Fisheries Research 93 (2008) 212–221

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Performance of deviance information criterion model selection in statistical catch-at-age analysis

Michael J. Wilberg*, James R. Bence

Quantitative Fisheries Center and Department of Fisheries and Wildlife, 13 Natural Resources, Michigan State University, East Lansing, MI 48824-1222, USA

ARTICLE INFO

Article history:
Received 18 December 2007
Received in revised form 23 April 2008
Accepted 24 April 2008

Keywords:
Stock assessment
Model averaging
Bayesian
Catchability

ABSTRACT

Most stock assessments involve fitting alternative models and selecting among them to provide management advice. Incorrect model specification can lead to unreliable population and mortality estimates, and methods to decide among assessment models so as to obtain reliable estimates are needed. We used Monte Carlo simulations to assess whether using deviance information criterion (DIC) model selection and averaging resulted in improved accuracy of important management quantities from statistical catch-at-age models. We challenged DIC with three estimation models (that differed in how they estimated catchability) and three scenarios of data accuracy and time-varying catchability. DIC usually selected the structurally appropriate model, and point estimates from the best model or the model average were relatively unbiased in that the average deviation from the true value was near zero. The distributions of point estimates about true values from DIC-based model averaging and from the best model (lowest DIC) were similar, perhaps because all of the estimation models were quite similar to the data-generating models. DIC seems to provide a useful metric to compare evidence in favor of alternative assessment models. This study is one of the first to evaluate the performance of DIC in models where the purpose is to predict unobserved quantities.

1. Introduction

Development of a fishery stock assessment often involves fitting alternative models and using what is thought to be the best among them to provide management advice (National Research Council [NRC] 1998). The "best" model is often selected by ad hoc criteria with unknown performance characteristics. Model selection is an area of importance because estimated quantities important for management, such as exploitable biomass, can be extremely sensitive to model structure (Punt and Hilborn, 1997; Patterson, 1999; McAllister and Kirchner, 2002). Common uncertainties in statistical catch-at-age (SCA) model structure include stock-recruitment relationships, selectivity functions, and assumptions linking fishery catch with abundance and effort (Patterson, 1999; McAllister and Kirchner, 2002). In some cases, results from several models have been reported to managers, but quantitative estimates of the relative likelihood a particular model was most "correct" have typically not been provided (McAllister and Kirchner, 2002).

Model selection criteria have been applied to SCA models, but previous applications have been limited in the types of models that could be compared. Helu et al. (2000) evaluated performance of Akaike's Information Criterion (AIC; Akaike, 1973) and Schwartz's Bayesian Information Criterion (BIC; Schwartz, 1978) to assess model selection in SCA models and found that AIC and BIC both performed well by selecting the candidate model that was the same as the data-generating model in most of their scenarios. Unfortunately, although AIC or BIC may perform well in some cases, their implementation is problematic when models differ in their random effects or hierarchical structures because the number of parameters in these models is not easy to determine (Burnham and Anderson, 2002; Spiegelhalter et al., 2002). Therefore, comparing structurally complex SCA models requires alternative model selection approaches that can account for random effects and priors on parameters.

The deviance information criterion (DIC) has been developed, in a Bayesian context, to select among complex hierarchical models where the number of effective parameters is not readily apparent (Spiegelhalter et al., 2002). Much like AIC and BIC, DIC selects among models by trading off goodness of fit and model complexity. DIC is a generalization of AIC and reduces to AIC in the case of a fixed effects model with diffuse priors (Spiegelhalter et al., 2002). However, DIC is particularly applicable to models with random effects or hierarchical structure because it estimates the effective number

* Corresponding author. Current address: Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, P.O. Box 38, Solomons, MD 20688, USA. Tel.: +1 410 326 7273; fax: +1 410 326 7318.
E-mail address: wilberg@cbl.umces.edu (M.J. Wilberg).

0165-7836/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.fishres.2008.04.010

of parameters rather than requiring the user to provide this. The "focus of prediction" of DIC is on the random effects, rather than the distributional parameters for the random effects, and for some assessment purposes this may be an advantage, as has been argued in other fields (Berg et al., 2004).

Although DIC has been applied in many studies (e.g., Barry et al., 2003; Kizilkaya and Tempelman, 2003), relatively few studies have evaluated the performance of DIC model selection (e.g., Cardoso and Tempelman, 2003; Kizilkaya and Tempelman, 2003, 2005; Berg et al., 2004; Ward, 2008). These studies found that DIC usually could be used to select the correct model (i.e., the model that generated the data) from the set of candidate models, and that the estimated number of effective parameters seemed reasonable for a given model.

Our objectives were to determine if using DIC, as a model selection criterion, resulted in choosing an appropriate model structure and level of complexity. Also, we wanted to evaluate whether using formal model selection methods provided more accurate point estimates of important fishery management quantities, such as fishing mortality rate and biomass in the last year. In this context, we recognize that although fishery stock assessments often have Bayesian aspects, point estimates from the assessments usually play a central role in informing management. To achieve these objectives, we designed a simulation study and challenged the model selection criteria with three estimation models and three scenarios of data accuracy and time-varying catchability.

2. Methods

We used Monte Carlo simulations to evaluate whether using DIC to select among or average over SCA model variants provided more accurate estimates of quantities used for management than an approach of using a single model structure in all cases. Our basic approach was to (1) generate simulated data sets from each of three different generating models; (2) apply three different stock assessment estimation models to each simulated data set and calculate DIC in each case; and (3) Use DIC to select the best model and to calculate results averaged over models (Fig. 1). Thus, each simulated data set led to four sets of estimates, namely the model average and those from each of the estimation models, which were then compared to the true values from the data-generating model. Our data-generating models differed in their relationship between fishing mortality and observed effort: (1) fishery catchability varying as white noise, (2) fishery catchability increasing a constant amount each year, and (3) fishing mortality as unrelated to observed effort. We chose these data-generating scenarios because the relative performance of different estimation models was likely to change over this range of conditions (Wilberg and Bence, 2006). Three estimation models were fitted to each of the 300 datasets (100 from each scenario). These estimation models contained different assumptions regarding fishery catchability: (1) catchability was modeled as white noise, (2) as a random walk, and (3) where catchability was effectively estimated as a free parameter for each year. While the number of simulations is small relative to most simulation studies because of the computationally intensive nature of these methods, we believe the sample sizes are large enough to show general trends in performance.

All models contained 15 years of data and eight age classes with the last age class representing all fish that age and older. Data-generating models were based on commercial fisheries for lake whitefish (*Coregonus clupeaformis*) in the upper Great Lakes, although nothing is unusual about the life history of lake whitefish that would suggest our results would not be broadly applicable. Symbols and equations defining the data-generating models and estimation models are presented in Tables 1 and 2. Equations are

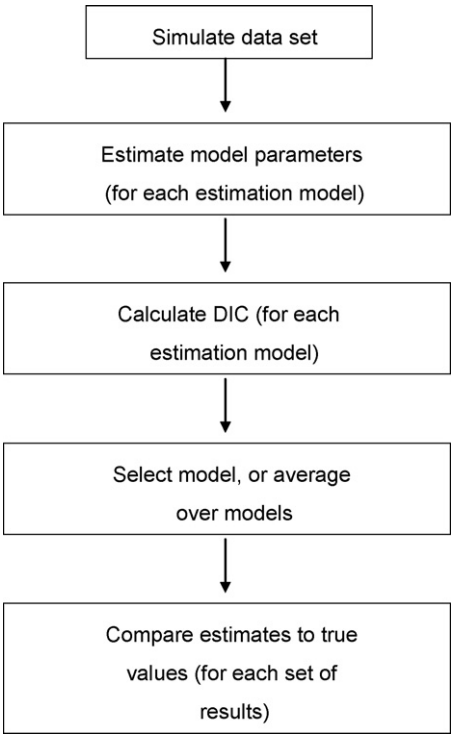

Fig. 1. Flow chart of simulation study to evaluate performance of deviance information criterion (DIC) in statistical catch-at-age stock assessments for one data-generating model. Three data-generating models were used in the study (see text for details), and the procedure was repeated 100 times for each data-generating model.

referred to in the text as Eq. T*x.y*, where *x* is the table number and *y* is the equation number within Table *x*. To simplify presentation, equivalent quantities and parameters in estimation and data-generating models are not differentiated except when they both appear in the same equation, in which case estimated quantities are denoted with a caret above the symbol (Table 3).

2.1. Data-generating model

The data-generating model described the population dynamics and created data sets of total fishery catch, the age composition of the fishery catch, total survey catch per unit effort (CPUE), the age composition of the survey, and fishery effort. To model population dynamics, we used an age-structured model that followed cohorts over time. Recruitment (abundance-at-age 1) was generated from a lognormal distribution with a coefficient of variation (CV) of 100%. Numbers-at-age in the first year were calculated assuming a stable age distribution with lognormal errors, where recruitment and mortality rates prior to the first year of the simulation were on average the same as in the first year Eq. (T2.1). Cohorts were tracked over time by applying a simple exponential mortality model Eq. (T2.2a); the last age class was treated as representing all fish age 8 and older Eq. (T2.2b). Biomass each year was the sum over ages of the product of age-specific abundance and mean mass-at-age Eq. (T2.3).

We used a separable model to generate fishing mortality rates (i.e., fishing mortality was the product of an age effect and a year effect). The total mortality rates were determined by the natural mortality rate and age-specific fishing mortality rates Eq. (T2.4). *M* was held constant across ages and years at 0.25. The instantaneous fishing mortality rate was a function of catchability, fishing effort, and age-specific selectivity Eq. (T2.5). We allowed fishing mortality to change over time by allowing fishery effort to change and

Table 1
Symbols and descriptions of variables for data-generating and estimation models

Symbol	Description	Value (if needed in the data-generating model)
$\bar{R}$	Average recruitment	1,000,000
$N_{y,a}$	Abundance by age and year	
B_y	Biomass	
$Z_{y,a}$	Total instantaneous mortality rate by age and year	
$F_{y,a}$	Instantaneous fishing mortality rate by age and year	
M	Instantaneous natural mortality rate	0.25
$s_{a,\mathrm{f}}$	Fishery age-specific selectivity	See Fig. 2
$s_{a,\mathrm{s}}$	Survey age-specific selectivity	See Fig. 2
E_y	Fishery effort	See Fig. 2
$q_{y,\mathrm{f}}$	Fishery catchability	
$\tilde{E}_y$	Observed fishery effort	
q_{s}	Survey catchability	0.0001
$\bar{q}_{\mathrm{f}}$	Mean fishery catchability	0.05
$C_{y,a}$	Expected fishery catch-at-age	
$I_{y,a}$	Expected survey catch-at-age	
$\tilde{C}_y$	Observed total fishery catch	
$\tilde{I}_y$	Observed total survey catch	
$u_{y,a,\mathrm{f}}$	Proportion of catch-at-age in fishery	
$u_{y,a,\mathrm{s}}$	Proportion of catch-at-age in survey	
w_a	Mean mass-at-age	0.16, 0.45, 0.82, 1.2, 1.55, 1.86, 2.11, 2.3
δ_y	Deviations for white noise catchability	
ε_y	Deviations for linear increase catchability	
ω_y	Deviations for random walk catchability	
a, b	Parameters for linear increase in catchability	0.032, 0.00225
f_y	Fishing intensity by year	
σ_γ	Standard deviation for $\log_e$ recruitment variation	1.0
σ_τ	Standard deviation for $\log_e$ fishery measurement error	0.1
σ_υ	Standard deviation for $\log_e$ of survey measurement error	0.2–0.8
σ_δ	Standard deviation for $\log_e$ catchability deviations for white noise	0.2
σ_ε	Standard deviation for $\log_e$ catchability deviations for linear increase	0.05
σ_ω	Standard deviation for $\log_e$ random walk catchability deviations	0.2

Table 2
Data-generating and estimation model equations

Population model equations		Application
$N_{1,a} = \bar{R}e^{-\sum_{a=1}^{a-1} Z_{1,a}+\gamma_a}; \gamma_a \sim N\left(0, \sigma_\gamma^2\right)$	(T2.1)	Generation
$N_{y+1,a+1} = N_{y,a}e^{-Z_{y,a}}$	(T2.2a)	Both
$N_{y+1,8} = N_{y,7}e^{-Z_{y,7}} + N_{y,8}e^{-Z_{y,8}}$	(T2.2b)	Both
$B_y = \sum_a N_{y,a}w_a$	(T2.3)	Both
$Z_{y,a} = M + F_{y,a}$	(T2.4)	Both
$F_{y,a} = q_y E_y s_a$	(T2.5)	Both
Catchability model equations		
White noise		Both
$\log_e q_{y,\mathrm{f}} = \log_e \bar{q}_{\mathrm{f}} + \delta_y; \delta_y \sim N\left(0, \sigma_\delta^2\right)$	(T2.6)	
Linear increase		Generation
$q_{y,\mathrm{f}} = a + b(y) + \varepsilon_y; \varepsilon_y \sim N\left(0, \sigma_\varepsilon^2\right)$	(T2.7)	
Random walk		Estimation
$\log_e q_{y+1,\mathrm{f}} = \log_e q_{y,\mathrm{f}} + \omega_y; \omega_y \sim N\left(0, \sigma_\omega^2\right)$	(T2.8)	
Freely estimate f_y (no catchability)		Estimation
$F_{y,a} = f_y s_{a,\mathrm{f}}$	(T2.9)	
Observation model equations		
$C_{y,a} = \frac{F_{y,a}}{Z_{y,a}}(1 - e^{-Z_{y,a}})N_{y,a}$	(T2.10)	Both
$\tilde{C}_y = e^{\tau_y}\sum_a C_{y,a};\ \tau_y \sim N(0, \sigma^2)$	(T2.11)	Both
$I_{y,a} = q_{\mathrm{s}} s_a N_{y,a}$	(T2.12)	Both
$\tilde{I}_y = e^{\upsilon_y}\sum_a I_{y,a};\ \upsilon_y \sim N(0, \sigma^2)$	(T2.13)	Both

The application column indicates whether the equation was used in the data-generating model (Generation), the estimation model (Estimation), or both.

by incorporating two processes of time-varying catchability (see below).

The overall level of fishing mortality varied among simulations. This was accomplished by multiplying the baseline effort (Fig. 2A) by a Uniform(1,2) number selected for each simulation. The baseline effort series was designed to produce an average level of F for fully selected ages approximately equal to M. Thus, this procedure led to F for fully selected ages varying among simulations between M and $2M$. For the white noise catchability and linearly increasing catchability scenarios observed effort equaled true effort. For the scenario with uninformative effort, the observed effort series was drawn as uniform random numbers between the minimum true effort (effort in year 1) and the maximum true effort (effort in year 8). The selectivity pattern for the fishery was dome shaped to simulate a gill net fishery (Fig. 2B).

We included two models for time-varying catchability, which caused SCA models to have variable performance (Wilberg and Bence, 2006). The $\log_e$ of catchability was modeled as white noise to simulate a fishery where catchability varied from year to year about a constant mean Eq. (T2.6), perhaps due to environmental effects. In the second scenario, catchability increased linearly over time with a small amount of white noise error Eq. (T2.7), which could represent learning by fishers or increases in gear efficiency. Both models were parameterized to have the same expected catchability (over the time series) and similar variances of $\log_e q_{\mathrm{f}}$. We achieved this by simulating data sets and adjusting the catchability parameters until the mean and variance of catchability were the same as in the white noise case. We used a value of 0.2 for the standard deviation of the $\log_e$ of catchability. This value is similar to estimates of the CV of catchability for commercial fisheries in New Zealand (Francis et al., 2003), but was less than median values of the CV of fishery CPUE estimated by Harley et al. (2001) for International Council for the Exploration of the Sea fisheries of 0.4–0.8. Note that the CV of fishery CPUE, as estimated by Harley et al. (2001), is an upper bound on the CV of catchability for those fisheries because it also reflects measurement error and variation in survey catchability.

Fishery catch was calculated with the Baranov catch equation (Eq. (T2.10); Quinn and Deriso, 1999). We multiplied total catch by a lognormal measurement error to calculate observed fishery catch Eq. (T2.11); the measurement error CV for fishery catch was about 0.1. Observed age compositions for the fishery catch were generated by drawing a random sample from a multinomial distribution of size 200 with proportions equal to the true proportions of catch-at-age in the fishery. We used this effective sample size because it indicates an informative age compositions within the range often achieved in real fisheries data (e.g., Crone and Sampson, 1998). Survey CPUE-at-age was calculated as the product of survey

Table 3
Objective function equations for statistical catch-at-age models

$L = \sum_i \ell_i$	(T3.1)	Objective function
$\ell_1 = \frac{1}{2\sigma_\tau^2}\sum_y (\log_e(\tilde{C}_y) - \log_e(\hat{C}_y))^2$	(T3.2)	Fishery catch
$\ell_2 = \frac{1}{2\sigma_\upsilon^2}\sum_y (\log_e(\tilde{I}_y) - \log_e(\hat{I}_y))^2$	(T3.3)	Survey catch-per-effort
$\ell_3 = -n_f \sum_y \sum_a u_{y,a,f} \log_e(\hat{u}_{y,a,f})$	(T3.4)	Proportion at age in the fishery catch
$\ell_4 = -n_s \sum_y \sum_a u_{y,a,s} \log_e(\hat{u}_{y,a,s})$	(T3.5)	Proportion at age in the survey catch
$\ell_5 = \frac{1}{2\sigma_q^2}\sum_y (\hat{\delta}_y)^2$	(T3.6)	White noise catchability
$\ell_5 = \frac{1}{2\sigma_q^2}\sum_y (\hat{\omega}_y)^2$	(T3.7)	Random walk catchability

catchability, abundance, and survey selectivity (Fig. 2B; Eq. (T2.12)), and observed survey CPUE was the product of total survey CPUE and a lognormal measurement error Eq. (T2.13).

As was the case for average fishing mortality, survey quality varied randomly among simulated datasets. This was accomplished by selecting the measurement error CV for each simulation from a Uniform(0.2,0.8) distribution. These levels of survey CV were selected because they provided contrast in performance of several estimation models (Wilberg and Bence, 2006). Catchability of the survey was constant over time. Observed survey age compositions were generated by drawing a random sample from a multinomial distribution of size 150 with proportions equal to the true proportions

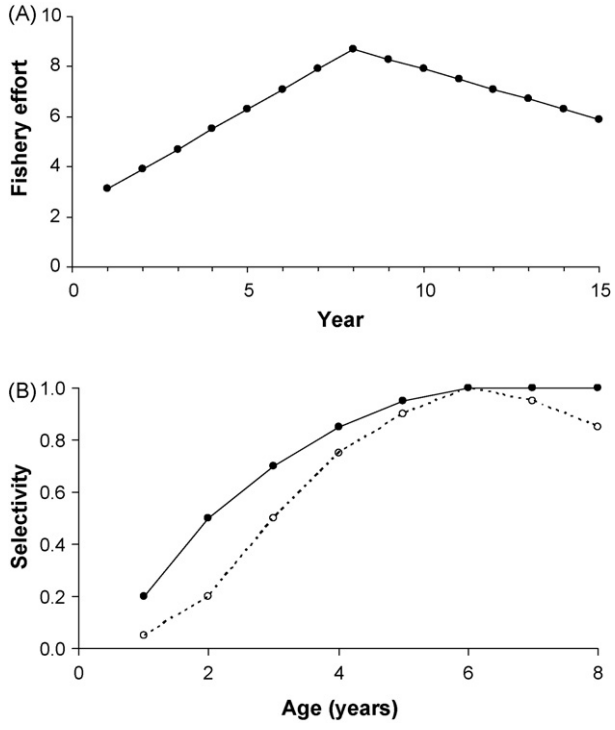

Fig. 2. Fishing effort (panel A) and selectivity patterns (panel B) for the fishery (dashed line) and survey (solid line) for data-generating models.

of CPUE at age calculated from Eq. (T2.12). We chose this effective sample size because it still reflects an informative age-composition, but also that survey age-composition data will often be based on fewer sampled trips than such data from fisheries, and this can be an important determinant of effective sample size (Crone and Sampson, 1998).

2.2. Estimation model

The estimation models were largely the same as the simulation models except for how catchability was estimated and how numbers-at-age in the first year and recruitments were handled. Common parameters among models included $N_{1,1} \ldots N_{15,1}$ (Recruitment), $N_{1,2} \ldots N_{1,8}$ (numbers-at-age in the first year), and $s_{1,f} \ldots s_{7,f}$ (fishery selectivity), $s_{1,s} \ldots s_{7,s}$ (survey selectivity) and q_s (survey catchability). All models had 52 unique estimated parameters. Parameterization of the models to reduce correlations among parameters is described in Appendix A. Numbers-at-age in the first year and recruitment for each year were estimated as parameters during the model fitting process. After the first year and age, abundance-at-age followed a standard exponential mortality model with the last age representing all fish that age and older Eqs. (T2.2a) and (T2.2b).

The total mortality rate ($Z_{y,a}$) was the sum of M and $F_{y,a}$ Eq. (T2.4); M was assumed known at 0.25 (the true value from the simulation models). Fishing mortality followed a separable model for all of our estimation models (Fournier and Archibald, 1982; Deriso et al., 1985; Methot, 1990). Fishery and survey selectivities were estimated as individual parameters by constraining the log of the age-specific selectivities to sum to zero. This method was used to reduce correlations among selectivity parameters. Estimation models contained three methods of estimating catchability: white noise, random walk, and no catchability (directly estimating fishing mortality). The first estimation model allowed $\log_e$ fishery catchability to vary with white noise about a constant mean Eq. (T2.6). The second estimation model allowed $\log_e$ fishery catchability to vary according to a random walk Eq. (T2.8). In our third estimation model, we estimated the fishing mortality rate for fully selected age classes as a parameter, and then applied the fishery selectivity to calculate age-specific fishing mortality rates Eq. (T2.9). This method does not use fishery effort as a data source. The estimation models also predicted proportions of fishery and survey catch-at-age.

2.3. Model fitting and convergence

We estimated model parameter values using a Bayesian approach as implemented in AD Model Builder version 6.0.2 (Otter Research Ltd., 2000). The objective function was the sum of the likelihood components and priors. Each component was the negative of the log likelihood for a single data source or an informative prior related to time-varying catchability Eq. (T3.1). Our estimation models assumed lognormal distributions of errors for total catch for the fishery Eq. (T3.2) and survey CPUE Eq. (T3.3) and multinomial distributions for age compositions of the fishery Eq. (T3.4) and the survey (Eq. (T3.5); Fournier and Archibald, 1982).

For estimation models that used fishery effort as a data source, fishing mortality was an explicit function of effort and catch was linked to abundance and fishery effort by estimating catchability coefficients. We assumed lognormal deviations for catchability in the white noise Eq. (T3.6), and random walk Eq. (T3.7) estimation models. This component in the objective function was a prior and penalizes large deviations from mean catchability (for the white noise model) or large year-to-year deviations (in the random walk model). We placed uninformative uniform priors on common

parameters among models and these priors were the same in each model.

Determining weights for different data sources is a challenging issue in SCA and other stock assessment models (Quinn and Deriso, 1999) and is usually an iterative process where different sets of weights are evaluated. We weighted data sources by the inverse of their variances for lognormal likelihood components and priors and by their effective sample sizes for multinomial likelihood components. Effective sample sizes and CVs of the fishery and survey catch and age compositions were set to their true values from the generating models. The standard deviation for the white noise and random walk catchability deviations (on the $\log_e$ scale) was assumed known at 0.2, which was approximately equal to the expected standard deviation in the data-generating models. This approach implicitly assumes that some process is available to the analyst to appropriately assess the information content of the different data sources.

The AD Model Builder implementation of Markov Chain Monte Carlo (MCMC) includes first obtaining the parameter values that maximize the posterior probability density and the associated asymptotic variance–covariance matrix, then using these parameter estimates as starting values for the MCMC chain. AD Model Builder uses a Metropolis–Hastings algorithm, which samples from a scaled multivariate normal distribution with variances and covariances proportional to the asymptotic variance–covariance matrix. We ran the MCMC chain for each model for 5 000 000 steps and saved values from every 100th step. In some cases, the models did not converge to a stable mixing distribution for at least 1 500 000 steps. Therefore, we used a burn-in period of 2 000 000 steps, which reduces the effect of starting values on the MCMC estimates (Gelman et al., 2004).

We used visual inspection to evaluate convergence of about 10% of the models fitted. From this subset, models that did not appear to have converged, as indicated by large differences in the means and variances of the posterior distribution among subsections of the MCMC chain, had low effective sample sizes of estimated biomass and fishing mortality in the last year (<250). Effective sample size was calculated according to the method of Thiebauz and Zwiers (1984), where effective sample size is a function of the number of saved MCMC steps minus the burn-in and the autocorrelation of values of the MCMC chain. Therefore, we considered results of a model adequate for use in the study if the effective sample size of estimated biomass in the last year was greater than 250. We used the effective sample size for biomass in the last year as our indicator for model adequacy because it combines information on many of the estimated parameters and often had the lowest effective sample size among parameters and other estimated quantities such as F or the value of the negative log likelihood.

We only evaluated performance of models for data sets where all the models passed our effective sample size criterion in order to avoid potential biases that could arise by some models consistently failing our criterion for difficult data sets and others simply producing poor estimates. Also, comparisons between model selection and model averaging may not be comparable if the number of models used for the average changes among data sets.

2.4. Model selection

DIC, like other information-theoretic model selection criteria, trades off a measure of model fit (average deviance) and a measure of model complexity (effective number of parameters; Spiegelhalter et al., 2002).

$$\mathrm{DIC} = \bar{D} + p_D \tag{1}$$

The average deviance, $\bar{D}$, for model j is an estimate of model adequacy and is calculated as

$$\bar{D} = \frac{1}{C}\sum_{c=1}^{C} -2\ \log_e p(\mathrm{data}|\theta_c) \tag{2}$$

where C is the number of MCMC steps saved minus the burn-in, and $\log_e p(\mathrm{data}|\theta_c)$ is the natural logarithm of the likelihood function (Spiegelhalter et al., 2002). Like with AIC and BIC, smaller DIC values indicate better models. The effective number of parameters is the difference between the average deviance and the deviance evaluated at the posterior mean parameter estimates,

$$p_D = \bar{D} - D(\bar{\theta}). \tag{3}$$

The model with the lowest DIC was considered the best model when we chose among estimation models applied to each simulated data set.

2.5. Model probabilities and model averaging

In many cases, model averaging provides superior predictive performance over using a single model selected by a model selection criterion because estimates from a single model ignore uncertainty in model selection (i.e., model selection uncertainty; Hoeting et al., 1999; Burnham and Anderson, 2002, 2004 and references therein). Bayesian theory would suggest averaging models based upon Bayes factors, but there are conceptual and practical difficulties, including difficulties in obtaining reliable estimates of Bayes factors from MCMC samples (Kass and Raftery, 1995) and concerns that Bayes factors may not consider the appropriate focus of estimation for many problems (Berg et al., 2004; van der Linde, 2005). Therefore, we calculated model averaged estimates of biomass and fishing mortality in the last year, weighting estimates from different models by weights derived from DIC differences (by adapting the method of Burnham and Anderson (2002) for AIC). These DIC-based weights were estimated by rescaled DIC differences among models,

$$P(M_a) = \frac{e^{-(\Delta_a/2)}}{\sum_b e^{-(\Delta_b/2)}}, \tag{4}$$

where $P(M_a)$ is the weight ("probability") for model a and Δ is the DIC difference between model a and the best model (for the best model Δ equals zero). This procedure rescales the DIC differences from the log scale to the normal scale. After Burnham and Anderson (2002) for AIC, we view these weights as probabilities that a given model would be best, among the candidate models, at predicting new data generated from the same process that produced the original data. However, this method of model averaging is ad hoc and does not produce equivalent posterior model probabilities as Bayes factors (Spiegelhalter et al., 2002). In addition to using these model probabilities in model averaging, we also present frequency distributions of them to provide insight into how definitive the DIC evidence was. The relative values of these model probabilities is directly related to differences in DIC values, with DIC differences of 2.0 and 4.0 corresponding to the model with the lower DIC being 2.7 and 8.7 times as probable as the alternative model, respectively. Burnham and Anderson (2002) proposed that models with AIC within 2.0 of the best model have substantial support, and models with AIC differences greater than 4.0 from the best model have considerably less support. Spiegelhalter et al. (2002) suggest that such rules of thumb work reasonably well for DIC.

2.6. Evaluation of estimation model performance

We determined how often the correct structural model was selected, even though there was not a truly correct model in the scenario with a linear increase in catchability or in the uninformative effort scenario. In the white noise case, the white noise estimation model was correct. The random walk model was considered the correct model in the linear increase scenario because it tended to perform better than other models in this scenario (Wilberg and Bence, 2006) and because it is designed to allow for gradual changes in catchability over time. In the case with uninformative effort data, the model that ignored fishery effort data was considered the correct model.

In stock assessments, estimated quantities in the last year are often most important for forecasting and management, and point estimates are almost universally used in managing fisheries. Therefore, we evaluated selection method performance by calculating the relative error (RE) of estimated biomass and average fishing mortality (for ages 4–8) in the last year.

$$\mathrm{RE} = \frac{\text{estimated} - \text{true}}{\text{true}} \times 100 \quad (5)$$

In Bayesian analyses, the result is a posterior distribution, and the mean of the posterior distribution is commonly used as a point estimate of the quantity of interest. We evaluated performance of models using the mean and the median of the posterior distribution to determine which provided a more accurate point estimate. We evaluated performance of DIC model selection, model averaging, and only using a single model with root mean square relative error (RMSE), which summarizes the variance and bias of model estimates. Although bias is essentially a frequentist concept, we believe it is a useful quantity to consider given the wide usage of point estimates in fisheries management.

3. Results

All three estimation models achieved our minimum effective sample size for 82% of the white noise data sets, 89% of the linear increase data sets, and 80% of the uninformative effort data sets. All three models failed our sample size diagnostic in 5% of cases, two models failed in 4%, and one model failed in 7% of cases. Most of the MCMC chains appeared to have converged to their stable mixing distribution within 10 000 steps. However, in several cases, the MCMC routine required nearly 2 000 000 steps as a burn-in period to reach a stable mixing distribution, and some chains did not seem to reach a stable mixing distribution within 5 000 000 steps.

Estimates of the effective number of parameters, p_D, were typically less than the actual number of estimated parameters, 52. The effective number of parameters for the estimation model with random walk catchability was the lowest with a mean of 47.1 and a range of 45.3–47.9. The estimation model with white noise catchability had the second fewest effective parameters with a mean of 48.4 (range 46.4–49.2). The estimation model that freely estimated fishing mortality for each year had the most effective parameters with a mean of 51.6 (range 48.6–52.4), which was quite close to the true number of estimated parameters. Some model fits had unrealistically low (e.g., 20) or negative estimates of the number of effective parameters. All of these models had extremely low effective sample sizes for several quantities (usually less than 250), and unrealistic estimates of the number of effective parameters seemed to indicate poor estimation performance. Only one model fit had an unrealistically low effective number of parameters and an effective sample size for biomass greater than 250; we excluded this data set from further consideration because of an unbelievable estimate of effective number of parameters.

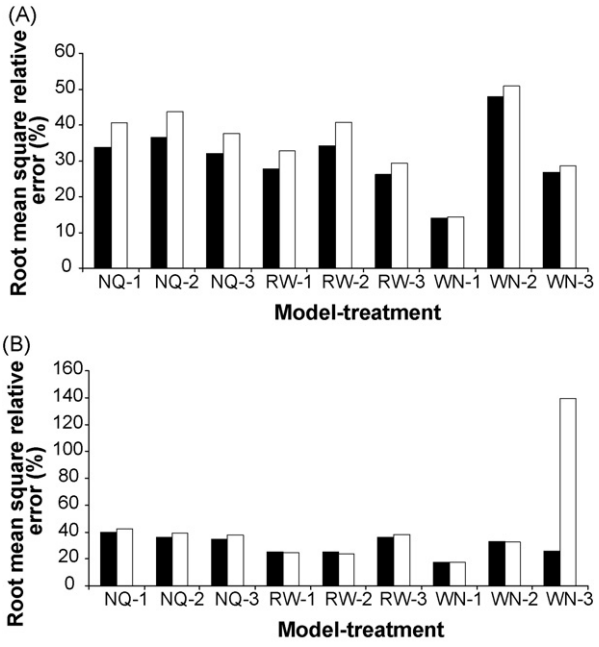

Fig. 3. Root mean square error of point estimates of biomass (A) and fishing mortality (B) in the last year of the simulation. Solid bars represent performance of the median of the posterior distribution as a point estimate and open bars represent the mean. NQ (no catchability), RW (random walk), and WN (white noise) represent estimation models and data-generating models are 1 (white noise catchability), 2 (linear increase in catchability), and 3 (uninformative effort).

The median of the posterior distribution usually provided more accurate point estimates of biomass and fishing mortality rate in the last year than the mean (Fig. 3). On average, the RMSE of the median was 4.4% lower than the mean for biomass and 13.6% lower for fishing mortality rate. The large difference in performance between the mean and median for the white noise model in the uninformative effort scenario was caused by a single bad estimate. If this case was removed, the median still had a slightly lower RMSE for fishing mortality (1.1% difference) than the mean. All subsequent results are based on medians as point estimates.

DIC usually selected the correct estimation model in each scenario. For data sets that were generated with white noise catchability, the white noise estimation model was selected 72% of the time. In the scenario with a linear increase in catchability, the random walk model was selected 99% of the time. In the uninformative effort scenario, the estimation model that ignored fishery effort was selected 88% of the time. However, model probabilities, based on DIC differences provided equivocal evidence in favor of the best model ($P(M_a) < 0.9$) in 100% of the white noise and about 75% of the linear increase in catchability scenarios (Fig. 4). In contrast, the estimation model with no catchability was more often strongly supported ($P(M_a) > 0.9$ 55% of the time) when effort data were uninformative (Fig. 4).

Using DIC and model averaging tended to provide relatively unbiased estimates (Fig. 5) and had similar RMSEs to the correct model (Fig. 6). Differences among model RMSEs varied among data-generating scenarios. The largest difference among model performance occurred in the white noise data-generating scenario, and the least difference among estimation model performance occurred in the scenario with uninformative effort. Somewhat surprisingly, when effort was uninformative, the estimation model that best represented the data generation process did not perform best in terms of RMSE. This appears to be a case where estimating additional parameters does not provide bet-

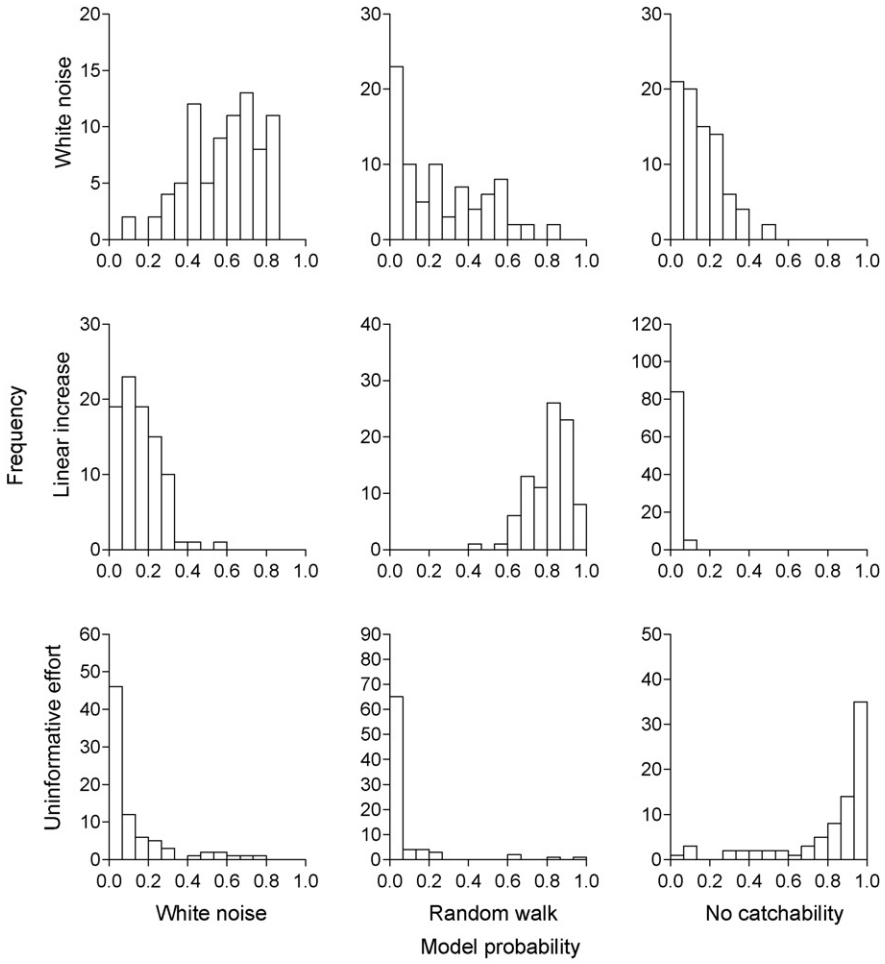

Fig. 4. Histograms of the model probabilities (see Eq. (4) for definition). Rows represent alternative data-generating scenarios, and columns represent estimation models. A probability of near 1.0 indicates strong evidence that a model is the best in the set of fitted models.

ter estimates, even though this leads to a more correct model specification.

4. Discussion

This study indicates that DIC can be used to choose models that are structurally appropriate given the data-generating model and provided performance characteristics similar to knowing the best model in advance. Using DIC for model selection and model averaging produced relatively unbiased and accurate point estimates of biomass and fishing mortality in the last year. While our results show promise for use of DIC in the stock assessment arena, we strongly suspect that prior information on the nature of the correct (or a robust) model often will provide large benefits. For example, Prager (2002) evaluated the procedure of selecting between logistic and generalized production models based on likelihood ratio tests when the true value of shape parameter used in the generating model was close to that for a logistic model. With realistic levels of observation error the model selection procedure performed poorly relative to just using the logistic model, because the logistic model was rejected only in cases where the estimated shape parameter of the generalized model was far from logistic. We agree with Prager (2002), that information-theoretic criteria for model selection would likely perform in a similar fashion. Prager (2002) discussed an alternative to assuming the logistic model, of using the generalized model with an informative prior for the shape parameter. Within the context of Bayesian analysis, another way to address this problem would be to formally assign prior model probabilities to the alternative models, and use Bayesian model averaging based on Bayes factors (e.g., McAllister and Kirchner, 2002). In contrast with use of Bayes factors, the model averaging procedure we used implies lower prior probability for more complex models because of the penalty imposed by the effective number of parameters, similar to a Bayesian interpretation of AIC (Burnham and Anderson, 2004).

In this study we made a number of simplifying assumptions, which included assuming M, variances, and effective sample sizes were known, constant, and correctly specified. These assumptions were pragmatic choices and are common to many simulation studies evaluating assessment methods (e.g., Yin and Sampson, 2004). In real assessments these quantities are estimates, and their true values are unlikely to be truly constant. Ideally, uncertainty in such quantities and the possibility that values vary would be acknowledged and accounted for in the real assessment methods and range of candidate models. The effect of these simplifying assumptions for our simulation results will be to generally overstate the accuracy of assessment results. While we have no specific reasons to suspect that these simplifying assumptions distorted our conclusions about the utility of DIC as a model selection method, we cannot rule out this possibility. We emphasize that DIC would not necessarily be useful in a real assessment situation if all the candidate models had severe violations of their assumptions.

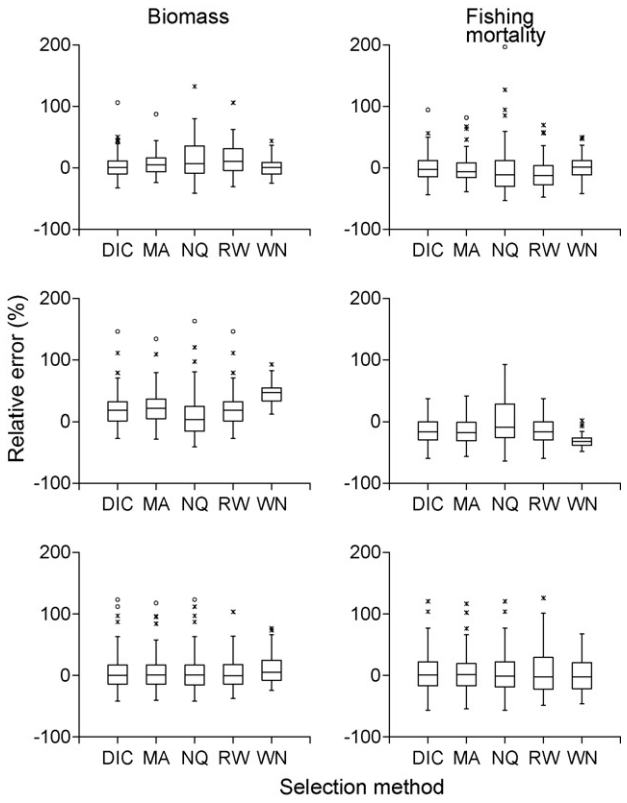

Fig. 5. Box plots of relative error of point estimates (medians) of biomass and average fishing mortality in the last year. The middle line indicates the median, the box indicates the interquartile range, and the whiskers indicate the range of observed values between the 25th or 75th percentile and 1.5 times the interquartile range; stars and circles indicate points outside 1.5 and 3.0 times the interquartile range, respectively. Estimation methods are indicated by DIC for deviance information criterion, MA for model average, NQ for the estimation model that did not estimate catchability, RW for the estimation model with random walk catchability, and WN for the estimation model with white noise catchability. The top row of panels displays performance under the white noise catchability data-generating model, the middle row indicates the linear increase scenario, and the bottom row indicates the scenario with uninformative effort.

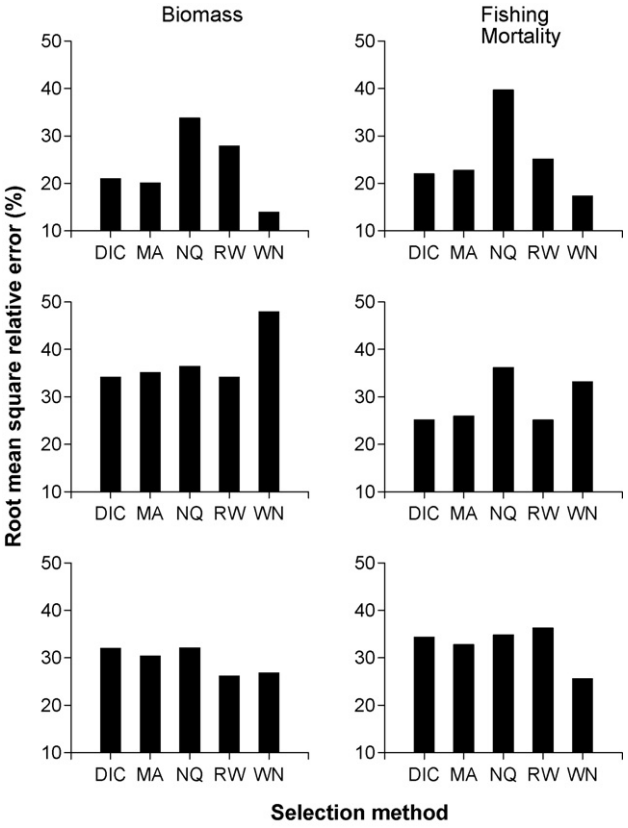

Fig. 6. Root mean square error of point estimates (medians) of biomass and fishing mortality in the last year of the simulation. Estimation methods are indicated by DIC for deviance information criterion, MA for model average, NQ for the estimation model that did not estimate catchability, RW for the estimation model with random walk catchability, and WN for the estimation model with white noise catchability. The top row of panels display performance under the white noise catchability data-generating model, the middle row indicates the linear increase scenario, and the bottom row indicates the scenario with uninformative effort.

Outliers can have a large influence on parameter estimates in stock assessments (Chen et al., 2000) and on the apparent evidence in favor of alternative models. Prager (2002) found that fitting of swordfish data with a generalized production model by nonlinear least squares suggested a shape parameter divergent from a logistic model; after outliers were removed, the 80% CI for the shape parameter overlapped a logistic shape. Furthermore, in this case the two models suggested qualitatively different stock status and management parameters. We suspect that the presence of non-informative outliers could generally pose a problem for model selection procedures. Removing different data points for use by different models, after initial fitting, as done by Prager (2002) is not compatible with model selection by DIC or related information criteria, although DIC could be used if one adopted a procedure of removing the same data points from use in all models. Alternatively, outliers could be treated using robust estimation methods that presume that data distributions are mixtures of informing and contaminant distributions (Chen and Fournier, 1999). However, theoretical justifications and most simulation validations of DIC, like ours, have assumed exponential family distributions. Performance of DIC when faced with alternatives, such as mixture distributions, has shown mixed results and is a continuing focus of research (van der Linde, 2005; Celeux et al., 2006).

For the conditions of our study, although DIC-based model averaging did not appear to provide more accurate point estimates than just selecting the best model, there was also no obvious penalty associated with the averaging. Model averaging did not produce less biased estimates as a rule in our study, but differences in performance between the best model and the model average were relatively small. In general, Burnham and Anderson (2002) found that model averaged estimates (based on AIC) were less biased than simply using estimates from the best model after model selection. A lack of benefits of model averaging in our study may be because the best models were the same as or quite similar to the data-generating models. Also, when evidence is strongly in favor of a single model, there is little difference between estimates from the best model and model averaged estimates, as in our scenario with uninformative effort. However, in real world applications it is unlikely that the estimation models will be as similar to the data-generating reality as in this study, although differences among model estimates can still be small because of relatively uninformative data (e.g., Wade et al., 2007). Model average estimates might sometimes provide a large increase in performance for point estimates in assessments, unlike what we saw in this study, especially in cases where alternative models suggest different conclusions about stock dynamics (e.g., McAllister and Kirchner, 2002). Often the primary benefit of model averaging has been to more accurately represent uncertainty in model results (Hoeting et al., 1999;

Burnham and Anderson, 2004), an aspect of model performance we did not evaluate. Model averaging based on DIC remains intuitively appealing, although its ad hoc nature argues for much more research on the performance of the method (Spiegelhalter et al., 2002).

The choice of a model selection criterion should depend on the objectives of the study and the "focus of prediction" (Celeux et al., 2006). DIC is a general model selection criterion, and was developed to perform well in prediction of future data sets generated by the same mechanism (Spiegelhalter et al., 2002; van der Linde, 2005). For many stock assessment applications, predictions of the most recent years may be of more importance for management than earlier years, but DIC, like AIC, BIC, and Bayes factors, treats each year's data the same. For applications where the most recent year's estimates are of primary importance, model selection methods that weight these most recent years may be preferable to other standard methods, but would need to be developed for the specific objective. For this reason, we chose to evaluate performance of DIC by comparing the accuracy of estimates in the last year. This study is one of the first to evaluate the performance of DIC in models where the purpose is to predict such unobserved quantities

Results from multiple stock assessment models are usually presented to decision makers as a form of sensitivity analysis (Punt and Hilborn, 1997; Patterson, 1999), and several interest groups may suggest or present alternative models of stock dynamics (McAllister and Kirchner, 2002). Yet, decision makers are usually not provided with quantitative rankings of how much better the best model is than alternative models. Our results suggest that DIC is potentially useful to provide quantitative advice for ranking, comparing, and integrating results of alternative models. However, DIC model comparisons are limited in the types of models that can be compared because the models must use the same dependent variables (Spiegelhalter et al., 2002). These data would include catch and CPUE time series and age compositions of the time series.

Certainly DIC should not be the only tool used for model selection. Factors such as model plausibility, sensitivity, examination of residual patterns, and retrospective bias should also be considered when choosing among models (NRC, 1998). However, DIC provides an objective metric to compare among alternative assessment models, especially in those that differ in their hierarchical structure or random effects, and does show promise for helping select among stock assessment models even when models are quite similar. Additionally, because DIC was able to select the correct structure of the underlying model, it could be a useful tool for estimating weights for alternative models of system dynamics for use in management strategy evaluation (Smith et al., 1999; Sainsbury et al., 2000; Rademeyer et al., 2007).

Acknowledgements

We thank Dan Hayes, Mike Jones, Rob Tempelman, Genny Nesslage, Mike Prager, and an anonymous reviewer for providing helpful comments that improved this manuscript. This work was supported in part by Michigan Sea Grant College Program Project R/GLF-50 under NA76RG0133 and NA76RG1145 from National Sea Grant, NOAA, U.S. Department of Commerce and from funds from the State of Michigan, Michigan Department of Natural Resources Fisheries Division Studies 236102 and 230713 with partial funding under the Great Lakes Fish and Wildlife Restoration Act administered by the U.S. Fish and Wildlife Service (F-80-R), the Great Lakes Fishery Commission, the International Association for Great Lakes Research, and Michigan State University. This is publication 2008–11 of the Quantitative Fishery Center at Michigan State University and Contrib. No. 4191 of the University of Maryland Center for Environmental Science Chesapeake Biological Laboratory.

Appendix A

The MCMC algorithm was very sensitive to parameter correlations greater than about 0.8. Under these conditions, the MCMC algorithm mixed very poorly and produced very "sticky" MCMC chains (i.e., chains with high autocorrelation). Therefore, we reparameterized aspects of the models to reduce these correlations. All parameters described below were estimated on the log scale. Two groups of parameters were highly correlated within each group: parameters that determined overall scale of population size, and selectivity parameters for the fishery and survey. Parameters that determine the overall scale of the population size included mean recruitment, mean abundance-at-age in year 1, fishery catchability (or mean F in the model that ignored fishery effort data), and survey catchability. In order to minimize correlation among these parameters, we parameterized the model by estimating the $\log_e$ of mean recruitment and a deviation from this for each of these other "scale-setting" parameters. The other parameters that had high correlations were selectivity at age for the fishery and the survey. To reduce these correlations, the models were parameterized to estimate deviations from a mean $\log_e$ selectivity that was forced to equal zero. This constraint serves to make the selectivity parameters identifiable and not confounded with the associated catchability (for fishery or survey), in the same way that the more usual approach of setting selectivity to 1.0 for a fully selected age (e.g., Fournier and Archibald, 1982) does.

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary, pp. 81–267.

Barry, S.C., Brooks, S.P., Catchpole, E.A., Morgan, B.J.T., 2003. The analysis of ring recovery data using random effects. Biometrics 59, 54–65.

Berg, A., Meyer, R., Yu, J., 2004. Deviance information criterion for comparing stochastic volatility models. J. Business Econ. Stat. 22, 107–120.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociol. Methods Res. 33, 261–304.

Cardoso, F.F., Tempelman, R.J., 2003. Bayesian inference on genetic merit under uncertain paternity. Genet. Sel. Evol. 35, 469–487.

Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M., 2006. Deviance information criteria for missing data models. Bayesian Anal. 4, 651–674.

Chen, Y., Breen, P.A., Andrew, N.L., 2000. Impacts of outliers and mis-specification of priors on Bayesian fisheries-stock assessment. Can. J. Fish. Aquat. Sci. 57, 2293–2305.

Chen, Y., Fournier, D., 1999. Impacts of atypical data on Bayesian inference and robust Bayesian approach in fisheries. Can. J. Fish. Aquat. Sci. 56, 1525–1533.

Crone, P.R., Sampson, D.B., 1998. Evaluation of assumed error structure in stock assessment models that use sample estimates of age composition. In: Funk, F., Quinn II, T.J., Heifitz, J., Ianelli, J.N., Powers, J.E., Schweigert, J.F., Sullivan, P.J., Zhang, C.I. (Eds.), Fishery Stock Assessment Models. Alaska Sea Grant College Program Report No. AK-SG-98-01. University of Alaska Fairbanks.

Deriso, R.B., Quinn II, T.J., Neal, P.R., 1985. Catch-age analysis with auxiliary information. Can. J. Fish. Aquat. Sci. 42, 815–824.

Fournier, D., Archibald, C.P., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39, 1195–1207.

Francis, R.I.C.C., Hurst, R.J., Renwick, J.A., 2003. Quantifying annual variation in catchability for commercial and research fishing. Fish. Bull. 101, 293–304.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis, 2nd edition. Chapman and Hall, Boca Raton.

Harley, S.J., Myers, R.A., Dunn, A., 2001. Is catch-per-unit-effort proportional to abundance? Can. J. Fish. Aquat. Sci. 58, 1760–1772.

Helu, S.L., Sampson, D.B., Yin, Y., 2000. Application of statistical model selection criteria to the Stock Synthesis assessment program. Can. J. Fish. Aquat. Sci. 57, 1784–1793.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 4, 382–417.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.

Kizilkaya, K., Tempelman, R.J., 2003. Cumulative *t*-link threshold models for genetic analysis of calving ease scores. Genet. Sel. Evol. 35, 489–512.
Kizilkaya, K., Tempelman, R.J., 2005. A general approach to mixed effects modeling of residual variances in generalized linear mixed models. Genet. Sel. Evol. 37, 31–56.
McAllister, M., Kirchner, C., 2002. Accounting for structural uncertainty to facilitate precautionary fishery management: illustration with Namibian orange roughy. Bull. Mar. Sci. 70, 499–540.
Methot, R.D., 1990. Synthesis Model: an adaptable framework for analysis of diverse stock assessment data. In: Low, L. (Ed.). Proceedings of the Symposium on Application of Stock Assessment Techniques to Gadids. International North Pacific Fisheries Commission Bulletin 50, pp. 259–277.
Otter Research Limited, 2000. An Introduction to AD Model Builder Version 4 for Use in Nonlinear Modeling and Statistics. Otter Research Ltd., Sidney, BC.
National Research Council (NRC), 1998. Improving Fish Stock Assessments. National Academy Press, Washington, DC.
Patterson, K.R., 1999. Evaluating uncertainty in harvest control law catches using Bayesian Markov chain Monte Carlo virtual population analysis with adaptive rejection sampling and including structural uncertainty. Can. J. Fish. Aquat. Sci. 56, 208–221.
Prager, M.H., 2002. Comparison of logistic and generalized surplus-production models applied to swordfish, *Xiphias gladius*, in the north Atlantic Ocean. Fish. Res. 58, 41–57.
Punt, A.E., Hilborn, R., 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Rev. Fish. Biol. Fish. 7, 35–63.
Quinn II, T.J., Deriso, R.B., 1999. Quantitative Fish Dynamics. Oxford University Press, New York.
Rademeyer, R.A., Plagányi, É.A., Butterworth, D.S., 2007. Tips and tricks in designing management procedures. ICES J. Mar. Sci. 64, 618–625.
Sainsbury, K.J., Punt, A.E., Smith, A.D.M., 2000. Design of operational management strategies for achieving fishery ecosystem objectives. ICES J. Mar. Sci. 57, 731–741.
Schwartz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Smith, A.D.M., Sainsbury, K.J., Stevens, R.A., 1999. Implementing effective fisheries-management systems—management strategy evaluation and the Australian partnership approach. ICES J. Mar. Sci. 56, 967–979.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B 64, 583–639.
Thiebauz, H.J., Zwiers, F.W., 1984. The interpretation and estimation of effective sample size. J. Climate Appl. Meteorol. 23, 800–811.
van der Linde, A., 2005. DIC in variable selection. Statistica Neerlandica 59, 45–56.
Wade, P.R., Watters, G.M., Gerodette, T., Reilly, S.B., 2007. Depletion of spotted and spinner dolphins in the eastern tropical Pacific: modeling hypotheses for their lack of recovery. Mar. Ecol. Prog. Ser. 343, 1–14.
Ward, E.J., 2008. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tool. Ecol. Model. 211, 1–10.
Wilberg, M.J., Bence, J.R., 2006. Performance of time-varying catchability estimators in statistical catch-at-age analysis. Can. J. Fish. Aquat. Sci. 63, 2275–2285.
Yin, Y., Sampson, D.B., 2004. Bias and precision of estimates from an age-structured stock assessment program in relation to stock and data characteristics. N. Am. J. Fish. Manage. 24, 865–879.