Flexible modeling for stock-recruitment relationships using Bayesian nonparametric mixtures

Kassandra Fronczyk[a], Athanasios Kottas *[a], Stephan Munch[b]

[a] Department of Applied Mathematics and Statistics, Baskin School of Engineering, 1156 High Street, University of California, Santa Cruz, CA 95064, USA
[b] Marine Science Research Center, Stony Brook University, Stony Brook, NY 11794, USA

Abstract: The stock and recruitment relationship is fundamental to the management of fishery natural resources. However, inferring stock-recruitment relationships is a challenging problem because of the limited available data, the collection of plausible models, and the biological characteristics that should be reflected in the model. Motivated by limitations of traditional parametric stock-recruitment models, we propose a Bayesian nonparametric approach based on a mixture model for the joint distribution of log-reproductive success and stock biomass. Flexible mixture modeling for this bivariate distribution yields rich inference for the stock-recruitment relationship through the implied conditional distribution of log-reproductive success given stock biomass. The method is illustrated with cod data from six regions of the North Atlantic, including comparison with simpler Bayesian parametric and semiparametric models.

Keywords: Dirichlet process; Log-reproductive success; Markov chain Monte Carlo; Multivariate normal mixtures; North Atlantic cod; Stock biomass.

1 Introduction

The stock-recruitment (S-R) relationship, i.e., the relationship between the stock size and the level of recruitment to that stock is a key aspect of fishery research with direct implications to the management of natural resources. With recent technological advances and geographical expansion, many fishery resources have been driven to low

*Corresponding author. Tel.: 831-459-5536 ; fax: 831-459-4829; E-mail: thanos@ams.ucsc.edu

and unproductive levels. In fact, in recent years the number of overexploited stocks has dramatically increased in most regions of the world (Garcia and de Leiva Moreno 2003). Consequently, there has been a substantial rise in concern about the basic biological sustainability of fishing. The S-R relationship is used to make decisions on the limits of sustainable fishing and is therefore fundamental to the management of fishery resources.

However, modeling and inferring S-R relationships is a challenging problem. Data are limited and measured with noise in both stock biomass and estimated recruitment; the S-R relationship is likely to be nonlinear over some ranges of stock sizes; and there are various plausible biological mechanisms that are consistent with very different functional relationships between stock and recruitment. In particular, although many parametric S-R models can be biologically derived, the relationship is usually disguised by environmental variability and difficult to determine with accuracy.

The Bayesian paradigm offers a natural modeling framework for S-R relationships. In general, Bayesian approaches enable incorporation of prior knowledge, have the capacity to model different sources of data uncertainty, and build inference from probabilistic modeling, which thus yields appropriate uncertainty quantification of point estimates that does not rely on asymptotic considerations. Indeed, parametric Bayesian techniques are increasingly utilized in fisheries and, more generally, in ecology; see, e.g., Punt and Hilborn (1997), McAllister and Kirkwood (1998), Clark (2005; 2007).

However, parametric approaches to S-R modeling, Bayesian and classical alike, may be too restrictive, since they rely on specific parametric forms for the S-R function. Depending on their assumptions, parametric models can produce very different inference and can be highly influenced by extreme observations.

Semiparametric and nonparametric estimation methods for S-R relationships avoid strong assumptions implied by parametric approaches, and are thus becoming increasingly popular. Classical nonparametric methods include: construction of the distribution of recruitment given stock biomass through nonparametric density estimators (Evans and Rice 1988); using generalized additive models to estimate the relationship of recruitment with spawning biomass and an environmental variable, such as sea surface temperature (Jacobson and MacCall 1995); fitting a locally weighted smoothing function with nonparametric regression and spline methods (Cook 1998); and using neural networks to estimate the S-R function (Chen and Ware 1999). A practical drawback of these methods involves uncertainty quantification for the estimates of the S-R function and of management reference points resulting from the estimated S-R relationship. This is a direct consequence of the fact that classical nonparametric estimation techniques do not involve probabilistic modeling of the underlying (condi-

tional) distribution of recruitment given stock biomass or of the corresponding joint distribution. Hence, by avoiding potentially suspect parametric distributional forms, i.e., by avoiding likelihood specification, they are inevitably limited to point estimation. When developed, error bounds depend heavily on asymptotic results, which are unreliable because of the small sample sizes typically available for S-R inference.

The impetus for a Bayesian nonparametric approach is the same with classical nonparametric methods, that is, to provide inference for the S-R relationship that avoids restrictive parametric assumptions. However, Bayesian nonparametric methodology utilizes a drastically different approach to inference. Instead of avoiding modeling the stochastic mechanism that generates the data, it treats the corresponding distribution as the unknown (infinite-dimensional) parameter, which is assigned a *nonparametric* prior that can support the space of all plausible distributions. Hence, Bayesian nonparametric methods combine the advantages of Bayesian modeling with the appeal of nonparametric inference. In particular, they provide data-driven, albeit model-based, inference, and, importantly, more reliable predictions than parametric models.

To our knowledge, the only application of Bayesian nonparametrics to modeling S-R relationships appears in Munch, Kottas, and Mangel (2005) and Patil (2007). This work is based on semiparametric modeling, using the standard regression setting with a normal distribution for recruitment on the log scale, and with a Gaussian process (GP) prior for the S-R function. Although, as shown in Munch et al. (2005), the GP semiparametric model outperforms traditional parametric models, it still includes potentially restrictive modeling aspects as it builds inference from a parametric recruitment distribution, a stationary prior model for the S-R function, and an implicit assumption that stock biomass is observed with negligible measurement errors.

Here, we propose a more general fully nonparametric Bayesian modeling approach for S-R relationships, which, while retaining a computationally tractable framework for inference, it obviates the need for the above assumptions. The approach is built from a mixture model for the joint distribution of log-reproductive success, $\log(R/S)$, and stock biomass, where R and S denote recruitment and stock size, respectively. (Note that regression approaches to modeling the S-R function typically use $\log(R/S)$ and S as the response and covariate, respectively.) Flexible mixture modeling for this bivariate distribution yields rich inference for the S-R relationship through the implied conditional distribution of $\log(R/S)$ given S. Key features of the model include its capacity to uncover both non-linear S-R relationships as well as non-standard shapes for the conditional density of log-reproductive success. We illustrate the utility of the proposed nonparametric model by applying it to cod data from six regions of the North Atlantic, and comparing inference results with the semiparametric GP model discussed

above, and the Ricker model, a commonly utilized parametric model.

The outline of the paper is as follows. Section 2 provides background on the North Atlantic cod data. Section 3 introduces the Bayesian nonparametric mixture model. In Section 4 we present the results from the North Atlantic cod data, including the model comparison study, and Section 5 concludes with a discussion. The appendices include technical details on implementation of the models.

2 Data Description

Cod are demersal marine fish that, historically, matured around 7 years and live for several decades (see, e.g., Barot et al 2004). Large females produce millions of eggs and the total number of eggs produced is roughly indexed by the total biomass of spawning individuals. Recruitment refers to the number of individuals that result from the spawning biomass in a given year. Survival from the egg stage to the age at recruitment is density dependent leading to a non-linear relationship between stock biomass and recruitment. We consider data on recruitment and stock biomass for cod from six North Atlantic regions: NE Arctic, Icelandic, Irish Sea, Faroe, Skagerrak, and West of Scotland. Each region has data dating from as far back as 1946 through 2004. Recruitment and biomass estimates were determined by virtual population analysis (Hilborn and Walters 1992) of commercial landings data and fishery independent research surveys (ICES 2005). Virtual population analysis is essentially an accounting procedure used to determine the number of fish that must have been in the population given the numbers that have been removed over time and an independent estimate of natural mortality. These data have been analyzed previously by many authors (see, e.g., Brander and Mohn 2004; Stige et al 2006; and further references therein).

As discussed in the Introduction, in the S-R regression setting, the response of interest is typically log-reproductive success, $y = \log(R/S)$, and the covariate $x = S$. Hence, the data comprises $\{(y_i, x_i) : i = 1, \ldots, n\}$, where y_i denotes the subsequent log-reproductive success from the spawning biomass x_i in year i. The data from the six regions (plotted in Figure 1) range in length from 27 to 59 years.

3 Bayesian Nonparametric Modeling for Stock-Recruitment Relationships

Section 3.1 presents the nonparametric mixture modeling approach for the S-R relationship. The methods for posterior inference are discussed in Section 3.2.

3.1 The mixture modeling approach

To develop a flexible inferential framework for S-R relationships, we propose a nonparametric mixture model for the joint density, $f(y, x)$, of log-reproductive success, $y = \log(R/S)$, and stock biomass, $x = S$. A flexible model for $f(y, x)$, which can, for instance, accommodate skewness, excess variability, and possible multimodalities, is key to our inferential objectives, since estimation and uncertainty quantification for the S-R relationship is based on the conditional density $f(y \mid x)$ corresponding to $f(y, x)$. In seeking flexible modeling and inference for densities, one is naturally led to mixture distributions. In our context, mixtures of (bivariate) normal densities provide a natural choice for modeling $f(y, x)$. General mixtures of normal densities can approximate any continuous density (e.g., Lo 1984), while at the same time, the structure of the normal distribution enables ready interpretation for key functionals of the mixture model (such as the conditional expectation $\mathrm{E}(y \mid x)$) as well as relatively easy implementation of inference techniques for the mixture model.

To motivate the proposed Bayesian nonparametric mixture model for $f(y, x)$, consider a parametric Bayesian model formulation for a finite mixture of M normal densities $\sum_{j=1}^{M} \pi_j \mathrm{N}_2(y, x; \mu_j, \Sigma_j)$, where $\mathrm{N}_2(\mu, \Sigma)$ will denote the bivariate normal density or distribution (depending on the context) with mean vector μ and covariance matrix Σ. Hence, each data point (y_i, x_i) arises from one of the M mixture components, which has a distinct mean vector and covariance matrix, that is, we consider the general version of location-scale normal mixtures. The component specific mixing parameters (μ_j, Σ_j) arise from a prior distribution, say, $G_0(\mu, \Sigma)$, possibly conditionally on hyperparameters; moreover, the mixture weights π_j are typically assigned a conjugate Dirichlet prior distribution. Note that the finite mixture model can be equivalently written as $\sum_{j=1}^{M} \pi_j \mathrm{N}_2(y, x; \mu_j, \Sigma_j) \equiv \int \mathrm{N}_2(y, x; \mu, \Sigma) \mathrm{d}G(\mu, \Sigma)$, where G is a discrete distribution with possible values (μ_j, Σ_j) and corresponding probabilities π_j, for $j = 1, ..., M$. Discreteness of the mixing distribution G is key as it enables clustering of the mixing parameters (μ_j, Σ_j) into a number of unique components ($< M$) associated with a corresponding grouping of the data observations. However, choosing the number of mixture components M is difficult, and inference methods for mixtures with random number of components are rather complex.

Bayesian nonparametric mixture modeling offers a powerful alternative where the parametric discrete distribution for G is replaced with a nonparametric prior that supports all mixing distributions. In this context, the Dirichlet process (DP) (Ferguson 1973) is the most commonly used nonparametric prior model for the random mixing

distribution G. We will write $\mathrm{DP}(\alpha, G_0)$ to denote the DP prior for G defined in terms of two parameters: a parametric centering (base) distribution G_0 (formally, the prior expectation of the DP, $\mathrm{E}(G) = G_0$), and a positive scalar parameter α, which controls the variability of G about G_0 with larger values of α resulting in realizations G that are *closer* to G_0. Arguably, the most useful definition of the DP is its constructive definition (Sethuraman 1994), according to which a random distribution G drawn from a $\mathrm{DP}(\alpha, G_0)$ prior has an almost sure representation as

$$G(\cdot) = \sum_{l=1}^{\infty} \omega_l \delta_{\boldsymbol{\theta}_l}(\cdot), \tag{1}$$

where δ_a denotes a point mass at a, the $\boldsymbol{\theta}_l = (\mu_l, \Sigma_l)$ are i.i.d. from G_0, and $\omega_1 = \zeta_1$, $\omega_l = \zeta_l \prod_{m=1}^{l-1}(1-\zeta_m)$ for $l \geq 2$, with ζ_l i.i.d. from Beta(1,α) (independently of the $\boldsymbol{\theta}_l$). Therefore, the DP generates (almost surely) discrete distributions with a countable number of possible values drawn from the base distribution G_0, and corresponding weights defined through latent Beta(1,α) variables based on the mechanism described above, which is referred to as *stick-breaking*. The stick-breaking process builds the weights by iteratively breaking off a portion of a stick of unit length. The first Beta draw ζ_1 defines the first weight, the second weight is defined by breaking a portion $\zeta_2(1-\zeta_1)$ from the remaining part of the stick, $1-\zeta_1$, and the process continues ad infinitum (note that $\sum_{l=1}^{\infty} \omega_l = 1$, almost surely). In practice, the number of effective weights is controlled by the value of α (or the hyperprior placed on α), in particular, small α values favor DP prior realizations for G that are effectively supported by a small number of point masses.

Hence, our proposed model for the joint density, $f(y, x)$, of log-reproductive success and stock biomass is given by a DP mixture of bivariate normals,

$$f(y, x; G) = \int \mathrm{N}_2(y, x; \mu, \Sigma)\mathrm{d}G(\mu, \Sigma), \quad G \sim \mathrm{DP}(\alpha, G_0). \tag{2}$$

We take $G_0(\mu, \Sigma; m, V, Q) = \mathrm{N}_2(\mu; m, V)\mathrm{IW}(\Sigma; v, Q)$. Here, $\mathrm{IW}(v, Q)$ denotes the inverse Wishart distribution for the 2×2 (positive definite) matrix Σ with density proportional to $|\Sigma|^{-(v+3)/2} \exp\{-0.5\mathrm{tr}(Q\Sigma^{-1})\}$. This choice for G_0 is convenient for implementation of posterior inference as it corresponds to the standard conditionally conjugate specification for the normal kernel mean and covariance parameters. As detailed in Section 3.2 and Appendix A, the model is completed with priors for α and for the hyperparameters (m, V, Q) of G_0.

We next discuss a truncated version of the DP mixture model which facilitates Markov chain Monte Carlo (MCMC) posterior simulation (e.g., Ishwaran and James

2001), as well as interpretation of (and inference for) the mean regression function arising from the DP mixture, a key functional with respect to inference for the S-R relationship. Specifically, using the DP stick-breaking definition in (1), we approximate the countable discrete mixing distribution G in (2) with a finite dimensional version defined as $G_N(\cdot) = \sum_{l=1}^{N} p_l \delta_{\boldsymbol{\theta}_l}(\cdot)$. As before, the $\boldsymbol{\theta}_l = (\mu_l, \Sigma_l)$, $l = 1, ..., N$, are i.i.d. from G_0. Here, the random weights, $\mathbf{p} = (p_1, ..., p_N)$, arise from a truncation of the stick-breaking process: with $V_1, \ldots, V_{N-1}$ i.i.d. from Beta$(1, \alpha)$, we define $p_1 = V_1$; $p_l = V_l \prod_{m=1}^{l-1}(1 - V_m)$, for $l = 2, ..., N-1$; and set $p_N = 1 - \sum_{l=1}^{N-1} p_l = \prod_{m=1}^{N-1}(1 - V_m)$. The induced joint density, $f(\mathbf{p}; \alpha)$, for the random weights corresponds to a generalized Dirichlet distribution given in Appendix A. Regarding the choice of N, an appropriate truncation value can be found by considering the behavior of the higher order weights ω_l in (1). For instance, $\mathrm{E}(\sum_{l=N}^{\infty} \omega_l \mid \alpha) = \{\alpha/(\alpha+1)\}^{N-1}$. Given a suitable tolerance level and a prior estimate for α, this expression provides the corresponding truncation value N. For the analysis of the North Atlantic cod data in Section 4, we used $N = 25$ (larger values of N did not change posterior inference results).

Using the truncated version G_N of G, the normal mixture model for the bivariate density of log-reproductive success and stock biomass can be expressed as $f(y, x; G_N) = \sum_{l=1}^{N} p_l \mathrm{N}_2(y, x; \mu_l, \Sigma_l)$. Moreover, $f(x; G_N) = \sum_{l=1}^{N} p_l \mathrm{N}(x; \mu_l^x, \Sigma_l^{xx})$ provides the marginal stock biomass density. Here, μ_l^x and Σ_l^{xx} are the mean and variance of the marginal normal distribution for x induced by the joint $\mathrm{N}_2(y, x; \mu_l, \Sigma_l)$ distribution. Hence, inference for the conditional density of log-reproductive success at any fixed value, x_0, of stock biomass is available through $f(y \mid x_0; G_N) = f(y, x_0; G_N)/f(x_0; G_N)$. Whereas traditional S-R models only allow for unimodal log-reproductive success densities, the proposed DP mixture model can capture general unimodal and multimodal shapes. As illustrated with the North Atlantic cod data in Section 4, the model yields full inference for conditional response densities with shapes that can change with different stock biomass values.

Inference for the S-R relationship can be obtained through any central feature of the conditional log-reproductive success distribution. We work with the mean regression function, $\mathrm{E}(y \mid x; G_N) = \{f(x; G_N)\}^{-1} \int y f(y, x; G_N) \mathrm{d}y$, which can be expressed as

$$\mathrm{E}(y \mid x; G_N) = \frac{1}{f(x; G_N)} \sum_{l=1}^{N} p_l \mathrm{N}(x; \mu_l^x, \Sigma_l^{xx}) \{\mu_l^y + \Sigma_l^{yx} \Sigma_l^{xx-1} (x - \mu_l^x)\}, \qquad (3)$$

where μ_l^y is the marginal mean for y, and Σ_l^{yx} is the covariance between y and x arising from the joint $\mathrm{N}_2(y, x; \mu_l, \Sigma_l)$ distribution. The conditional mean log-reproductive success can also be written as $\mathrm{E}(y \mid x; G_N) = \sum_{l=1}^{N} q_l(x) \{\mu_l^y + \Sigma_l^{yx} \Sigma_l^{xx-1} (x - \mu_l^x)\}$, where $q_l(x) = p_l \mathrm{N}(x; \mu_l^x, \Sigma_l^{xx})/f(x; G_N)$, for $l = 1, ..., N$, that is, a locally weighted mixture

of linear regressions. Because the weights $q_l(x)$ depend on stock biomass values, they provide local structure to $\mathrm{E}(y \mid x; G_N)$, and thus, when suggested by the data, yield a non-linear shape for the conditional mean log-reproductive success. The coefficients of the component specific linear regressions are given by the parameters of the corresponding $\mathrm{N}_2(\mu_l, \Sigma_l)$ distribution, in particular, the sign of the slope, $\Sigma_l^{yx}/\Sigma_l^{xx}$, depends on the correlation of the l-th normal mixture component.

For the illustration with the North Atlantic cod data (Section 4.1), we used a fairly noninformative prior specification resulting in roughly constant (in x) prior means for $\mathrm{E}(y \mid x; G_N)$ with wide prior uncertainty bands. Our motivation was to allow the data to drive the shape of the estimated S-R relationship, and also to demonstrate that there is significant prior to posterior learning under the proposed nonparametric mixture model even with the small to moderate available sample sizes for the cod data. However, if one wishes to favor in the prior specific shapes for the conditional mean log-reproductive success, this is possible given the structure of $\mathrm{E}(y \mid x; G_N)$. For instance, the Ricker model (Ricker 1954) is one of the most commonly used among the several biologically derived parametric S-R models. The basic Ricker model is built from the assumption that early life mortality is a linearly decreasing function of spawners (e.g., Quinn and Deriso 1999). Under the $(x = S, y = \log(R/S))$ scale, the Ricker model postulates a decreasing linear S-R relationship, $\mathrm{E}(y \mid x) = \beta_0 - \beta_1 x$ with $\beta_1 > 0$ (more details are given in Section 4.2 where the DP mixture model is shown to outperform the Ricker model for the cod data). Hence, if for a particular application, it is of interest to incorporate to the DP mixture model prior beliefs about Ricker-type density dependence, this can be accomplished by, for instance, favoring small α values in the prior along with negative correlations for the Σ_l matrices. Note that the implications of any particular prior choice can be readily studied by computing prior realizations for the conditional mean log-reproductive success using (3). The key feature of the nonparametric mixture model is that the S-R function is not strictly restricted to the linear shape (even when such a shape is favored in the prior), and thus non-standard S-R relationships can be uncovered in the posterior inference results when supported by the data.

3.2 Posterior inference

The standard hierarchical model formulation for the data $= \{(y_i, x_i) : i = 1, \ldots, n\}$ involves mixing parameters $(\tilde{\mu}_i, \tilde{\Sigma}_i)$, $i = 1, ..., n$, such that the (y_i, x_i), given $(\tilde{\mu}_i, \tilde{\Sigma}_i)$, are independent $\mathrm{N}_2(y_i, x_i; \tilde{\mu}_i, \tilde{\Sigma}_i)$, with the $(\tilde{\mu}_i, \tilde{\Sigma}_i)$, given G, arising i.i.d. from G. Under the truncation approximation, G_N, to G, latent configuration variables $\mathbf{L} =$

$(L_1, \ldots, L_n)$ are introduced to break the finite mixture approximation to the countable DP mixture model. In particular, each L_i takes a value in $\{1, \ldots, N\}$ with $L_i = l$ signifying that $(\tilde{\mu}_i, \tilde{\Sigma}_i) = (\mu_l, \Sigma_l)$, for $i = 1, \ldots, n$ and $l = 1, \ldots, N$. Hence, including the configuration variables and replacing G with $G_N \equiv (\mathbf{p}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$, the model becomes

$$\begin{aligned}
(y_i, x_i) \mid \boldsymbol{\theta}, L_i &\overset{ind}{\sim} \mathrm{N}_2(y_i, x_i; \mu_{L_i}, \Sigma_{L_i}),\ i = 1, \ldots, n \\
L_i \mid \mathbf{p} &\overset{i.i.d.}{\sim} \sum_{l=1}^{N} p_l \delta_l(\cdot),\ i = 1, \ldots, n \\
\mathbf{p} \mid \alpha &\sim f(\mathbf{p}; \alpha) \\
\boldsymbol{\theta}_l \mid m, V, Q &\overset{i.i.d.}{\sim} G_0(\boldsymbol{\theta}_l; m, V, Q),\ l = 1, \ldots, N. \qquad (4)
\end{aligned}$$

Regarding the model hyperparameters, we place a gamma(a_α, b_α) prior on α, and take $\mathrm{N}_2(a_m, B_m)$, $\mathrm{IW}(a_V, B_V)$, and $\mathrm{W}(a_Q, B_Q)$ priors for m, V, and Q, respectively, where $\mathrm{W}(a_Q, B_Q)$ is a Wishart distribution for the 2×2 (positive definite) matrix Q with density proportional to $|Q|^{(a_Q - 3)/2} \exp\{-0.5\mathrm{tr}(QB_Q^{-1})\}$ (with $a_Q \geq 2$). Specification of the hyperprior parameters is discussed in Appendix A.

We use blocked Gibbs sampling (e.g., Ishwaran and James 2001) to obtain the posterior distribution $p(\mathbf{p}, \boldsymbol{\theta}, \mathbf{L}, \alpha, m, V, Q \mid \text{data})$ corresponding to model (4). The blocked Gibbs sampler updates parameters using draws from standard distributions, the details of which are given also in Appendix A.

Each posterior sample for $(\mathbf{p}, \boldsymbol{\theta})$ provides a posterior realization for G_N directly through its definition, $\sum_{l=1}^{N} p_l \delta_{(\mu_l, \Sigma_l)}$. Then, for any specified point (y_0, x_0) in the (log-reproductive success, stock biomass) space, we can obtain posterior realizations of the joint mixture density for log-reproductive success and stock biomass, $f(x_0, y_0; G_N)$, the marginal stock biomass density, $f(x_0; G_N)$, and the conditional log-reproductive success density $f(y_0 \mid x_0; G_N)$, using again the definition of these densities as discussed in Section 3.1. Moreover, by evaluating expression (3), we obtain posterior samples for the conditional mean log-reproductive success at any desired set of stock biomass values, from which point and interval estimates for the S-R function can be produced.

Finally, a key inferential objective when modeling S-R relationships is estimation of reference points that are used in fisheries management decisions. To illustrate such inference under the DP mixture modeling approach, we consider an important reference point, the unfished biomass, B_0, as can be derived under a simple model for the stock dynamics. In particular, following the approach in Munch et al. (2005), we assume that stock biomass on an annual time step arises through $S_{t+1} = S_t - (\nu + \psi)S_t + R(S_t)$, where ν and ψ are the annual fractions of the population removed by natural and

fishing mortality, respectively, and $R(S_t)$ denotes recruitment as a function of biomass at year t. Therefore, at equilibrium and in the absence of fishing, B_0 is the solution to $R(S) = \nu S$. Hence, B_0 corresponds to the stock biomass value that satisfies restriction $\log(R(S)/S) = \log(\nu)$, and thus under the joint DP mixture model for $(\log(R/S), S)$, the density for B_0 is given by $f(x \mid y = \log(\nu); G_N)$. The posterior distribution for this conditional density can be obtained as above, working in this case with the joint mixture density and the marginal log-reproductive success density. As with the conditional log-reproductive success density, if the situation warrants, the model can uncover non-standard shapes in the B_0 density.

4 Data Illustrations

Section 4.1 presents results from the analysis of the North Atlantic cod data under the DP mixture modeling approach of Section 3. In Section 4.2, we consider comparison with simpler parametric and semiparametric S-R models.

4.1 Results for the North Atlantic cod data

For each of the North Atlantic regions, we fit the DP mixture model in (4) to the corresponding cod data. Figure 1 plots point estimates (posterior means) and 95% interval estimates for the mean S-R regression function in (3). The model uncovers the diverse conditional mean log-reproductive success relationships for each of the six regions. The NE Arctic, Icelandic, and Faroe regions have roughly negative linear relationships. Each of them shows evidence of a slight curvilinear curve within a small portion of the range of stock biomass values, in particular, for smaller biomass values for the NE Arctic and Icelandic regions and for higher values of the Faroe region. The posterior uncertainty bands for the NE Arctic and Icelandic regions are tighter than those of the Faroe region. The Irish Sea region has an unusual non-linear relationship with interval estimates that widen near the extremes of the data, while the Skagerrak region has an approximately quadratic shape, with the exception of the curve near the maximum stock biomass level, and uncertainty bands that are evenly spread throughout the stock biomass range. The last region, West of Scotland, has a relatively weak signal; in fact, it corresponds to the smallest sample size with only 27 data points.

To illustrate inference for the conditional density of log-reproductive success, we consider the NE Arctic and West of Scotland regions, and in Figure 2, show posterior

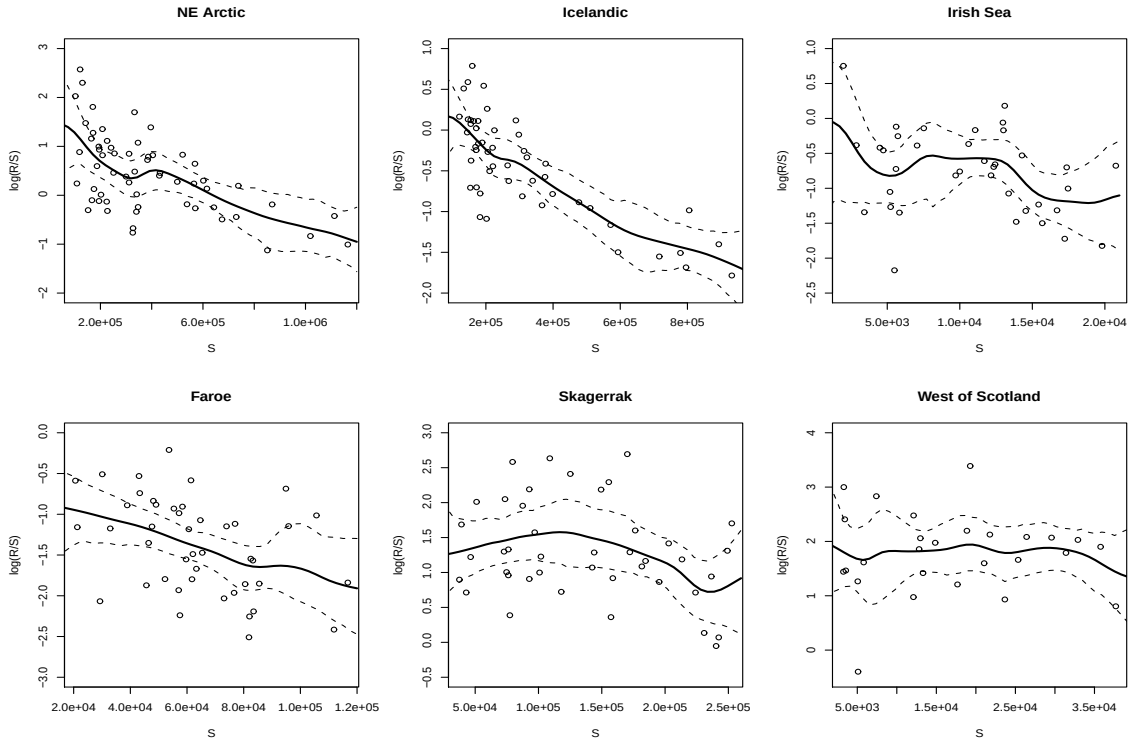

Figure 1: For each of the North Atlantic regions, posterior mean (solid lines) and 95% interval estimates (dashed lines) for the conditional mean log-reproductive success, overlaid on plot of the data for stock biomass (S) and log-reproductive success ($\log(R/S)$). The label in each panel indicates for each region the time interval in years with available data.

point and 95% interval estimates for the conditional response density at four fixed stock biomass values. At lower biomass levels, the log-reproductive success densities for the NE Arctic region depict obvious departures from normality; at $S = 200,000$, the density is bimodal, whereas at $S = 350,000$, it has a heavy left tail. Inspection of the data from this region suggests that this is a plausible feature for the log-reproductive success distribution rather than an artifact of the flexible prior model. The higher values of stock biomass result in more standard unimodal response densities. The log-reproductive success densities from the West of Scotland region are unimodal across the range of stock biomass values, with roughly the same amount of dispersion. Evidently, such density shapes would be successfully uncovered by traditional parametric models. Hence, a key feature of the nonparametric DP mixture model is that it has the capacity to capture both non-standard log-reproductive success density shapes as well as non-linear S-R relationships. And, as importantly, when the data do not support such

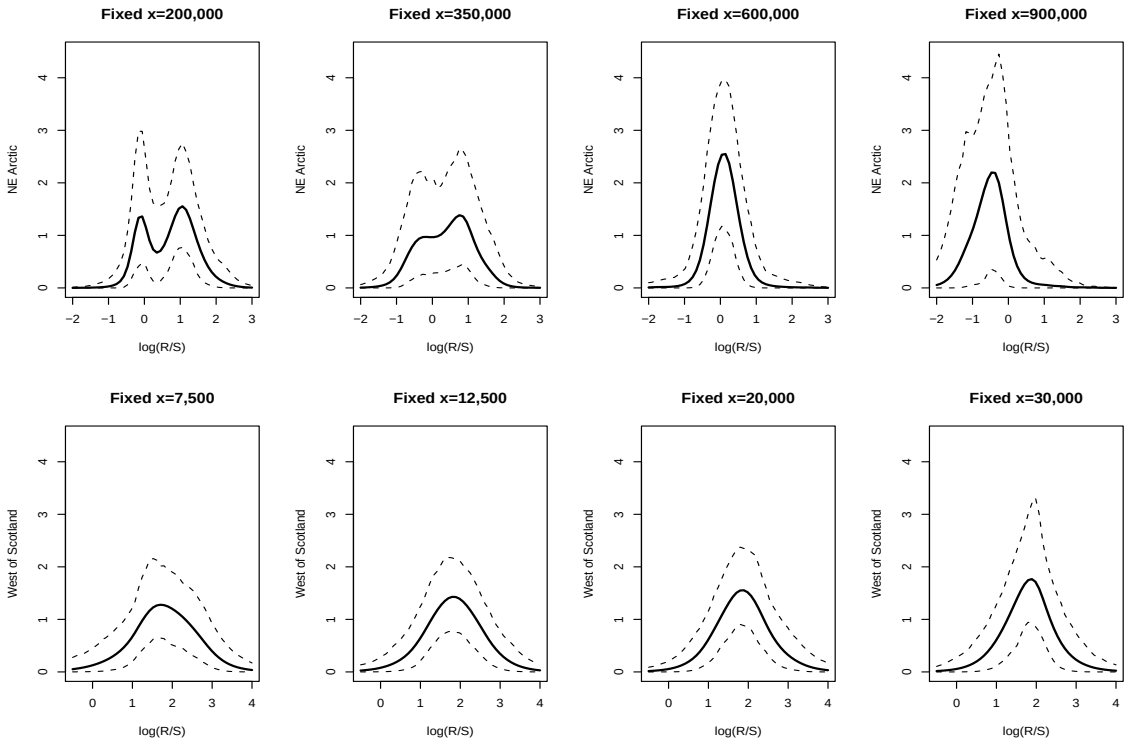

Figure 2: For the NE Arctic region (top row) and the West of Scotland region (bottom row), posterior mean (solid lines) and 95% interval estimates (dashed lines) of the conditional density of log-reproductive success ($\log(R/S)$) given four fixed stock biomass values (indicated in the label of each panel).

non-standard features, it yields inferences that are comparable with the ones that would result from commonly utilized parametric S-R models.

For each of the six regions, Figure 3 provides posterior point and 95% interval estimates of the density for unfished biomass, B_0, obtained as discussed in Section 3.2. In keeping with prior research on cod (e.g., Eero et al. 2007), we set the natural mortality rate $\nu = 0.2$. All the results are consistent with the data as becomes clear by comparing Figures 3 and 1 (noting where value $\log(0.2)$ lies in the panels of the latter figure). The DP mixture model yields multimodal B_0 posterior density estimates for the NE Arctic, Icelandic, and Irish Sea regions. On the other hand, it results in more conventional estimates in the Faroe region. For both the Skagerrak region and the West of Scotland region, $\log(0.2)$ is outside the range of observed log-reproductive success values. This is reflected in the point estimates for the B_0 density, which are very dispersed with wide associated uncertainty bands.

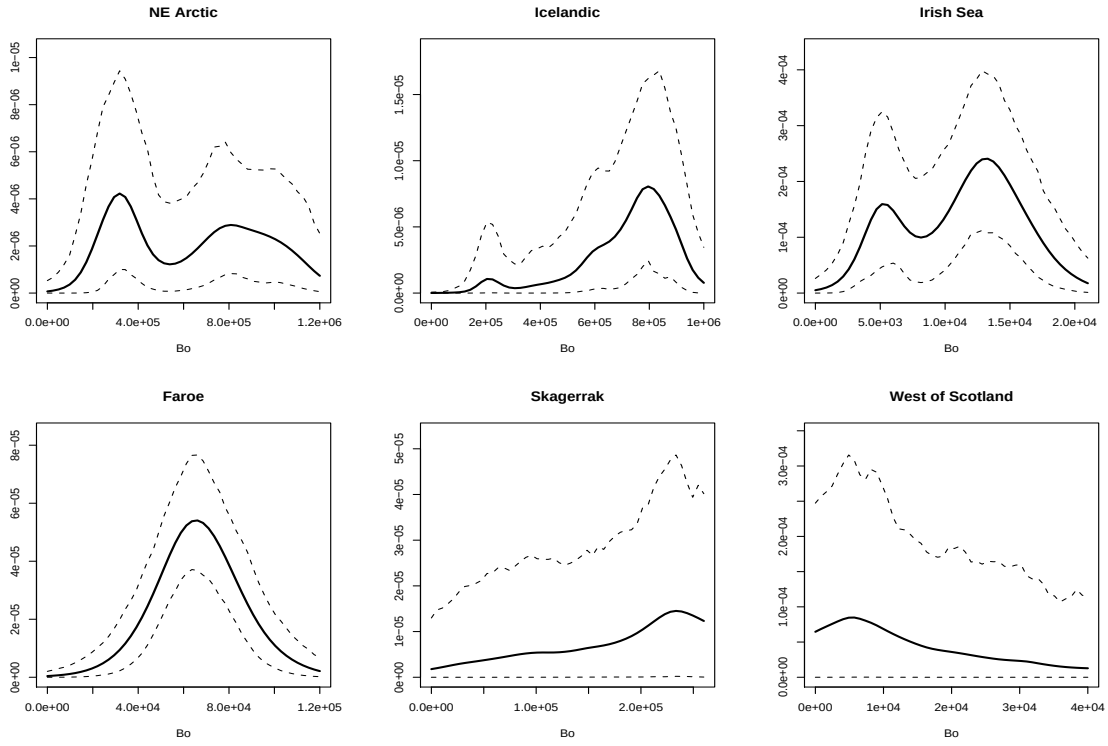

Figure 3: For each of the North Atlantic regions, posterior mean (solid lines) and 95% interval estimates (dashed lines) for the density of unfished biomass (B_0) corresponding to natural mortality rate $\nu = 0.2$.

4.2 Comparison study

Here, we consider model comparison working with the standard parametric Ricker model (discussed in Section 3.1), and with a semiparametric model, which is simpler than the fully nonparametric DP mixture model. We report results based on the data from the NE Arctic and Irish Sea regions.

We perform a Bayesian analysis of the Ricker model, $R_i = aS_i \exp(-bS_i)v_i$, where $a > 0$ and $b > 0$, and the v_i are independent multiplicative errors, typically, assumed to arise from a lognormal distribution. Hence, the Ricker model can be transformed to a linear regression model of $\log(R/S)$ on S, i.e.,

$$y_i = \beta_0 - \beta_1 S_i + \varepsilon_i, \;\; \varepsilon_i \mid \sigma^2 \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2), \quad i = 1, ..., n$$

where $y_i = \log(R_i/S_i)$, $\beta_0 = \log(a)$, and $\beta_1 = b > 0$. We assign a normal prior to β_0, a gamma prior to β_1, and an inverse gamma prior to σ^2. MCMC sampling for $(\beta_0, \beta_1, \sigma^2)$ is straightforward as discussed in Appendix B. Inference for the S-R relationship follows

directly from the posterior draws for parameters β_0 and β_1.

We also fit a Bayesian semiparametric model using a Gaussian process (GP) prior for the S-R function. Specifically,

$$y_i = h(S_i) + \varepsilon_i, \;\; \varepsilon_i \mid \sigma^2 \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2), \quad i = 1, ..., n$$

where the S-R function, $h(\cdot)$, is assigned a $\mathrm{GP}(\mu(S), C(S, S'))$ prior (see, e.g., Neal 1998, on Bayesian GP regression). We take a constant mean function, $\mu(S) = \mu$, and an isotropic exponential covariance function given by $C(S, S') = \tau^2 \exp\{-\phi|S - S'|\}$, where τ^2 is the GP variance, and $\phi > 0$ controls how rapidly the correlation decreases with distance between biomass values. We place a normal prior on μ, inverse gamma priors on σ^2 and τ^2, and a uniform prior on ϕ.

The GP regression model adds flexibility relative to traditional S-R models by placing a nonparametric prior on the S-R function rather than assuming a specific parametric form. It was studied by Munch et al (2005) in the context of modeling S-R relationships, though the scale used there involved $y = \log R$ and $x = \log S$. Appendix B provides details on prior specification and MCMC posterior simulation for the GP model parameters as well as on posterior predictive inference for function $h(\cdot)$.

Figure 4 shows posterior mean and 95% interval estimates for the S-R function: $\beta_0 - \beta_1 S$ under the parametric Ricker model (left column); $h(S)$ under the semiparametric GP model (middle column); and $\mathrm{E}(\log(R/S) \mid S; G_N)$ under the nonparametric DP mixture model (right column). To ensure a fair comparison, the priors were chosen such that *a priori* the conditional densities of log-reproductive success given stock biomass are comparable across the three models, and, in fact, corresponding to a fairly non-informative prior specification. For instance, for all three models and both regions, the prior estimate (prior mean) for conditional mean log-reproductive success was roughly constant (in S) around 0, and the associated prior interval estimates were including values for $\log(R/S)$ in the range from -4 to 4.

In the NE Arctic region, the Ricker model is roughly equivalent to both the GP and the DP mixture model. The GP regression curve generates slightly more variability than the DP model, but both include a small cubic trend at the smaller stock biomass values. However, the Irish Sea region gives noticeably different regression curves using the Ricker model and either of the GP or the DP mixture model. The Ricker model is unable to depart from linearity and is sensitive to the outlying observations. The GP model produces a more accurate regression curve than the Ricker model, with an approximately negative slope and modest curvature. The associated uncertainty bands have roughly the same width across the range of biomass values, regardless of the amount of data in the surrounding area. This is partly due to the fact that the

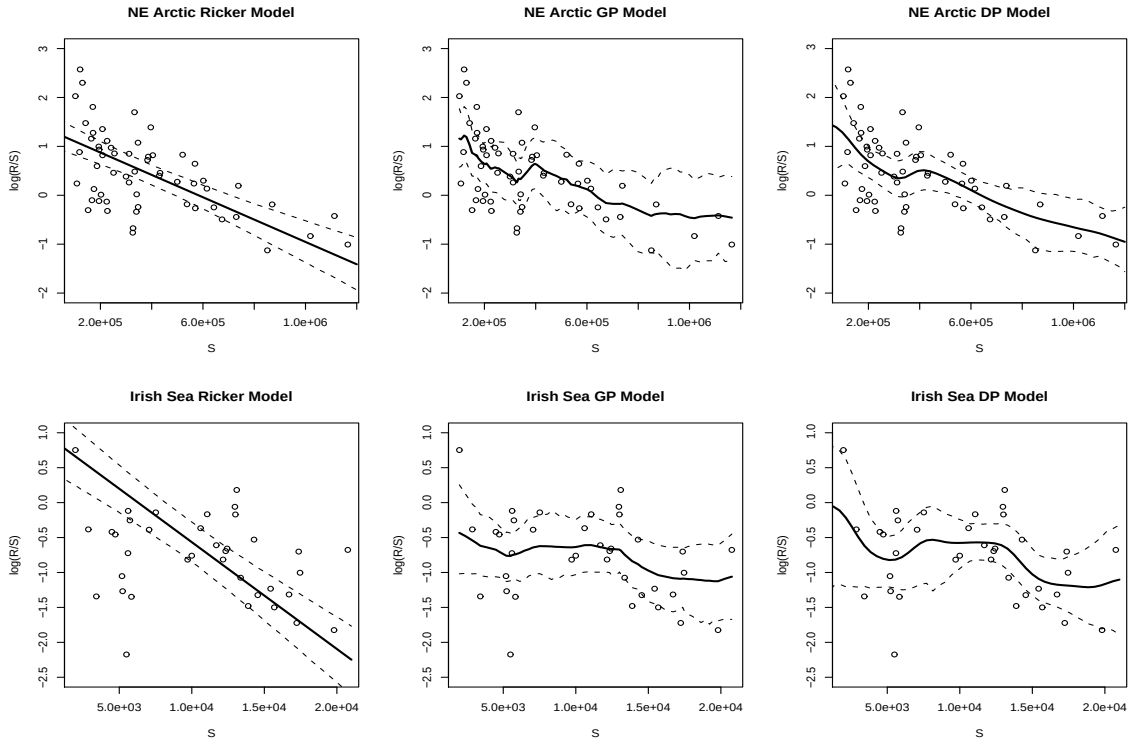

Figure 4: For the NE Arctic region (top row) and the Irish Sea region (bottom row), posterior mean (solid lines) and 95% interval estimates (dashed lines) for the S-R regression curve based on the parametric Ricker model (left column), the semiparametric Gaussian process model (middle column), and the DP mixture model (right column). Each panel includes a plot of the corresponding data for stock biomass (S) and log-reproductive success ($\log(R/S)$).

Irish Sea region data are spread relatively evenly through the range of stock sizes, but it can also be attributed to the GP prior stationary covariance function. The DP mixture model does not rely on stationarity assumptions and therefore uncertainty bands become wide or narrow based on how much data information is available about the particular biomass area.

Regarding inference for the log-reproductive success response distribution, the nonparametric model clearly outperforms the other two models. Both of the parametric and semiparametric models are built from the standard additive regression setting treating the stock biomass covariate as fixed, and assuming a normal response density with constant variance. Even if the error distribution is replaced with more flexible parametric families, neither of the two models would be able to recover the density shapes for the NE Arctic region (Figure 2) revealed by the DP mixture model.

A further key advantage of the DP mixture model is with regard to inference for the unfished biomass management reference point. Although the Ricker and GP models produce, in principle, a posterior distribution for B_0, the resulting inference is not as flexible as under the DP mixture model and, in fact, may not be attainable in practice. The Ricker model can be solved analytically for B_0, in particular, $B_0 = (\beta_0 - \log(\nu))/\beta_1$. However, this expression is not guaranteed to be positive, may result in values that are far outside the range of the data, and yields typically a unimodal distribution. For the GP model, B_0 corresponds to the value of stock biomass at which $h(S) = \log(\nu)$. Therefore, for each posterior realization of the $h(S)$ curve, B_0 can be obtained by finding values S_0 and S_1 such that $h(S_0) < \log(\nu) < h(S_1)$ and interpolating between these grid points. The GP model is unable to provide results when the curve does not attain the value of $\log(\nu)$, as in the case of the Skagerrak and West of Scotland regions.

To formally compare the three models, we use a version of the minimum posterior predictive loss criterion from Gelfand and Ghosh (1998). The criterion favors the model, m, that minimizes the predictive loss measure,

$$D(m) = P(m) + G(m) = \sum_{i=1}^{n} \mathrm{Var}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data}) + \sum_{i=1}^{n} \{y_i - \mathrm{E}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})\}^2.$$

Here, $\mathrm{E}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$ and $\mathrm{Var}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$ are the mean and variance, respectively, under model m, of the posterior predictive distribution for *replicated* response $y_{\mathrm{new},i}$ with associated stock biomass x_i. Hence, $G(m)$ is a goodness-of-fit term, whereas $P(m)$ acts as a penalty term for model complexity measured through posterior predictive variability; note that models that are either too complex or too simple will yield relatively large posterior predictive variances. Therefore, $D(m)$ achieves a balance between predictive uncertainty (term $P(m)$) and fidelity to observations (term $G(m)$), without the need to explicitly specify the number of model parameters, which is not clear-cut for Bayesian semiparametric or nonparametric models. Computing $D(m)$ for the Ricker and GP models is straightforward as discussed in Appendix B. Under the DP mixture model, estimation of $D(m)$ is based on expectations with respect to the conditional posterior predictive distribution of log-reproductive success given stock biomass; details are given in Appendix A.

Table 1 reports results for $D(m)$ under the three models. The goodness-of-fit term $G(m)$ is comparable across the three models for the NE Arctic region data, but the penalty terms in the Ricker and GP models are drastically larger than the DP mixture model. For the Irish Sea region data, the Ricker model performs poorly in both the goodness-of-fit and the penalty term. The GP model produces a smaller goodness-of-fit term, but a substantially larger penalty term than the DP mixture model. Under

Table 1: For the NE Arctic and Irish Sea regions, estimates of the goodness-of-fit term, $G(m)$, penalty term, $P(m)$, and the posterior predictive loss measure, $D(m)$, under the parametric model ("Ricker"), the semiparametric model ("GP"), and the nonparametric mixture model ("DP").

	NE Arctic			Irish Sea		
	Ricker	GP	DP	Ricker	GP	DP
$G(m)$	25.174	20.366	23.031	21.951	6.529	11.673
$P(m)$	43.551	54.979	26.159	34.259	28.833	12.571
$D(m)$	68.725	75.345	49.190	56.210	35.362	24.244

the minimum posterior predictive loss criterion, the Ricker model is slightly favored over the GP model for the NE Arctic region data, whereas the GP model fares better than the Ricker model with the more complicated data in the Irish Sea region. For both regions, the DP mixture model outperforms both the parametric Ricker and semiparametric GP models.

5 Discussion

We have presented a Bayesian nonparametric method to flexibly model the stock-recruitment (S-R) relationship, which is a key aspect of fishery research. A distinguishing feature of the modeling approach is that it incorporates uncertainty in both the log-reproductive success and the stock biomass values. The corresponding joint distribution was modeled with a nonparametric Dirichlet process (DP) mixture. The implied conditional distribution of log-reproductive success given stock biomass values can be used for a wide range of practically important inferences. In particular, working with cod data from six regions of the North Atlantic, we illustrated the capacity of the model to uncover non-linear S-R relationships in the presence of small to moderate sample sizes. The DP mixture model also yields full inference for conditional log-reproductive success densities with shapes that can change with different values in the biomass space. Moreover, by reversing the order of conditioning, i.e., working with the conditional distribution of stock biomass given values of log-reproductive success specified through natural mortality rates, we obtained inference for unfished biomass, an important management reference point. Again, the DP mixture model resulted in flexible estimation for unfished biomass with appropriate quantification of the associ-

ated posterior uncertainty. Particular emphasis was placed on comparison of inference results with simpler parametric and semiparametric models from the fisheries literature. The DP mixture model outperformed those models for the North Atlantic cod data, based on both empirical (graphical) comparison and more formal comparison, using a minimum posterior predictive loss criterion.

We note that multivariate normal DP mixture priors, of the form in (2), have been applied in various settings following the work of Müller, Erkanli, and West (1996) on multivariate density estimation and curve fitting. However, the scope of inference has been typically limited to posterior point estimates, obtained through posterior predictive densities, $\mathrm{E}(f(y, x; G) \mid \text{data})$. This is especially restrictive for regression applications where posterior predictive densities can only provide approximations to, say, $\mathrm{E}\{f(y \mid x; G) \mid \text{data}\}$ and $\mathrm{E}\{\mathrm{E}(y \mid x; G) \mid \text{data}\}$, which are the natural point estimates for the conditional response density and the regression function, respectively. As demonstrated in this paper, proper point estimates, and more importantly, associated uncertainty quantification require the posterior of the random mixing distribution G (or its truncation approximation G_N). Earlier applications of this curve fitting approach to regression settings are reported in Taddy and Kottas (2009; 2010).

We also note certain potential limitations of the proposed mixture modeling approach with regard to its application to inference for S-R relationships. First, the approach is built from a mixture model for the joint distribution of log-reproductive success, $\log(R/S)$, and stock biomass, S, and thus, from a regression perspective, this is a setting where the definition of the response variable includes the covariate. The spurious correlations that may arise as a result of this scale for the S-R data are challenging to address under standard parametric modeling with, say, a normal response distribution for $\log(R/S)$. This problem is alleviated under the proposed approach due to the flexibility of the DP mixture model which allows inferences to adapt to the distributional shape and correlation structure that may be induced by the $(\log(R/S), S)$ scale. Moreover, because the DP mixture model is developed for the joint distribution for $(\log(R/S), S)$, the resulting inferences for the S-R relationship incorporate both measurement error uncertainty (due to the estimation error in the reported values for R and S) and process error uncertainty (due to the stochastic nature of the relationship between R and S). However, the two sources of uncertainty can not be formally distinguished, since the model does not involve any structural assumptions about measurement error in R and S. Following the work of Ludwig and Walters (1981), the literature includes approaches for S-R estimation incorporating error in measurement; however, this work is based on parametric assumptions for the S-R function and related distributions. Finally, although the need to model S-R relationships with time

series methods has long been recognized (Walters 1985) and related aproaches have been proposed (e.g., Millar and Meyer 2000), others have noted that time series bias is small for productive stocks (Myers and Barrowman 1995). Indeed, the majority of the fisheries literature continues to employ standard regression techniques. Our objective in this paper is to compare with this literature and to offer a more general modeling strategy.

An extension of both methodological and practical interest involves building a non-parametric hierarchical model for the largely varying S-R relationships for cod over the North Atlantic regions. The hierarchical model must be sufficiently flexible to capture the diverse S-R relationships across the different regions, while incorporating available spatial information about neighboring regions. In this regard, an important covariate is sea surface temperature, which is measured at a grid within each region. Therefore, it can be incorporated in the model formulation by averaging over the whole region, or more generally, by integrating an underlying spatial model over the grid of each region. Results from this research will be reported in a future article.

Acknowledgements
The work of A. Kottas and K. Fronczyk was supported in part by NSF grant DEB-0727543; the work of S. Munch was supported in part by NSF grant DEB-0727312. The authors wish to thank a reviewer for useful comments that led to improved presentation of the material in the paper.

Appendix A: Prior specification and posterior simulation for the Dirichlet process mixture model

Here, we present the details of the MCMC posterior simulation algorithm for the DP mixture model presented in Section 3. An approach to prior specification for the model hyperparameters is also discussed.

Prior specification: The approach taken to specifying the priors for the DP hyperparameters is to select hyperprior parameter values such that the resulting mixture covers the support of the underlying distribution.

In particular, the approach is based on a small amount of prior information, using

rough prior guesses at the center, say, c_x and c_y, and range, say, r_x and r_y, of stock biomass and log-reproductive success values, respectively. Let R be the 2×2 diagonal matrix with diagonal elements $(r_y/4)^2$ and $(r_x/4)^2$, which are rough prior estimates for the variability in log-reproductive success and biomass values. For a default prior specification method, we consider a single component of the mixture model, $\mathrm{N}_2(\cdot; \mu, \Sigma)$, which is the limiting case of the DP mixture for $\alpha \to 0^+$. Under this version of the model, the marginal prior mean and covariance matrix for the data are given by a_m and $(v-3)^{-1}a_Q B_Q + (a_V - 3)^{-1}B_V + B_m$, respectively. Hence, we set $a_m = (c_y, c_x)^T$, and use matrix R to specify each of the components in the prior covariance above. To this end, the degrees of freedom, v, of the inverse Wishart distribution of Σ in G_0, and the corresponding parameters a_V and a_Q in the priors of V and Q are set at twice the dimension of the data vector, i.e., at 4. Note that 4 is the integer value of a_V that yields finite expectation, and at the same time, the largest possible dispersion in the prior for V. Moreover, smaller values of $(v-3)^{-1}a_Q$ result in more dispersed priors for Q. Finally, R is split evenly between the three marginal prior covariance components to determine the diagonal matrices B_Q, B_V, and B_m.

The DP prior precision parameter, α, controls the number, n^*, of distinct mixture components (e.g., Escobar and West 1995). In particular, for moderate to large sample sizes, a useful approximation to the prior expectation $\mathrm{E}(n^* \mid \alpha)$ is given by $\alpha \log\{(\alpha + n)/\alpha\}$. This expression can be averaged over the gamma(a_α, b_α) prior for α to obtain $\mathrm{E}(n^*)$, thus selecting a_α and b_α to agree with a prior guess at the expected number of distinct mixture components.

MCMC posterior inference: To sample from the posterior, $p(\mathbf{p}, \boldsymbol{\theta}, \mathbf{L}, \alpha, m, V, Q \mid \text{data})$, of the DP mixture model in (4), we use a version of the blocked Gibbs sampler (Ishwaran and Zarepour 2000; Ishwaran and James 2001).

Let $\mathbf{z}_i = (y_i, x_i)$, $i = 1, \ldots, n$, and denote the n^* distinct values in the vector of configuration variables, $\mathbf{L} = (L_1, \ldots, L_n)$, by $L_1^*, \ldots, L_{n^*}^*$. Moreover, let $M_j^* = |\{i : L_i = L_j^*\}|$, $j = 1, \ldots, n^*$, and $M_l = |\{L_i : L_i = l\}|, l = 1, \ldots, N$.

The updates of the $\boldsymbol{\theta}_l = (\mu_l, \Sigma_l)$, $l = 1, ..., N$, depend of the value of l. Specifically, for any $l \notin \{L_j^* : j = 1, \ldots, n^*\}$, we have $p(\boldsymbol{\theta}_l \mid \mathbf{L}, m, V, Q, \text{data}) = g_0(\boldsymbol{\theta}_l; m, V, Q)$, where g_0 is the density of the DP prior base distribution G_0. Thus, in this case, we draw $\mu_l \mid m, V \sim \mathrm{N}_2(m, V)$ and $\Sigma_l \mid Q \sim \mathrm{IW}(v, Q)$. For $l = L_j^*$, $j = 1, \ldots, n^*$,

$$p(\boldsymbol{\theta}_{L_j^*} \mid \mathbf{L}, m, V, Q, \text{data}) = g_0(\boldsymbol{\theta}_{L_j^*}; m, V, Q) \prod_{\{i: L_i = L_j^*\}} \mathrm{N}_2(\mathbf{z}_i; \boldsymbol{\theta}_{L_j^*}).$$

Therefore, for any $j = 1, ..., n^*$, we extend the Gibbs sampler to draw from the posterior

full conditionals for $\mu_{L_j^*}$ and $\Sigma_{L_j^*}$. The former is bivariate normal with mean vector $(V^{-1}+M_j^*\Sigma_{L_j^*}^{-1})^{-1}(V^{-1}m+\Sigma_{L_j^*}^{-1}\sum_{\{i:L_i=L_j^*\}}\mathbf{z}_i)$ and covariance matrix $(V^{-1}+M_j^*\Sigma_{L_j^*}^{-1})^{-1}$. The latter is given by an $\text{IW}(v+M_j^*, Q+\sum_{\{i:L_i=L_j^*\}}(\mathbf{z}_i-\mu_{L_j^*})(\mathbf{z}_i-\mu_{L_j^*})^T)$ distribution.

The posterior full conditional for each L_i is a discrete distribution, $\sum_{l=1}^N \tilde{p}_{li}\delta_l(\cdot)$, with updated weights $\tilde{p}_{li} \propto p_l \text{N}_2(\mathbf{z}_i;\boldsymbol{\theta}_l)$, $l=1,...,N$.

The posterior full conditional for $\mathbf{p}$ is proportional to $f(\mathbf{p};\alpha)\prod_{l=1}^N p_l^{M_l}$. Here, $f(\mathbf{p};\alpha)$ is the joint prior for $\mathbf{p}$, conditionally on α, induced by the truncated stick-breaking construction given in Section 3.1. Specifically, $f(\mathbf{p};\alpha)$ corresponds to a special case of the generalized Dirichlet distribution,

$$f(\mathbf{p};\alpha) = \alpha^{N-1}p_N^{\alpha-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1}\times\cdots\times(1-\sum_{l=1}^{N-2}p_l)^{-1}.$$

Hence, up to its normalizing constant, the full conditional for $\mathbf{p}$ can be written as

$$p_1^{(M_1+1)-1}\times\cdots\times p_{N-1}^{(M_{N-1}+1)-1}p_N^{(\alpha+M_N)-1}(1-p_1)^{(\alpha+\sum_{l=2}^N M_l)-[(M_2+1)+(\alpha+\sum_{l=3}^N M_l)]}\times \\ \cdots\times(1-\sum_{l=1}^{N-2}p_l)^{(\alpha+M_{N-1}+M_N)-[(M_{N-1}+1)+(\alpha+M_N)]}$$

and thus, can be recognized as a generalized Dirichlet distribution with parameters $(M_1+1, M_2+1, \ldots, M_{N-1}+1)$ and $(\alpha+\sum_{k=2}^N M_k, \alpha+\sum_{k=3}^N M_k, \ldots, \alpha+M_N)$. Using the constructive definition of the generalized Dirichlet distribution, the vector $\mathbf{p}$ can be generated by drawing latent $V_1^*,\ldots,V_{N-1}^*$, where the V_l^*, $l=1,...,N-1$, are independent $\text{Beta}(M_l+1, \alpha+\sum_{k=l+1}^N M_k)$, and setting $p_1=V_1^*$; $p_l=V_l^*\prod_{m=1}^{l-1}(1-V_m^*)$, $l=2,\ldots,N-1$; and $p_N=1-\sum_{l=1}^{N-1}p_l$.

Finally, standard updates can be used for the DP hyperparameters (m,V,Q) and α. Denote their corresponding priors (discussed in Section 3.1) by $\pi(m)$, $\pi(V)$, $\pi(Q)$ and $\pi(\alpha)$. Then, the joint full conditional for the DP base distribution hyperparameters,

$$p(m,V,Q \mid \boldsymbol{\theta},\mathbf{L},\text{data}) \propto \pi(m)\pi(V)\pi(Q)\prod_{j=1}^{n^*} g_0(\boldsymbol{\theta}_{L_j^*};m,V,Q).$$

Thus, m has a bivariate normal posterior full conditional with mean vector $(n^*V^{-1}+B_m^{-1})^{-1}(V^{-1}\sum_{j=1}^{n^*}\mu_{L_j^*}+B_m^{-1}a_m)$ and covariance matrix $(n^*V^{-1}+B_m^{-1})^{-1}$. Also, the full conditional for V is an inverse Wishart with a_V+n^* degrees of freedom and scale matrix $B_V+\sum_{j=1}^{n^*}(\mu_{L_j^*}-m)(\mu_{L_j^*}-m)^T$. And Q has a Wishart posterior full conditional with degrees of freedom a_Q+n^*v and scale matrix $(B_Q^{-1}+\sum_{j=1}^{n^*}\Sigma_{L_j^*}^{-1})^{-1}$.

Moreover, $p(\alpha \mid \mathbf{p}) \propto \pi(\alpha)f(\mathbf{p};\alpha) \propto \alpha^{(N+a_\alpha-1)-1}e^{-\alpha(b_\alpha-\log(p_N))}$, i.e., the posterior full conditional for α is given by a gamma distribution with shape parameter $a_\alpha+N-1$

and rate parameter $b_\alpha - \log(p_N) = b_\alpha - \sum_{m=1}^{N-1} \log(1 - V_m^*)$.

Estimation of the model comparison criterion: To compute the predictive loss measure, $D(m)$, used in Section 4.2 for formal model comparison, we need to estimate $\mathrm{E}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$ and $\mathrm{Var}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$. This requires expectations with respect to the posterior predictive distribution, under model m, for replicated response $y_{\mathrm{new},i}$ with associated stock biomass value x_i. Recall from Section 3.1 the notation for partitioning the mean vector and covariance matrix of the $\mathrm{N}_2(y, x; \mu_l, \Sigma_l)$ distribution. Then, under the DP mixture model, the expressions for the first and second posterior predictive moments given each stock biomass value, x_i, $i = 1, \ldots, n$, are as follows:

$$\begin{aligned}
\mathrm{E}(y \mid x_i, \mathrm{data}) &= \{p(x_i \mid \mathrm{data})\}^{-1} \int y p(y, x_i \mid \mathrm{data}) \mathrm{d}y \\
&= \{p(x_i \mid \mathrm{data})\}^{-1} \int \sum_{l=1}^{N} p_l \mathrm{N}(x_i; \mu_l^x, \Sigma_l^{xx}) \{\mu_l^y + \Sigma_l^{yx} \Sigma_l^{xx-1} (x_i - \mu_l^x)\} \times \\
&\qquad\qquad p(\boldsymbol{\theta}, \mathbf{p} \mid \mathrm{data}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{p} \\
\mathrm{E}(y^2 \mid x_i, \mathrm{data}) &= \{p(x_i \mid \mathrm{data})\}^{-1} \int y^2 p(y, x_i \mid \mathrm{data}) \mathrm{d}y \\
&= \{p(x_i \mid \mathrm{data})\}^{-1} \int \sum_{l=1}^{N} p_l \mathrm{N}(x_i; \mu_l^x, \Sigma_l^{xx}) [\{\mu_l^y + \Sigma_l^{yx} \Sigma_l^{xx-1} (x_i - \mu_l^x)\}^2 + \\
&\qquad\qquad \Sigma_l^{yy} - \Sigma_l^{yx} \Sigma_l^{xx-1} \Sigma_l^{xy}] p(\boldsymbol{\theta}, \mathbf{p} \mid \mathrm{data}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{p}
\end{aligned}$$

where $p(x_i \mid \mathrm{data}) = \int \sum_{l=1}^{N} p_l \mathrm{N}(x_i; \mu_l^x, \Sigma_l^{xx}) p(\boldsymbol{\theta}, \mathbf{p} \mid \mathrm{data}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{p}$. The $\mathrm{E}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$ are obtained from Monte Carlo integration of $\mathrm{E}(y \mid x_i, \mathrm{data})$ and the $\mathrm{Var}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$ are calculated with Monte Carlo integration of $\mathrm{E}(y^2 \mid x_i, \mathrm{data}) - \{\mathrm{E}(y \mid x_i, \mathrm{data})\}^2$.

Appendix B: Posterior simulation for the parametric and semiparametric models of the comparison study

Here, we describe MCMC fitting for the parametric Ricker model and the semiparametric GP model used in the comparison study presented in Section 4.2.

Parametric model: For the Ricker model, denote by a_0 and b_0 the mean and variance of the normal prior for β_0, and by a_σ and b_σ the shape and rate parameters of the inverse gamma prior for σ^2. The model can be fitted with an MCMC algorithm based on a normal full conditional for β_0 with mean $(b_0 \sum_{i=1}^{n} y_i + b_0 \beta_1 \sum_{i=1}^{n} S_i + a_0 \sigma^2)/(n b_0 + \sigma^2)$ and variance $(b_0 \sigma^2)/(n b_0 + \sigma^2)$, and an inverse gamma full conditional for σ^2 with

shape parameter $a_\sigma + 0.5n$ and rate parameter given by $b_\sigma + 0.5\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 S_i)^2$. Moreover, we update β_1 with a random walk Metropolis-Hastings step, using a normal proposal distribution on the logarithmic scale.

The elements of the predictive loss criterion, $\mathrm{E}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$ and $\mathrm{Var}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$, $i = 1, \ldots, n$ (see Section 4.2 and Appendix A), are estimated using draws from the posterior predictive distribution at each S_i, $i = 1, \ldots, n$, which is given by $p(y_{\mathrm{new},i} \mid \mathrm{data}) = \int \mathrm{N}(y_{\mathrm{new},i}; \beta_0 - \beta_1 S_i, \sigma^2) p(\beta_0, \beta_1, \sigma^2 \mid \mathrm{data})\, \mathrm{d}\beta_0 \mathrm{d}\beta_1 \mathrm{d}\sigma^2$. Each of these predictive distributions is readily sampled using the posterior draws for $(\beta_0, \beta_1, \sigma^2)$.

Semiparametric model: Turning to the GP regression model, let $\mathrm{N}(m, s^2)$ be the prior for μ, denote by inverse-gamma(a_τ, b_τ) and inverse-gamma(a_σ, b_σ) the priors for τ^2 and σ^2, respectively, and let $(0,b_\phi)$ be the support of the uniform prior for ϕ. The mean of the normal prior for μ is set at a prior guess on the center of the log-reproductive success values, and the variance is set equal to $10(r_y/4)^2$, where r_y is a prior guess at the range of the log-reproductive success values. We also set $a_\tau = a_\sigma = 2$ resulting in inverse gamma priors for τ^2 and σ^2 with infinite prior variances. The rate parameters of both priors are also specified through $10(r_y/4)^2$. Under the exponential correlation function for the GP prior, $3/\phi$ is the *range of dependence*, that is, the distance between stock biomass values, $d = |S - S'|$, such that the correlation between $h(S)$ and $h(S')$ is approximately 0.05. The prior for ϕ is based on the maximum difference between stock biomass values, $d_{\max} = \max|S - S'|$, choosing b_ϕ such that $0.01 d_{\max} = 3/b_\phi$.

Regarding posterior simulation, the model can be fitted with a Gibbs sampler based on standard updates for all parameters except ϕ. Let $\mathbf{y} = (y_1, ..., y_n)$, $\boldsymbol{\eta}$ be the n-dimensional vector with $\eta_i = h(S_i)$, $i = 1, \ldots, n$, I_n the identity matrix of dimension n, and $\mathbf{1}_n$ the vector of dimension n with each of its element equal to 1. Then, the posterior full conditionals are given by:

$$
\begin{aligned}
\boldsymbol{\eta} \mid \mu, \sigma^2, \tau^2, \phi, \mathrm{data} &\sim \mathrm{N}_n\left((\sigma^{-2} I_n + C^{-1})^{-1}(\sigma^{-2}\mathbf{y} + \mu C^{-1}\mathbf{1}_n), (\sigma^{-2} I_n + C^{-1})^{-1}\right) \\
\mu \mid \boldsymbol{\eta}, \tau^2, \phi &\sim \mathrm{N}\left(\frac{\mathbf{1}_n^T C^{-1}\boldsymbol{\eta} + m s^{-2}}{\mathbf{1}_n^T C^{-1}\mathbf{1}_n + s^{-2}}, \frac{1}{\mathbf{1}_n^T C^{-1}\mathbf{1}_n + s^{-2}}\right) \\
\sigma^2 \mid \boldsymbol{\eta}, \mathrm{data} &\sim \text{inverse-gamma}\left(a_\sigma + 0.5n, b_\sigma + 0.5\sum_{i=1}^{n}(y_i - \eta_i)^2\right) \\
\tau^2 \mid \boldsymbol{\eta}, \mu, \phi &\sim \text{inverse-gamma}\left(a_\tau + 0.5n, b_\tau + 0.5(\boldsymbol{\eta} - \mu\mathbf{1}_n)^T H^{-1}(\boldsymbol{\eta} - \mu\mathbf{1}_n)\right) \\
\phi \mid \boldsymbol{\eta}, \mu, \tau^2 &\propto |C|^{-1/2} \exp\{-0.5(\boldsymbol{\eta} - \mu\mathbf{1}_n)^T C^{-1}(\boldsymbol{\eta} - \mu\mathbf{1}_n)\} 1(0 < \phi < b_\phi)
\end{aligned}
$$

where H is the observed correlation matrix with elements $H_{ij} = \exp\{-\phi|S_i - S_j|\}$, and $C = \tau^2 H$ is the observed covariance matrix. The parameter ϕ is updated with a Metropolis-Hastings step using a normal proposal distribution on the logarithmic scale.

Posterior predictive inference for the S-R relationship, $h(S)$, is obtained through the n values in vector $\boldsymbol{\eta}$ augmented with a set of M new stock biomass values, $\tilde{S} = (\tilde{S}_1, \ldots, \tilde{S}_M)$. Let $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, ..., \tilde{\eta}_M)$, where $\tilde{\eta}_j = h(\tilde{S}_j)$, $j = 1, ..., M$. The vector $\tilde{\boldsymbol{\eta}}$, conditionally on $\boldsymbol{\eta}$ and the GP hyperparameters, follows an M-variate normal distribution with mean vector $(\mu \mathbf{1}_M + C^{Mn}C^{-1}(\boldsymbol{\eta} - \mu \mathbf{1}_n))$ and covariance matrix $C^{MM} - C^{Mn}C^{-1}(C^{Mn})^T$, where $C_{ij}^{Mn} = \tau^2 \exp\{-\phi|\tilde{S}_i - S_j|\}$, and $C_{ij}^{MM} = \tau^2 \exp\{-\phi|\tilde{S}_i - \tilde{S}_j|\}$. Therefore, a posterior realization for $h(S)$ is obtained through $\{\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}}\}$ at each iteration of the MCMC using the currently imputed parameter values.

The $\mathrm{E}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$ and $\mathrm{Var}^{(m)}(y_{\mathrm{new},i} \mid \mathrm{data})$, $i = 1, \ldots, n$, needed to calculate the posterior predictive loss criterion, can be computed through the mean and variance, respectively, of replicated log-reproductive success values corresponding to each observed stock biomass level. For each observed stock biomass value S_i, $i = 1, ..., n$, the replicated responses are sampled from the associated posterior predictive distribution given by $p(y_{\mathrm{new},i} \mid \mathrm{data}) = \int \mathrm{N}(y_{\mathrm{new},i}; \eta_i, \sigma^2) p(\eta_i, \sigma^2 \mid \mathrm{data})\, \mathrm{d}\eta_i \mathrm{d}\sigma^2$.

References

Barot S, Heino M, O'Brien L, Dieckmann U (2004) Long-term trend in the maturation reaction norm of two cod stocks. Ecological Applications 14:1257-1271

Brander K, Mohn R (2004) Effect of the North Atlantic Oscillation on recruitment of Atlantic cod (Gadus morhua). Canadian Journal of Fisheries and Aquatic Sciences 61:1558-1564

Chen DG, Ware DM (1999) A neural network model for forecasting fish stock recruitment. Canadian Journal of Fisheries and Aquatic Sciences 56:2385-2396

Clark JS (2005) Why environmental scientists are becoming Bayesians. Ecology Letters 8:2-14

Clark JS (2007) Models for Ecological Data: An Introduction. Princeton University Press

Cook RM (1998) A sustainability criterion for the exploitation of North Sea cod. ICES Journal of Marine Science 55:1061-1070

Eero M, Köster FW, Plikshs M, Thurow F (2007) Eastern Baltic cod (*Gadus morhua callarias*) stock dynamics: extending the analytical assessment back to the mid-1940s. ICES Journal of Marine Science 64:1257-1271

Escobar M and West M (1995) Bayesian density estimation and inference using mix-

tures. Journal of the American Statistical Association 90:577-588

Evans GT, Rice JC (1988) Predicting recruitment from stock size without the mediation of a functional relation. Journal du Conseil - Conseil International pour l'Exploration de la Mer 44:111-122

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Annals of Statistics 1:209-230

Garcia SM, de Leiva Moreno JI (2003) Global Overview of Marine Fisheries. In Responsible Fisheries in the Marine Ecosystem. Sinclair M, Valdimarsson, G (eds) FAO and CABI Publishing. pp 1-24

Gelfand A, Ghosh S (1998) Model choice: A minimum posterior predictive loss approach. Biometrika 85:1-11

Hilborn R, Walters CJ (1992) Quantitative Fisheries Stock Assessment: Choice, Dynamics, and Uncertainty. Chapman and Hall, New York

ICES (2005) Report of the ICES Advisory Committee on Fishery Management, Advisory Committee on the Marine Environment and Advisory Committee on Ecosystems, 2005. ICES Advice. Vol. 1 No. 11

Ishwaran H, James LF (2001) Gibbs Sampling for Stick-Breaking Priors, Journal of the American Statistical Association 96:161-73

Ishwaran H, Zarepour M (2000) Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models. Biometrika 87:371-390

Jacobson LD, MacCall AD (1995) Stock-recruitment models for Pacific sardine (Sardinops sagax). Canadian Journal of Fisheries and Aquatic Sciences 52:566-577

Lo AY (1984) On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. The Annals of Statistics 12: 351-357

Ludwig D, Walters CJ (1981) Measurement errors and uncertainty in parameter estimates for stock and recruitment. Canadian Journal of Fisheries and Aquatic Sciences 38: 711-720

McAllister MK, Kirkwood GP (1998) Bayesian stock assessment: a review and example application using the logistic model. ICES Journal of Marine Science 55:1031-1060

Millar RB, Meyer R (2000) Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings within-Gibbs sampling. Applied Statistics 49:327-342

Müller P, Erkanli A, West M (1996) Bayesian curve fitting using multivariate normal mixtures. Biometrika 83:67-79

Munch SB, Kottas A, Mangel M (2005) Bayesian nonparametric analysis of stock-recruitment relationships. Canadian Journal of Fisheries and Aquatic Sciences 62:1808-1821

Myers RAM, Barrowman NJ (1995) Time series bias in the estimation of density dependent mortality in stock-recruitment models. Canadian Journal of Fisheries and Aquatic Sciences 52:223-232

Neal RM (1998) Regression and classification using Gaussian process priors. In Bayesian statistics 6: Proceedings of the sixth Valencia international meeting. J M Bernardo, J O Berger, A P Dawid, A F M Smith (eds) Oxford University Press. pp 475-501

Patil A (2007) Bayesian Nonparametrics for Inferences of Ecological Dynamics. Ph.D. Dissertation, University of California, Santa Cruz

Punt A, Hilborn R (1997) Fisheries stock assessment and decision analysis: The Bayesian approach. Reviews in Fish Biology and Fisheries 7:35-65

Quinn TJI, Deriso RB (1999) Quantitative Fish Dynamics. Oxford University Press, New York

Ricker WE (1954) Stock and recruitment. Journal of the Fisheries Research Board 11:559-623

Sethuraman J (1994) A constructive definition of Dirichlet priors. Statistica Sinica 4:639-650

Stige LC, Ottersen G, Brander K, Chan KS, Stenseth NC (2006) Cod and climate: effect of the North Atlantic Oscillation on recruitment in the North Atlantic. Marine Ecology Progress Series 325:227-241

Taddy M, Kottas A (2009) Markov Switching Dirichlet Process Mixture Regression. Bayesian Analysis 4:793-816

Taddy M, Kottas A (2010) A Bayesian nonparametric approach to inference for quantile regression. Journal of Business and Economic Statistics 28: 357-369

Walters CJ (1985) Bias in the estimation of functional relationships from time series data. Canadian Journal of Fisheries and Aquatic Sciences 42:147-149