NRC Research Press

ARTICLE

Spatial semiparametric models improve estimates of species abundance and distribution

Andrew Olaf Shelton, James T. Thorson, Eric J. Ward, and Blake E. Feist

Abstract: Accurate estimates of abundance are imperative for successful conservation and management. Classical, stratified abundance estimators provide unbiased estimates of abundance, but such estimators may be imprecise and impede assessment of population status and trend when the distribution of individuals is highly variable in space. Model-based procedures that account for important environmental covariates can improve overall precision, but frequently there is uncertainty about the contribution of particular environmental variables and a lack of information about variables that are important determinants of abundance. We develop a general semiparametric mixture model that incorporates measured habitat variables and a nonparametric smoothing term to account for unmeasured variables. We contrast this spatial habitat approach with two stratified abundance estimators and compare the three models using an intensively managed marine fish, darkblotched rockfish (*Sebastes crameri*). We show that the spatial habitat model yields more precise, biologically reasonable, and interpretable estimates of abundance than the classical methods. Our results suggest that while design-based estimators are unbiased, they may exaggerate temporal variability of populations and strongly influence inference about population trend. Furthermore, when such estimates are used in broader meta-analyses, such imprecision may affect the broader biological inference (e.g., the causes and consequences of the variability of populations).

Résumé : Des estimations exactes de l'abondance sont essentielles au succès de la conservation et de la gestion. Si les estimateurs d'abondance stratifiés classiques fournissent des estimations non biaisées de l'abondance, ces estimateurs peuvent être imprécis ou entraver l'évaluation de l'état et de la tendance de la population si la répartition des individus est très variable dans l'espace. Si des procédures basées sur des modèles qui tiennent compte d'importantes covariables environnementales peuvent améliorer la précision globale, il y a souvent une incertitude associée à la contribution de différentes variables environnementales et un manque d'information sur les variables qui sont d'importants déterminants de l'abondance. Nous avons développé un modèle de mélange semi-paramétrique général qui incorpore des variables mesurées de l'habitat et un terme de lissage non paramétrique pour tenir compte des variables non mesurées. Nous comparons cette approche d'habitat spatial à deux estimateurs d'abondance stratifiés à la lumière d'observations sur un poisson marin faisant l'objet d'une gestion intensive, le sébaste tacheté (*Sebastes crameri*). Nous démontrons que le modèle d'habitat spatial produit des estimations de l'abondance plus précises, interprétables et raisonnables du point de vue biologique que les méthodes classiques. Nos résultats donnent à penser que, si les estimateurs basés sur la conception de l'échantillonnage sont non biaisés, ils peuvent exagérer la variabilité temporelle des populations et influencer fortement l'inférence concernant la tendance démographique. En outre, quand ces estimations sont utilisées dans des métaanalyses plus larges, cette imprécision pour avoir une incidence sur l'inférence biologique élargie (p. ex. les causes et conséquences de la variabilité des populations). [Traduit par la Rédaction]

Introduction

Accurate assessment of population status and trend are fundamental to the successful conservation and management of species. Imprecise or biased estimates of population biomass or abundance may cause managers to fail to take actions when warranted or induce changes to management when none are required. Historically, resource managers have relied on classical, design-based sampling methods, such as stratified randomized sampling (Cochran 1977). If abundance is driven by habitat variables, explicitly accounting for these variables should provide more precise estimates of abundance than design-based approaches that ignore habitat information. Recent efforts have shown how habitat covariates can be integrated with distance sampling and tag-resighting procedures to improve abundance estimates of terrestrial vertebrates (Royle et al. 2013; Sillett et al. 2012).

The use of habitat variables in abundance estimation has a long history in terrestrial ecosystems, but in marine ecosystems basic spatial habitat information has been lacking until the past decade or so. Examples include high-resolution bathymetry and a variety of oceanographic variables (e.g., Sbrocco and Barber 2013). While habitat data have informed the location of marine reserves (Ward et al. 1999) and the identification of vulnerable and important biogenic habitats (Krigsman et al. 2012), marine habitat and population dynamics are rarely integrated. Furthermore, in both terrestrial and marine settings, species–habitat associations are usually assessed based on simple overlap and focus on identifying areas of high quality habitat for a given species or community (Johnson et al. 2013). Most habitat studies combine data collected over multiple years, so among-year differences in abundance and distribution are ignored (Boyce et al. 2002; Johnson et al. 2013). In

Received 23 September 2013. Accepted 25 June 2014.

Paper handled by Associate Editor Andrew Cooper.

A.O. Shelton, E.J. Ward, and B.E. Feist. Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, 2725 Montlake Blvd E, Seattle, WA 98112, USA.
J.T. Thorson. Fisheries Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, 2725 Montlake Blvd E, Seattle, WA 98112, USA.
Corresponding author: Andrew Olaf Shelton (e-mail: ole.shelton@noaa.gov).

Can. J. Fish. Aquat. Sci. **71**: 1–12 (2014) dx.doi.org/10.1139/cjfas-2013-0508

Published at www.nrcresearchpress.com/cjfas on 8 July 2014.

contrast, managers of marine resources tends to focus on time-series trends and typically ignore or integrate across spatial and habitat variation (but see Maunder et al. 2006; Lewy and Kristensen 2009; Rooper and Martin 2012). A typical approach is to estimate abundance by creating a grid in the ocean along latitude and longitude lines and estimate a mean density of fish within each grid cell (Brynjarsdóttir and Stefánsson 2004). With newly available spatially explicit habitat and marine survey data, the link between habitat and abundance surveys provide the potential to improve the accuracy of abundance estimates.

There are methodological and computational challenges for integrating habitat data into abundance estimates. First, as surveys are repeated over time, the time-series nature should be maintained. Surveys typically done in freshwater and terrestrial systems (e.g., distance sampling, mark–recapture) are prohibitive in many marine systems due to the magnitude of fish populations and the difficulty in sampling oceanic habitats. In marine ecosystems, most spatio-temporally explicit samples are taken from survey vessels and only rarely are the same locations sampled repeatedly over time. This makes traditional time-series approaches difficult to implement. Second, while spatially explicit marine covariates such as depth, temperature, and benthic substrate are generally available, other important habitat covariates are unmeasured or logistically unfeasible to measure. For example, the distribution of habitat-forming species such as corals and sponges remain unknown outside of localized areas (e.g., Krigsman et al. 2012). This requires models that explicitly account for abundance over time and can integrate both measured and unmeasured habitat variables.

We describe a spatial model for estimating abundance time-series of an exploited fish (darkblotched rockfish, *Sebastes crameri*) along the continental margin of the northeast Pacific. We compare this model, which takes advantage of newly available habitat data, with two existing approaches that rely on classical stratification methods. Our methodology maintains biological interpretability of model parameters yet accounts for poorly understood relationships between unobserved covariates and fish occurrence and abundance via a semiparametric model. Relative to classical stratification approaches, our spatial model provides information about abundance trends on a finer spatial grain and is robust with relatively small sample sizes. We show that our model's structure increases the precision of within- and among-year abundance estimates.

Methods

Statistical models

All three of our statistical modeling approaches share the same basic form, and use the same data. We write the response, catch in biomass of a species at position s_{iy}, as $Z(s_{iy})$, where i indexes the observation and y indexes year.

Design-based model

The classical nonparametric design-based estimate for biomass in region j and year y is calculated as the sample mean of catch divided by area swept for each tow, $Y(s_{iy})$,

$$D_{jy} = A_j n_{jy}^{-1} \sum_i^{n_{jy}} \frac{Z(s_{iy})}{Y(s_{iy})} \tag{1}$$

where A_j and n_{jy} are the total area and number of observations, respectively. The variance of each region is calculated as sample variance divided by sample size (Cochran 1977). While the design-based model provides an unbiased estimate of overall biomass, the statistical efficiency of this estimator is dependent upon the distribution of Z. In many ecological data sets (Thorson et al. 2011), Z can have a large number of zero observations with strong positive skew and wide tails (high kurtosis), leading to unbiased but imprecise estimates of biomass.

Stratified delta model

Species distributions with large proportions of zeros motivate the use of mixture distributions that can break the observed catches into two sub-models: a model describing the presence or absence of a species, and a model for the distribution of catches conditioned on the presence of fish (Stefánsson 1996; Maunder and Punt 2004). This approach has gained favor for its flexibility in modeling zero observations in sampling data (Martin et al. 2005). We write the random variable $Z(s_{iy})$ as a mixture distribution conditioned on model parameters. First, let $B(s_{iy})$ be a binary variable that equals 1 if the species is present and 0 otherwise, $B(s_{iy})|\phi(s_{iy}) \sim \text{Bernoulli}(\phi(s_{iy}))$. Here $\phi(s_{iy})$ is the probability of catching at least one individual; therefore, the probability of catching zero fish is $1 - \phi(s_{iy})$. Then conditionally,

$$\begin{aligned} &Z(s_{iy})\,\big|\,B(s_{iy}) = 0 \sim \delta_0 \\ &Z(s_{iy})\,\big|\,B(s_{iy}) = 1, \mu(s_{iy}), \psi \sim \text{Gamma}(\mu(s_{iy}), \psi) \end{aligned} \tag{2}$$

where δ_0 is the Dirac distribution at 0 and the second line provides the distribution of nonzero biomass conditional on the presence of the species. While a range of distributions could be used to model positive catches, we adopted a gamma distribution, $\text{Gamma}(\mu(s_{iy}), \psi)$ (Maunder et al. 2006; Lewy and Kristensen 2009) parameterized in terms of its mean, μ, and coefficient of variation, ψ[1]. The models are known in the fisheries literature as delta generalized linear mixed models (δ-GLMM; Stefánsson 1996; Maunder and Punt 2004) and more generally as hurdle models (Ver Hoef and Jansen 2007). The two components of the mixture are assumed to be independent. We used a logit-link function for the probability of occurrence and a log-link for the positive component, so the general form δ-GLMM is

$$\begin{aligned} \text{logit}(\phi_j(s_{iy})) &= \mathbf{X_1}(s_{iy})\boldsymbol{\beta_1} + \gamma_1 Y(s_{iy}) \\ \log(\mu_j(s_{iy})) &= \mathbf{X_2}(s_{iy})\boldsymbol{\beta_2} + \gamma_2 \log(Y(s_{iy})) \end{aligned} \tag{3}$$

with the subscripts making explicit that the covariates in the probability of occurrence model need not be identical to the covariates of the positive model. Here γ is an offset parameter that controls for variation in the area sampled by each survey trawl observation. We fix $\gamma_1 = \gamma_2 = 1$. All covariates in this model are based on strata defined by latitudinal breaks and depth, so the $\mathbf{X}$s are design matrices with categorical covariates. We hereafter refer to this model as the strata model.

Spatial habitat delta model

In our spatial δ-GLMM (hereafter, habitat model), all of the habitat variables are treated as continuous surfaces, so each observed trawl location has a distinct set of covariates associated with each observed point. The important distinction between the two delta models is that for the strata model the expected probability of occurrence and expected biomass are identical for each observation within predefined categories, whereas for the habitat model each point has a unique expectation. To incorporate unobserved habitat variables, we extend eq. 3 to include multivariate normal spatially smooth terms, $\boldsymbol{w}$, yielding a semiparametric model,

[1] The expected value of this parameterization is $E[x] = \mu$ and coefficient of variation, $CV[x] = \psi$. This gamma density can be connected to the more familiar $\text{Gamma}(\alpha, \beta)$ parameterization by substituting $\alpha = \psi^{-2}$ and $\beta = (\mu\psi^2)^{-1}$.

$$\begin{aligned}\text{logit}(\phi_j(s_{iy})) &= \boldsymbol{X_1}(s_{iy})\boldsymbol{\beta_1} + \gamma_1 Y(s_{iy}) + \boldsymbol{w_1}(\boldsymbol{s}) \\ \log(\mu_j(s_{iy})) &= \boldsymbol{X_2}(s_{iy})\boldsymbol{\beta_2} + \gamma_2 \log(Y(s_{iy})) + \boldsymbol{w_2}(\boldsymbol{s})\end{aligned} \tag{4}$$

The parameter $\boldsymbol{w}$ represents a spatially smooth surface that allows for locations to deviate from the predicted value driven by habitat variables alone. The habitat model also includes a fixed year effect that allows both the occurrence and positive models to shift up or down among years. For consistency among models we assumed identical effort offsets.

For the habitat model, an important consideration is how to model and interpret the spatial term $\boldsymbol{w}$. We model $\boldsymbol{w}$ as a smooth spatial surface (Cressie and Wikle 2011), $\boldsymbol{w} \sim \text{MVN}(0, \boldsymbol{C_w}(d, \theta, \sigma^2))$, where $\boldsymbol{C_w}(d, \theta, \sigma^2)$ is a covariance matrix with parameter θ that controls the correlation between points as a function of distance, d. We assume the spatial variance, σ^2, is homogeneous and use an isotropic exponential correlation model (Cressie and Wikle 2011). In practice $\boldsymbol{w}$ provides a flexible means to account for factors not explicitly accounted for by the habitat covariates (Latimer et al. 2009; Finley et al. 2009).

We construct our spatial correlation matrix to avoid confounding spatial and temporal components of variation. We only allow samples collected within a given year to have spatial covariance (i.e., the covariance is separable; Cressie and Wikle 2011). The covariance matrix, $\boldsymbol{C_w}(d, \theta, \sigma^2)$, is block-diagonal with elements comprised of year-specific spatial covariance matrices,

$$\boldsymbol{C_w} = \begin{bmatrix} \boldsymbol{C}_{2003} & 0 & \cdots & 0 \\ 0 & \boldsymbol{C}_{2004} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \boldsymbol{C}_{2011} \end{bmatrix} \tag{5}$$

where $\boldsymbol{C}_{2003}$ is the covariance matrix for observations in 2003, for example. This structure allows the annual spatial covariances to be independent, accounting for the possibility for interannual changes in species distribution to generate a distinct clustering pattern for each year. While we initially attempted to estimate a distinct θ and σ^2 for each year, this proved to be difficult so for all final estimates we used a single θ and a single σ^2 shared across all years. We note that an alternative approach would be to model yearly variation in θ and σ^2 using hierarchical structures. Because the two components of the mixture distribution are independent, the parameters θ and σ^2 for the two components are also estimated independently.

Due to the computational challenges that arise in the estimation of spatial models, we used predictive process models to reduce model dimension and abbreviate computing time (Banerjee et al. 2008; Latimer et al. 2009; Finley et al. 2009). Briefly, the predictive process approach develops an approximation of the full covariance matrix $\boldsymbol{C_w}$ using a much smaller covariance matrix. To do this, we establish a new set of points that are interspersed with the observed locations. These locations are known as knots, and the number of knots is much smaller than the number of observations. For the statistical model, we have to estimate a spatial component for each knot location, $\boldsymbol{w}^*$, instead of a spatial component for each observation. We estimate a spatial covariance matrix among the knots and predict the value of spatial effects at the observed points from the knots. The key advantage of introducing the knots is that we only have to calculate the inverse of the covariance matrix of the knots; because the length of $\boldsymbol{w}^*$ is much less than the length of $\boldsymbol{w}$, the computational savings are substantial. We used Bayesian methods to implement all of our models. Model implementation and other technical details are described in the Appendix.

Application to darkblotched rockfish

Darkblotched rockfish are a long-lived species (max. age >100 years) that range from Alaska to southern California, but they are most abundant from southern British Columbia to northern California (Love et al. 2002). They are found at depths of 100–600 m and individuals tend to migrate to deeper waters as they mature and age (Nichol 1990). Adults tend to rest on soft substrate near cobble and boulders. Since the 1950s, darkblotched rockfish have been fished commercially and were declared a species of conservation concern in 2000 (Gertseva and Thorson 2013). Between 2003 and 2011 (and continuing), darkblotched rockfish were sampled in fishery-independent trawl surveys (~750 trawls annually from mid-May to late October) as part of a larger effort by NOAA's Northwest Fisheries Science Center (NWFSC; Bradburn et al. 2011). For each trawl, the number and biomass of all fish are recorded, along with average depth and bottom water temperature. Detailed descriptions of the sampling design, gear, and sampling protocols used for this survey are found elsewhere (Keller et al. 2012). Note that because bottom trawls selectively sample fish (e.g., very small fish fit through the trawl mesh and will not be observed), we are estimating the biomass of fish that can be observed in the trawls, not necessarily the entire population. We compiled information on benthic sediment grain size and the location of rock outcrops from NMFS (2013).

We fit the three models to the darkblotched rockfish survey data using all 2003–2011 trawl surveys that occurred north of Point Conception, California (34.5°N latitude, 5090 tows in total). To illustrate the consequences of the different modeling assumptions for biomass, we compared three fisheries regions off the coasts of Washington and Oregon, USA (regions A, B, and C; Fig. 1). For the strata model, we treated strata and year effects as fixed in both probability of occurrence and positive components and interactions between strata and year as random (Thorson and Ward 2013).

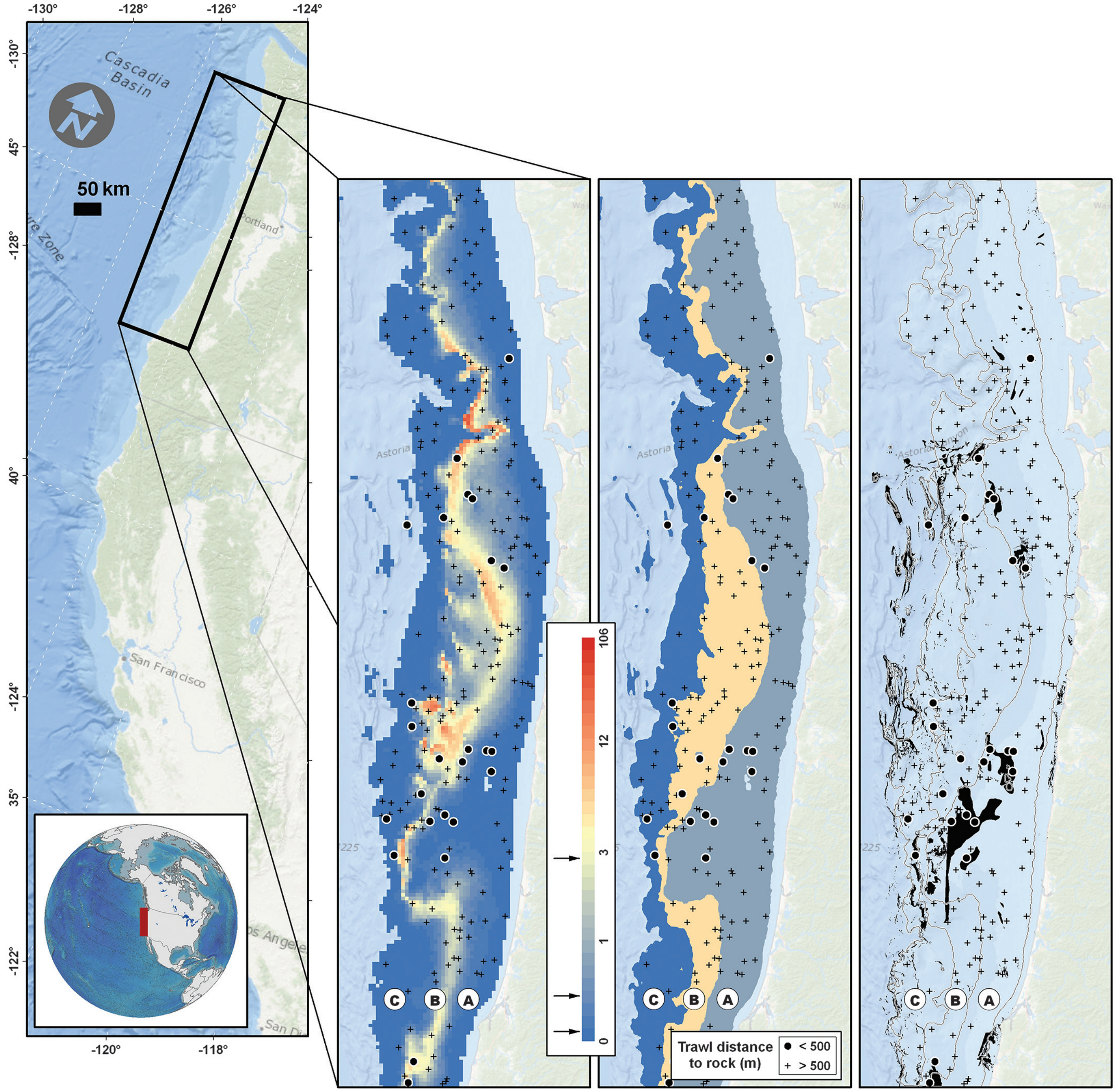
For the habitat model, we estimated the occurrence and positive model with four potential habitat variables: depth, sediment grain size, distance to nearest rock outcrop at least 1 ha in area, and bottom temperature. We considered linear and quadratic terms for habitat variables, but we did not consider interactions among covariates (see Appendix A).

Following model estimation, we generated an estimate of biomass for each model for the three regions off the Oregon and Washington coasts. For the design model, we calculated the estimated darkblotched biomass for each of the three target regions (eq. 1). For the strata model, we calculated the posterior density of darkblotched for each region, λ_{jy}, which has expected value, $E[\lambda_{jy}] = E[\phi_{jy}]E[\mu_{jy}]$. We calculated total biomass and its credible interval by expanding the biomass per hectare to the total number of hectares in each region. For the habitat model, we generated a posterior density surface for the entire region by calculating the predicted density for each grid cell on a regular 2 km × 2 km grid. For comparison, we provide biomass estimate and credible intervals for regions A, B, and C using the density predictions from the 2 km × 2 km grid and expanding the densities to the total area of each region.

Results and discussion

For our focal zone from 43 to 47.5°N latitude, we show strong differences in predicted densities for the stratified and habitat model in 2007 (Fig. 1; see also Fig. A1). The habitat model shows how considerable variation may exist within a region and how this variation is averaged over in the strata model (Fig. 1; results for the design model are very similar). Despite the variation within the strata, estimates of mean occupancy and biomass results are very similar between the design, strata, and habitat models in 2007 (Fig. 2; Fig. A2). For estimated total biomass, the three models provide distinct time series of total biomass in each region

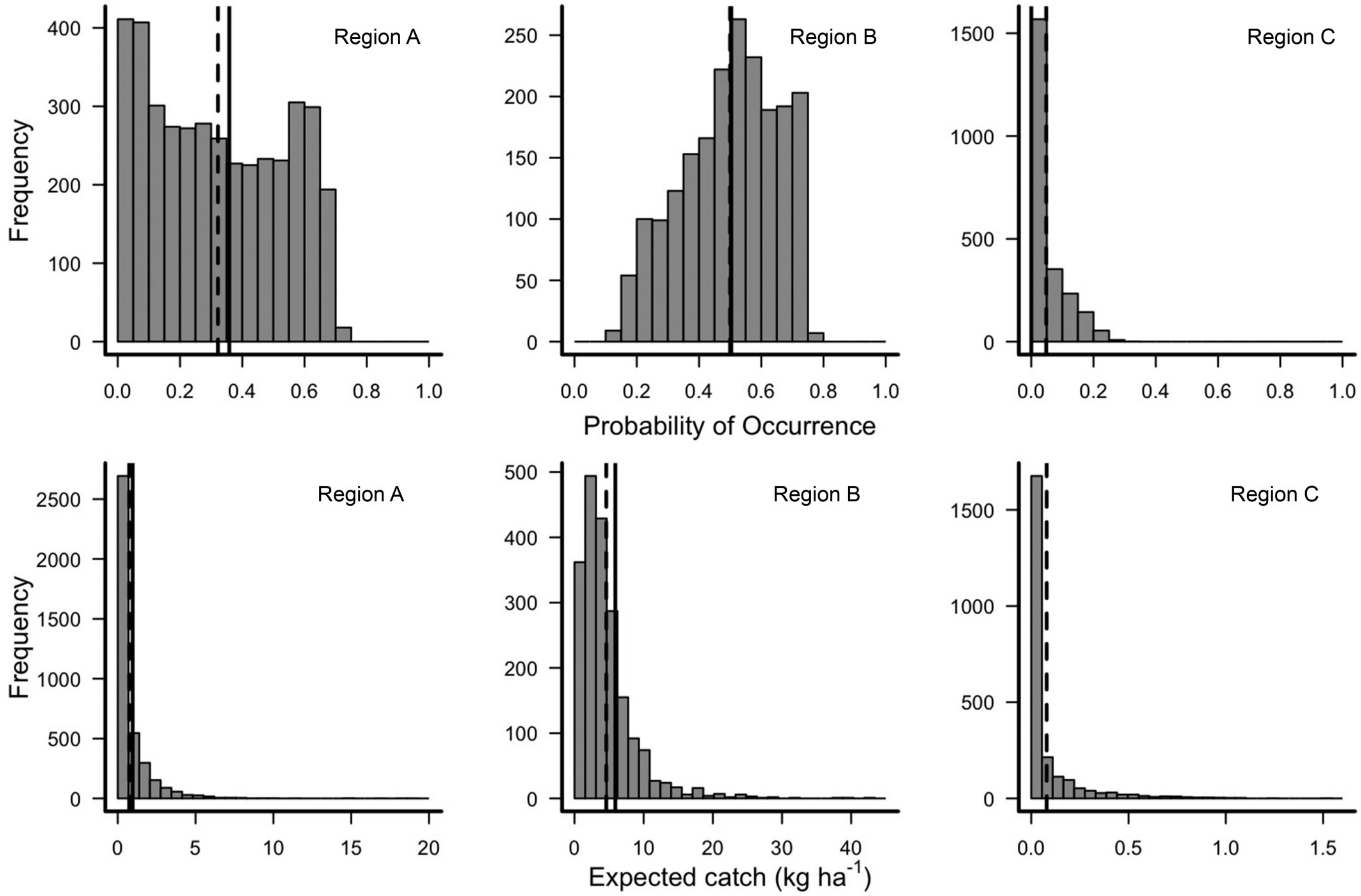
Fig. 1. Predicted darkblotched rockfish density (kg·ha^{-1}) in 2007. Crosses (+) indicate the location of survey trawls >500 m from rocky substrate, dots (•) indicate trawl location near rocky substrate (<500 m). (Left) Expected kilograms of darkblotched rockfish per hectare from the habitat model. Mean of the posterior predictive distribution is shown for the centroid of each 2 km × 2 km grid cell. Faint lines outline the regions used in the strata based models. (Center) Expected catch per hectare from the strata model in the three statistical regions. Arrows in the legend bar indicate the expected catch for the three regions. (Right) Black patches are rocky substrate. Faint lines delineate statistical region boundaries.



(Fig. 3). Because we lack distance-sampling and mark–recapture data to adjust both our occupancy and biomass estimates, our results for both model components must be considered proportional to the true biomass of fish available to the trawl gear (Royle et al. 2013). With regard to the uncertainty within a given year, the habitat model provides the most precise estimates of biomass for all three strata, whereas the design-based estimator is the least precise (average CV for region A: design = 0.51, strata = 0.38, habitat = 0.36; Table 1). The difference in among-year CV between the three models is smallest in region B, which has the most observations (Tables 1, 2).

The most striking difference among models is the decreased temporal variability in the habitat model relative to the design or strata models. Both the design and strata models show dramatic year-to-year fluctuations in biomass. In region B, for example, the strata model estimates >90% decrease from 2003 to 2004, a 12-fold increase from 2004 to 2005, before declining again by 90% from 2005 to 2006 (Fig. 3). In contrast, the habitat model shows some

Fig. 2. Results for the δ-GLMM for darkblotched rockfish. Top row: Histogram of distribution of predicted probability of occurrence for 2 km × 2 km grid cells in 2007 for regions A (left; *N* = 3966 prediction locations), B (center; *N* = 1972), and C (right; *N* = 2477). Bottom row: Histogram of expected catch for the three strata. In all panels, the dashed line shows the mean for the habitat model and the solid line the mean from the stratified model. Expected catch for the strata model is undefined for region C (see Appendix A). Note the very long tails in the abundance models.



temporal variation but all biomass estimates in our time series fall between 1900 and 5300 mt for region B — less than a 3-fold range (Fig. 3), and this level of interannual variability is more consistent with the results of the most recent stock assessment (Gertseva and Thorson 2013).

Throughout this paper we have avoided discussion of model selection to compare the three candidate models. This is partially due to the difficulty is performing Bayesian model selection among models that make fundamentally different assumptions about the data. Furthermore, estimation problems for the strata and design models make it difficult to perform formal model comparison. If the model has not converged or are poorly estimated, model selection is affected. This occurs regularly in stratified models when there are few positive observations in a given strata (as is the case for region C here). The options for addressing this problem include redefining the strata or changing the sampling design. Both options are time intensive and, in the case of the multi-species NWFSC trawl survey, this would likely involve developing a different set of strata for each species. Thus using a habitat model has the added advantage of avoiding subjective strata boundaries by using a smooth habitat surface that does not depend upon subjective boundaries.

By itself, the estimated biomass time series does not provide the final word on the status of darkblotched rockfish. A full stock assessment of darkblotched rockfish incorporates information about age- and length-structure of the population, fisheries catch and effort, and the catchability of the population to make conclusions about the status of the stock (Gertseva and Thorson 2013). However, even in this case, improved biomass estimates could improve the estimates of other biological parameters and final determination of stock status.

Understanding the cause of among-year variation in biomass estimates is crucial because it determines how informative data are regarding population trends. Differences among models suggest that much of the strata model among-year variation is caused by random assignment of sampling locations combined with strong habitat associations of darkblotched rockfish. As sample locations vary among years, survey tows fall on “good” darkblotched habitat in some years and “poor” habitat in other years. For example, in 2007, 16 surveys tows (13%) in region A occurred within 500 m of a rocky substrate, whereas in 2008 only 8 (7%) occurred within 500 m of rocky bottom (Fig. 1; Fig. A1). Darkblotched rockfish are negatively associated with rocky habitat (see Appendix A), so the location of trawls will strongly affect biomass estimates. The extreme skew of the biomass observations is an important determinant of temporal variation; there are many catches near zero but a small number of extreme observations (1026 tows with >0 catch; median catch = 1.25 kg but 51 observations >50 kg, 18 observations >100 kg, and 3 observations >1000 kg).

Rare, large catches known as extreme catch events (Thorson et al. 2011) pose a particular challenge for the strata model because a small number of above-average observations exert high statistical leverage and cause the expected density in a given region to increase disproportionately. Indeed, extreme catches (>100 kg)

Fig. 3. Comparison of abundance estimates for the three models and three regions: A (top), B (middle), and C (bottom). Points show posterior median, interquartile range, and 95% credible intervals estimates for the habitat and strata models while the design model shows mean, interquartile range, and 95% confidence intervals. Note that the *y* axis varies substantially among panels.

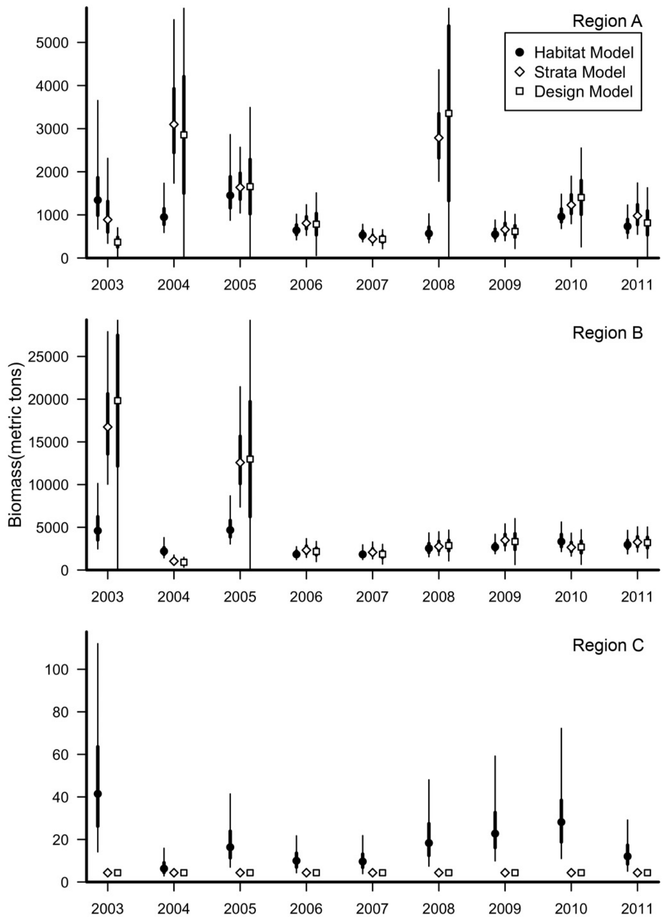


Table 1. Average coefficient of variation of total biomass by strata for 2003–2011.

	Model		
Region	Design	Strata	Habitat
A	0.51	0.38	0.36
B	0.41	0.33	0.33
C	n/a	n/a	0.59

Note: Values are the across year average of the coefficient of variations.

Table 2. Number of survey trawls by year for each strata, with number of trawls with nonzero biomass in parentheses.

	Region		
Year	A	B	C
Depth:	55–183 m	183–549 m	549–1280 m
2003	45 (9)	40 (24)	60 (0)
2004	79 (19)	49 (29)	26 (0)
2005	119 (37)	51 (21)	51 (0)
2006	112 (43)	63 (35)	42 (0)
2007	123 (49)	71 (36)	46 (0)
2008	111 (34)	62 (29)	43 (0)
2009	100 (33)	78 (38)	41 (0)
2010	114 (40)	62 (30)	38 (0)
2011	105 (22)	85 (40)	49 (0)

occurred in region A in 2004 and 2008 and catches of more than 1000 kg occurred in 2003 and 2005 in region B. These observations correspond to the largest differences between the habitat, strata, and design models (Fig. 3). The habitat model is less sensitive to these extreme events because different locations have distinct expected densities based on their habitat variables and spatial nonparametric term. Thus, in the presence of extreme catches, the strata model adjusts by dramatically changing the biomass estimate, but the habitat model accounts for it via its environmental variables and spatial correlations. Observation of variation above the rate associated with habitat variables is accounted for through the spatial variable ***w***. In the case of extreme observations, the spatial contribution for that location is very large, but since we assume that ***w*** is a smooth spatial surface, only the biomass estimates at points near to this extreme observation are affected. In contrast, the strata and design models assume all data are arising from a single shared process, with all locations providing exchangeable samples of a single mean. Thus, rare extreme observations will have a large effect on total biomass estimates.

Additionally, the habitat model accounts for variation by changing the biological interpretation of model parameters in eq. 2. In the strata model, all observations within a stratum are considered exchangeable, independent samples of a single shared mean, μ, so ψ represents the CV of the single mean within a strata. In the habitat model, ψ represents the CV for the observations at a particular location — each trawl has a distribution of possible outcomes whose mean is driven by the habitat covariates and the spatial location — and so the ψ describes the variation in outcomes from a single trawl observation. This addition of an observation error term aids in accounting for extreme observations.

Finally, our analysis has substantial implications for an entire class of biological questions built on the analysis of abundance time series in terrestrial and aquatic systems. Many meta-analytic analyses treat published time series of abundance derived from design-based estimators and treat such model output as data in subsequent analyses (Myers et al. 1999; Knape and de Valpine 2012). We note that if darkblotched rockfish are any indication, reasonable modeling structures may yield radically different empirical patterns of temporal variability of populations. For example, the California Cooperative Oceanic Fisheries Investigations (CalCOFI) survey has focused on understanding fluctuations of larval abundance and how complex dynamic models can be used to explain large temporal variation in larval abundance (e.g., Hsieh et al. 2005*a*, 2005*b*). All of these analyses are derived from time series of design-based estimates of abundance. The estimators ignore environmental and oceanographic variables associated with larval sampling. We suggest that improved precision for abundance estimates may be possible if these surveys utilize available habitat information and that different estimators could potentially alter the conclusions of many such meta-analytic analyses.

Acknowledgements

The authors thank M. Bellman, N. Cadigan, V. Gertseva, Y. Jiao, P. Levin, M. McClure, C. Romsos, I. Stewart, I. Taylor, N. Tolimieri, W. Wakefield, and an anonymous reviewer for providing thoughtful discussions and comments on earlier versions of this manuscript. This work was partially supported by NOAA's NMFS Habitat Assessment Improvement Plan grant. The authors would also like to thank the entire survey team for the NWFSC shelf-slope survey.

References

Banerjee, S., Gelfand, A.E., Finley, A.O., and Sang, H. 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. B Stat. Method. **70**: 825–848. doi:10.1111/j.1467-9868.2008.00663.x.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., and Schmiegelow, F.K.A. 2002. Evaluating resource selection functions. Ecol. Model. **157**: 281–300. doi:10.1016/S0304-3800(02)00200-4.

Bradburn, M.J., Keller, A.A., and Horness, B.H. 2011. The 2003 to 2008 US West Coast bottom trawl surveys of groundfish resources off Washington, Oregon, and California: estimates of distribution, abundance, length, and age composition. NOAA, NMFS, Northwest Fisheries Science Center.

Brynjarsdóttir, J., and Stefánsson, G. 2004. Analysis of cod catch data from Icelandic groundfish surveys using generalized linear models. Fish. Res. **70**: 195–208. doi:10.1016/j.fishres.2004.08.004.

Cochran, W.G. 1977. Sampling techniques. John Wiley & Sons Inc., New York.

Cressie, N., and Wikle, C.K. 2011. Statistics for spatio-temporal data. John Wiley & Sons Inc., New York.

Finley, A.O., Sang, H., Banerjee, S., and Gelfand, A.E. 2009. Improving the performance of predictive process modeling for large datasets. Comp. Stat. Data Analy. **53**: 2873–2884. doi:10.1016/j.csda.2008.09.008.

Gertseva, V., and Thorson, J.T. 2013. Status of the darkblotched rockfish resource off the continental U.S. Pacific Coast in 2013. NOAA, NMFS, Northwest Fisheries Science Center.

Hsieh, C.-H., Reiss, C., Watson, W., Allen, M.J., Hunter, J.R., Lea, R.N., Rosenblatt, R.H., Smith, P.E., and Sugihara, G. 2005*a*. A comparison of long-term trends and variability in populations of larvae of exploited and unexploited fishes in the Southern California region: a community approach. Prog. Oceanogr. **67**: 160–185. doi:10.1016/j.pocean.2005.05.002.

Hsieh, C.-H., Glaser, S.M., Lucas, A.J., and Sugihara, G. 2005*b*. Distinguishing random environmental fluctuations from ecological catastrophes for the North Pacific Ocean. Nature, **435**: 336–340. doi:10.1038/nature03553.

Johnson, A.F., Jenkins, S.R., and Hiddink, J.G. 2013. Linking temperate demersal fish species to habitat: scales, patterns and future directions. Fish Fish. **14**(3): 256–280. doi:10.1111/j.1467-2979.2012.00466.x.

Keller, A.K., Wallace, J.R., Horness, B.H., Hamel, O.S., and Stewart, I.J. 2012. Variations in eastern North Pacific demersal fish biomass based on the U.S. west coast groundfish bottom trawl survey (2003–2010). Fish. Bull. **110**: 205–222.

Knape, J., and de Valpine, P. 2012. Are patterns of density dependence in the Global Population Dynamics Database driven by uncertainty about population abundance? Ecol. Lett. **15**: 17–23. doi:10.1111/j.1461-0248.2011.01702.x.

Krigsman, L.M., Yoklavich, M.M., Dick, E.J., and Cochrane, G.R. 2012. Models and maps: predicting the distribution of corals and other benthic macro-invertebrates in shelf habitats. Ecosphere, **3**:art3. http://dx.doi.org/10.1890/ES11-00295.1.

Latimer, A.M., Banerjee, S., Sang, H., Jr., Mosher, E.S., and Silander, J.A., Jr. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecol. Lett. **12**: 144–154. doi:10.1111/j.1461-0248.2008.01270.x. PMID:19143826.

Lewy, P., and Kristensen, K. 2009. Modelling the distribution of fish accounting for spatial correlation and overdispersion. Can. J. Fish. Aquat. Sci. **66**(10): 1809–1820. doi:10.1139/F09-114.

Love, M.S., Yoklavich, M., and Thorsteinson, L. 2002. The rockfishes of the northeast Pacific. University of California Press, Berkeley and Los Angeles, Calif.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., and Possingham, H.P. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol. Lett. **8**: 1235–1246. doi:10.1111/j.1461-0248.2005.00826.x.

Maunder, M.N., and Punt, A.E. 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. **70**: 141–159. doi:10.1016/j.fishres.2004.08.002.

Maunder, M.N., Hinton, M.G., and Bigelow, K.A. 2006. Developing indices of abundance using habitat data in a statistical framework. Bull. Mar. Sci. **79**: 545–559.

Myers, R.A., Bowen, K.G., and Barrowman, N.J. 1999. Maximum reproductive rate of fish at low population sizes. Can. J. Fish. Aquat. Sci. **56**(12): 2404–2419. doi:10.1139/f99-201.

National Marine Fisheries Service (NMFS). 2013. Groundfish essential fish habitat synthesis: a report to the Pacific Fishery Management Council. NOAA, NMFS, Northwest Fisheries Science Center, Seattle, Wash.

Nichol, D.G. 1990. Life history examination of darkblotched rockfish (*Sebastes crameri*) off the Oregon coast. Masters thesis, Oregon State University.

Rooper, C.N., and Martin, M.H. 2012. Comparison of habitat-based indices of abundance with fishery-independent biomass estimates from bottom trawl surveys. Fish. Bull. **110**: 21–35.

Royle, J.A., Chandler, R.B., Gazenski, K.D., and Graves, T.A. 2013. Spatial capture-recapture models for jointly estimating population density and landscape connectivity. Ecology, **94**: 287–294. doi:10.1890/12-0413.1. PMID:23691647.

Sbrocco, E.J., and Barber, P.H. 2013. MARSPEC: ocean climate layers for marine spatial ecology. Ecology, **94**: 979. doi:10.1890/12-1358.1.

Sillett, T.S., Chandler, R.B., Royle, J.A., Kéry, M., and Morrison, S.A. 2012. Hierarchical distance-sampling models to estimate population size and habitat-specific abundance of an island endemic. Ecol. Appl. **22**: 1997–2006. doi:10.1890/11-1400.1.

Stefánsson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. **53**: 577–588. doi:10.1006/jmsc.1996.0079.

Thorson, J.T., and Ward, E.J. 2013. Accounting for space–time interactions in index standardization models. Fish. Res. **147**: 426–433. doi:10.1016/j.fishres.2013.03.012.

Thorson, J.T., Stewart, I.J., and Punt, A.E. 2011. Accounting for fish shoals in single- and multi-species survey data using mixture distribution models. Can. J. Fish. Aquat. Sci. **68**(9): 1681–1693. doi:10.1139/f2011-086.

Ver Hoef, J.M., and Jansen, J.K. 2007. Space-time zero-inflated count models of Harbor seals. Environmetrics, **18**: 697–712. doi:10.1002/env.873.

Ward, T.J., Vanderklift, M.A., Nicholls, A.O., and Kenchington, R.A. 1999. Selecting marine reserves using habitats and species assemblages as surrogates for biological diversity. Ecol. Appl. **9**: 691–698. doi:10.1890/1051-0761(1999)009[0691:SMRUHA]2.0.CO;2.

Appendix A

Methods

Details of the habitat model

Covariate selection

We used only habitat data for covariates that were available for the entire spatial domain of the trawl survey. We used available data layers from the Essential Fish Habitat (EFH) Phase 1 report primarily (NMFS 2013). We included depth, sediment grain size, bottom temperature, and distance to nearest rocky habitat in our initial model. Distance to rock was calculated using the Nearest Features tool (Jenness Enterprises, v. 3.8b) in ESRI ArcView (v. 3.2a) to calculate the distance from each of the trawl survey sites to the nearest rock habitat patch. Rock was defined as any grid cell in the substrate type data layer with a value of 1 or 4. We only used rocky patches greater than 1 ha in area. All habitat covariates were centered before model estimation. While we expect many of the habitat attributes of the trawl locations such as depth and bottom type to be constant across the entire trawl time series, we know that other factors that we could not include are also affecting fish populations. For example, the total biomass of a particular species may be changing over time — declining due to fishing pressure or poor recruitment or increasing due to fishing restrictions or favorable oceanographic conditions. Therefore, we also included the option of estimating a fixed categorical value for each year. Adding such a year effect allows for the probability of occurrence and overall abundance to vary across the time series. We do not allow for interactions between the categorical year effects, and thus we assumed a constant effect of habitat variables across years and only allowed for a discrete shift up or down between years. Recall, however, that the spatial effect, $\boldsymbol{w}$, allows for the deviation from this overall habitat mean to vary spatially among years.

We initially considered including information about biogenic habitats, but our initial survey of available data for biogenic habitats concluded that the data were too limited in quality and their spatial extent to be included in our model.

We do not include any region-specific categorical variables in our model. Because using such categorical variables require delineating boundaries between regions, and they generate discontinuities in the occurrence and abundance along region boundaries. We wished to avoid creating these boundaries and any discussion about the division of the coast into discrete regions. We avoid arbitrarily imposing a spatial structure for species–habitat relationships by using the spatial component of the model to provide flexibility in accounting for continuous latitudinal variation.

We only considered models using the main effects of the habitat covariates and did not consider any interactions among the covariates. This small number of parameter ensures that the parameters maintain biological interpretability. The small model dimension also reduces the likelihood of model over-fitting and reduces the importance of performing extensive cross-validation testing to avoid overfitting.

After exploratory analysis, we elected to transform depth and distance to rock outcrop before their inclusion in the models. We $\log_e$-transformed depth and square-root transformed the distance to the nearest rock outcrop. The transformation of depth improved the explanatory value of depth, and the square-root transformation was effective at increasing the contrast between locations that are in close proximity to rocky outcrop, reducing the statistical leverage of points at great distances from any rock outcrop.

Computational issues for spatial models

Spatial data present a series of computational problems. In particular, when the number of observations gets large, standard procedures to estimate parameters in a point process model become computationally difficult and slow (Banerjee et al. 2008). The computational issues are entirely driven by the spatial term, $\boldsymbol{w}$. Problems arise because the covariance matrix $\boldsymbol{C_w}$ is large and not diagonal. Because estimation involves calculating the matrix inverse of $\boldsymbol{C_w}$, computation can be exceedingly slow. This is known as the large N problem in spatial statistics (Banerjee et al. 2008; Cressie and Wikle 2011) and remains a difficult problem even when explicit matrix inversion is replaced with fast linear solvers.

A number of approaches based on approximating the covariance matrix have been proposed to speed the computation of spatial models (Royle and Wikle 2005; Banerjee et al. 2008). We employ the predictive process modeling approach to improve model computation speed. A thorough discussion of predictive process approach can be found elsewhere (Banerjee et al. 2008; Latimer et al. 2009; Finley et al. 2009), so we only outline the methods here.

Briefly, the predictive process approach develops an approximation of the full covariance matrix $\boldsymbol{C_w}$ using a much smaller covariance matrix. To do this, we establish a new set of points that are interspersed with the observed locations. These locations are known as knots, and the number of knots is much smaller than the number of observations. For the statistical model, we have to estimate a spatial component for each knot location, $\boldsymbol{w}$*, instead of a spatial component for each observation. We estimate a spatial covariance matrix among the knots and predict the value of spatial effects at the observed points from the knots. The key advantage of introducing the knots is that we only have to calculate the inverse of the covariance matrix of the knots; because the length of $\boldsymbol{w}$* is much less than the length of $\boldsymbol{w}$, the computational savings are substantial. We employ the "modified" predictive process model described by Finley et al. (2009). This modified model contains an adjustment parameter estimates to avoid bias in the estimation of spatial parameters. Bayesian estimation of $\boldsymbol{w}$ also allows for uncertainty in estimates of the parameter θ.

The use of predictive process models requires the consideration of two additional model aspects. The number of knots needs to be specified and the location of knots needs to be determined. Using a smaller number of knots will speed computation time but result in a smoother, less rugose spatial surface compared with a model that uses the raw data (Banerjee et al. 2008). Following some preliminary exploration, we used 150 knots for the probability of occurrence model. For the positive component, we used 91 knots. To determine the knot locations we selected a single set of knot locations using a k-means clustering algorithm on all years of observations simultaneously (via the k-means function in R). We then used this single set of knot locations for each year in the model estimation.

We used a block-diagonal covariance matrix to avoid confusing temporal and spatial variation (see eq. 5). Initially we considered model structures in which σ_y^2 was allowed to vary among years as well as models in which it was constant among years (i.e., $\sigma_y^2 = \sigma^2$). Estimating a distinct variance for each year did not greatly change model predictions because estimated yearly variances were very similar among years. However, attempting to estimate yearly variance slowed model convergence and mixing without dramatically improving model fit. Thus in the final runs we always used a single, shared σ^2. Similarly, we estimate a single θ for all years in the final model runs. These model assumptions force the scale of spatial aggregation to be similar in all years, but allows the location of spatial aggregations to vary.

Table A1. A list of the habitat covariates included in the darkblotched rockfish statistical model.

Habitat covariates	Forms included in the model
Depth (m)	Log(depth) Log(depth)^2
Bottom temperature (°C)	Bottom temperature (Bottom temperature)^2
Sediment grain size (Φ scale; Krumbein and Sloss 1963)	Grain size (Grain size)^2
Distance to nearest rocky outcrop (km)	(km)$^{0.5}$

Estimation of the habitat model

General descriptions for estimating predictive process models can be found in Banerjee et al. (2008), Finley et al. (2009), and Latimer et al. (2009). Freely available software for parameter estimation using predictive process models and associate can be found in the R package spBayes on the CRAN website (http://cran.us.r-project.org/). Excellent tutorials are also available there. We implemented our own code in R to make use of the block-diagonal covariance matrix (eq. 5) and speed computation for our specific situation.

MCMC details

Priors

A key component of Bayesian models is the specification of prior distributions for the parameters. By tradition, noninformative priors have been used in most ecological and fisheries applications. Table A2 summarizes the prior distributions for the parameters. We used diffuse multivariate prior distributions for the regression parameters, conjugate inverse-gamma distributions for σ^2, and uniform distributions for θ and ψ. We constrained the scale parameter θ to the range {20 1000} for darkblotched rockfish in the probability of occurrence model based on visual inspection of the spacing of trawl survey locations with the intention of precluding the possibility of estimating spatial structure that is at a finer scale than the survey data. Because the abundance part of the model only includes nonzero observations and thus comprises a smaller subset of the data, the density of observations decreased and the distance between observations increased. Therefore, we used $\theta \sim$ Unif (50, 1000) for the positive model.

Full conditional distributions

We are interested in calculating the posterior density for the parameters and latent states given the observed data. Let $z_1(\boldsymbol{s})$ represent the observed presence–absence data of the model, then the full posterior for the presence–absence component can be written,

$$p[\boldsymbol{\phi}(s), \boldsymbol{w}^*, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{z}_1(s)] \propto p[\boldsymbol{z}_1(s) | \boldsymbol{\phi}(s)] p[\boldsymbol{\phi}(s) | \boldsymbol{w}^*, \boldsymbol{\beta}, \ \theta, \sigma^2] \\ p[\boldsymbol{w}^* | \theta, \sigma^2] p[\boldsymbol{\beta}, \theta, \sigma^2] \tag{A.1}$$

with the right hand side showing how the posterior can be factored into four components. We can write a similar model for the abundance model. We estimated $\boldsymbol{\phi}$ and $\boldsymbol{w}$* as latent states. To understand the value of the nonparametic term, $\boldsymbol{w}$*, we also ran a set of nonspatial habitat models (i.e., $\boldsymbol{w} = 0$) for comparison. The nonspatial habitat model is a regression model that assumes all of the observations are independent. These nonspatial habitat models do not involve estimating $\boldsymbol{w}$*, θ, or σ^2. For models with and without $\boldsymbol{w}$* we use a mix of Gibbs and Metropolis–Hastings sampling steps to estimate parameters (Gelman et al. 2013). To the nonspatial habitat models we added a small, fixed amount of pure

Table A2. Prior distributions used in the habitat statistical model.

Parameter	Probability of occurrence	Abundance
$\boldsymbol{\beta}$	Multivariate normal (0, $100^2\boldsymbol{I}$)	Multivariate normal (0, $100^2\boldsymbol{I}$)
σ_y^2	Inverse-Gamma (3, 1)	Inverse-Gamma (0.75, 0.5)
θ	Uniform (20, 1000)	Uniform (20, 1000); Uniform (50, 1000)
Ψ	n/a	Uniform (0.1, 5)

Note: **I** is the identity matrix.

error to the model to ease MCMC sampling: e.g., the probability of occurrence is then $\text{logit}(\phi(s_i)) = \boldsymbol{X}(s_i)\boldsymbol{\beta} + \varepsilon_i$, where ε_i are independent and $\varepsilon_i \sim N(0, \tau^2)$. The positive model had an analogous form. For the probability of occurrence and abundance models, we set $\tau^2 = 0.01$.

Due to a large number of models under consideration, we initially ran a single MCMC chain for each model. For the models that appeared to best match the data, we ran subsequent MCMC chains from dispersed starting point to verify convergence to a single stationary distribution. For probability of occurrence models, visual inspection of chains suggested strata models converged relatively slowly but had decent mixing properties. Thus we ran a very long burn-in chain of 100 000 iterations and a monitoring chain of 50 000 iterations. While both of these chain lengths were excessive, the chain length removed any questions about model convergence. We then used the ending values from the strata model to initiate the habitat models. The main advantage of initiating the habitat model at the converged nonspatial habitat model was that it provided values of the latent variables ϕ that were relatively close to each observation. The estimates of $\boldsymbol{\beta}$ often changed substantially with the addition of the spatial surface described by $\boldsymbol{w}$*, but reasonable starting points for ϕ greatly improved model convergence speed. Because habitat models ran much more slowly and the parameter values were already near their stationary distribution, we ran a 20 000 iteration burn-in and a 10 000 iteration monitoring run. In most cases, the MCMC chains for the spatial model converged but mixed relatively slowly (i.e., MCMC draws from the stationary distribution were highly autocorrelated). Therefore, we performed three independent MCMC runs for the probability of occurrence model and combined the output.

The positive model had better MCMC characteristics overall. We used a burn-in of 30 000 and a monitored MCMC of 50 000 iterations for the strata model and a burn-in of 5000 and monitored MCMC of 10 000 iterations for the habitat model. We ran multiple chains to confirm convergence.

Model selection for the habitat model

In this section, we discuss how we compare among the possible spatial habitat models using posterior predictive scoring rules. Generally, we are interested in identifying models that make good predictions. A way of formalizing this desire for good predictions is to say that we want to maximize the predicted probability of observing the value of a new data point, z_{new}, given our previously observed data and our estimated parameters. For notational simplicity, let $\boldsymbol{\Theta}$ be the estimated parameters and latent variables in the model. Thus, a good model would be one that provides a large value of $p(z_{\text{new}}|\boldsymbol{z}, \boldsymbol{\Theta})$. Proper rules for comparing a data value z_{new} with its predictive distribution involve the logarithm of the height of $p(z_{\text{new}}\,|\boldsymbol{z}, \boldsymbol{\Theta})$, or $\log(p(z_{\text{new}}\,|\boldsymbol{z}, \boldsymbol{\Theta}))$ (Gneiting and Raftery 2007; Krnjajić et al. 2008; Draper and Krnjajić 2010; Draper 2013). This metric of predictive quality is known as the log-score (LS).

Ideally, we would estimate $\log(p(z_{\text{new}}\,|\boldsymbol{z}, \boldsymbol{\Theta}))$ via cross-validation; we would exclude some set of our observations from our model estimation procedure and predict those excluded values. This suggests we would need to run a number of MCMC models for each covariate and each run would have a different set of data points excluded from model estimation (e.g., Draper and Krnjajić 2010; Shelton et al. 2012; Draper 2013). In practice, this is impractical due to the long computing times for models estimated with MCMC. Fortunately, with reasonably large sample sizes, we can use what is known as the full sample log-score that will approximate the cross-validation derived log-score (Krnjajić et al. 2008; Draper and Krnjajić 2010; Draper 2013). For each draw of the MCMC, g, we calculate the predicted probability of each observed data point, i, then

$$\text{(A.2)} \qquad \text{LS} = \sum_{g=1}^{G}\sum_{i=1}^{n}\log\Big[p\Big(z_i\big|z, \Theta^g\Big)\Big]$$

here n is the number of observations and G is the number of MCMC iterations. Larger log-scales indicate a higher overall match between prediction and observations. An alternative scoring criterion would be to divide the right side of eq. A.2 by n to provide log-scores on a per observation basis.

For the habitat model, we ran a series of models (nonspatial and spatial) that included various combinations of predictor variables and selected the models that maximized LS. For both the probability of occurrence and positive models we included a fixed year effect. We also inspected the posterior distribution of parameters and confirmed that the regression parameters for the final models were centered away from 0. In all cases, models that included the spatially smooth term $\boldsymbol{w}$ were strongly preferred over nonspatial models.

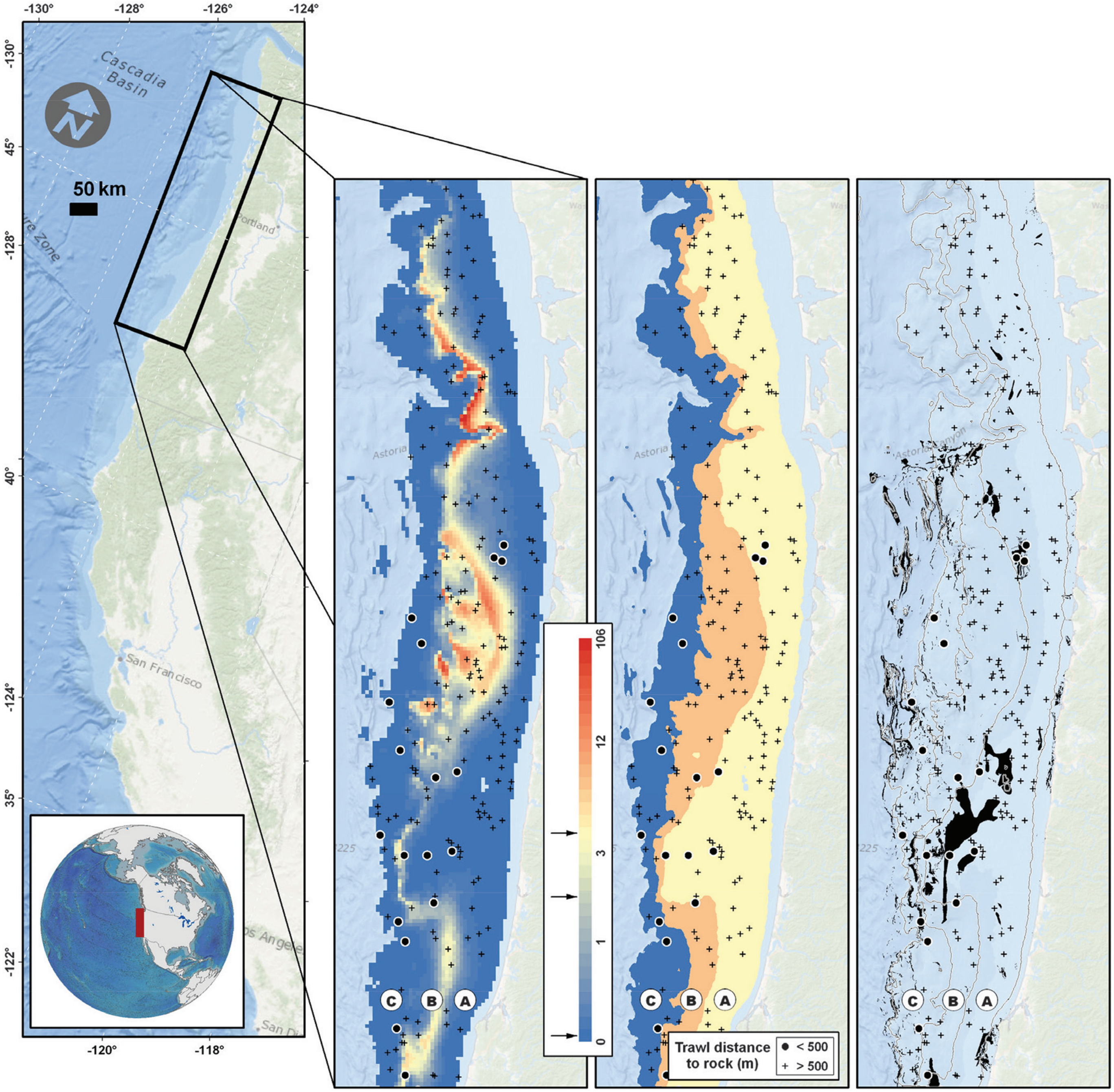
Constructing prediction maps

After producing posterior distributions for model parameters and the spatial latent variables at the knot locations, we used draws from the joint posterior distribution to generate predictive map for probability of occurrence and for abundance. These two surfaces correspond to a surface for ϕ and a surface for $\boldsymbol{\mu}$, respectively. These two surfaces can be combined to provide a surface for the expected value of catch, ***E[Z]***.

We first generated a gridded (2 km × 2 km) coast-wide map of the model spatial domain. The north–south extents of the domain approximated the US border, while the shoreline and seaward boundaries were defined by a vector shoreline geospatial data layer (NOAA 2001), and the 1600 m isobath (3-arcsecond grain, (~86 m) NOAA 2003), respectively. We created the 2 km × 2 km gridded polygon data layer using Generate Regular Points in ArcMap, which is a Hawth's Tools ArcGIS tool that runs in ArcMap (v. 9.3.1). We overlaid this gridded domain with the four habitat covariate data layers and calculated the corresponding values for each of the grid cells. Since the covariates were continuous variables, each was expressed as an area weighted mean for each of the grid cells.

We use $\boldsymbol{s_0}$ to denote the predicted grid centers along the coast. For depth, sediment grain size, and distance to rock outcrop, the covariate values at each location were consistent across years. For bottom temperature, we did not have a direct measure for each of the 2 km × 2 km grid cells, so we used the trawl survey site bottom temperature data to interpolate a gridded surface of bottom temperature for each year (2003–2011). We used the kriging command ESRI ARC/INFO grid (v. 9.2) to interpolate bottom temperature. We interpolated bottom temperature on a 1 km × 1 km grid for each year of the trawl survey data using the following kriging parameters: model domain polygon used as "barrier cover"; SPHERICAL semivariogram model for kriging method; maximum of 12 neighboring input sample points and, 100 km search radius to select neighboring points. We also used these interpolated bottom temperature data layers to fill in missing bottom temperature in 272 of the bottom trawl survey sites.

Fig. A1. Predicted darkblotched rockfish density (kg·ha^{-1}) in 2008. Crosses (+) indicate the location of survey trawls at least 500 m from rocky substrate, dots (•) indicate trawl location near rocky substrate (<500 m). (Left) Expected catch per hectare in kilograms of darkblotched rockfish from the habitat model in 2008. Mean of the posterior predictive distribution is shown for the centroid of 2 km × 2 km cells. Faint lines outline the areas used in the strata based models. (Center) Expected catch per hectare from the strata model in the three statistical areas. Arrows in the legend bar indicate the expected catch for the three areas. (Right) Black patches delineate the location of rocky substrate in the areas. Faint lines delineate statistical region boundaries.

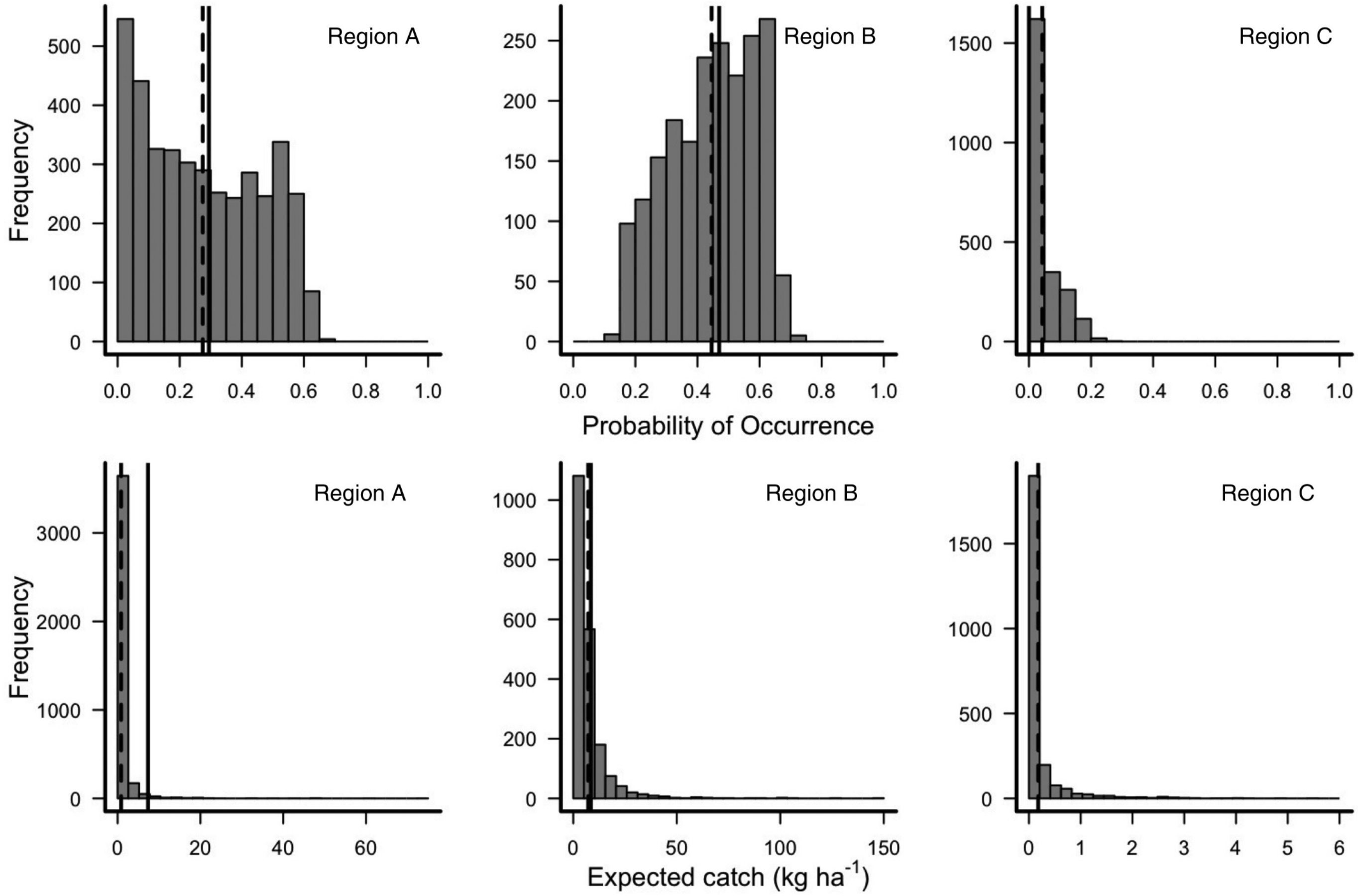


We used a slightly different approach for calculating distance to the nearest rocky habitat patch for the 2 km × 2 km gridded data layer. Calculating the distance from the centroids of each of the ~43 000, 2 km × 2 km grid cells to the nearest edge of each of the rocky habitat patches exceeded the capabilities of the Nearest Features tool that we used in generating the covariates for each of the bottom trawl survey sites, so we used the NEAR command in ESRI ARC/INFO (v. 9.2), which is a more robust software package.

Each year modeled had a distinct offset (intercept) corresponding to a coast-wide change in the probability of occurrence or abundance. Given these maps, we can generate predicted values for $\boldsymbol{s_0}$. For the probability of occurrence model, we generate predicted values at point $\boldsymbol{s_0}$ for the gth draw from the posterior,

$$\text{(A.3)} \qquad \operatorname{logit}(\phi^g(s_0)) = \mathbf{X}_1(\boldsymbol{s}_0)\boldsymbol{\beta}_1^g + \gamma_1 Y_1 + c^T(\boldsymbol{s_0}, \theta^g)((\sigma^2)^g C^*(\theta^g))^{-1}\boldsymbol{w}^{*g} + \boldsymbol{\eta}(\boldsymbol{s})$$

Fig. A2. Results for the δ-GLMM for darkblotched rockfish. Top row: Histogram of distribution of predicted probability of occurrence for 2 km × 2 km grid cells in 2008 for regions A (left; N = 3966 prediction locations), B (center; N = 1972), and C (right; N = 2477). Bottom row: Histogram of expected catch for the three strata. In all panels, the dashed line shows the mean for the habitat model and the solid line the mean from the stratified model. Note the very long tails in the abundance models.



where the first term on the right side is the predicted value from the fixed habitat covariates, the second is the effort offset, and the third term is the linear interpolation of the spatial effect at each predicted point from the sampled knot locations (i.e., the standard kriging projection). The fourth term, $\boldsymbol{\eta}(\boldsymbol{s})$, is a bias correction term for the spatial that arises as project from a small number of knots to the observed locations (see Finley et al. 2009). Specifically, $\boldsymbol{\eta}(\boldsymbol{s}) \sim N(\mathbf{0}, \Sigma)$ where $\Sigma = \mathrm{Diag}(\boldsymbol{C}(\boldsymbol{s}, \boldsymbol{s}) - \boldsymbol{c}^{\boldsymbol{T}}(s, \theta)\boldsymbol{C}^{*-1}(\theta)\boldsymbol{c}(s, \theta))$ and $\boldsymbol{C}^*$ is the correlation matrix for knot locations, $\boldsymbol{C}$ is the correlation matrix among observations, and $\boldsymbol{c}$ is a matrix describing the covariance between the prediction points and the knot locations (see Banerjee et al. 2008; Finley et al. 2009). Importantly, $\boldsymbol{\eta}(\boldsymbol{s})$ is a vector of independent normal random variables that accounts for the averaged bias underestimation over the observed locations (Finley et al. 2009). The addition of $\boldsymbol{\eta}(\boldsymbol{s})$ makes this a modified predictive process model (Finley et al. 2009). An analogous model was constructed for the positive component of the model. For all predictions we use an effort offset of 1 ha (0.01 km^2) swept for prediction (i.e., $\gamma_1 = 1$).

Each draw of the posterior distribution could thus provide predicted value of logit($\boldsymbol{\phi}$) (or for the positive portion of the model, log($\boldsymbol{\mu}$)) at each predicted location. Each component can then be back-transformed to generate a map of predicted probability of occurrence (bounded by 0 and 1) or the expected biomass caught.

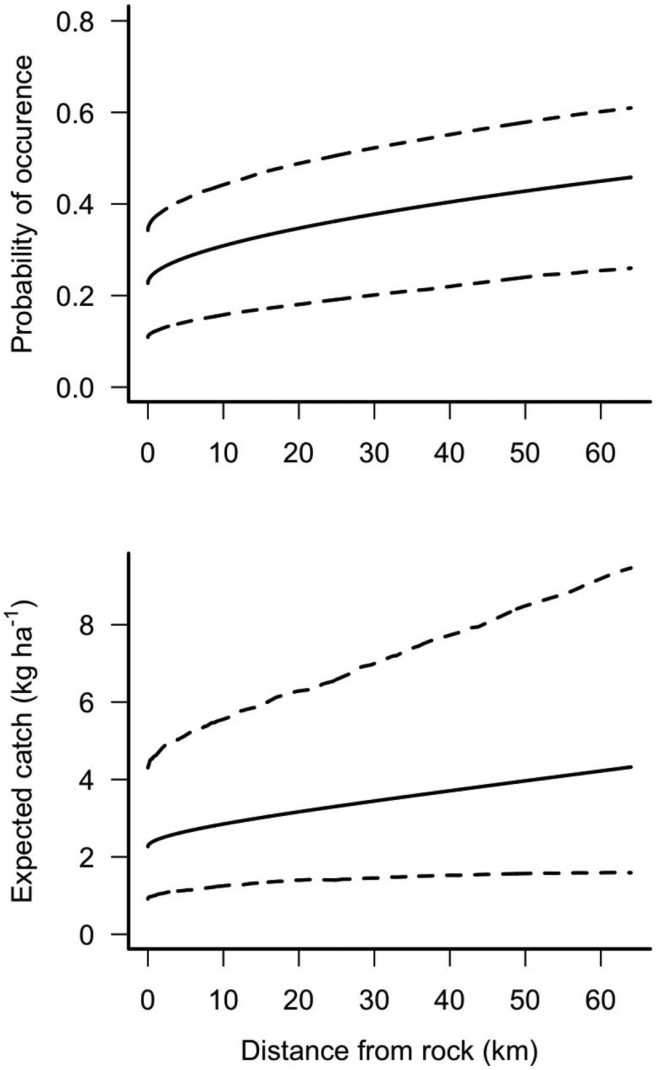
To save computing time, we selected 500 evenly spaced draws from the joint posterior distribution, produced a prediction from each of the 500 posterior draws. We then calculated the mean, median, and credible intervals for each prediction location. As an aside, because the fixed and spatial components of the above model are additive, it is also possible to produce a map derived exclusively using the habitat covariates. This can be thought of as a predictive map for species occurrence based exclusively on the measured habitat characteristics unmodified by unobserved habitat covariates.

Discussion and results

Both design and strata models have a very difficult time estimating parameters for strata that have few positive observations. In our example, region C has no positive observations of darkblotched for any year (Table 2). As a result, the design and strata models cannot estimate the abundance portion of the model. The design based model estimates 0 biomass in region C each year while the strata model estimates tiny biomass with unreasonable uncertainty bounds, which are determined by the upper and lower bounds on the prior for positive abundance (mean biomass 0.001 mt, SE = 10^{11}). Given that darkblotched rockfish have been observed to 900 m depth (Love et al. 2002), the actual number of darkblotched rockfish in this strata is low but nonzero as suggested by the habitat model.

The disparity between the strata and habitat model is surprising. In general, we expect the use of habitat covariates to provide nearly equivalent estimates to the strata model as long as sampling is random with respect to the habitat covariates (Cochran 1977). In our case, discrepancies between density estimates arise primarily because the habitat model includes information on the location of rocky substrate while the stratified model does not and

Fig. A3. Marginal effect of the distance to rock on the probability of occurrence (top) and expected catch (bottom). The solid line shows the mean effect and the dashed line the 90% credible interval. Curves are plotted for a depth of 180 m and bottom temperature of 7 °C.



because survey locations are not random with respect to rocky habitats (Fig. 1; Fig. A1). Rocky outcrops are undersampled by the survey because trawl gear is damaged by rocky substrate. Across the entire time series for regions A, B, and C, 6.1% of the area is considered rocky habitat, yet only 3.0% of trawls occur on rocky bottom. Furthermore, darkblotched rockfish occurrence and abundance both are estimated to decline in near rocky outcrops (Fig. A3). The net result is that the habitat model estimates smaller abundances on rocky substrate while the strata model assumes these rocky "untrawlable" habitats are identical to the sampled locations, resulting in higher probability of occurrence and abundance estimates in the strata model.

To date, generalized additive models (GAMs) are the most frequently used tool exploring species–habitat relationships (Wood 2006; Valavanis et al. 2008; Johnson et al. 2013). GAMs are often advocated based on their ability to estimate occupancy and abundance as a complex, nonlinear function of measured covariates. While GAMs allow for complex nonlinear structures, a major drawback to their use is that estimated parameters may lack biological interpretation. Furthermore, when covariates are allowed to interact with spatial locations in complex ways (e.g., when latitude and longitude are included as terms in the model), GAMs have the ability to match virtually any pattern present in the data, causing concerns about model overfitting and the true predictive power of such models (Telford and Birks 2005; NMFS 2013), though cross-validation methods can reduce this concern (Wood 2006; Banerjee et al. 2008). Thus, while the flexibility and complexity of GAMs may improve the match between observations and data, they do not necessarily translate into improved biological insight. However, in a Bayesian context, GAMs and GLMMs share many attributes and are closely connected (Wood 2006; Paciorek 2007; Banerjee et al. 2008; Cressie and Wikle 2011). Our use of Bayesian methodology and use of predictive process models provides direct descriptions of parameter and abundance uncertainty (Royle and Wikle 2005; Banerjee et al. 2008; Swanson et al. 2012).

References

Banerjee, S., Gelfand, A.E., Finley, A.O., and Sang, H. 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. B Stat. Method. **70**: 825–848. doi:10.1111/j.1467-9868.2008.00663.x.

Cochran, W.G. 1977. Sampling techniques. John Wiley & Sons Inc., New York.

Cressie, N., and Wikle, C.K. 2011. Statistics for spatio-temporal data. John Wiley & Sons Inc.

Draper, D. 2013. Bayesian model specification: heuristics and examples. *In* Bayesian theory and applications. *Edited by* P. Damien, P. Dallaportas, N.G. Polson, and D.A. Stephens. Clarendon Press, Oxford. pp. 409–431.

Draper, D., and Krnjajic, M. 2010. Calibration results for Bayesian model specification. Bayesian Analysis, **1**: 1–43.

Finley, A.O., Sang, H., Banerjee, S., and Gelfand, A.E. 2009. Improving the performance of predictive process modeling for large datasets. Comp. Stat. Data Analy. **53**: 2873–2884. doi:10.1016/j.csda.2008.09.008.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. 2013. Bayesian data analysis. 3rd ed. CRC press.

Gneiting, T., and Raftery, A.E. 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. **102**: 359–378. doi:10.1198/016214506000001437.

Johnson, A.F., Jenkins, S.R., and Hiddink, J.G. 2013. Linking temperate demersal fish species to habitat: scales, patterns and future directions. Fish Fish. **14**(3): 256–280. doi:10.1111/j.1467-2979.2012.00466.x.

Krnjajić, M., Kottas, A., and Draper, D. 2008. Parametric and nonparametric Bayesian model specification: a case study involving models for count data. Comp. Stat. Data Analy. **52**: 2110–2128. doi:10.1016/j.csda.2007.07.010.

Krumbein, W.C., and Sloss, L.L. 1963. Stratigraphy and sedimentation. W.H. Freeman, San Francisco.

Latimer, A.M., Banerjee, S., Sang, H., Jr., Mosher, E.S., and Silander, J.A., Jr. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecol. Lett. **12**: 144–154. doi:10.1111/j.1461-0248.2008.01270.x. PMID:19143826.

Love, M.S., Yoklavich, M., and Thorsteinson, L. 2002. The rockfishes of the northeast Pacific. University of California Press, Berkeley and Los Angeles, Calif.

National Marine Fisheries Service (NMFS). 2013. Groundfish essential fish habitat synthesis: a report to the pacific fishery management council. NOAA NMFS Northwest Fish. Sci. Center, Seattle, Wash.

National Oceanic and Atmospheric Administration (NOAA). 2001. U.S. vector shoreline derived from NOAA nautical charts. NOAA's Ocean Service, Office of Coast Survey (OCS), Silver Spring, MD.

National Oceanic and Atmospheric Administration (NOAA). 2003. U.S. coastal relief model — Northwest Pacific. National Geophysical Data Center, NESDIS, NOAA, U.S. Department of Commerce, Boulder, CO.

Paciorek, C.J. 2007. Computational techniques for spatial logistic regression with large data sets. Comp. Stat. Data Analy. **51**: 3631–3653. doi:10.1016/j.csda.2006.11.008.

Royle, J.A., and Wikle, C.K. 2005. Efficient statistical mapping of avian count data. Environ. Ecol. Stat. **12**: 225–243. doi:10.1007/s10651-005-1043-4.

Shelton, A.O., Dick, E.J., Pearson, D., Ralston, S., and Mangel, M. 2012. Single-species landings and uncertainty estimates from multi-species fisheries landings data: hierarchical Bayesian models for California groundfish fisheries. Can. J. Fish. Aquat. Sci. **69**(2): 231–246. doi:10.1139/f2011-152.

Swanson, A.K., Dobrowski, S.Z., Finley, A.O., Thorne, J.H., and Schwartz, M.K. 2012. Spatial regression methods capture prediction uncertainty in species distribution model projections through time. Glob. Ecol. Biog. **22**: 242–251. doi:10.1111/j.1466-8238.2012.00794.x.

Telford, R.J., and Birks, H.J.B. 2005. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. Quat. Sci. Rev. **24**: 2173–2179. doi:10.1016/j.quascirev.2005.05.001.

Valavanis, V.D., Pierce, G.J., Zuur, A.F., Palialexis, A., Saveliev, A., Katara, I., and Wang, J. 2008. Modelling of essential fish habitat based on remote sensing, spatial analysis and GIS. Hydrobiologia, **612**: 5–20. doi:10.1007/s10750-008-9493-y.

Wood, S.N. 2006. Generalized additive models: an introduction with R. Chapman & Hall.