Extracts from SESAME: A Simulation-Estimation Stock Assessment Model Evaluation Project Focused on Large Pelagic Species

D S Kolody
P C Jumppanen
D G Ricard
J R Hartog
A L Preece
T Polacheck

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

Abstract

The following working paper consists of excerpts from a draft of a large report:

Kolody, D.S., Jumppanen P.C., Ricard, D.G., Hartog, J.R., Preece, A.L., and Polacheck, T. 2004. SESAME: a simulation-estimation stock assessment model evaluation project focused on large pelagic species. CSIRO Marine Laboratories Report 241.

Only the Executive Summary, Conclusions and Recommendations are reproduced in this document. The study had a primary emphasis on exploring the application of various stock assessment models to simulated data from operating models parameterized to resemble alternative plausible representations of SBT fisheries. In many cases, the simulations included specifications that are poorly represented by most stock assessment models, but were identified to be plausible interpretations of SBT system dynamics. We attempt to make inferences about the statistical properties of the assessment models in terms of bias, variance and robustness of MPD estimators (and a minor exploration of uncertainty quantification using the inverse Hessian multi-variate normal approximation for approximating confidence limits). The study also presents some results from the Standing Committee on Tuna and Billfish Methods Working Group simulation-estimation study using a yellowfin tuna simulator. We encourage interested readers to consult the full document (an electronic version will be available from the lead author at the 2004 CCSBT SAG/SC, and can soon be downloaded from http://www.csiro.marine.au/cmr_pubs/reports/index.html).

1 CONTENTS

2 EXECUTIVE SUMMARY

The SESAME[1] (Simulation-Estimation Stock Assessment Model Evaluation) project was undertaken to provide insight about model formulation for pelagic fisheries assessment, and to consider the policy implications for Regional Fisheries Management Organizations (RFMOs) with respect to scientific advice provided from these models. Sophisticated stock assessment models currently attempt to integrate many different types of data into a single coherent framework that describes the population dynamics and estimates the impacts of fishing. These inferences are usually used to make recommendations to managers to assist in the attainment of management objectives. Pelagic fisheries data typically includes total catch in mass or numbers, frequency distributions of catch-at-length, -mass or -age, fishing effort, and, in some case, tag releases and recaptures. The relatively complicated integrative models that are used for these assessments have a number of potentially attractive features, but there are a number of issues related to the statistical properties of these models, and technical issues related to the implementation, that need further consideration. We identified several problems that were potentially important for the stock assessment of large pelagic fisheries, and simulated the assessment modelling process in an attempt to understand the relative importance of the different issues. Different modelling approaches were compared, and we make a range of recommendations based on the results.

The southern bluefin tuna (SBT) fishery provided the main emphasis for this study, in part because of the range of stock assessment models that have been applied to this species in recent years, and the absence of objective methods for synthesizing inferences across models. However, the SBT life history, fishery and data characteristics share many features with other regional Australian fisheries, particularly the tropical pelagic tunas and billfishes. A second major component of SESAME involved participation in the Standing Committee on Tuna and Billfish Methods Working Group (SCTB-MWG). This latter project involved collaboration with a number of international scientists with interests in the assessment of Pacific Ocean tuna fisheries other than SBT. The SCTB-MWG project was complementary to the work undertaken with our simulated SBT system, because it emphasized a different set of priorities, including the spatial dynamics of the fish population. The MWG project focused on a fishery simulator developed at the Secretariat of the Pacific Community Oceanic Fisheries Programme (SPC-OFP), and parameterized to represent plausible yellowfin tuna (YFT) dynamics in the Western and Central Pacific Ocean (WCPO). We include some preliminary results from the MWG project here, but the MWG is planning a more comprehensive analysis.

Both the SESAME SBT and SCTB-MWG YFT studies involved simulation-estimation methods for evaluating assessment models. In principle, this is a simple

[1] This project was developed under a proposal initially titled "Evaluation of complex population models used for the assessment and management of migratory fish stocks" and was re-christened Simulation-Estimation Stock Assessment Model Evaluation (SESAME) to avoid confusion with the mathematical definition of “complexity” that relates to systems that exhibit emergent behaviour, and is not directly relevant to this project.

concept in which operating models are defined to simulate the dynamics of fisheries systems including data collection. These operating models tend to be considerably more detailed than any stock assessment model and may include plausible processes that have not been, or cannot be, reliably quantified in the real world. Population models of the sort used in actual stock assessments are applied to the simulated data, and the quality of inferences are evaluated by comparing the assessment model estimates with the known values from the operating model. By repeating this process numerous times and with different assumptions, the statistical properties of the models (including estimator bias, variance and robustness to assumption violations) can be described and compared. In practice, there are a number of reasons why this methodology is not straightforward. There are purely technical issues related to the vast amount of data to be handled, computational time constraints and the difficulty in reliably automating complicated non-linear function minimization. And there are conceptual difficulties relating to the specification of operating models and assessment models, and the flow of information between the two (i.e. inevitably, subjective assumptions must be made in assessment models, and models with better assumptions should generally perform better, but how do we simulate the probability of analysts making good subjective assumptions?). We approached this study from the perspective of applied stock assessment practitioners, trying to understand what sort of limitations that we currently have, and the types of errors that we can expect to have made in the recent past. However, we did not attempt to simulate the whole assessment process. We evaluated various models under various conditions, but did not attempt to simulate the types of decisions that are normally undertaken when conflicting model results are observed in a real assessment.

We examined a range of assessment models, though not all were applied to every operating model scenario. The simplest models included Fox and Schaefer age-aggregated production models and Age-Structured Production Models (ASPMs). For the SESAME SBT scenarios, the more complicated models included the Statistical Catch-at-Age/Length Integrated Analysis (SCALIA) models originally developed for SBT assessment, and our application of MULTIFAN-CL. The SCTB-MWG YFT study involved application of several models (MULTIFAN-CL, A-SCALA and ADAPT-VPA) by individuals from numerous fisheries institutions, in addition to those applied as part of SESAME.

In undertaking this study, we had to strike a balance between examining many scenarios for general trends and identification of potentially troublesome situations, or looking at relatively few scenarios in detail, attempting to understand exactly why assessment models perform the way they do. The initial stages of the study suggested that the complicated assessment models often have unanticipated interactions between components that are not easy to explain, and different analysts have somewhat different views on what the important features are for evaluation. As a result, we opted for a more superficial overview of the types of problems that we might expect and present an archive of results from which further inferences might be gained. Our synthesis includes a number of observations relating to both general and fairly specific issues. Many of our conclusions are not entirely new, but there are few studies that have attempted to demonstrate and quantify assessment model performance as comprehensively as SESAME. In the report, we provide specific insights relevant to the assessment of SBT (and note that these issues are also applicable to the conditioning of operating models used for the evaluation of

Management Procedures). Conclusions and recommendations of more general relevance include the following:

1. The complicated integrative stock assessment models seem to provide reasonable inferences (and better than simpler models) when the model structural assumptions and data are good.

2. We found the assessment modelling estimation errors to often be larger than expected, particularly when operating models were parameterized with "difficult" (less than ideal, but not implausible) characteristics. The "best" point estimates were frequently very biased, and often highly variable, when assessment models were repeatedly applied to stochastic realizations from a given operating model. Some system characteristics (e.g. stock recruitment curve, natural mortality, temporal variability in catchability of the primary relative abundance index) usually could not be reliably estimated from the fisheries data that are generally available. Some inferences (e.g. current biomass relative to biomass at some historical point in time, recruitment trends prior to the last few years) were generally more reliable.

3. Inferences from complicated assessment models often tend to be sensitive to arbitrary assumptions. The model behavior can be misleading in ways that we would probably not anticipate without simulation testing. Simpler models often seem to provide more robust estimates than the complicated models when certain types of assumption violation are present.

4. Our attempts to estimate statistical uncertainty using the multivariate-normal approximation (from the inverse Hessian matrix at the mode of the likelihood-based objective function) were not very successful (i.e. the estimated confidence intervals were usually too narrow and did not encompass the known operating model values with the expected frequency).

5. We believe that there is scope for improving the statistical properties of these models, including the statistical uncertainty estimation conditional on the assessment model being "reasonably correct". Improvements might include: restructuring the likelihood function (e.g. using robust likelihood terms and random effects models) or applying bias correction methods. Uncertainty estimation would presumably be improved by using Bayesian posteriors and/or boot-strapping methods (the latter having the attractive feature that they are less sensitive to errors in likelihood functions). However, we fear that statistical improvements will probably never entirely resolve the fundamental problem that these models generally require too many arbitrary assumptions. For the time being, we recommend that scientific advice should place greater emphasis on the expression of model uncertainty rather than statistical uncertainty conditional on the model being correct. Research into methods for expressing uncertainty across models also should be continued. Similarly, diagnostic methods for comparing models should be evaluated in a simulation context, to illustrate the limitations that might be expected.

6. The age-aggregated production models, Fox in particular, performed better than expected under a range of circumstances. In the SESAME SBT

simulations, the Fox model generally performed as well as or better than the SCALIA models that estimated natural mortality, and seemed to be robust to some of the problems that produced bad behavior in the SCALIA models. The preliminary results from the SCTB MWG YFT study suggested that the Fox model performed as well as or better than the SCALIA and MULTIFAN-CL models for most or all of the operating model scenarios (in terms of relative biomass estimates). We found the YFT results particularly surprising, and question whether the operating model specifications provided adequate diversity to challenge the assessment models.

7. We were not left with a good impression of (at least our implementation of) age-structured production models. In both simulated SESAME SBT and SCTB-MWG YFT applications, they were prone to numerical problems, and generally required unrealistically good prior knowledge to yield performance comparable with the more complicated models.

8. Relative abundance indices (standardized CPUE) are likely the most important input for fitting most pelagic fisheries stock assessment models. The simple age-aggregated models seemed to describe the simulated YFT dynamics as well as the complicated models, while ignoring several auxiliary types of data (but this was less evident in the SBT simulations), presumably in part because the effort-fishing mortality relationship was very good. Temporal trends in catchability for the relative abundance indices produced serious problems for all assessment models in the SBT simulations, and attempts to estimate catchability variability were not very successful (despite reasonably good auxiliary data). This strongly suggests that effort standardization (or development of fishery-independent surveys), and quantification of uncertainty in abundance indices, needs to be one of the highest priorities for any stock assessment.

9. We would encourage a greater diversity of simulation testing to cover a broader range of problems that regularly challenge stock assessment analysts, including alternative exploitation histories, spatial dynamics, biological characteristics, and data characteristics. These studies would probably benefit from explicit consideration of several problems that we encountered here, related to the definition of plausible operating models, the handling of prior information that may be available to analysts, and the actual criteria selected for evaluating model performance.

Additional conclusions and research recommendations pertaining to the interface of science and management are described below.

Overall, this study leaves us with a deeper appreciation of the limitations of assessment modelling. This position of healthy skepticism seems to be growing in popularity among fisheries scientists in recent years, as exemplified in the words of Schnute and Richards (2001): *"Recent failures of important fish stocks give mathematical models a poor reputation as tools for fisheries management ... We recommend that modelers remain skeptical, expand their knowledge base, apply common sense, and implement robust strategies for fisheries management."* This theme underpins our advice for managers and policy makers with respect to pelagic

fisheries stock assessment modelling (a non-technical summary of issues relevant to managers is appended to the report):

1. Considerable uncertainty is inevitable with current methods of stock assessment. It is important that managers and assessment scientists continue to decrease their focus on "best" point estimates, and embrace the stock assessment uncertainty. We recommend that model structural uncertainty should be explored with primary importance, while statistical uncertainty conditional on the model being "correct" should be secondary (unless the inferences are robust to the major plausible structural uncertainties). The complicated integrative models are useful for expressing the uncertainty about the stock status and implications of management actions, while simple models do not have sufficient structural flexibility for achieving this (although, in many cases, the simple models may yield point estimates of comparable quality to the complicated models).

2. Assessment scientists and managers should work together to identify methods for managing the fishery that are robust to the major underlying and foreseeable uncertainties. Formal Management Procedure (MP) development (or Management Strategy Evaluation) is growing in popularity and seems to represent a promising method for achieving this objective. MPs have a distinct advantage in that they quantify the risk of the combined assessment and management, within a feedback control system (classical assessments generally assume a pre-determined pattern of future catch or effort in fishery projections, which is not an adequate representation of how effective fisheries management generally works). MPs are also evaluated using performance measures that should be readily defined from management objectives (whereas assessment model evaluation such as we have undertaken in SESAME, might include many estimators that are largely irrelevant, depending on the type of management decisions that are required). In an MP context, the complicated assessment models would play an important role in conditioning the operating model used to simulate the uncertainty in future fishery dynamics, and should play a role in monitoring the performance of the MP at periodic intervals. In this manner, there would be no need for a comprehensive application of the complicated integrative models every time that a management decision is required. Simple models, or even data-based stock status indicators often seem to provide an excellent basis for making short-medium term decisions once they are "tuned" to be robust to the major uncertainties identified in the operating models. However, it still remains to be seen whether operating models can be reliably specified to adequately represent most fisheries systems.

3. Management decisions should focus on reference points that can be reliably estimated to the extent possible. e.g. MSY has a convenient theoretical interpretation, but if we cannot estimate it, it might not be of much practical use. In contrast, we seem to have more success estimating relative biomass, which suggests that the 1980 biomass rebuilding target in the CCSBT might provide a reasonably quantifiable target.

4. As the emphasis on stock assessment shifts from the traditional provision of advice, toward the development of management strategies that are robust to uncertainty, there needs to be an increase in the amount of interaction between scientists, managers and industry. Without effective communication of industry priorities and management objectives, scientists are likely to impose their own value judgments into the process and potentially constrain the range of options under consideration inappropriately. Similarly, managers will need to become conversant with the concepts of uncertainty quantification and risk, to participate in the exploration of alternative management decisions (e.g. it will be important to be able to trade-off objectives of optimizing expected performance as opposed to providing a reasonable degree of robustness to unlikely events). The complicated models provide useful tools for these discussions, but they will never eliminate the difficult decisions that have to be taken to resolve conflicting management objectives.

5. A greater reliance on complicated models will probably require an increase in technically competent staff and resources for fisheries assessment. However, in the case of MPs, despite an initial increase in resources, an MP should be relatively easy to implement in subsequent years. Intensive reviews of operating models should only be required at periodic intervals, as management objectives change, unanticipated events occur, or substantially new data becomes available with which to evaluate the MP performance.

6. While there is an increasing recognition that more effort needs to be spent on quantifying fisheries model uncertainty, the methods for doing this are currently rather ad hoc, and would benefit from many avenues of research. Simulation-estimation studies evaluate the performance limits and data requirements of models in a known setting. Retrospective analyses evaluate the consistency of a given assessment model as data accumulates over time. Meta-analyses combine experience across fisheries systems. Goodness-of-fit diagnostics help decide when a model structure is incompatible with the data. While we are optimistic of the benefits of the shift toward uncertainty quantification, we also recognize that there is potentially a risk of over-emphasizing uncertainty, such that in the context of pre-cautionary management, this could lead to unreasonable loss of economic opportunity. Identifying the appropriate balance in uncertainty quantification remains a major challenge.

7. The quality of assessment model performance and uncertainty quantification increases as data improves. No amount of statistical wizardry or computational power can overcome the fundamental limitations of poor data. Data collection programs should strive for continual improvement (e.g. for the SBT fishery, direct ageing information should be collected and efforts should continue to find reliable fishery-independent abundance indices). However, not all data are equally informative, and given finite resources, there should be prioritization of data collection programs. Simulation studies are an important tool for providing guidance to this prioritization. In the quest for better data, it is often not recognized that a measure of the actual error associated with the data is also desirable (e.g. statistical models usually require assumptions about the relative reliability of catch length sampling, but formal analyses rarely

underpin these assumptions). If advice is expected with regard to fundamentally new objectives (e.g. ecosystem management), then there will probably be requirements for fundamentally new data (e.g. through fishery-independent observational studies).

5.13 CONCLUSIONS AND RECOMMENDATIONS

The following points attempt to summarize our main inferences in relation to the project objectives as defined in the Introduction.

1) Evaluate the performance of Statistical Catch-at-Age/Length Integrated Analysis (SCALIA) models in relation to the advice and stock status parameters needed for the formulation of management policies, with particular emphasis on the SBT fishery.

- The SESAME simulations indicate that the complicated integrative stock assessment models can provide reasonable inferences about stock dynamics under the right conditions, but there can also be large inferential errors even when the data are unrealistically good, and assessment model assumptions correspond closely to the true underlying dynamics of the system. The assessment model with the specification that we might expect to be the best on the basis of the individual model components does not necessarily yield the best average performance, presumably due to subtle inconsistencies that inevitably arise in model abstraction, complicated interactions among model terms and limitations to the information content of the available data. Model performance degrades considerably as data quality decreases, and when operating model dynamics deviate from assessment model assumptions in plausible ways. These simulations are qualitatively consistent with our observations in real assessment applications, in which inferences tend to be sensitive to arbitrary model assumptions.

- The inevitable model sensitivity leads us to support the view that the provision of stock assessment advice should be focused on illustrating the major uncertainties in the system and developing robust management strategies for coping with this uncertainty. It is unlikely that any single stock assessment model specification can meet the demands of this objective. However, integrative modelling frameworks that have the structural flexibility to admit the potentially important characteristics of the fishery provide the best tool with which this can be attempted. Formal Management Procedure development represents a promising method with which robust fisheries management might be achieved, and we expect that this approach will continue to become more popular in the future.

2) Evaluate performance of assessment models with respect to:

I. Stock and recruitment relationship estimation

- The SBT simulations suggested that the stock recruitment relationship is difficult to estimate, even with seemingly good data, substantial contrast in SSB and the known functional form of the relationship. The majority of SCALIA models were generally able to distinguish high productivity from low on average, but there was generally an under-estimation bias. The precision was not encouraging, especially when substantial recruitment auto-correlation

was present; such that we would not be surprised if the point estimates were very bad in any individual application. Our applications of MULTIFAN-CL to the SBT scenarios suggested a strong over-estimation of productivity. The quality of the MPD steepness estimates deteriorated as the data quality decreased and plausible assumption violations were introduced.

- The SBT simulations suggested that SCALIA models quantified the recruitment variability reasonably well (empirical CV slightly low, and auto-correlation slightly high) even if the input variance was poorly specified. However, substantial auto-correlation in the operating model resulted in a substantial under-estimation of the recruitment variability.

- The assumption of a (somewhat) incorrect stock recruitment relationship did not make much difference to the limited number of assessment model inferences that we were able to evaluate. However, this was a very limited test, and we would not expect this to be true in general.

II. Catch under-reporting biases

- The SBT simulation trials indicated that a consistent 20% catch under-reporting bias in any single fishery (juvenile, longline feeding or longline spawning) might not have a large effect on the assessment results (relative to some of the other factors explored). We expect that a temporal trend in the magnitude of the reporting bias would have been more realistic and problematic (particularly if CPUE from the affected fishery is used as a relative abundance index), but this was not examined.

III. Age estimation from cohort-slicing vs: Catch-at-Length

- The SBT simulations suggested that, when data are very good, age estimation from cohort-slicing results in some unsurprising errors in recruitment estimation (high variance in the estimates of individual recruitment events, and inflated auto-correlation in the recruitment deviations, relative to catch-at-length models). But we could not conclude that the biomass and management-related estimates were any worse than similar catch-at-length models. Performance differences between catch-at-length models and cohort-sliced catch-at-age models were less evident under the more difficult assessment conditions. However, given current computing power and modelling methods, it is not clear why one would prefer to use cohort-slicing.

- In the SBT applications, MULTIFAN-CL did not seem to perform as well as the similarly parameterized SCALIA models, and we suspect that part of this might be due to the fact that MULTIFAN-CL was not using the direct-ageing data that was available. For long-lived species, we expect that direct age estimation data will always be much more informative than size data.

- It was not obvious that large, truly random, catch-at-length samples (1000) were more informative than small samples (50), perhaps in part due to subtle differences between the operating model dynamics and assessment model assumptions. We note that this is not a justification for reducing catch-at-

length sampling programs, because it is very difficult to obtain unbiased fishery length samples without an extensive program. However, this result might suggest that catch-at-length representation in the assessment models can be improved.

IV. Unrecognized changes in SBT length-at-age

- Assessment models that relied on catch-at-length data suffered from serious estimation biases when the length-at-age distribution of the simulated SBT stocks changed in the early part of the time series (but was assumed constant in the assessment model). The effect was negligible for the models that did not use the catch-at-length data. The potential implications should be explored explicitly in the next assessment at the CCSBT-SAG.

V. Fishery selectivity assumptions

- We found that the assessment performance was surprisingly unaffected by the SBT operating model scenarios with systematic temporal variability in selectivity. A sudden sustained shift in longline selectivity does cause predictable estimation errors for assessment models that assume that it is constant, but estimating selectivity variability can account for the change reasonably well. However, we did not test if this remains true when multiple fisheries change their selectivity simultaneously. Conversely, in the MWG YFT simulations, we made a limited attempt to estimate selectivity temporal variability to compensate for the absence of spatial structure in the assessment model, and this was not very successful.

- We simulated a form of size selective fishing mortality in the SBT fishery, and found that the implications were negligible for the assessment models that used age-based selectivity. More troublesome size selective mortality scenarios could undoubtedly be defined, but we consider this to be a low priority for SBT.

VI. Fishery catchability (reliability of CPUE as a relative abundance index)

- Most of the complicated assessment model specifications had serious problems in the SBT simulations when the main longline fishery had an increasing catchability trend (including different variations of SCALIA and MULTIFAN-CL). The problem was more serious than expected given the magnitude of the trend, and suggests some curious model interaction; possibly with the tagging data. The production models and SCALIA model without tagging data were the least affected. Other forms of temporal variability in catchability posed less problem for the assessments.

- The simulations suggest that the relative abundance index is probably the most important data in all of the scenarios examined. There is probably limited capacity for reliably estimating trends in catchability for the main relative abundance index within these models (at least with the data history available for SBT). This strongly suggests that quantification of uncertainty in the relative abundance indices should be a major focus in any stock assessment.

VII. Spatial structure of the fish population and fishing fleet

- We relied on the spatially dis-aggregated SPC-OFP YFT simulations to make inferences about likely spatial effects in pelagic fisheries assessment. The results from this study are still under investigation under the direction of the SCTB MWG. Our preliminary results suggest that the Fox model seemed to provide performance as good as, or better than, the complicated models (MULTIFAN-CL and SCALIA) in most cases. The SCALIA models performed the worst when it was assumed that the relationship between effort and fishing mortality was not very reliable, but simply giving higher weight to the effort data seemed to bring SCALIA performance into line with the other models. These results support our assertion that the relative abundance index is the driving factor in these models, and that catchability trends are difficult to estimate. Given the apparent success of the Fox model using global nominal CPUE as a relative abundance index, we question whether the YFT simulator was appropriately parameterized to test interesting spatial issues.

VIII. Uncertainty Quantification

a. Estimator Performance

- This is addressed under Objective 1, and I – VII above.

b. Statistical Uncertainty Estimation (conditional on a model)

- The confidence intervals estimated by the SCALIA model (calculated from the inverse Hessian multi-variate normal approximation) did not encompass the true quantities from the operating model with the expected frequency (i.e. confidence intervals were much too narrow), even for the most well-behaved operating model. We expect that this effect will be even greater for real stock assessment applications, because assessment model assumptions will generally not be as good as these test conditions. Other methods of uncertainty estimation might be more successful, but we expect that the performance of approaches that are dependent on the interpretation of the objective function as a true likelihood will usually be limited by substantial biases in the estimators.

c. Model Uncertainty

- This study suggests that assessment model inferences are often likely to be sensitive to inevitable and arbitrary model assumptions, and this is consistent with experience in many real stock assessment situations. We consider that the representation of model uncertainty is more important than the expression of statistical uncertainty conditional on the model being correct. Formal methods for approaching this issue need further development, but we would prefer to see an ad hoc representation of model uncertainty than an elegant expression of statistical uncertainty that fails to admit a broad range of alternative interpretations that are consistent with the data.

d. Assessment Uncertainty and Fisheries Management

- This is addressed under Objective 6 below.

3) Compare the performance of SCALIA models with simpler age-aggregated and age-structured production models, and MULTIFAN-CL.

- The age-aggregated production models (particularly Fox) yielded results that were better than expected in most cases. In the SBT simulations, the Fox model was usually better than at least some of the more complicated models (e.g. SCALIA models that attempted to estimate natural mortality), and seemed to be robust to some assumption violations (e.g. unrecognized changes in the length-at-age distributions over time). From the preliminary results that we have available from the YFT study, it appears that the Fox model was comparable to, or better than, both SCALIA and MULTIFAN-CL in terms of relative biomass trend estimation in most operating model scenarios. Despite these apparent successes, we do recognize serious limitations in the usefulness of these models, particularly for quantifying uncertainty.

- We were not left with very good impressions of the Age-Structured Production Models that we explored. They were prone to an implementation error in most of the YFT applications. The stochastic recruitment version did not converge reliably in automated applications. The deterministic recruitment version performed well in many of the SBT simulations, but only when provided with excellent prior knowledge of both natural mortality and fishery selectivity. Implementing stochastic recruitment and additional external analyses to estimate selectivity detracts from the simplicity that was part of the underlying appeal of these simple models.

- The SCALIA models probably performed the best of all the assessment models for the SBT simulations when the data were very good and assumptions adequately satisfied. However, the SCALIA models were more sensitive to some assumption violations than the production models (temporal variability in length-at-age, catchability trend), and did not perform well when natural mortality was estimated. The SCALIA models were generally not as successful as the age-aggregated models and MULTIFAN-CL for the YFT simulations. A large part of this performance discrepancy appears to be related to the analyst assumptions about the relationship between effort and fishing mortality rather than fundamental problems in the general methodology.

- We recognize that MULTIFAN-CL is at the forefront of single species assessment model development in most respects, but would not yet want to see it universally adopted, if it meant the cessation of development of alternatives. Our limited exploration with the simulated SBT data suggested there are currently some features that are not well suited for SBT applications (e.g. inability to use catch-at-age data, although this is reportedly being addressed; inability to input time-dependent length-at-age relationships). We were not able to conclude from the SCTB MWG study whether migration dynamics can be reliably estimated, or what the data requirements would be for this to be possible (this may be addressed further at SCTB 17).

4) Participate in the Standing Committee on Tuna and Billfish Methods Working Group project designed to evaluate assessment models using a Western and Central Pacific Ocean yellowfin tuna fishery simulator developed by the Secretariat of the Pacific Community Oceanic Fisheries Programme.

- As part of SESAME, we applied various age-aggregated and age-structured production models and different SCALIA specifications to the simulated SPC-OFP YFT data in 2002 and 2003. We provide some preliminary results from these simulations (including conclusions above), but a more comprehensive synthesis is proposed for SCTB 17 in 2004.

5) Provide advice on the appropriateness and implications of these models for the provision of stock status advice in an RFMO context on SBT specifically, and tuna in general.

- It is probably inevitable that technically complicated models will be used to underpin scientific advice for most major pelagic RFMOs soon and for the foreseeable future. This implies that sufficient numbers of technically competent scientific staff will be required to run and interpret these models. However, mere adoption of these models is not likely to result in substantially improved advice to managers. Sophisticated models cannot make up for poor quality data, lack of informative contrast in the fishery history, or the need for arbitrary assessment model assumptions. However, we do think that these models provide a powerful tool for expressing uncertainty about the plausible states of the fishery that are consistent with the data.

- Management Procedures (MPs or Management Strategy Evaluation) might represent one of the best methods for defining and achieving management objectives that are robust to the major uncertainties about the status and future production potential of the fishery. This may include the use of complicated integrative assessment models in the role of operating models for simulating fishery dynamics. This has been the approach adopted by the CCSBT, and it seems to be moving in a positive direction. The results of the SESAME study are supportive of the directions taken in the development of the operating model for SBT Management Procedures. We observe that the CCSBT MP operating model was the result of explicit exploration of many sources of uncertainty, MP behavior was tested for robustness to the key uncertainties, and the final population representation encompassed the variability in the key structural uncertainties of several model specifications, to the extent possible given pragmatic time constraints. However, we do note with some dismay, that as of June 2004, the CCSBT had not yet reached final agreement on a set of operating models for testing candidate MPs, despite having an initial model implementation completed in Sep 2002. This approach is potentially a powerful tool for effective management, but cannot be expected to resolve disagreements about management objectives.

- For the CCSBT-SAG 2004, we recommend that models in the form of the MP operating model, or SCALIA, should form the main focus of model-based

assessments, and we encourage the exploration of structural extensions, as time allows. The results here suggest (and support previous recommendations) that additional attention should be given to the interpretation of CPUE as a stationary relative abundance index, consideration of the effects of historical changes to the SBT length-at-age distribution on the spawning grounds, and the exploration of alternative functional forms of the stock recruitment relationship. We note that the latter effect did not seem to be very important in the SESAME SBT simulations, but it seems to be an important issue and we are not confident that it was tested under suitably representative conditions. We expect that a changed emphasis in some of our modelling assumptions, and the addition of 3 additional years of data might lead to a substantially changed view of uncertainty as currently expressed in the SBT operating model.

- It is possible that the expectations placed upon complicated integrative models might continue to increase as sustainable fisheries legislation proliferates. We do not currently understand how well we can represent spatial dynamics in assessment models, or the data requirements for successful parameter estimation. This may become increasingly important in the design of spatial management strategies. Advice on multi-species trophic interactions may be expected soon, and there may be attempts to estimate these effects within these models. We would caution that considerably more testing would be required before we would have much confidence in the results. However, we also note that some forms of robust management might be achievable even in the absence of reliable stock assessment methods.

6) Provide a non-technical description of the key scientific issues and critical assumptions in SCALIA assessments that managers will have to deal with in negotiations and formulation of policy in the CCSBT and other tuna RFMOs.

- We have attempted to write the main text of this report with a minimum of equations and technical language, such that it should be reasonably accessible to most people with a background in fisheries, and a non-technical summary is appended to the main report.

Recommendations for future Research

- We have found these simulation studies revealing about the limitations that we might reasonably expect in our assessment modelling endeavors, and would like to see additional studies of this type with a broader range of participants, assessment models and operating models. It would be worth attempting to further improve our understanding of the relative importance of different population features (e.g. spatial structure vs: age structure) in different systems. Similarly, it would be worth trying to improve our understanding of the relative importance of different types of data. e.g. if the relative abundance index is truly the dis-proportionately important data under-pinning these assessments, it should also be the main focus for analytical effort and uncertainty quantification. Some sort of accessible repository for simulated

data sets would provide a useful means with which assessment modellers could benchmark their model performance.

- We would like to see more work done to evaluate assessment model diagnostics as they might be applied in a real stock assessment (i.e. examination of the quality of agreement between predictions and observations). Throughout SESAME, we were applying assessment models in an automated fashion, such that the results could not be interpreted with the benefit of common sense, experience and auxiliary information that would normally be expected in real stock assessment applications. We largely ignored this issue by framing the objectives in terms of the evaluation of particular models, as opposed to an evaluation of an actual assessment. An assessment generally involves the application of several models, usually with some attempt to choose among them, or integrate across them (based on fit to the objective function or otherwise). There are many possible approaches for examining the quality of fit between model predictions and observations, and the degree of statistical rigor varies. Given our general skepticism about the literal interpretation of the objective function as a true likelihood, it is not clear how useful these diagnostics are. But a formal expert system probably could be devised that would help to avoid some of the most serious assessment modelling errors.

- This study suggests that we can usually expect model uncertainty to exceed statistical uncertainty estimated conditional on the model being correct. However, there is a perception, particularly among statisticians, that major methodological improvements can still be made in assessment modelling. We think it is worth exploring the most promising avenues, including, 1) making likelihood functions more statistically "correct", 2) formally incorporating more of the model uncertainty within an integrated framework, 3) making the objective functions more robust to common assumption violations, and 4) developing an approach for dealing with conflicting inferences among different components of the data. We would also like to see a more comprehensive comparison of different methods for estimating statistical uncertainty. New developments would be particularly welcomed if they demonstrated performance improvements when evaluated against operating models that are suitably challenging to illustrate many of the difficult features that seem to afflict most real-life stock assessment situations.

- There should be more effort spent developing and evaluating robust management procedures. This will presumably involve improving methods for translating assessment uncertainty into operating models, developing creative solutions for controlling the distribution of fishing effort, balancing conflicting management objectives and expressing risks that cannot be reliably quantified. Ultimately, we expect that many of the problems of assessment modelling might plague MP development, but changing the emphasis from parameter estimation to management outcomes might focus modelling effort in more productive directions.

6 ACKNOWLEDGEMENTS

Many people made valuable contributions to the SESAME project directly or indirectly. Thanks to Marinelle Basson and Geoff Tuck for undertaking the onerous task of reviewing the mighty tome, and providing helpful discussion along the way. Thanks to Scott Cooper and Tim Jones for providing support for the design and maintenance of the databases used for SESAME. Many of the ideas explored were developed after discussions with John Sibert, John Hampton, Rob Campbell, Ray Conser, and other members of the Standing Committee on Tuna and Billfish Methods Working Group and the Commission for the Conservation of Southern Bluefin Tuna Scientific Committee. Thanks to Marc Labelle for providing details of the YFT simulations and MULTIFAN-CL results so that we could present a preliminary analysis of SCTB-MWG results here. Thanks to Toni Cracknell for guiding us through some boggy bits and Lea Crosswell for cover design. Agriculture, Fisheries and Forestry Australia (AFFA) and the Australian Fisheries Management Authority (AFMA) are gratefully acknowledged for providing financial support for the project.