FISH and FISHERIES

Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC

Arni Magnusson[1,2], *André E Punt*[1] & *Ray Hilborn*[1]

[1]School of Aquatic and Fishery Sciences, University of Washington, PO Box 35520, Seattle, WA 98195, USA; [2]Marine Research Institute, Skulagata 4, PO Box 1390, 121 Reykjavik, Iceland

Abstract

Fisheries management depends on reliable quantification of uncertainty for decision-making. We evaluate which uncertainty method can be expected to perform best for fisheries stock assessment. The method should generate confidence intervals that are neither too narrow nor too wide, in order to cover the true value of estimated quantities with a probability matching the claimed confidence level. This simulation study compares the performance of the delta method, the bootstrap, and Markov chain Monte Carlo (MCMC). A statistical catch-at-age model is fitted to 1000 simulated datasets, with varying recruitment and observation noise. Six reference points are estimated, and confidence intervals are constructed across a range of significance levels. Overall, the delta method and MCMC performed considerably better than the bootstrap, and MCMC was the most reliable method in terms of worst-case performance, for our relatively data-rich scenario and catch-at-age model, which was not subject to substantial model misspecification. All three methods generated too narrow confidence intervals, underestimating the true uncertainty. Bias correction improved the bootstrap performance, but not enough to match the performance of the delta method and MCMC. We recommend using MCMC as the default method for quantifying uncertainty in fisheries stock assessment, although the delta method is the fastest to apply, and the bootstrap is useful to diagnose estimator bias.

Keywords Bias correction, bootstrap, delta method, MCMC, stock assessment, uncertainty

Correspondence:
Arni Magnusson, Marine Research Institute, Skulagata 4, PO Box 1390, 121 Reykjavik, Iceland
Tel: +354 575 2000
Fax: +354 575 2001
E-mail: arnima@hafro.is

Received 21 Aug 2011
Accepted 23 Mar 2012

DOI: 10.1111/j.1467-2979.2012.00473.x

Introduction

Which method performs best?

Fisheries management relies not only on point estimates of key quantities, such as biomass and harvest rate, but also on the uncertainty about these estimates. The uncertainty can be used to convey likely outcomes resulting from different management decisions, or incorporated into management strategy evaluation to find a long-term harvest strategy that performs well in face of uncertainty.

When estimating measurement uncertainty, fisheries scientists generally choose the statistical method they are most familiar with, or one that has become traditional for a particular stock. Three commonly used methods that will be evaluated here are the delta method, the bootstrap, and Markov chain Monte Carlo (MCMC) simulation. These methods have been shown to perform well with simple models, when all assumptions are met (Oehlert 1992; Efron and Tibshirani 1993; Gelman *et al.* 2004). In this study, we ask the question: given a typical age-structured stock assessment model and simulated datasets, which method performs best?

Patterson *et al.* (2001) provide a thorough review of uncertainty methods and describe three paradigms for evaluating uncertainty in stock assessment: frequentist, likelihood, and Bayesian inference. For the purposes of fisheries stock assessment, the theoretical difference between these paradigms is often ignored in practice (Restrepo *et al.* 2000; Patterson *et al.* 2001; Gavaris and Ianelli 2002; Hilborn 2003), and the methods are all used to express the plausible range of estimated quantities. In the strict frequentist sense, a confidence interval is a probabilistic statement about the proportion of such intervals that would cover the true parameter value in repeated experiments (Neyman 1937; Casella and Berger 2002). This frequentist statement treats the interval limits as random and the parameter as fixed, in the context of repeated experimental trials, and is therefore quite meaningful in a simulation study like this one, but it does not directly answer the relevant questions for environmental decision-making (Ellison 1996; Punt and Hilborn 1997; Ascough *et al.* 2008). Bayesian inference, on the other hand, treats the interval limits as fixed and the parameter as random, leading to an intuitive statement about the probability that the true parameter value lies in the interval. The Bayesian interval is sometimes called a 'credible interval' (Casella and Berger 2002), a 'posterior interval' (Gelman *et al.* 2004), or simply a 'confidence interval' (Hilborn and Mangel 1997; Clark

2005) when the theoretical difference is considered of secondary importance, as is the case in this study.

For the purposes of this study, an uncertainty method is considered to perform well when it generates x% confidence intervals for estimated quantities that contain the true value approximately x% of the time. The method should generate neither too narrow intervals that underestimate uncertainty, nor too wide intervals that overestimate uncertainty.

The delta method was introduced by Cramér (1946) and popularized in ecological modelling by Seber (1973). Most applications of the delta method in stock assessment (e.g. Booth and Quinn 2006; Trzcinski *et al.* 2006; McGarvey *et al.* 2007) use the AD Model Builder programming framework to automate the computation of the required partial derivatives (Schnute *et al.* 1998; Fournier *et al.* 2012). The bootstrap was introduced by Efron (1979) and popularized by Efron and Tibshirani (1993). Early applications of the bootstrap in stock assessment include Mohn (1993) and Punt and Butterworth (1993). Variations of the bootstrap are outlined by Patterson *et al.* (2001), citing Gavaris and Van Eeckhaute (1998) as the current recommended bootstrap method for stock assessment. MCMC simulation of probability distributions was introduced by Metropolis *et al.* (1953) and Hastings (1970), and popularized in fisheries circles by Gelman *et al.* (1995). The potential usefulness of MCMC in stock assessment was described by McAllister and Ianelli (1997) and Punt and Hilborn (1997), with early applications including Punt and Kennedy (1997), Virtala *et al.* (1998), and Patterson (1999).

Patterson *et al.* (2001) list five desirable properties of methods quantifying uncertainty. They should be (i) based on statistical distributions derived from data rather than arbitrarily chosen distributions, (ii) unbiased, (iii) accurate, (iv) use few distributional assumptions and be robust to misspecifications of such assumptions, and least importantly (v) easy to understand and implement. They mention that the bootstrap and MCMC have become more common than the delta method in fisheries stock assessment to avoid restrictive distributional assumptions. Hilborn (2003) noted that the use of the bootstrap has faded in recent years, as Bayesian methods have grown in popularity, because of their intuitive probability statements and theoretical and technical progress in this field of computational statistics. The bootstrap has been described as an automatic processor for frequentist inference, with MCMC as its Bayesian counterpart (Efron 2000).

Previous comparison studies

There are mainly two approaches to compare the performance of uncertainty methods, either using real stock assessment data or using simulated data. With real data, one can compare the estimated uncertainty for each method and speculate why differences occur. With simulated data, one knows the true value of the estimated quantities and can therefore quantitatively judge the performance of each method. A simulation study can use a relatively complex operating model to generate the simulated datasets and a simpler assessment model to fit those datasets, or use the same model to violate fewer assumptions.

Mohn (1993) compared the delta method and bootstrap, fitting an age-structured model to actual cod data. Retrospective analysis was used to approximate the true estimated values, showing that the delta method tended to underestimate uncertainty. Gavaris (1999) also compared the delta method and the bootstrap, fitting an age-structured model to haddock data. The bootstrap distribution indicated skewed uncertainty about stock abundance, implying that the delta method with a symmetric Gaussian distribution would be inappropriate for statistical inference. Patterson (1999) compared the bootstrap and MCMC, fitting an age-structured model to herring data and noted that MCMC generated wider confidence intervals than the bootstrap. Gavaris *et al.* (2000) compared the delta method, the bootstrap, and MCMC and analyzing data from three stocks using two age-structured models. The uncertainty methods gave somewhat different results, but no clear or consistent trends emerged. Booth and Quinn (2006) compared the delta method and MCMC, fitting a simple age-structured model to monkfish data. The two methods gave similar results when non-informative Bayesian priors were used for MCMC, and the study highlighted how prior information can be incorporated to decrease uncertainty when using MCMC. Mohn (2009) compared the delta method, bootstrap, and MCMC, fitting an age-structured model to cod data. The bootstrap generated considerably wider confidence intervals than the delta method and MCMC, and the author pointed out that the bootstrap might be overestimating measurement uncertainty.

Fewer studies have used simulated data to compare the performance of uncertainty methods. Punt and Butterworth (1993) compared the delta method and the bootstrap, using an age-structured operating model and a simpler biomass-dynamic assessment model. The methods worked equally well, as long as some bootstrap pitfalls were avoided. Restrepo *et al.* (2000) compared the delta method, bootstrap, and MCMC, fitting age-structured assessment models to a simulated dataset. The delta method and bootstrap performed marginally better than MCMC in their study, and bias-correction methods proved beneficial.

This study

Overall, previous comparison studies have not identified which uncertainty method performs best. They have highlighted the strengths and weaknesses of each method and provided useful recommendations regarding their implementation. This study revisits the question with previous recommendations in mind, using a modern statistical catch-at-age model both to simulate and to analyze data that are known to be informative (Magnusson and Hilborn 2007). The study also benefits from greater computing power than was available a decade ago, allowing a more rigorous experimental design that involves a larger number of simulated datasets and population trajectories.

The working hypothesis is that all three methods work perfectly, for example, that 90% confidence intervals for a reference point contain the true value 90% of the time. This hypothesis is not going to be accepted or rejected, but the delta method, bootstrap, and MCMC will be rated in terms of how accurate the probabilistic statement is.

Methods

First, we define a set of true population parameters and generate stochastic datasets, using an operating model based on age-structured population dynamics. The performance of three uncertainty methods is then evaluated, with respect to how accurately they report the uncertainty about reference points. The simulation procedure (Fig. 1) is repeated 1000 times, using 10 different recruitment scenarios so the results do not depend on a particular population trajectory. The operating model first outputs the resulting reference point values, and then applies random observation noise to the assessment data

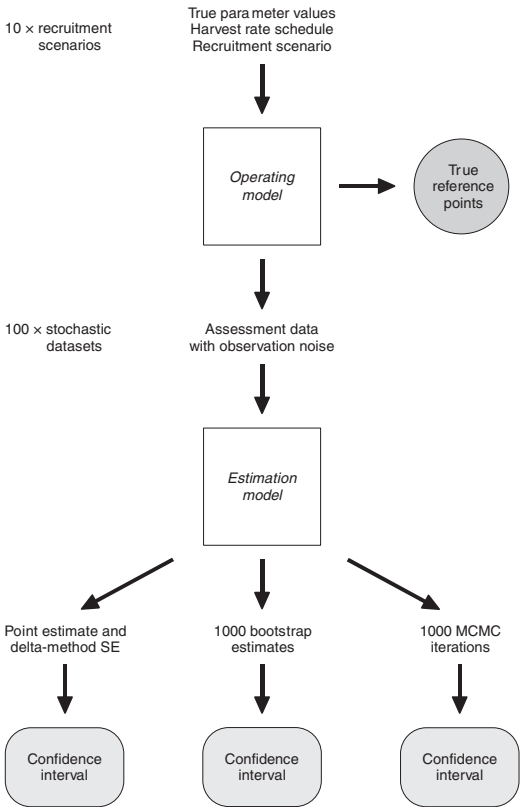

Figure 1 The simulation procedure. Arrows indicate the process for a single run, and replications indicate how the study consists of multiple runs.

that are used as input for the estimation model. Finally, the confidence interval for each reference point is evaluated using the delta method, bootstrap, and MCMC, and compared with the 'true' reference points.

Operating model

The operating model is age-structured and follows the parametrization of the Coleraine generalized population model (Hilborn *et al.* 2003). The population dynamics of this operating model are described in detail by Magnusson and Hilborn (2007). There are 10 age classes, including a plus group, and 20 years of data, nominally referred to as 1985–2004, and the biology and fishery characteristics (see Supporting Information, Figure S1, Tables S1 and S2) are based on Atlantic cod (*Gadus morhua,* Gadidae).

Each dataset includes landings, a survey abundance index, commercial catch at age and survey

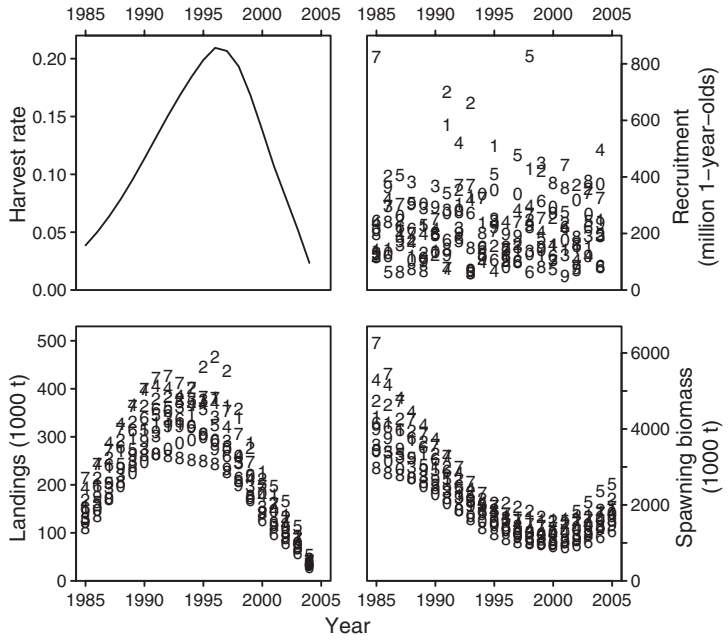

Figure 2 Harvest rate, recruitment, landings, and biomass in the operating model. The plotting symbols identify recruitment scenarios 1–10.

catch at age. The landings are assumed to be known exactly, but the commercial and survey catch-at-age data and the abundance index are subject to random observation error. The datasets are based on 10 recruitment scenarios that are generated randomly (Table S3), and within each scenario there are 100 stochastic datasets with different realizations of observation noise. The level of recruitment variability (lognormal $\sigma_R = 0.6$), observation noise for the abundance index (lognormal $\sigma_I = 0.2$), and observation noise for the commercial (multinomial ${}_Cn = 50$) and survey catch-at-age (multinomial ${}_Sn = 50$) are similar to those used in assessments of Icelandic cod (ICES 2003). All the scenarios follow the same harvest rate schedule, but the recruitment pattern leads to 10 different landings and biomass trajectories (Fig. 2).

The survey abundance index is proportional to the biomass vulnerable to the survey in the middle of the fishing year,

$$I_t = q \sum_a {}_S S_a N_{t,a} w_a e^{-M/2} \times \exp({}_I\varepsilon_t), \qquad (1)$$

where I_t is the observed abundance index at time t, q is the catchability coefficient, ${}_S S_a$ is survey selectivity at age a, $N_{t,a}$ is population size, w_a is body weight, M is the natural mortality rate, and ${}_I\varepsilon_t \sim N(0, \sigma_I)$ is observation noise. The commercial catch-at-age data are provided to the assessment model in the form of proportions at age. These proportions are generated assuming that sampling is multinomial,

$${}_C P_{t,a} \sim \text{Multinom}\left({}_C n, \frac{{}_C S_a N_{t,a}}{\sum_a {}_C S_a N_{t,a}}\right) \Big/ {}_C n, \qquad (2)$$

where ${}_C P_{t,a}$ is the observed catch at age and ${}_C n$ is the multinomial sample size. Survey catch-at-age data are generated in the same way.

Estimation model

The estimation model is a statistical catch-at-age model (Fournier and Archibald 1982) implemented using Coleraine and has the same parametrization as the operating model. It would therefore fit the data perfectly, if it was not for the observation noise both in the survey abundance index and in the commercial and survey catch-at-age data. The parametrization allows the commercial selectivity curve to decline at the oldest ages, but the survey selectivity curve is correctly assumed to be asymptotic. Some of the estimated parameters, including natural mortality rate M, stock-recruitment steepness h, and declining right-hand commercial selectivity are known to be correlated and problematic to estimate (Magnusson and Hilborn 2007). Wide bounds (Table S2) are assigned to all estimated parameters so as not to impose any major con-

straints on the parameter values. The estimation model is given the correct (i.e. operating model) value for recruitment variability, $\sigma_R = 0.6$.

The objective function for the estimation model is the sum of four components. The first three relate to observed data, and the last component is a penalty on deviations from Beverton-Holt recruitment. The abundance index is assumed to be lognormally distributed, the robust normal likelihood for proportions (Fournier *et al.* 1990) is assumed for the commercial and survey catch-at-age data, while the recruitment deviates are assumed to be lognormal. The magnitude of observation error for the abundance index is estimated using maximum likelihood, while the effective sample sizes for the commercial and survey catch-at-age data are estimated using the approach of McAllister and Ianelli (1997). The same age-composition sample size is assumed for all years, calculated as the median of estimated annual effective sample sizes.

Reference points

Six reference points are evaluated as potential management quantities of interest: B_{current} (current spawning biomass), u_{current} (current harvest rate), Depletion (depletion level, B_{current} relative to virgin spawning biomass), maximum sustainable yield (MSY), $B_{\text{current}}/B_{\text{MSY}}$ (B_{current} relative to B_{MSY}) and Surplus (current surplus production). These reference points describe the current stock status and potential yield, and are described in detail by Magnusson and Hilborn (2007). MSY and B_{MSY} are defined as the long-term average catch and spawning biomass when the harvest rate is set to an optimal value, u_{MSY}. Surplus production is defined as the last year's catch, plus the resulting change in vulnerable biomass.

The true reference point values from the operating model vary between recruitment scenarios (Table 1), except u_{current} that is predefined (Fig. 2, Table S3), and MSY that depends only on R_0, h, M, and commercial selectivity. The true MSY value is in all cases 203 thousand t, with harvest rate $u_{\text{MSY}} = 0.154$ and spawning biomass $B_{\text{MSY}} = 1270$ thousand t.

Evaluating uncertainty

The three methods used to quantify uncertainty start with the same input, the simulated datasets. Equation (3) summarizes how each method generates a probability distribution that is used to construct confidence intervals,

$$\begin{aligned} & y \xrightarrow[\text{delta}]{\text{model}} \hat{\theta}, \widehat{\text{SE}}_{\hat{\theta}} \xrightarrow{\text{Norm}} p(y \mid \theta) \\ & y \xrightarrow{\text{model}} \hat{\theta} \xrightarrow{\text{bootstrap}} y_1^*, y_2^*, \ldots, y_B^* \xrightarrow{\text{model}} \\ & \quad \hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^* \xrightarrow[\text{biascorr}]{\text{density}} p(y \mid \theta) \\ & y \xrightarrow[\text{MCMC}]{\text{model}} \theta_1, \theta_2, \ldots, \theta_T \xrightarrow{\text{density}} p(\theta \mid y) \end{aligned} \tag{3}$$

where y denotes the observed data, θ is a vector of parameters (and derived quantities), the ^ symbol indicates an estimate of a parameter or derived quantity, $\widehat{\text{SE}}_{\hat{\theta}}$ is the estimated standard error of $\hat{\theta}$, y_b^* is a bootstrap dataset, $\hat{\theta}_b^*$ is a bootstrap estimate, and θ_t is an MCMC iteration. The sampling distribution $p(y|\theta)$ and posterior distribution $p(\theta|y)$ are used to generate confidence intervals at any given confidence level.

Delta method

The estimation model uses automatic differentiation (Griewank and Corliss 1991; Fournier *et al.* 2012) to evaluate the Hessian matrix and hence the approximate variance–covariance matrix for the

Table 1 True reference point values from the operating model for each recruitment scenario. B_{current}, MSY, and Surplus are expressed in thousands of tonnes. u_{current} and MSY are 0.023 and 203, respectively, for all 10 recruitment scenarios.

Reference point	Recruitment scenario									
	1	2	3	4	5	6	7	8	9	10
B_{current}	1904	2156	1611	1793	2537	1318	1960	1704	1484	1802
Depletion	0.479	0.543	0.405	0.451	0.639	0.332	0.493	0.429	0.374	0.454
$B_{\text{current}}/B_{\text{MSY}}$	1.499	1.697	1.268	1.411	1.997	1.038	1.543	1.341	1.168	1.418
Surplus	83	315	158	164	166	198	245	300	84	227

MSY, maximum sustainable yield.

estimated parameters. The delta method (Seber 1973), which assumes that both estimation bias and the quadratic terms of the Taylor series are negligible, is then used to estimate the variance of each derived quantity,

$$\widehat{\mathrm{SE}}_{\hat{g}} = \sqrt{\sum_i \sum_j \widehat{\mathrm{Cov}}\left(\hat{\theta}_i, \hat{\theta}_j\right)\left(\frac{\partial g}{\partial \theta_i}\right)\left(\frac{\partial g}{\partial \theta_j}\right)}, \quad (4)$$

where g is a derived quantity, such as a reference point, that is a function of some estimated parameters θ_1, θ_2,..., θ_n. The symmetric confidence interval for g is then:

$$\left[\hat{g} - z_{\alpha/2}\widehat{\mathrm{SE}}_{\hat{g}},\ \ \hat{g} + z_{\alpha/2}\widehat{\mathrm{SE}}_{\hat{g}}\right], \quad (5)$$

The reference points B_{current} and MSY are log-transformed for the purpose of applying the delta method, because the uncertainty about these quantities can be expected to be closer to lognormal than normal (Mohn 1993; Patterson *et al.* 2001), and exploratory bootstrap and MCMC runs indicated that this was the case. Although surplus production is also measured in biomass units, it is not log-transformed, as exploratory results showed fairly symmetric distributions, and because surplus production can be negative when weak cohorts are entering the fishable stock.

Bootstrap

A parametric model-conditioned approach is used to generate 1000 bootstrap datasets for each simulated dataset. In their simulation study, Punt and Butterworth (1993) found that 100 bootstrap datasets was adequate for variance estimation, but 1000 bootstrap datasets are used here, because more replicates are needed to estimate quantiles than variance. The bootstrap is parametric with residuals sampled from estimated probability distributions, and model-conditioned in that the residuals are not applied to the observed data but to predictions from the model fit to the original data (Efron and Tibshirani 1993). The parametric bootstrap was chosen because it is probably what would be used in practice for this particular assessment model, as there is no straightforward way to resample residuals for the catch-at-age data when they are proportions. The bootstrap survey abundance index is

$$I_t^* = \hat{I}_t \times \exp({}_I\varepsilon_t^*), \quad {}_I\varepsilon_t^* \sim N(0, \hat{\sigma}_I^2), \quad (6)$$

where I_t^* is the bootstrap datum for year t, $\hat{I}_t$ is the predicted index for year t from the model fit to the original dataset, ${}_I\varepsilon_t^*$ are bootstrap residuals, and $\hat{\sigma}_I$ is the estimated magnitude of observation error. The bootstrap commercial catch at age is

$${}_CP_{t,a}^* \sim \mathrm{Multinom}\left({}_C\hat{n},\ {}_C\hat{P}_{t,a}\right)/{}_C\hat{n}, \quad (7)$$

where ${}_CP_{t,a}^*$ are the bootstrap data, ${}_C\hat{n}$ is the estimated effective sample size, and ${}_C\hat{P}_{t,a}$ is the model-predicted commercial catch at age for year t.

The estimation model is fitted to each of the 1000 bootstrap datasets, resulting in 1000 bootstrap estimates for each parameter and derived quantity. A bias-correction factor is then applied, which has been shown to lead to more accurate confidence intervals (Efron 2003). In fisheries stock assessment, Gavaris and Van Eeckhaute (1998) and others have used the BCa algorithm (bias correction and acceleration, Efron and Tibshirani 1993) with the acceleration coefficient set to zero. Acceleration relates to the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ, so zero acceleration implies that the standard error of $\hat{\theta}$ is the same for all θ. The algorithm then simplifies to

$${}_{BC}\vec{\hat{\theta}}^* = \hat{\Omega}^{-1}\left[\Phi\left(2\Phi^{-1}[\hat{\Omega}(\hat{\theta})] + \Phi^{-1}(\vec{\alpha})\right)\right], \quad (8)$$

where ${}_{BC}\vec{\hat{\theta}}^*$ is a vector of bias-corrected bootstrap estimates in ascending order, $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\hat{\Omega}(x) = \#\{\hat{\theta}^* < x\}/B$ is the empirical cumulative distribution function of the bootstrap estimates $\hat{\theta}^*$, while $\Phi^{-1}(\cdot)$ and $\hat{\Omega}^{-1}(\cdot)$ are the corresponding inverse functions, B is the number of bootstraps, and $\vec{\alpha}$ is a vector of probability levels $1/B$, $2/B$, ..., B/B.

The bias-correction algorithm compares the bootstrap estimates of a given quantity to the original point estimate. If the median of the bootstrap estimates is above or below the original point estimate, it is seen as an indication of a biased estimator. As the original point estimate was subject to the same bias, the algorithm corrects for the bias by transforming the bootstrap estimates (Fig. 3). The algorithm performs no transformation if the median of the bootstrap estimates is the same as the original point estimate. It is also worth noting that the bias-corrected bootstrap estimates are always within the range of the 'raw' bootstrap estimates. The resulting confidence interval may be narrower or wider.

The algorithm fails in the rare case when the bias is so extreme that all the bootstrap estimates are above or below the original point estimate. In these cases, the $\hat{\Omega}(\hat{\theta})$ term, the proportion of bootstrap

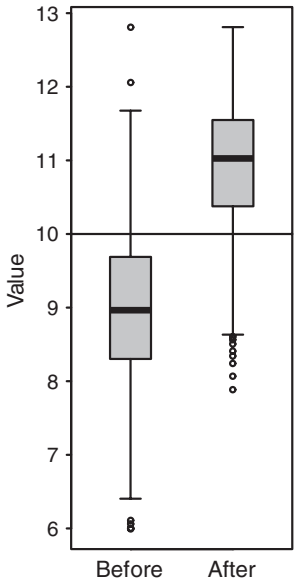

Figure 3 Effect of bias correction on bootstrap estimates. In this hypothetical example, the bootstrap estimates (left boxplot) are lower than the point estimate from the original data (horizontal line). The resulting bias-corrected bootstrap estimates (right boxplot) take into account that the original point estimate was subject to the same bias.

estimates that are below the original point estimate, is 0 or 1, resulting in expressions such as $\hat{\Omega}^{-1}[\Phi(-\infty+\infty)]$, which is not mathematically defined. To avoid this problem, a robust algorithm is used, where the $\hat{\Omega}(\hat{\theta})$ term is bounded between 0.1 and 0.9:

$$ {}_{BC}\vec{\theta}^{*} = \hat{\Omega}^{-1}\Big[\Phi\Big(2\Phi^{-1}[\max\{0.1, \min\{0.9,\ \hat{\Omega}(\hat{\theta}\})] \ +\ \Phi^{-1}(\vec{\alpha})\Big)\Big]. \tag{9} $$

These safety bounds guarantee a valid interval, but without the safety bounds, 5 of 6000 bias-corrected bootstrap intervals were undefined. The bias demonstrated in Fig. 3 corresponds to $\hat{\Omega}(\hat{\theta}) = 0.84$, so the safety bounds at 0.1 and 0.9 are irrelevant for that example. The computer code for the robust bias-correction algorithm is provided in Appendix 1. The bias-corrected bootstrap confidence interval is calculated as:

$$ \Big[\frac{\alpha}{2}\text{ quantile from }{}_{BC}\vec{\theta}^{*},\ \Big(1-\frac{\alpha}{2}\Big)\text{ quantile from }{}_{BC}\vec{\theta}^{*}\Big]. \tag{10} $$

MCMC

Markov chain Monte Carlo simulation is used to approximate the posterior distribution of estimated parameters and reference points. The simulation method is Metropolis-Hastings with an adaptive multivariate normal jumping distribution (Gelman *et al.* 2004; Fournier *et al.* 2012).

All model parameters are assigned uniform priors based on their bounds (Table S2), except the deviations about Beverton-Holt stock-recruitment relationship have a lognormal prior. The MCMC simulation is run for 1 million iterations and then thinned, keeping every 1000th iteration. Convergence of the estimated reference points is diagnosed using the *coda* package (Plummer *et al.* 2006), adopting an autocorrelation threshold of 0.1, Geweke threshold of 1.96, and Heidelberger-Welch threshold of 0.05. If any criteria are not met, the MCMC chain is extended to a maximum of 10 million iterations, still thinning to 1000 iterations, to reduce autocorrelation and stabilize the distribution quantiles. This proved to be necessary for a few hundred model runs, owing to unstable model convergence as can be expected when simultaneously estimating correlated parameters such as natural mortality rate M, stock-recruitment steepness h, and declining right-hand selectivity (Magnusson and Hilborn 2007).

The MCMC confidence interval is calculated as

$$ \Big[\frac{\alpha}{2}\text{ quantile from }\theta_1,\theta_2,\ldots,\theta_T,\ \Big(1-\frac{\alpha}{2}\Big)\text{ quantile from }\theta_1,\theta_2,\ldots,\theta_T\Big], \tag{11} $$

where θ_1, θ_2, …, θ_T are the iterations retained from the MCMC chain.

Results

We first examine the performance of confidence intervals at the 90% confidence level, then broaden the analysis to all confidence levels, and finally examine the sensitivity of the results to changes to assumptions. Results are shown for both bias-corrected and 'raw' (non-bias-corrected) bootstrap confidence intervals to evaluate whether and how much the bias correction improves the bootstrap performance.

90% confidence level

A total of 24 000 confidence intervals are analyzed at the 90% confidence level (four uncertainty methods, six reference points, 10 recruitment scenarios, and 100 stochastic datasets for each recruitment scenario). Before summarizing, it is useful to

look at an example set of confidence intervals (Fig. 4), where the uncertainty method is MCMC, the reference point is current surplus production, and the recruitment scenario is 10. For a 90% confidence level, one would expect around 90% of the confidence intervals to cover the true value (227 thousand t), so the coverage probability in this example, 93 of 100, is slightly higher than the nominal value of 90.

Looking across all uncertainty methods, reference points, and recruitment scenarios, the coverage probability is usually lower than the target (Table 2), with 216 of 240 combinations having coverage probabilities below 90. The example described previously, with a coverage probability of 93, can be found in the lower right-hand corner of Table 2. The coverage probabilities vary considerably between recruitment scenarios and the purpose of including ten scenarios instead of only one is to prevent the results from depending on a particular recruitment history.

The overall trends emerge after averaging over recruitment scenarios (Table 3), with the delta method, bootstrap, and MCMC all showing coverage probabilities <90, that is, the methods lead to 90% confidence intervals that cover the true value <90% of the time. Overall, the delta method and MCMC perform better than the bias-corrected bootstrap, with mean coverage probabilities of 73.0, 72.5 and 64.1, respectively. The performance of the bootstrap is considerably poorer before bias correction, with a mean coverage probability of 57.4. The delta method outperforms the other methods in evaluating the uncertainty about the current biomass, current harvest rate, depletion, and surplus production, but performs poorly for $B_{current}/B_{MSY}$. MCMC performs better than the delta method and bootstrap for MSY and $B_{current}/B_{MSY}$, and its mean coverage probability is above 60 for all reference points. The bias-corrected bootstrap has similar or lower coverage probabilities than the delta method and MCMC, including a particularly low coverage probability of 45.6 for MSY. Bias correction generally improves the bootstrap performance, although it reduces the coverage probability from 71.2 to 65.6 for $B_{current}/B_{MSY}$. On the other hand, bias correction leads to a substantial increase in coverage probability for current harvest rate, from 44.9 to 66.5.

All confidence levels

When the analysis is repeated for different confidence levels (Fig. 5, Table S4), the results confirm the trends for the 90% confidence level. The delta method, bootstrap, and MCMC show coverage probabilities that are consistently lower than expected at all confidence levels. The general pattern is that the delta method and MCMC perform better than the bootstrap, the main exception being $B_{current}/B_{MSY}$, where the delta method performs considerably worse than MCMC

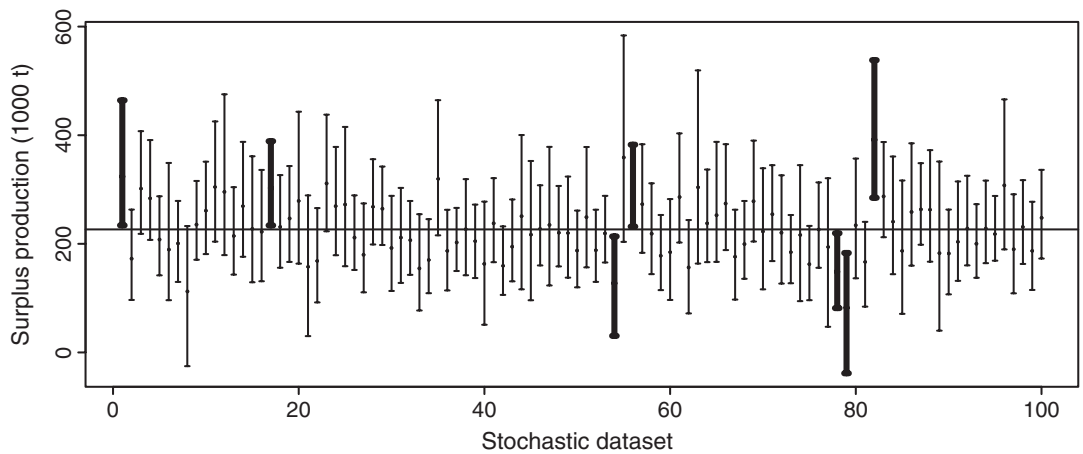

Figure 4 Example results, showing 90% confidence intervals for surplus production, from Markov chain Monte Carlo analysis of 100 stochastic datasets for recruitment scenario 10. Seven confidence intervals (thick lines) of one hundred do not cover the true value (horizontal line). In this example, the coverage probability is 93.

Table 2 Coverage probability for 90% confidence intervals for each uncertainty method, reference point, and recruitment scenario. The non-bias-corrected bootstrap is referred to as 'raw' and bias-corrected bootstrap as 'bootstrap'. Ideally, the coverage probability at this confidence level should be 90.

Method	Reference point	Recruitment scenario									
		1	2	3	4	5	6	7	8	9	10
Delta	$B_{current}$	80	70	75	77	73	80	78	70	72	72
	$u_{current}$	78	71	73	74	68	78	73	70	71	71
	Depletion	79	54	66	74	77	79	67	83	70	80
	MSY	72	19	100	96	94	99	61	99	100	87
	$B_{current}/B_{MSY}$	37	54	51	47	42	50	63	39	57	44
	Surplus	70	95	90	86	81	94	89	93	71	94
Raw	$B_{current}$	71	45	65	58	57	78	42	59	49	61
	$u_{current}$	52	28	53	44	46	71	18	52	36	49
	Depletion	61	17	46	58	62	79	5	76	50	68
	MSY	20	1	49	47	28	100	6	98	78	37
	$B_{current}/B_{MSY}$	85	62	49	75	72	70	78	73	65	83
	Surplus	44	94	82	67	67	86	75	83	20	96
Bootstrap	$B_{current}$	62	66	66	66	61	68	74	69	61	57
	$u_{current}$	65	65	69	72	60	78	62	77	59	58
	Depletion	67	63	71	73	64	69	39	71	73	59
	MSY	30	6	73	56	42	58	27	76	59	29
	$B_{current}/B_{MSY}$	61	71	74	66	63	63	56	57	87	58
	Surplus	68	77	81	78	70	80	79	85	67	83
MCMC	$B_{current}$	66	72	61	71	66	69	68	66	69	69
	$u_{current}$	63	64	55	66	59	67	66	62	65	63
	Depletion	73	72	66	74	71	64	53	61	71	83
	MSY	81	30	89	98	79	98	78	93	99	87
	$B_{current}/B_{MSY}$	70	84	66	78	62	53	76	36	74	73
	Surplus	75	94	90	88	78	90	85	91	66	93

MCMC, Markov chain Monte Carlo; MSY, maximum sustainable yield.

Table 3 Coverage probability for 90% confidence intervals for each uncertainty method and reference point, averaged across recruitment scenarios. The non-bias-corrected bootstrap is referred to as 'raw' and bias-corrected bootstrap as 'bootstrap'. Ideally, the coverage probability at this confidence level should be 90.

	Delta	Raw	Bootstrap	MCMC
$B_{current}$	74.7	58.5	65.0	67.7
$u_{current}$	72.7	44.9	66.5	63.0
Depletion	72.9	52.2	64.9	68.8
MSY	82.7	46.4	45.6	83.2
$B_{current}/B_{MSY}$	48.4	71.2	65.6	67.2
Surplus	86.3	71.4	76.8	85.0
Average	73.0	57.4	64.1	72.5

MCMC, Markov chain Monte Carlo; MSY, maximum sustainable yield.

and the bootstrap. The delta method performs slightly better than MCMC for current biomass, current harvest rate, and depletion at confidence levels higher than 50%, but the two methods perform equally well for MSY and surplus production. On the whole, the bias-corrected bootstrap performs poorer than the delta method and MCMC, particularly for MSY. Bias correction leads to improved performance of the bootstrap, with the exception of $B_{current}/B_{MSY}$, and the improvement is especially noticeable for current harvest rate. MCMC has the most consistent performance for the various reference points (Fig. 5). Its coverage probability is always close to that for the best-performing method, and it shows no conspicuous failures, unlike the bootstrap for MSY and the delta method for $B_{current}/B_{MSY}$.

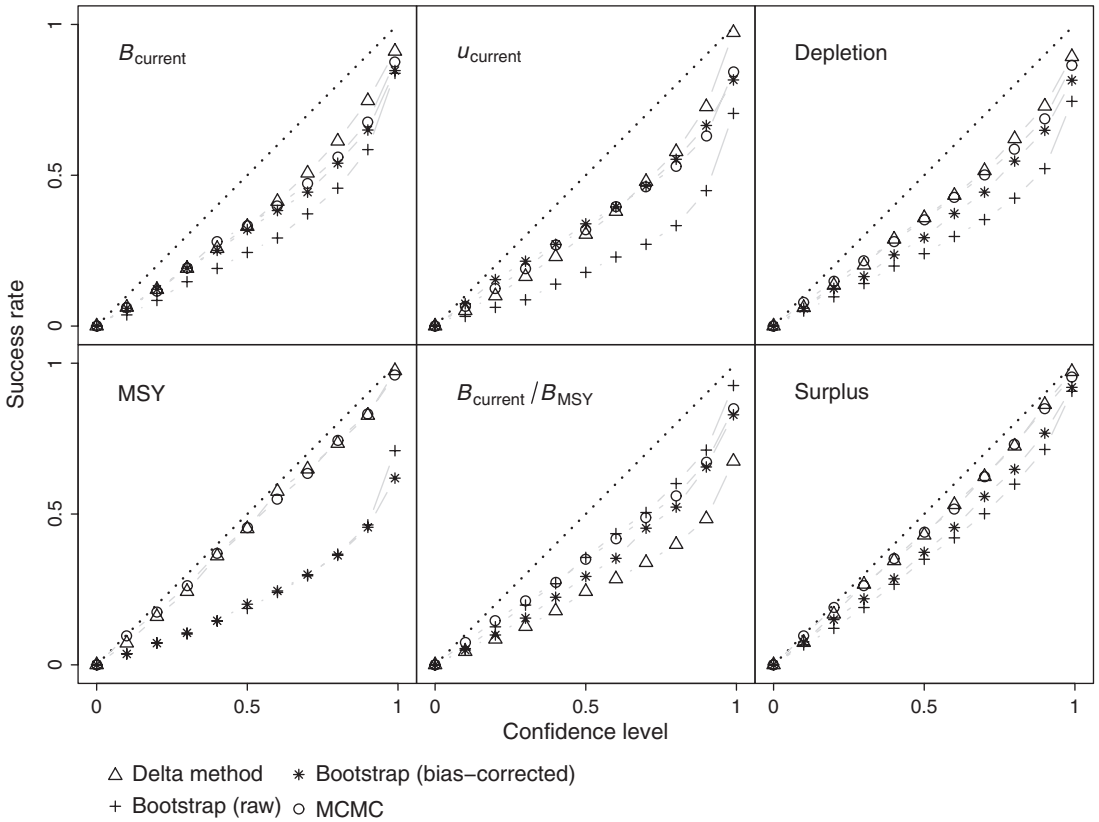

Figure 5 Coverage probability for confidence intervals by uncertainty method and reference point, evaluated at several confidence levels (0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 99%).

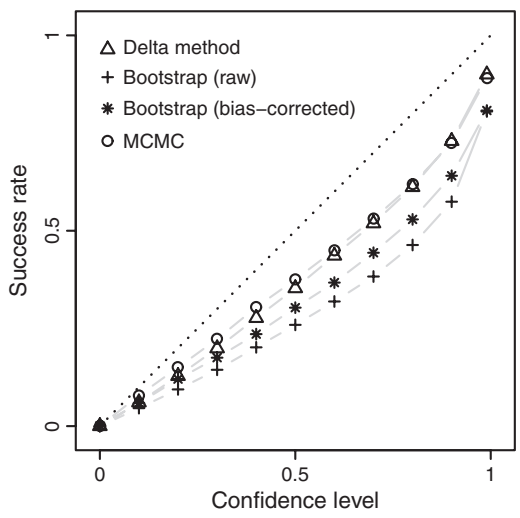

Figure 6 Coverage probability for confidence intervals for each uncertainty method averaged across all six reference points, evaluated at several confidence levels (0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 99%).

When the results are averaged across the six reference points (Fig. 6), the delta method and MCMC show similar performance, substantially better than the bias-corrected bootstrap. At the 50% confidence level, MCMC has a mean coverage probability of 38, the delta method has 35 and the bootstrap 30. At the 95% confidence level, the delta method and MCMC have a mean coverage probability of 80, while the bootstrap has 71 (Fig. 6, Table S4).

Sensitivity analysis

Four analyses are used to examine what factors may lead to the low coverage probabilities (Fig. 6). The first analysis assumes that the estimation method 'knows' the true magnitude of observation error, the second analysis uses a multinomial catch-at-age likelihood, the third assumes that the estimation method 'knows' the bias of estimated reference points, and the fourth combines all of the above.

These analyses are only conducted for the delta method owing to computational demands. Finally, a supplementary sensitivity analysis (Table S5) indicates that the overall results would not change very much if more recruitment scenarios would be included in the study.

Known magnitude of observation error

The observation noise in the simulated datasets is generated using lognormal $\sigma_I = 0.2$ and multinomial ${}_Cn = 50$ and ${}_Sn = 50$, but this magnitude of observation error is often underestimated by the estimation model. The iteratively estimated $\hat{\sigma}_I$ is often too low, the median estimate being 0.186, while ${}_C\hat{n}$ and ${}_S\hat{n}$ are often too high, with median estimates 52 and 53. This leads to narrower confidence intervals, which could explain the low coverage probabilities. When the estimation model uses the true values for σ_I, ${}_Cn$ and ${}_Sn$, the coverage probability of the delta method improves only marginally, from 72.9 to 74.5 at the 90% confidence level (Fig. 7, left panel).

Multinomial catch-at-age likelihood

The operating model generates catch-at-age data under the assumption of multinomial sampling, but the estimation model uses the Fournier robust normal likelihood for proportions. This introduces a model misspecification, which could explain the low coverage probabilities. When the estimation model assumes a multinomial catch-at-age likelihood instead of the robust normal likelihood for proportions, the coverage probability of the delta method improves noticeably, from 72.9 to 78.4 at the 90% confidence level (Fig. 7, center panel).

Known bias

Each reference point is estimated with some bias. The median of the 1000 point estimates compared with the true value, median $(\hat{\theta} - \theta)/\theta$, is −0.13 for current biomass, +0.22 for current harvest rate, −0.20 for current depletion level, +0.20 for MSY, +0.05 for $B_{\text{current}}/B_{\text{MSY}}$, and +0.14 for current surplus production. When the delta-method confidence intervals are shifted to correct for the median bias of each reference point, the coverage probability improves noticeably, from 72.9 to 78.8 at the 90% confidence level (Fig. 7, right panel).

Combined effect

When the estimation model assumes a multinomial catch-at-age likelihood, given the true values for σ_I, ${}_Cn$ and ${}_Sn$, and the confidence intervals are then shifted to correct for the median bias of each reference point, the coverage probability improves considerably, from 72.9 to 82.6 at the 90% confidence level.

Discussion

Confidence intervals are too narrow

The delta method, bootstrap, and MCMC all produced confidence intervals that did not cover the true value as often as the nominal confidence level implies (Fig. 6). The three methods are well established in the statistical literature, widely used, and have been shown to perform well for simple models, when all assumptions are met (Seber 1973; Efron and Tibshirani 1993; Gelman *et al.* 2004). The purpose of this study was to examine how well they

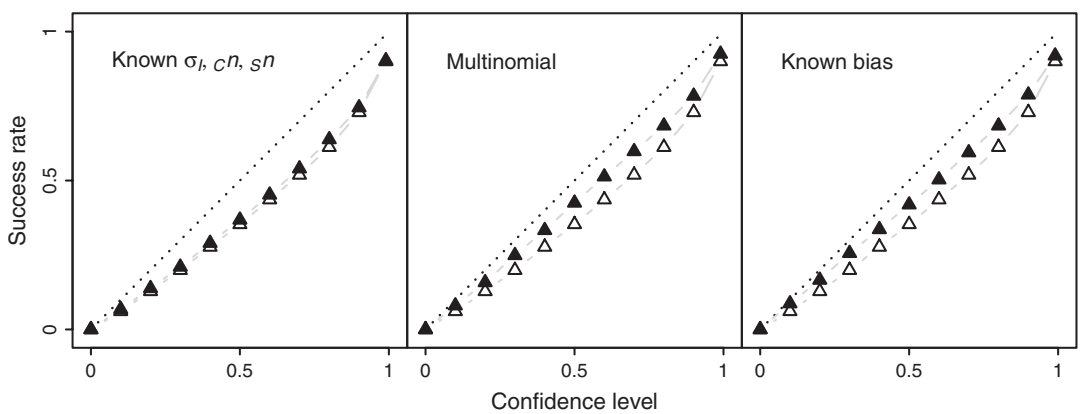

Figure 7 Coverage probabilities for the sensitivity tests. White triangles indicate the base case delta method (same as Fig. 6), and black triangles indicate the outcomes of each sensitivity test.

perform with a typical stock assessment model of medium complexity, when most assumptions are met. Generally, the performance of all statistical methods degrades with increased model complexity, as non-linearity and correlated parameter estimates undermine the assumptions and lead to estimation bias (Seber and Wild 1989). The optimization method also becomes less likely to find the global minimum. In this study, sensitivity analysis showed that even after correcting for estimation bias, the delta-method 90% confidence intervals still covered the true value <80% of the time.

The expectation was that the methods would show some inaccuracy, because of model complexity, but not necessarily that the confidence intervals would be too narrow. It would seem, *a priori*, just as possible that some of the methods might generate confidence intervals that covered the true value more often than the nominal confidence level implies.

This study is based on a scenario that is known to be informative (Magnusson and Hilborn 2007), where the stock is first fished down and then allowed to rebuild (Fig. 2), with standardized surveys and age data from the start of the fishery. Furthermore, the data are generated using the same dynamics as the estimation model, and landings, body weight, maturity, and recruitment variability are known without error. When analyzing real data, we can expect model error and process variability to lead to considerably more bias, and therefore, confidence intervals that are even less likely to cover the true value. The notion that statistical methods in stock assessment tend to underestimate the real extent of uncertainty is reflected in the literature (Hilborn and Mangel 1997; Punt and Hilborn 1997; Patterson *et al.* 2001; Gavaris and Ianelli 2002) and demonstrated here in a setting where one would expect the methods to perform well. In a recent meta-analysis of multiple assessment models fitted to 17 stocks off the United States West Coast, Ralston *et al.* (2011) found that model specification error can be expected to be considerably greater than the estimation error.

Delta method and MCMC perform better than bootstrap

The delta method and MCMC provided better confidence intervals than the bootstrap on average (Fig. 6). For example, at the 90% confidence level, the delta-method intervals covered the true reference point value 73.0% of the time, MCMC 72.5% and bias-corrected bootstrap 64.1%. Although the intervals from all three methods were generally too narrow, the delta method and MCMC were considerably closer to attaining the nominal confidence level.

It is somewhat surprising to see how well the delta method performed, compared with the bootstrap and MCMC. Automatic differentiation (Griewank and Corliss 1991; Fournier *et al.* 2012) facilitates the use of the delta method with complex models, where derived quantities are not simple functions of estimated parameters, by applying automated algorithms to compute the partial derivatives. In application, the delta method is orders of magnitude faster than the computationally intensive bootstrap and MCMC methods, which can be a major advantage for iterative simulations, complex models, and/or large datasets (Maunder *et al.* 2009).

The delta method has been shown to perform about as well as the bootstrap for stock assessment (Punt and Butterworth 1993; Restrepo *et al.* 2000), or slightly worse (Mohn 1993; Gavaris 1999). A simulation study comparing the delta method, bootstrap, and MCMC (Restrepo *et al.* 2000) found that the delta method and bootstrap performed about as well, but MCMC performed poorer. The present study's ranking of the uncertainty methods is therefore quite different from the results of previous simulation studies. The contradictory results are most likely due to model differences; the previous studies used relatively simple biomass-dynamic models and ADAPT, with fewer estimated parameters, fewer objective function components, and more restrictive assumptions than the statistical catch-at-age model used here. Variations in the implementation of the delta method, bootstrap and MCMC can also affect their performance (Patterson *et al.* 2001; Gelman *et al.* 2004; Givens and Hoeting 2005). Finally, the studies vary in terms of whether they compare confidence intervals or variance estimates, and whether they focus on the uncertainty about model parameters or reference points.

Bias correction improves bootstrap performance

Overall, the bootstrap performed considerably better with bias correction than without it (Fig. 6), with the mean coverage probability at the 90% confidence level increasing from 57.4 to 64.1. This

shift of 6.7, compared with a shift of 32.6 that would bring it to ideal performance, amounts to around 20% improvement. This performance improvement did not, however, apply uniformly across all reference points (Fig. 5), ranging from particularly beneficial for current harvest rate and depletion, to slightly detrimental for MSY and $B_{current}/B_{MSY}$.

The estimation of current harvest rate was subject to greater bias than the other reference points. It is therefore reassuring to see that bias correction was most beneficial for that reference point, effectively correcting for the positive bias. It is also reassuring to see a similar performance gain for negatively biased reference points, such as current depletion. When bias correction does not lead to improved performance, it is because the perceived bias, that is, the difference between the bootstrap estimates and the point estimate, does not reflect the true estimation bias.

This study evaluates the performance of the BCa bias-correction method for the bootstrap using zero acceleration. Alternative approaches include ABC (approximate bootstrap confidence), ABCq, and various ways to estimate the BCa acceleration coefficient (Efron and Tibshirani 1993). The implementation used here is recommended in the current fisheries stock assessment literature (Patterson *et al.* 2001; Gavaris and Ianelli 2002), and the results from this study support that recommendation.

Other findings

Why did the bootstrap perform so poorly for MSY? This reference point was positively biased, mainly due to a positive bias in the estimated stock-recruitment steepness parameter h. This bias can be expected in statistical catch-at-age models when the stock-recruitment steepness is estimated, unless the data include years of extremely low abundance, and the natural mortality rate and selectivity of older fish are known (Magnusson and Hilborn 2007; Conn *et al.* 2010). Despite this bias, the delta method and MCMC performed very well for MSY, providing confidence intervals of appropriate width at any given confidence level (Fig. 5).

The delta method showed unusually low coverage probability for $B_{current}/B_{MSY}$, compared with the other reference points. The most likely reason for this is that the assumption of symmetric Gaussian uncertainty is not appropriate for this ratio statistic. There are many transformations that could be used for each reference point, and it is beyond the scope of this study to explore all possibilities. Logarithm and square-root transformation are ruled out if the original quantity can be negative, such as, surplus production, as are logit and probit transformation when the quantity can exceed 1.0, such as depletion level and $B_{current}/B_{MSY}$. Transforming reference points has an important effect on the performance of the delta method, but transforming model parameters can improve the performance of the bootstrap and MCMC as well (Efron and Tibshirani 1993; Gelman *et al.* 2004). The use of statistical transformations in stock assessment models is a topic worthy of further investigations.

The sensitivity analysis showed that the estimation model performed noticeably better when multinomial likelihood was used for catch at age, instead of the default Fournier robust normal likelihood for proportions. As the operating model uses multinomial random draws to generate the catch at age data, this sensitivity test quantifies the model error introduced by likelihood misspecification. The Fournier likelihood is designed to be more robust than the multinomial likelihood when observed data are subject to greater variability than statistical theory predicts (Fournier *et al.* 1990) and has been shown to perform well when that is the case (Ernst 2002). This study, on the other hand, shows that the Fournier likelihood does not perform as well as the multinomial likelihood when the data are random draws from the multinomial distribution. The Fournier likelihood is not a generalization of the multinomial that allows greater variance, but rather a hybrid between normal and multinomial that explicitly downweights two kinds of outliers: ages with few observations and predictions that are far from the observations. We recommend using the Fournier likelihood to analyze real fisheries data, and use it in this study to represent a typical estimation model in stock assessment.

The additional analyses also examined the impact of biased reference points, and how much of the total error is because of bias, as opposed to too narrow confidence intervals. Magnusson and Hilborn (2007) described what kinds of biases can be expected when estimating reference points, depending on the fishing history, model assumptions, and available data. As the fishing history and estimation model analyzed here were selected from

that study, the biases were known beforehand. When these biases were corrected for each reference point, the performance of the delta method improved about one-third towards ideal performance. When analyzing real fisheries data, the total error cannot be partitioned in this way, because bias can only be evaluated when the true value is known.

Recommendations

The overall performance trends suggest that MCMC is the most reliable of the three uncertainty methods, given the dataset and catch-at-age assessment method. Both the delta method and the bootstrap performed poorly for one or more reference points, while MCMC was always close to the best-performing method. When time and resources allow, we recommend using more than one method to evaluate uncertainty, to see whether they lead to markedly different conclusions. If only one method is to be used, it seems that MCMC is the least likely to severely misrepresent uncertainty. All three methods, however, have a strong tendency to underestimate the uncertainty.

On the average, the delta method performed well compared with the computationally intensive bootstrap and MCMC methods and can be recommended for quick evaluation of uncertainty while exploring a variety of modelling options, before applying the bootstrap and/or MCMC to selected model runs. The delta method may also be useful when confidence intervals are required in a large number of simulations. In this study, the delta-method calculations were around 1000 times faster than the bootstrap and MCMC. Management strategy evaluation (Butterworth and Punt 1999; De Oliveira *et al.* 2008) is a common application where this can be relevant. Possible transformations of model parameters and reference points should be explored when using the delta method.

One advantage of the bootstrap is that it can detect bias in the estimation model, thus providing valuable diagnostic information for the modeller (Haddon 2003). We recommend applying bias correction when using the bootstrap, having seen around 20% overall performance improvement in this study. That said, the bootstrap was generally outperformed by the delta method and MCMC, in spite of bias. It would be interesting to see a similar performance comparison of uncertainty methods where the estimators are more biased than here. Another potential advantage of the bootstrap is that computations can be split into parallel threads, thus taking less time than computing a very long MCMC chain.

Each method for evaluating uncertainty is based on a particular set of assumptions. It appears that the choice between frequentist methods, such as the delta method and bootstrap, and Bayesian methods, such as MCMC, is not the most important decision for the modeller. In this study, for example, the overall performance of the delta method and MCMC was more similar than that of the bootstrap. It is at least as important that the modeller considers, and preferably tests, the sensitivity of the results to specific assumptions within a method. The effects of different transformations for the delta method have been discussed earlier, and Patterson *et al.* (2001) describe several bias-correction methods. Choices within the bootstrap include parametric vs. non-parametric, model-conditioned vs. non-conditioned, and a variety of bias-correction methods (Efron and Tibshirani 1993). In Bayesian inference, the choice of prior distributions can be important, and different algorithms to approximate the posterior probability have their strengths and weaknesses (Gelman *et al.* 2004). The same estimation model can often be expressed as frequentist or Bayesian with few or no modifications, as is done in this study. The trend in the current statistical literature (e.g. Kass 2011) has been to deflate the frequentist-Bayesian debate, focusing instead on the assumptions that relate models to data. When frequentist and Bayesian procedures does lead to very different conclusions, the choice should primarily be based on their performance with simulated data.

Although we have limited the analysis to the delta method, bootstrap, and MCMC, other methods can also be used to evaluate uncertainty in stock assessment. Sampling-importance resampling (SIR) can be used to simulate Bayesian posterior distributions instead of MCMC, but both methods should lead to the same distribution if run long enough (Gelman *et al.* 2004). When stock assessment models include more than a dozen parameters, MCMC is more computationally efficient than SIR (McAllister *et al.* 1994; Punt and Hilborn 1997). Profile likelihood (Edwards 1992; Hilborn and Mangel 1997) is a straightforward method to evaluate the uncertainty about estimated parameters, but it is problematic to generate the profile likelihoods for derived quantities such as reference points and future projections. Finally, adjunct Monte Carlo can be

used to diagnose the consequences of changing the value of a fixed parameter, such as natural mortality rate or the shape of a stock-recruitment function (Patterson *et al.* 2001).

The main limitation of a study such as this one is that the conclusions are based on one particular estimation model and one artificial suite of data. Our goal in the experimental design was to use a typical stock assessment model of medium complexity, with generic groundfish data scenarios that are known to be rather informative (Magnusson and Hilborn 2007). Many stock assessments use simpler or more complex models that are conceptually and analytically related to the statistical catch-at-age model used here. In cases where the data and models are fundamentally different from what we used, perhaps involving species interaction or migration between areas, we can recommend using this study's simulation framework to investigate the performance of candidate uncertainty methods.

Acknowledgements

This research was supported by the Fulbright Program and the American-Scandinavian Foundation. We thank Jim Bence, Jim Ianelli, Steve Martell, Mark Maunder, Anders Nielsen, Jon Schnute, John Skalski, and two anonymous reviewers for insightful discussions and comments that improved the manuscript.

References

Ascough II, J.C., Maier, H.R., Ravalico, J.K. and Strudley, M.W. (2008) Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling* **219**, 383–399.

Booth, A.J. and Quinn II, T.J. (2006) Maximum likelihood and Bayesian approaches to stock assessment when data are questionable. *Fisheries Research* **80**, 169–181.

Butterworth, D.S. and Punt, A.E. (1999) Experiences in the evaluation and implementation of management procedures. *ICES Journal of Marine Science* **56**, 985–998.

Casella, G. and Berger, R.L. (2002) *Statistical Inference*, 2nd edn. Duxbury, Pacific Grove, CA.

Clark, J.S. (2005) Why environmental scientists are becoming Bayesians. *Ecology Letters* **8**, 2–14.

Conn, P.B., Williams, E.H. and Shertzer, K.W. (2010) When can we reliably estimate the productivity of fish stocks? *Canadian Journal of Fisheries and Aquatic Sciences* **67**, 511–523.

Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

De Oliveira, J.A.A., Kell, L.T., Punt, A.E., Roel, B.A. and Butterworth, D.S. (2008) Managing without best predictions: the Management Strategy Evaluation framework. In: *Advances in Fisheries Science. 50 Years on from Beverton and Holt* (eds A. Payne, J. Cotter and T. Potter). Blackwell, Oxford, pp. 104–134.

Edwards, A.W.F. (1992) *Likelihood*, 2nd edn. Johns Hopkins University Press, Baltimore, MD.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. (2000) The bootstrap and modern statistics. *Journal of the American Statistical Association* **95**, 1293–1296.

Efron, B. (2003) Second thoughts on the bootstrap. *Statistical Science* **18**, 135–140.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY.

Ellison, A.M. (1996) An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* **6**, 1036–1041.

Ernst, B. (2002) *An investigation on length-based models used in quantitative population modeling*. PhD thesis, University of Washington, 150 pp.

Fournier, D. and Archibald, C.P. (1982) A general theory for analyzing catch at age data. *Canadian Journal of Fisheries and Aquatic Sciences* **39**, 1195–1207.

Fournier, D.A., Sibert, J.R., Majkowski, J. and Hampton, J. (1990) MULTIFAN: a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*). *Canadian Journal of Fisheries and Aquatic Sciences* **47**, 301–317.

Fournier, D.A., Skaug, H.J., Ancheta, J. *et al.* (2012) AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27**, 233–249.

Gavaris, S. (1999) Dealing with bias in estimating uncertainty and risk. *NOAA Technical Memorandum* NMFS-F/SPO-40, 46–50.

Gavaris, S. and Ianelli, J.N. (2002) Statistical issues in fisheries' stock assessments. *Scandinavian Journal of Statistics* **29**, 245–271.

Gavaris, S. and Van Eeckhaute, L. (1998) Assessment of haddock on eastern Georges Bank. *CSAS Research Document* 98/66, 75 pp.

Gavaris, S., Patterson, K.R., Darby, C.D. *et al.* (2000) Comparison of uncertainty estimates in the short term using real data. *ICES CM* 2000/V:03, 30 pp.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995) *Bayesian Data Analysis*. Chapman and Hall, London.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*, 2nd edn. CRC, Boca Raton, FL.

Givens, G.H. and Hoeting, J.A. (2005) *Computational Statistics*. Wiley, Hoboken, NJ.

Griewank, A. and Corliss, G.F. (eds) (1991) *Automatic Differentiation of Algorithms: Theory, Implementation, and Application*. SIAM, Philadelphia, PA.

Haddon, M. (2003) To be Bayesian or to bootstrap: what is the risk? In: *Towards Sustainability of Data-Limited Multi-Sector Fisheries* (eds S.J. Newman, D.J. Gaughan, G. Jackson *et al.*). Department of Fisheries, Perth, WA, pp. 98–104.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hilborn, R. (2003) The state of the art in stock assessment: where we are and where we are going. *Scientia Marina* **67**(S1), 15–20.

Hilborn, R. and Mangel, M. (1997) *The Ecological Detective: Confronting Models with Data*. Princeton University Press, Princeton, NJ.

Hilborn, R., Maunder, M., Parma, A., Ernst, B., Payne, J. and Starr, P. (2003) Coleraine: a generalized age-structured stock assessment model. User's manual version 2.0. *University of Washington Report* SAFS-UW-0116, 54 pp.

ICES (2003) Report of the North-Western Working Group: Icelandic cod. *ICES CM* 2003/ACFM 24, 144–227.

Kass, R.E. (2011) Statistical inference: the big picture. *Statistical Science* **26**, 1–9.

Magnusson, A. and Hilborn, R. (2007) What makes fisheries data informative? *Fish and Fisheries* **8**, 337–358.

Maunder, M.N., Schnute, J.T. and Ianelli, J.N. (2009) Computers in fisheries population dynamics. In: *Computers in Fisheries Research* (eds B.A. Megrey and E. Moksness), 2nd edn. Springer, New York, NY, pp. 337–372.

McAllister, M.K. and Ianelli, J.N. (1997) Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. *Canadian Journal of Fisheries and Aquatic Sciences* **54**, 284–300.

McAllister, M.K., Pikitch, E.K., Punt, A.E. and Hilborn, R. (1994) A Bayesian approach to stock assessment and harvest decisions using the sampling/importance resampling algorithm. *Canadian Journal of Fisheries and Aquatic Sciences* **51**, 2673–2687.

McGarvey, R., Feenstra, J.E. and Ye, Q. (2007) Modeling fish numbers dynamically by age and length: partitioning cohorts into ''slices''. *Canadian Journal of Fisheries and Aquatic Sciences* **64**, 1157–1173.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Mohn, R.K. (1993) Bootstrap estimates of ADAPT parameters, their projection in risk analysis and their retrospective patterns. *Canadian Special Publication in Fisheries and Aquatic Sciences* **120**, 173–184.

Mohn, R. (2009) The uncertain future of assessment uncertainty. In: *The Future of Fisheries Science in North America* (eds R.J. Beamish and B.J. Rothschild). Springer, New York, NY, pp. 495–504.

Neyman, J. (1937) Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A* **236**, 333–380.

Oehlert, G.W. (1992) A note on the delta method. *The American Statistician* **46**, 27–29.

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. *Canadian Journal of Fisheries and Aquatic Sciences* **56**, 208–221.

Patterson, K., Cook, R., Darby, C. *et al.* (2001) Estimating uncertainty in fish stock assessment and forecasting. *Fish and Fisheries* **2**, 125–157.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**(1), 7–11.

Punt, A.E. and Butterworth, D.S. (1993) Variance estimates for fisheries assessment: their importance and how best to evaluate them. *Canadian Special Publication in Fisheries and Aquatic Sciences* **120**, 145–162.

Punt, A.E. and Hilborn, R. (1997) Fisheries stock assessment and decision analysis: the Bayesian approach. *Reviews in Fish Biology and Fisheries* **7**, 35–63.

Punt, A.E. and Kennedy, R.B. (1997) Population modeling of Tasmanian rock lobster, *Jasus edwardsii*, resources. *Marine and Freshwater Research* **48**, 967–980.

Ralston, S., Punt, A.E., Hamel, O.S. *et al.* (2011) A meta-analytic approach to quantifying scientific uncertainty in stock assessments. *Fisheries Bulletin* **109**, 217–231.

Restrepo, V.R., Patterson, K.R., Darby, C.D. *et al.* (2000) Do different methods provide accurate probability statements in the short term? *ICES CM* 2000/V:08, 18 pp.

Schnute, J.T., Richards, L.J. and Olsen, N. (1998) Statistics, software, and fish stock assessment. In: *Fishery Stock Assessment Models* (eds F. Funk, T.J. Quinn II, J. Heifetz *et al.*). Sea Grant Program, Fairbanks, AK, pp. 171–184.

Seber, G.A.F. (1973) *The Estimation of Animal Abundance and Related Parameters*. Griffin, London.

Seber, G.A.F. and Wild, C.J. (1989) *Nonlinear Regression*. Wiley, Hoboken, NJ.

Trzcinski, M.K., Mohn, R. and Bowen, W.D. (2006) Continued decline of an Atlantic cod population: how important is gray seal predation? *Ecological Applications* **16**, 2276–2292.

Virtala, M., Kuikka, S. and Arjas, E. (1998) Stochastic virtual population analysis. *ICES Journal of Marine Science* **55**, 892–904.

Appendix 1

Bootstrap bias correction

The following R function was implemented for this study to apply BCa bootstrap bias correction with zero acceleration, robust to extremely biased cases (Equation 9).

```
BCboot <- function(thetastar, thetahat, bounds=c(0.1,0.9))
#############################################################################
###                                                                       #
### Function: BCboot                                                      #
###                                                                       #
### Purpose:  Apply bias correction to bootstrap estimates                #
###                                                                       #
### Args:     thetastar is a vector of bootstrap estimates                #
###           thetahat is a point estimate from original data             #
###           bounds is a vector of lower and upper limits to handle extremely #
###             biased cases                                              #
###                                                                       #
### Notes:    BCa with zero acceleration                                  #
###           Based on bcanon() in package 'bootstrap' by Tibshirani      #
###           See Efron and Tibshirani (1993, pp. 184-186), Gavaris and Van  #
###             Eeckhaute (1998, p.10), Gavaris (1999, p. 47)             #
###                                                                       #
### Returns:  Vector of bias-corrected bootstrap estimates                #
###                                                                       #
#############################################################################
{
  B <- length(thetastar)
  alpha <- (1:B) / B
  lower <- bounds[1]
  upper <- bounds[2]

  z0 <- qnorm(max(lower, min(upper, sum(thetastar<thetahat)/B)))
  zalpha <- qnorm(alpha)
  newalpha <- pnorm(2*z0 + zalpha)
  Omegainv <- approx(alpha, sort(thetastar), newalpha, rule=2)$y
  bias.corrected <- Omegainv[rank(thetastar)]

  return(bias.corrected)
}
```

Supporting Information

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

Figure S1. Age-specific characteristics of the operating model: survey selectivity, commercial selectivity, maturity, and weight.

Table S1. Age-specific weight (kg) and maturity (proportion) used in the operating and estimation model.

Table S2. Parameter values used in the operating model, along with bounds used in the estimation model.

Table S3. Annual harvest rate and recruitment used in the operating model.

Table S4. Coverage probability for confidence intervals by uncertainty method and reference point, evaluated at several confidence levels.

Table S5. Coverage probability for 90% confidence intervals, when the computations are repeated while leaving out one recruitment scenario at a time.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.