{null}}\right)} \quad \text{Eq. 14}$$

To calculate the odds of correct classification under a null model, $Odds_{null}$, we consider a null model in which the true and predicted presence/absence (1/0) states are given by Bernoulli random variables with probabilities p_1 (equal to the global prevalence of the species) and c (equal to the optimal cutoff probability from ROC curve analysis), respectively. Then the null odds of correct classification are:

$$Odds_{model} = \frac{p_{above} \cdot \hat{p}_{TP} + p_{below} \cdot \hat{p}_{TN}}{p_{above} \cdot \hat{p}_{FP} + p_{below} \cdot \hat{p}_{FN}} \quad \text{Eq. 15}$$

For a given set of cross-validation error rates (\hat{p}_{TP} , \hat{p}_{TN} , \hat{p}_{FP} , and \hat{p}_{FN}), the minimum and maximum possible values of the *NLOR* are:

$$Odds_{null} = \frac{(1-p_1) \cdot c + p_1 \cdot (1-c)}{(1-p_1) \cdot (1-c) + p_1 \cdot c} \quad \text{Eq. 16}$$

The scaled $NLOR$, $SNLOR$, is calculated so that $SNLOR=0$ at the minimum possible value of the $NLOR$ and $SNLOR=1$ at the maximum possible value of the $NLOR$:

$$Odds_{model}^{\min} = \min\left(\frac{\hat{p}_{TP}}{\hat{p}_{FP}}, \frac{\hat{p}_{TN}}{\hat{p}_{FN}}\right), Odds_{model}^{\max} = \max\left(\frac{\hat{p}_{TP}}{\hat{p}_{FP}}, \frac{\hat{p}_{TN}}{\hat{p}_{FN}}\right) \quad \text{Eq. 17}$$

Values of $SNLOR$ closer to 0 indicate model predictions that have relatively high odds of being correct compared to a null model (high certainty), whereas values closer to 1 indicate model predictions that have relatively low odds of being correct compared to a null model (low certainty). Relative uncertainties were calculated in this way for the Stage I trend, Stage I residual, and the final Stage I model, using the cross-validation ROC curve cutoff c and cross-validation error rates (\hat{p}_P , \hat{p}_N , \hat{p}_{FP} , and \hat{p}_{FN}) determined from the ROC analysis of trend, residual, and final Stage I predictions, respectively. Below, the final Stage I relative uncertainty is denoted $\sigma^{I,rel}(x,y)$, and is equal to the value of $SNLOR$ for the final Stage I model for species/group i at location (x,y) .

6.A.11.2. Stage II

Relative uncertainty of Stage II trend, residual, and final model predictions were calculated as the ratio of prediction variances to the appropriate error variance (trend prediction variance: total sample variance minus residual variogram sill; residual variance: residual variogram sill; final prediction variance: total sample variance). Below, the final Stage II relative uncertainty is denoted $\sigma^{II,rel}(x,y)$.

6.A.11.3. Stage IxII

The relative uncertainty of final Stage IxII model predictions was calculated by combining the relative uncertainties of final Stage I and Stage II models as follows:

$$\sigma_i^{IxII,rel}(x,y) = p_i(x,y) \cdot [\sigma_i^{II,rel}(x,y)] + (1 - p_i(x,y)) \cdot \sigma_i^{I,rel}(x,y) \quad \text{Eq. 18}$$

The rationale behind equation 18 is that the Stage II relative uncertainty applies if the species is present (which is true with probability $p_i(x,y)$), whereas the Stage I relative uncertainty applies if the species is absent (which is true with probability $[1 - p_i(x,y)]$).

6.A.12. Model evaluation and uncertainty calibration

In addition to the standard GLM effect tests and diagnostics, model predictive performance was evaluated in and out of the training set using a variety of error statistics, error plots and ROC curve analysis. As a final summary of model performance in cross-validation and aid to the reader in interpreting relative uncertainty values for the final Stage IxII model, an uncertainty calibration plot was produced. For each location in the holdout set, the model developed from training data was used to predict the value at that location, and the magnitude of the difference between actual and predicted values (absolute error) was plotted versus the Stage I x II relative uncertainty value (Appendix 6.C.). Robust linear loess smoothing lines (Burkey, 2009) are plotted to show how actual out-of-set average prediction errors relate to parametric relative uncertainty estimates. Separate lines are plotted for overall error, and error when the species or group was present (since most species are relatively rare in any given survey, presences are harder to predict than absences). Similar relative uncertainty calibration plots are produced for Stage I predictions (presence/absence).

Uncertainty calibration plots, ROC analyses, error statistics, and other model evaluation diagnostics are included in the diagnostic tables in the main report, in Appendix 6.C., and in Online Supplement 6.2.

6.A.13. Combination of seasonal climatological maps to produce annual climatological maps

For each species and species group i , seasonal maps of climatological SPUE (Stage IxII predictions) were combined to produce annual maps as follows:

$$E\{Z_i(x,y)\}^{\text{annual}} = \sum_{\text{all seasons } j} E\{Z_i(x,y)\}_j \quad \text{Eq.19}$$

Using the laws of probability and the expectation operator (Ross, 2007), this procedure can be shown to yield an unbiased estimate of the SPUE prediction for the entire year, given that (1) each seasonal model prediction is the unconditional expected value of SPUE, $Z_i(x,y)$, and, (2) the seasons are defined as non-overlapping and together cover the entire climatological year. These two conditions are true by definition.

Annual integrated presence probability maps were produced by combining the seasonal climatological presence probability predictions (Stage I predictions), assuming statistical independence of the seasonal probabilities. Given 4 seasons, there are 15 possible ways in which a species or group can be present in at least one season. Represented as four digit binary codes, these are: 1000, 0100, 0010, 0001, 1100, 1010, 1001, 0110, 0011, 0101, 1110, 1011, 1101, 0111, 1111. The probabilities of each of these outcomes was summed to produce the annual integrated presence probability, $p_i(x,y)^{\text{annual}}$, which is equivalent to the annual climatological site occupancy probability for species/group i each location (x,y) .

To estimate the relative uncertainty associated with each annual map, the weighted average of the corresponding seasonal relative uncertainty maps was calculated, using the frequencies of occurrence of the species in each season as weights. For the annual SPUE map the relative uncertainty is given by:

$$\sigma_i^{I,rel}(x,y)^{\text{annual}} = \sum_{\text{all seasons } j} \left[\frac{\sigma_i^{I,rel}(x,y) \cdot f_{i,j}}{\sum_j f_{i,j}} \right] \quad \text{Eq. 20}$$

For the annual integrated presence probability map relative uncertainty, the relative uncertainty is given by:

$$\sigma_i^{I,rel}(x,y)^{\text{annual}} = \sum_{\text{all seasons } j} \left[\frac{\sigma_i^{I,rel}(x,y) \cdot f_{i,j}}{\sum_j f_{i,j}} \right] \quad \text{Eq. 21}$$

It can be shown that these relative uncertainties are monotonically related to the variance of the annual prediction error, but this relationship will not necessarily be linear for two reasons (Ross, 2007):

1. The relative uncertainty of Stage I predictions is based on a log-odds ratio, and,
2. Seasonal estimates of $Z_i(x,y)$ may not be uncorrelated, and therefore summation of variances, unlike summation of expected values, is not necessarily a linear operator.

Thus we rely on uncertainty calibration plots (plots of cross-validation error vs. relative uncertainty, Section 6.A.12.) to interpret the precise meaning of the relative uncertainty value for each species/group annual model.

6.A.14. Summary and implications of model assumptions

The seasonal predictive modeling approach described above makes a number of assumptions. To the extent these assumptions are violated, accuracy of predictions and uncertainty estimates may suffer. In this section we briefly review the major assumptions and their implications. The degree to which violations of model assumptions affect the performance of any given seasonal model can be assessed by considering the cross-validation performance statistics described in 6.A.12 and reported in the main text diagnostic tables, Appendix 6.C, and Online Supplement 6.2.

Important general assumptions:

- *Stationarity of pattern over time within seasons and among years*

Statistically, stationarity in this context means that the region-wide mean, variance, and spatial structure of abundance and occurrence patterns do not change over the time period we studied. Ecologically, stationarity implies that the ecosystem has not undergone any fundamental shifts in patterns and processes (e.g., climate trends, ocean climate regime shifts, introduced species, changes in patterns

of human activities like fishing). If this assumption is violated, temporal variation will show up as non-spatially structured error (“white noise”) in the model result. Model parameters and predictions may also be biased (cross-validation errors will not be centered at 0). The predicted spatial pattern may be an amalgam of different patterns that occurred at different time periods (e.g., “smearing” of hotspots that moved from year to year). If there are major changes in the underlying processes, the model will also be less generalizable to other time periods.

- *Stationarity of environmental predictor climatologies*

The use of long-term climatologies of time-varying environmental predictors (such as SST and stratification), assumes that the long-term seasonal mean spatial patterns of these variables have not changed over time. Major changes in the underlying environmental patterns and processes will make the model less generalizable to other time periods.

- *Unbiased year-to-year sampling (no temporal effect included)*

If the sampling pattern is non-random within seasons and/or across years, GLM parameter estimates and parametric uncertainties could be biased and inaccurate. This problem will be exacerbated if the assumptions of temporal stationarity of predictors and response are also violated. The Manomet survey was conducted on ships of opportunity, so samples were not random in space or time; therefore some biases due to unbalanced effort are expected.

- *Perfect detectability; freedom from other kinds of sample bias*

To the extent that a given species or species group is not perfectly detectable by the sampling protocol, relative occurrence and abundance indices will be biased compared to true abundance and occurrence values. Predictions from this model should be considered relative, rather than absolute, estimates of occurrence and abundance. In addition to detectability, similar biases can result from attraction of certain species to the survey platform (boats). Finally, systematic study biases may exist in the types of species that were recorded. For example, we found very few records of passerines in the Manomet dataset, even though there is evidence of offshore sightings of these species from other sources. These and other birds that are rare but not absent in the offshore may require other survey and modeling approaches if they are of conservation concern.

- *Constant relationship between sampling effort, relative indices of occurrence and abundance, and true values of occurrence and abundance*

Not only are species unlikely to be perfectly detectable, the relationship between our relative indices of occurrence and abundance and the true values of occurrence and abundance could vary in time and space, depending on differences in observers, weather conditions, animal behavior, etc. Such variation introduces an un-accounted for source of measurement error into data.

Important Stage I assumptions

- *Binomial distribution and logit link function*

To the extent that these distributional assumptions are violated, trend predictions may be biased and parametric confidence intervals inaccurate.

- *Use of receiver operating characteristic (ROC) curve optimal cutoff analysis to classify residuals from the trend model*

Use of the ROC classifier may introduce bias into the final presence probability estimates at the expense of balancing overall sensitivity and specificity.

- *Symmetry assumption for misclassification probability field*

Misclassification of absences as presences may not show the same spatial pattern as misclassification of presences as absences; if that is the case, then model predictions may be biased and the model may perform better for one type of misclassification than for others, even though parametric uncertainty estimates are the same.

Important Stage II assumptions

- *Normality and linearity of Box-Cox transformed predictors and responses in the Stage II trend model*

We assume that the Box-Cox transform in Stage II is sufficient to achieve normality of residual variances

and linearity of underlying response-predictor relationships. Since the underlying seabird relative abundance data are based on counts (divided by transect area to create a quasi-continuous density estimate), this requires that we assume the continuous Box-Cox transformed Gaussian distribution used to represent non-zero relative abundance is an adequate approximation to the underlying discrete probability distribution. The appropriateness of these assumptions is difficult to test directly and the reader should rely on cross-validation performance statistics to judge the extent to which these assumptions were approximately correct.

- *Trans-Gaussian assumption in the Stage II residual (geostatistical) model*
Simple Kriging also assumes approximate normality; therefore the adequacy of the Box-Cox transformation to achieve normality of the residual distribution is also important to the accuracy of the kriging prediction (especially the validity of the kriging variance).
- *Back-transform issues (extrapolation of the CDF tail)*
When back-transforming Stage II predictions, we have arbitrarily cut off the upper end of the distribution at 110% of the data maximum, which may not always be appropriate. This is only expected to influence the highest predicted values.

Important Stage IxII assumptions

- *Separability of abundance and presence/absence patterns*
We have assumed that abundance is conditionally independent of presence/absence (that is, abundance can be modeled independently of presence probability). If this assumption is violated, then the Stage IxII estimates will be biased. The direction of this bias will depend on the sign of the dependence, and on the Box-Cox transformation parameter. The degree of bias in predictions can be assessed (and corrected for) by examining cross-validation bias statistics in the diagnostic tables.

Important assumptions of annual maps

- *Seasonal estimates of expected SPUE, $Z_i(x,y)$, are uncorrelated with each other.*
If seasonal estimates of SPUE are positively correlated with each other, then the summation of unconditional expected values will still be correct but the relationship between actual prediction error and the predicted relative uncertainty value will be affected. The cross-validation uncertainty calibration plots should be used as a guide to the true relationship between relative uncertainty and prediction error for each annual model.
- *Seasonal estimates of presence probability $p_i(x,y)$ are independent of each other.*
If presence probabilities are not independent from season to season, then the integrated annual presence probability maps will over or underestimate annual site occupancy probability, depending on the sign of the dependence.