

R00/0833

2 EXECUTIVE SUMMARY

The SESAME₁ (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 ageaggregated 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 overemphasizing 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).