Implications for management procedure evaluation: the mechanical update and further exploration of the operating model

M Basson
T Polacheck
D Kolody
A Preece
J Hartog

Prepared for the CCSBT 5th Meeting of the Stock Assessment Group (SAG5) and the 9th Meeting of the Scientific Committee (SC9)
6-11 and 13-16 September 2004, Seogwipo KAL Hotel in Jeju, Korea

Table of Contents

List of Figures

Abstract

The operating model which has been used in the evaluation of MPs was conditioned with data up to 2000, and for a few data components up to 2001. In addition to doing assessments with data up to 2003 this year, the CCSBT decided to also update the operating model and condition the model with data up to 2003. The data series was not only extended, but the historic data (pre-2001) has also changed. This appears to have had a strong effect on the fit and parameter estimates of the operating model.

We consider the implications of these changes for the management procedure evaluation process. Since the operating model was used as an assessment model (CCSBT-ESC/0409/23), extensive further exploration of the model with the new data was undertaken. Implications from those results are also discussed.

Introduction

The CCSBT scientific committee has been working on the development of a management procedure over the past several years. A major part of the work is the evaluation and comparison of the performance of candidate management procedures. This has been done with the operating model (OM) which was developed for this task. Results of evaluations considered at the April 2004 MP Workshop were based on the OM generating a weighted combination of a range of plausible scenarios (e.g. different productivity and mortality values etc.). This is referred to as the reference case. The reference case was considered to represent our best 'assessment' of the likely range of stock dynamics.

New information made available during the Third Meeting of the Management Procedure Workshop (April 2004) highlighted a few possible problems with the reference set (quoted from the Report of the Third Meeting of the Management Procedure Workshop):

- The reference case model predicts with near certainty that SSB and CPUE will decline under current catches, a prediction that is inconsistent with the advice on stock status provided in 2001 through 2003.
- It appears that the operating model may be too certain in its predictions: recruitments estimated for 1998-1999 show unrealistically narrow confidence bounds in MCMC trials.
- Predictions of CPUE declines are inconsistent with new **nonstandardized** CPUE data for 2001 and 2002, which fall above the 95% confidence bounds estimated from the reference set[1].
- It was pointed out that the lack of small fish in the Japanese LL fishery in 2003 and recent drops in the Indonesian catch are likely to be inconsistent with model predictions.

In spite of these discrepancies and problems in model structure, the meeting concluded that the reference case model proved to be valuable as a tool for evaluating performance of different MPs. The report goes on: "In particular, the results provide an adequate

[1] It was noted that this may change once the medians of the four standardised CPUE series for 2001 onward become available.

representation of trade-offs between different performance statistics associated with the different candidates" (i.e. candidate management procedures). The problems identified above, meant that the estimates of probabilities of meeting different stock rebuilding targets can only be considered to be reliable in relative terms. Nonetheless, the meeting agreed that the estimates of probabilities based on the reference case represented our best estimates of the relative risks involved in following different management options.

The Workshop recommended that the operating model (reference case) be updated with the new data up to 2003, and re-evaluated for consistency of the operating model predictions with the assessment results. In particular, the SAG should consider whether the current operating model structure, updated with the new data, is an adequate model to use for computing probabilities of meeting rebuilding targets under different candidate MPs.

The mechanical update

Updating the reference case with new data up to 2003 was called "the mechanical update" (MU). During the data exchange it became clear, however, that it was not just a simple matter of adding the new data points. There were some errors in the data originally provided to the MP in the early stages of its development. There were other changes to the data which are discussed more fully in CCSBT-ESC/0409/27. The main point to note here is that the historic data also contained some changes compared to what was used for the reference case as it stood in April 2004.

It was clear from the results of the MU (information by e-mail, Haist 2004) that the mechanical update did not provide an adequate fit to the new dataset. There were apparently strong residual patterns in CPUE, particularly for the 'omega'=0.75 runs[2] (low75, med75 and high75). It is interesting to note that it's only for the 'low1' case (low steepness, omega=1) that the estimated spawning biomass and recruitment series are similar for the reference set and the revised set (up to 2000/01). The differences between these runs for all the other models show that the updated historic data have a strong effect on results.

The panel also concluded that the performance of the MU was unsatisfactory, and identified the following problems (briefly summarised here) with the mechanical update (CCSBT-ESC/0409/42):

1) Natural mortality seemed unrealistically high and with unrealistically low CVs
2) Steepness estimates were very near the upper bound in three of the six scenarios that compose the reference set
3) MCMC runs that start with the steepness parameter at the upper bound moved away from the bound but did not explore the full interval on steepness
4) Selectivity for Indonesian fishery was estimated to be monotonically increasing from age 10 through to age 30.
5) The error bounds on the low recruitment estimates for year classes 2000-2001 appeared to be unrealistically narrow given that they must be based upon rather little data.

[2] Recall that omega defines the relationship between the cpue index and density. A value of 1 indicate a linear relationship.

Results of 'Panel_tag' and 'Panel_notag' runs

The panel had very limited time for analyses and particularly for evaluating MCMC runs which take around 36 hours to run. Two runs are, however, presented in CCSBT-ESC/0409/42 with the suggestion that they could form the basis for a new reference set. The document does not indicate or suggest how one might treat the two runs. For example, whether they should be combined with equal weights, different weights, and how weights should be obtained. In any case, it is worth commenting on those two runs, namely 'panel_tag' and 'panel_notag' in the context of their potential use to evaluate the performance of MPs.

The first question is whether those runs managed to solve or alleviate the problems identified in the mechanical update. Results for panel_tag show that:

- the estimate for natural mortality was M10= 0.17 (CV=0.08) and MCMC runs show a posterior distribution somewhat clustered at the upper bound (0.2)
- the error bounds on the low recruitment estimates for year classes 2000-2001 still appear to be unrealistically narrow.

This model does not suffer from the problem of increasing selectivity in the Indonesian fishery because selectivity has been forced to be constant after age 22.

The 'panel_notag' run forced natural mortality to be low by specifying a prior: $M10 \sim N(0.1, 0.03^2)$. Results show that:

- although the MPD estimate is 0.09, the MCMC runs have a mode at ~ 0.12-0.13 and the posterior is a narrow distribution toward the top end of the prior (i.e. the model still favours the higher M10 values)
- the error bounds on the low recruitment estimates for 2000 and 2001 still appear to be unrealistically narrow.

The selectivity for the Indonesian fishery is dome-shaped for this model.

In defining the reference set, it is important to decide what is actually estimable within the operating model. Thus, if it is decided that there is actually sufficient information to estimate natural mortality, it is important not to reject or adjust model structure simply to achieve some notion of what the values should be – particularly if this can have strong interactions with other aspects of the model. In this regard we note that although notions of demography and biology strongly suggest that values of adult mortality of 0.17 for an animal which matures at around 12 and lives up to 40 years or more, are unrealistic, there is little basis on which to argue that such values are not possible. In the context of the OM model structure, the data in fact strongly support a higher value if we assume it can be estimated. With regard to the choice of a reference set, which is primarily aimed at providing high robustness for testing MPs, it is therefore important to be careful about how the priors (especially those determining the relative weights in the objective function) are adjusted or fiddled with to simply achieve notions of realism when there is little information on which to base one's notion of realism. It is also important to give careful consideration to what is actually estimable, taking into account that the objective function is not a true likelihood.

The two runs raise a further question regarding the treatment of steepness and mortality. When the reference set was constructed, it was argued that within a steepness range (low,

med or high), mortality could be handled with MCMC. It was agreed that steepness, however, should only be handled with MCMC within each range, and a weighted combination of the different models should be formed. (In fact, the omega parameter was set to the value midway between the two extremes of 0.75 and 1.0, previously used.) The MCMC runs for panel_tag and panel_notag integrated over both steepness and M. The draft report suggests that this was primarily for practical reasons. Definition of a revised reference set should explicitly consider whether there is now reason to believe (from the new data, for example) that it would be appropriate to treat steepness and mortality in this way or not.

Issues raised by recent assessment runs with the OM

The operating model was used as an assessment model with the new dataset (see CCSBT-ESC/0409/23), and results from these analyses are relevant and potentially useful in guiding the development and choice of a revised reference set. This is discussed under four headings: Low CVs on recent recruitment, apparent high M10, MCMC treatment of steepness and M, and other issues. References to assessment results in the section below are to CCSBT-ESC/0409/23.

Low CVs on recent recruitment

There is little doubt that recent recruitments should be more poorly estimated than past recruitment because of the limited amount of information on recent compared to past recruitment (the notion of 'following' a cohort in the data). There is also, of course, the issue of data reliability or representativeness which interacts with the amount of data and assumptions about the consistency and smoothness of selectivity over age and time (i.e. within the modelling framework data can be made more or less "noisy" by varying the frequency of selectivity changes and the weights given to selective penalties) Nevertheless, noisy data for 10 years of a cohort could lead to as large a CV as less noisy data for 5 years of a cohort. However, one would not expect the model to be estimating recent recruitment with CVs that are substantially lower than those for any of the past recruitments unless "measurement" errors are dominating the CV estimates and this has been substantially reduced overtime. Of particular concern is the apparently low CV of the 2000 and 2001 recruitments.

One needs to be careful about inferring CVs from graphical output for which the expected value exhibit a large range of values. Even with constant CVs, the error bars for low values can give the appearance of being unrealistic small. Thus, two runs done as part of assessment investigations illustrate that the CVs of recruitment do in fact show an increase from about 1990 (Figure 1). In one example the values for 2000 and 2001 are 24% and 34% respectively, and in the other example, they are 31% and 43%. (These CVs are based on the estimates of recruitment and their standard deviations generated by ADMB's 'std' reports. It is not clear whether they include uncertainty in the stock-recruit parameters or not).

For the period where there are data to inform recruitment estimates, the CVs are generally between 10 and 15%. In relative terms therefore, estimated CVs for recruitment deviations in 2000 and 2001 are not that unreasonable compared to others in the time-series. Nonetheless, it would be important to consider implications for evaluation of MPs (i.e. projections) and decide which years' CVs of recruitment should be taken from estimates and which should be based on 'average' empirical, or other, considerations.

In relative terms, estimated CVs for recruitment deviations in 2000 and 2001 are not that unreasonable compared to others in the time-series. Nonetheless, it would be important to consider implications for evaluation of MPs (i.e. projections) and decide which years' CVs of recruitment should be taken from estimates and which should be based on 'average' empirical, or other, considerations.

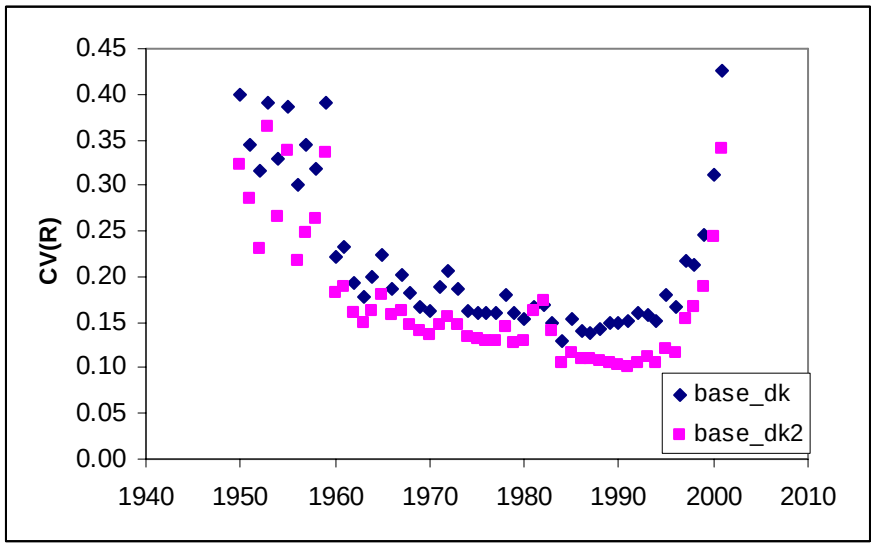

Figure 1 Examples of recruitment CVs for two runs 'base_dk' and 'base2_dk' in CCSBT-ESC/0409/23. The last two points on the right-hand side of the plot are for 2000 and 2001.

Estimation of steepness and natural mortality

The problem of rigorous model comparison and the evaluation of goodness of fit in the absence of a proper likelihood remains. On the basis of extensive assessment explorations and results we continue to have doubts about the ability of the model to estimate steepness and mortality (particularly both simultaneously). This will need to be considered when deciding on a revised reference set for MP evaluation.

Other issues

- Ages 4-30 vs 8-12 in CPUE fitting

Assessment results showed sensitivity to the age range over which catchability is assumed to operate when fitting to aggregated CPUE. We considered ages 4-30 and ages 8-12 in assessment runs since these two cases appeared to represent the 'extremes' in previous conditioning analyses. During MP trials, it was considered that the two options were equally plausible.

- Uncertainty in CPUE

Many of the assessment runs estimated σ_q at the lower bound of 0.1. This implies a surprisingly low CV. Runs based on the individual CPUE series also had relatively low estimates, for example, estimates of σ_q ranged between 0.107 to 0.187 for the 'base_dk' runs and between 0.125 and 0.276 for the 'base2_dk' runs. The empirical value was used in the projections to reflect what is considered to be a more realistic level of uncertainty (see Table

'Projection Code' in CCSBT/0409/42). Nonetheless, the low estimate could be interacting with other parameters or quantities which might reduce the uncertainty in projections. A related issue is the choice of median CPUE for conditioning. Although the individual CPUE series appear to be very similar, and although assessment results are quite similar, projections are quite different for the assessments using different CPUE series (see CCSBT-ESC/0409/23). Although it may still make sense to use the median in a management procedure (decision rule), this choice for conditioning may not be ideal, and may be losing some potentially important uncertainty.

Broader Implication for the Definition and Use of A Reference Set

The definition of a reference set in the Management Procedure Development process was linked to the integrated approach chosen for evaluating management procedures across a range of operating model scenarios. This reference set is meant to represent not only our "best" estimate of the range of future possible scenarios, but also "meaningful" and our "best" estimates of the absolute probability for each of the different scenarios included in it (and, implicitly, also the probability for those scenarios which are excluded). Given the approach that the SC has adopted, the recommendation and selection of an MP requires two fundamental decisions – the choice of a decision rule and the choice of a tuning level. Within the integrated approach adopted for evaluating MPs across a range of operating model scenarios, the selection of a tuning level requires that the results from the reference set tuning be used as measures of the "true" (or absolute) probability of achieving management objectives and the associated risks. Unless the results are treated as such, they contain little or no basis for deciding a tuning level. As we have seen in the results to date, the choice of a tuning level will be critical to overall performance and has large implications for near-term TAC levels. A high degree of sensitivity has been observed in the operating model, particularly to three years of additional data, revisions of data series, structural model uncertainty (e.g. A8-12 versus A4-30) and meta-priors (e.g. relative sample sizes). This high degree of sensitivity in the operating model raises questions about the feasibility of defining a robust, stable, reliable and yet accurate enough reference set that can be used with an integrated approach to provide absolute estimates of the probabilities and risks associated with the choice of a specific decision rule and tuning level.

This question of feasibility is worth considering with respect to some specific examples where the operating model has shown a high degree of sensitivity. The operating mode predictions about the current status (i.e. the conditioning results) are somewhat sensitive to the choice of ages 4 to 30 or ages 8-12 for using predicting the relationship between CPUE and abundance. More importantly, in the context of MP development is the sensitivity of the projection results under both constant catch and the various candidate MPs (see above and REF). At the 2003 SAG, the 4-30 and 8-12 age ranges were considered "equally plausible". There seems little empirical basis for selecting either one. A pragmatic, but simplistic solution would be to incorporate both the A4-30 and A8-12 cases into the reference set and give them equal weight. Alternatively, one could also use the value of the objective function to weight the two options assuming that the objective function can be treated as a true likelihood (although we know it is not). However, this would beg the question that the A4-30 and A8-12 options were only two out of a large range of possible and potentially "equally" plausible sets of hypotheses about how changes in selectivity reflect changes in catchability in standardized CPUE series. Consideration and inclusions of other hypotheses could be

expected to have substantial effects on the probability being estimated with the reference set and an MP. Simply excluding alternatives would be arbitrary, but given the lack of understanding and direct information on the interactions between selectivity and catchability it is not clear that we know how to capture the range of uncertainty – much less know the relative plausibility to assign to a potentially large range of alternatives.

The sensitivity of the projection results to the different standardized CPUE series raises even larger feasibility questions. In this case, the alternative hypotheses that have been considered and go into conditioning of the operating model is the median CPUE series. The median represents a rather limited set of hypotheses for CPUE standardization and assumptions about the relative SBT density in areas with no fishing effort. The four series that go into the definition of the median all provide very similar time series for the changes in relative abundance. The amount of difference among them is "trivially" small relative to the range of uncertainty in our ability to assess the status of the SBT stock. In fact, conditioning results for the different series yield only small differences in the estimates of the time series of recruitment and spawning stock biomass. However, the resulting projections can be dramatically different (see results in CCSBT-ESC/0409/23). This contrast between lack of sensitivity in the conditioning results and high sensitivity in the projections, violates an underlying principle or assumption that went into the selection of the reference set operating model. The definition process implicitly assumed that when only small differences were observed in the conditioning results for a range of options (particularly when it was associated with small differences in the input values) then those options could be collapsed into a single option for projections. By implication, it was assumed that inclusion of the alternative(s) and the weights assigned to them would have no effect on the results. The sensitivity exhibited to the different CPUE series clearly suggests that this is not always or necessarily true. Moreover, given that the different CPUE series are essentially alternative input data in the model, the resulting value of the objective function does not provide any basis for weighting the different alternatives. Thus, the basis for selecting the median of the four series (i.e. that it would not make any difference) does not provide a basis, in this case, for defining a reference set. (Note that it may still be feasible to use the median in a management procedure/decision rule.) Without more extensive exploration and testing it is not clear whether the bases for a number of other decisions are still valid or not.

The range of scenarios within the current (or an alternative, reasonably defined) reference set are likely to cover the reasonable possible future scenarios in terms of productivity, recruitment variability, data and other uncertainties. However, the estimated probabilities from a reference set of projections (e.g. the probability that B2020/B2002 > 1) will be sensitive to the perhaps "unresolvable" issues illustrated above. As such, the problem is more how to weight the scenarios rather than determining the range of scenarios that need to be included. This reflects the decision the SC made on how it would integrate across operating model scenarios. This has resulted in the emphasis on the reference set being to provide "best" estimates of probabilities. This in turn has shifted the focus from attempting to achieve "robust"/"acceptable" performance to attempting to achieve "optimal" performance out of the management procedure. The results from updating and refining the reference set, and from assessments, suggest that further discussion and consideration is needed on how best to evaluate and integrate results from testing management procedures across the range of operating models scenarios in the SBT context.

References

Haist, V. 2004. Notes on current SBT MP release, including the "Mechanical Update" of the SBT conditioning model with revised data through 2003. (Projection code version 10). Electronic file: Mechanical UPdate.doc July 2004.