CSIRO MARINE LABORATORIES
Report 241

SESAME: a simulation-estimation stock assessment model evaluation project focused on large pelagic species

Dale Kolody
Paavo Jumppanen
Daniel Ricard
Jason Hartog
Ann Preece
Tom Polacheck

CSIRO

June 2004

SESAME:

A Simulation-Estimation Stock Assessment Model Evaluation Project

focused on large pelagic species

Australian Government
Department of Agriculture, Fisheries and Forestry

CSIRO
MARINE RESEARCH

Australian Government
Australian Fisheries Management Authority

JUNE 2004

AFFA-FRRF Project No. 01/2438, and AFMA-MIRF Project No. R00/0833

■ Dale Kolody ■ Paavo Jumppanen ■ Daniel Ricard ■ Jason Hartog ■ Ann Preece ■ Tom Polacheck

SESAME: A Simulation-Estimation Stock Assessment Model Evaluation Project focused on Large Pelagic Species

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

Published by CSIRO Marine Research

DISCLAIMER
The authors do not warrant that the information in this report is free from errors or omissions. The authors do not accept any form of liability, be it contractual, tortious or otherwise, for the contents of this report or any consequences arising from its use or any reliance placed upon it. The information, opinions and advice contained in this book may not relate to, or be relevant to, a reader's particular circumstances. Opinions expressed by the authors are the individual opinions of those persons and are not necessarily those of the publisher or research provider.

Bibliography.
ISBN 1 876996 64 1.

1. Fish stock assessment - Australia. 2. Fishery resources - Australia. 3. Tuna fisheries - Australia. 4. Billfish fisheries - Australia. I. Kolody, Dale, S. II. CSIRO. Marine Research. III. Title. (Series : Marine research report (CSIRO. Marine Research) ; no. 241).

639.27780994

Preferred way to cite this publication:

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.

This is an electronic publication – limited paper copies printed by Piper Printing.

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).

3 INTRODUCTION

Fisheries stock assessment models continue to become increasingly complicated in an attempt to provide an ever more realistic representation of population dynamics and data collection processes, but it is not known whether the inferences obtained are actually improving our understanding of fish stocks and the quality of advice provided for fisheries management. This project was initiated to help understand the relationship between the type and amount of data available for the (single species) assessment of pelagic fisheries, and the quality of inferences that result when assessed with a range of models. The emphasis in this project is on Southern Bluefin Tuna (SBT) and to a lesser extent Yellowfin Tuna (YFT), two highly migratory pelagic species of economic importance to Australia that are shared with other regional fishing nations. Using a range of fishery simulations, we attempt to identify situations where modern assessment methods are likely to go wrong and where simpler methods perform well. We also make recommendations for regional fisheries managers and policy developers to consider in the application, interpretation and allocation of resources with respect to these models.

3.1 RATIONALE

With cheap computing power and efficient software, it is possible to model a diverse range of system characteristics that are expected to be important for understanding the abundance and distribution of exploited fish. The MULTIFAN-CL (e.g. Fournier et al. 1998, Hampton and Fournier 2001) development team has been at the forefront in attempting to describe tuna populations using data in the actual units of observation to the extent possible, including a flexible spatial resolution. This is theoretically attractive in that it allows the integration of several types of data into a single analysis, with a minimum of intermediate processing steps (e.g. catch-length frequency distributions can be used directly in the model objective function, where traditional VPA approaches would have required age estimates; dynamics of tagged fish are directly incorporated in the model, whereas independent analysis of tagging data might have been applied previously). This has the further advantage of allowing the integration of all the statistical uncertainty into a single coherent framework. MULTIFAN-CL is becoming the main assessment software for the tropical tunas of the Western and Central Pacific Ocean (WCPO), and similar models are being implemented by other Regional Fisheries Management Organizations (RFMOs) (e.g. A-SCALA in the Inter-American Tropical Tuna Commission IATTC (Maunder and Watters 2003); SCALIA and others in the Commission for the Conservation of Southern Bluefin Tuna CCSBT (e.g. Kolody and Polacheck 2001)).

Despite all of the attractive features of these sophisticated models, there are a number of scientific and technical concerns that need to be considered before embracing these models in all circumstances. These models usually estimate dynamics (i.e. hindcast estimates of system attributes) that correspond very well with observations. However, this is achieved in part because the models contain hundreds or even thousands of "free" parameters, and this leads to concerns that over-parameterization leads to over-fitting. There are usually a number of arbitrary and untestable assumptions required in stock assessment models to produce results that are consistent with prior perceptions about stock dynamics. When complicated models take many hours to fit, this can impede the exploration of alternative, but equally plausible assumptions that lead to different interpretations of stock dynamics, and this is especially a concern if assessments are conducted only during relatively short meetings when participants are assembled. The technical expertise required to develop and understand these models is considerable, and might have serious implications for how fisheries stock assessments are approached in the future.

Computer simulations provide the best means of evaluating and comparing the inferential performance of different stock assessment models. There is a recognized trade-off between model complexity and the reliability of parameter estimates (e.g. Walters 1986). In general, a complicated model potentially provides a more realistic structure for describing relevant features of the system, and should in turn provide less biased estimates than a simple model. However, more complicated models tend to have greater estimation variance, so the optimal model structure for a given estimation problem will generally be at some intermediate level of complication where the bias and variance trade-off favourably. Unfortunately, for fisheries systems, there is usually no way of knowing exactly where the trade-off is, because one never knows the true value of the feature of interest. Probably the best method for evaluating assessment models involves repeatedly simulating fisheries systems in which the data characteristics are “known”, and fitting different assessment models to the simulated data. This simulation-estimation approach is frequently used for stock assessment model evaluation (e.g. NRC 1998, ICES 1993) and is the basis of this work.

These concerns are relevant for Southern Bluefin Tuna (SBT) and provided much of the impetus for the SESAME project. The CCSBT Stock Assessment Group (SAG) failed to reach concensus on the status of the stock (particularly future productivity) in the late 1990s (e.g. Polacheck 2002), despite agreement on a general modelling approach. An independent assessment was tabled in the late 1990s, using one of the newer and highly parameterized assessment models (originally tabled as Hilborn and Butterworth 1996; this eventually evolved into Butterworth et al. 2003). In 2001, the stock status advice was based on an informal synthesis of results from a range of assessments, including the traditional ADAPT VPA (Hiramatsu and Tsuji 2001, Polacheck et al. 2001), age-aggregated and age-structured production models (Butterworth and Plaganyi 2001), an independent, extended implementation of Butterworth et al. (2003) (Polacheck and Preece 2001), a hybrid approach merging features of Butterworth et al. (2003) and MULTIFAN-CL (Kolody and Polacheck 2001), and a length-based VPA (Kurota et al. 2001). The SAG recognized serious model sensitivities in the traditional ADAPT VPA, and recognized that other models seemed to have better behavior (CCSBT 2001). While the model results were generally similar in their gross features, there were also conflicting inferences about the sustainability of current catches. The fact that catch levels could have been on either side of a critical limit magnified the perception of differences among models because 20 year projections at current catch levels yielded widely divergent stock size estimates. At the time, it was recognized that there is no general method for objectively synthesizing the results across models or ranking the performance of the models. It was hoped that by simulating the effects of key issues, the SESAME project would help to identify key sensitivities and guide model formulation issues in the future. Most of these SBT assessment issues are also directly relevant for the conditioning of the SBT operating model (Haist et al. 2002) that is being used for the evaluation of candidate Management Procedures (MPs). Many results have broader relevance for the tropical pelagic fisheries of interest to Australia, and stock assessment in general.

3.2 OBJECTIVES

The following list of objectives is paraphrased from the original SESAME proposal, and expanded to include relevant concerns that arose during the course of the project. Background for a number of the specific topics is provided in the following sections.

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.

2) Evaluate assessment models with respect to:

I. Stock and recruitment relationship estimation

II. Catch under-reporting biases

III. Age estimation from cohort-slicing, vs: catch-at-length

IV. Unrecognized changes in SBT length-at-age

V. Fishery selectivity assumptions

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

VII. Spatial structure of the fish population and fishing fleet

VIII. Uncertainty Quantification
   A. Estimator Performance
   B. Statistical Uncertainty Estimation (conditional on a model)
   C. Model Uncertainty
   D. Assessment Uncertainty and Fisheries Management

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

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.

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.

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.

3.3 Background to Specific Stock Assessment Issues for SBT and other regional tuna stocks

The topic areas introduced below were primarily motivated by actual SBT assessment issues, including the conditioning of operating models for the testing of candidate Management Procedures.

3.3.1 Objective I - Stock Recruitment Relationship Estimation

It is notoriously difficult to reliably quantify the relationship between fish stock size and recruitment for most fisheries for a number of reasons, including:

1) There are several functional forms for the relationship, that can be justified from population dynamics theory (e.g. Beverton-Holt, Ricker; with or without depensation at low stock sizes, etc.), but there are usually not sufficient data to distinguish which of them is more appropriate.

2) High variability in recruitment makes it difficult to identify the function with a limited number of observations.

3) There is often poor observational contrast – if the spawning stock biomass has not changed substantially over time, estimation usually requires a substantial extrapolation into unobserved regions of the relationship, regardless of how many observations exist.

4) The estimation procedure should account for uncertainty in both stock size and recruitment (i.e. some form of the Errors in Variables estimation paradigm should be invoked, as opposed to regression with one dependent variable).

5) Time series structure – the factors that drive recruitment variability (e.g. effective fecundity, spatial distribution effects on larval survival etc) are often driven by highly auto-correlated processes that reduce the effective number of observations, and potentially obscure the functional form of the SR (e.g. truly auto-correlated errors can be indistinguishable from systematic lack-of-fit),

6) Non-stationarity - past behaviour might not provide a useful indicator of future behaviour if the recruitment regime has undergone some fundamental change (e.g. an oceanographic effect or fishery affecting an ecologically-related species can change the trophodynamics of the target species).

Despite these problems, for the SBT stock (and most others), there is a precedence of attempting to estimate stock recruitment relationships. Without some method of estimating future recruitment, it is impossible to quantify future fishing impacts on the population (except for short term projections of long-lived species in which new recruits might form a minor part of the impacted population). Future projections, including those in operating models used for Management Procedure development, and many reference point calculations, (e.g. MSY) require some recruitment assumption. The SBT situation is perhaps atypical among tuna species, in that there is a strong indication that recruitment and spawning stock biomass have both strongly declined, and it is reasonable to assume that the two are in part related by a causal mechanism. However, it is also possible that effects independent of spawning biomass might be at least partially responsible (e.g. environmental regime shifts).

In many cases, assessment results include quantities that are dependent on stock recruitment relationships, even though the data are not sufficient to estimate them. Recent assessments for all the major tuna species of the WCPO include a stock recruitment relationship (e.g. Hampton and Kleiber 2003, Hampton et al. 2003, Langley et al. 2003), and any MSY calculations are dependent on some recruitment assumption. However, recognizing the difficulty in estimating the relationship, these analyses generally assume a Beverton Holt functional form and assign prior probability assumptions about the degree of compensation (steepness) in the relationship.

As part of the SESAME project, we explicitly attempted to examine how well the stock recruitment curves could be estimated for situations roughly resembling SBT. Preliminary results were presented to the CCSBT as Kolody and Jumpannen (2003). Specific questions that we attempted to address included:

1) If we are correct in our assumption of a Beverton-Holt functional relationship, how well can we actually estimate the steepness of the curve, and other quantities of interest for stock assessment?

2) What are the assessment implications of mis-specifying auto-correlation and the assumed variability of the recruitment deviations ?

Exploration of actual SBT assessment model fittings (in the context of operating model conditioning for CCSBT Management Procedure development) suggested that the relatively well-defined portion of the stock recruitment curve is linear, but not necessarily incompatible with high productivity (Polacheck et al. 2003c). If a Beverton-Holt curve is imposed, the linearity is only compatible with a very unproductive stock. This raises the question:

3) If recruitment is actually directly proportional to SSB up to a maximum level beyond which recruitment remains constant (a double-linear “hockey stick” function), then how well would this situation be approximated by a Beverton-Holt relationship, and what impact would this mis-specification have on the other assessment inferences?

The SCTB-MWG YFT simulation study did not appear to be designed to explicitly test the reliability of stock recruitment curve estimation, because (as we currently understand it) the underlying functional relationship was the same in all operating model scenarios. However, we do make some observations about the consistency of

SCALIA in estimating steepness in the YFT applications, and a comparison of how well SCALIA and the production models estimate MSY.

3.3.2 Objective II - Catch Under-Reporting Biases

Determination of total fishery removals was identified as one of the main priorities of the CCSBT Scientific Research Program (CCSBT 2000). There may be many problems estimating fishery-related mortality, e.g. due to extrapolation from incomplete sampling coverage, or unrecorded mortality due to discards or illegal fishing. In the case of SBT, there has always been concern about the quality of the catch statistics from non-CCSBT fishing fleets, and non-retention and mis-reporting by member nations. If changes in relative abundance cannot be properly linked to the total fishery removals, it is usually assumed that the reliability of a quantitative stock assessment will be severely degraded. In this study we attempted to address the following questions:

1) If there is a 10% or 20% under-reporting bias in one of the fisheries, how will this affect the assessment results?

2) How does the effect of the reporting bias differ if it is present in the juvenile purse seine, longline feeding grounds or longline spawning grounds fishery?

3) Can we minimize the impact of a substantial observation error in the catch component of the assessment model by allowing statistical catch uncertainty, or estimating natural mortality? Is there a negative implication of allowing total catch observation error when the catch actually is well described?

3.3.3 Objective III - Age Estimation from Cohort-Slicing vs: Catch-at-Length

Most stock assessment models for long-lived species represent the age-structure of the population and are ideally suited to integrate age composition data from catch samples. However, it is often technically difficult and expensive to estimate the ages of catch, and in many cases, the lengths (or mass) of fish are extensively sampled instead. In the case of SBT, length and mass samples are available for the majority of the Japanese and Australian fisheries historically, but only the Indonesian spawning ground fishery has substantial numbers of directly aged fish (via otolith annulli counts), and these have only been available since the 1990s. SBT assessment models have dealt with the absence of age data in different ways. Age-aggregated production models ignore the age composition data. The most common approach has been to use cohort-slicing to estimate the age-structure of the catch from the length frequency distribution. This is reasonably reliable for younger ages, but becomes less reliable with older fish because the length-at-age overlaps to a large extent. SBT ages 13+ are generally aggregated because of this effect. The third approach involves working with catch length frequency distributions directly. The assessment model attempts to get a good agreement between predicted and observed CL, thus eliminating the need for the intermediate processing step and not biasing the age composition due to the systematic errors in cohort slicing (although the related errors in converting from mass to length have been ignored to date). We were interested in examining the following questions related to these issues:

1. How do the models that use catch length and age frequency distributions compare with the age-aggregated models?

2. How do age-structured models that use cohort-slicing compare with those that use catch-at-length prediction?

In addition to these specific questions, we also make comparisons regarding assessment performance given different age and length sample sizes, and different assumptions about effective sample sizes (i.e. it is common to artificially downweight the actual number of samples in an assessment, to reflect the non-random nature of the sampling and/or to compensate for structural assumption violations known to be in the assessment model).

3.3.4 Objective IV - Unrecognized Changes in SBT Length-at-Age

The catch length frequency distributions on the SBT spawning grounds appears to have changed between the 1950s and the 1990s. Polacheck et al. (2003a) suggest that this could be due to a number of factors, including 1) differences in selectivity between the early and subsequent fisheries, 2) a sustained change in recruitment or mortality that resulted in a disproportionately small number of older fish on the spawning grounds in the early fishery, 3) sampling or measurement biases in the early fishery, or 4) a change in the length-at-age characteristics of the SBT population. The latter hypothesis is consistent with a density dependent effect resulting from intra-specific foraging competition. It is known that length-at-age of juveniles changed substantially between the 1960s and 1980s, so it is conceivable that similar changes have occurred in the adult population, but there are no data with which this can be directly examined.

An unrecognized change in the length-at-age distribution might have important implications for the assessment. Cohort-slicing requires a length-at-age distribution and will be misleading if this distribution is wrong. In principle, catch-at-length prediction could be used to estimate the length-at-age distribution (via estimation of multiple growth curves), but this has never been attempted in SBT assessment. And it is probably not worth attempting for SBT because the early spawning grounds fishery only includes the older fish with limited "modal progressions" that can be used to distinguish cohorts and quantify growth in younger fish. Thus catch-at-length models will also be adversely affected if a growth change is not recognized. To test this effect, we specified an operating model with a shift in the mean length at age in the earliest part of the fishery, and illustrate the likely assessment implications of assuming that growth has not changed.

3.3.5 Objective V - Fishery Selectivity Assumptions

Stock assessment models treat fishery selectivity in different ways. Selectivity refers to the combined processes which determine the manner in which the fishery catch age/size composition differs from the relevant fish population. Different gear types (or methods of deployment of the same gear) fishing in the same times/areas generally catch fish of different ages/sizes. Deployment of the same gear in different times/areas can also produce different catch composition because the fish are heterogeneously distributed; equivalently, consistent gear deployment in the same times/area will result in a selectivity change if the fish distribution changes. Fishers often intentionally change their targeting practices in response to changing market conditions or management interventions. In stock assessment models, there is usually a trade-off in the assumed relationship between the variability in selectivity over time, and the observation errors in the catch-at-age/size. At one extreme, the ADAPT-VPA interprets the catch-at-age to be known exactly, and estimates changing selectivity every time-step. At the other extreme, a purely separable VPA assumes that selectivity is constant over time, and interprets lack of agreement between predicted and observed catch age/size distributions as entirely observation error. Butterworth et al. (2003) were the first to illustrate intermediate interpretations that recognize both process and observation errors for the SBT fishery. In addition to dis-aggregating catch data into relatively homogeneous fishing fleets, they estimated temporal variability in selectivity using a random walk time series model. This is the approach adopted in SCALIA. The specific selectivity issues that we attempted to address in the SBT simulations included:

1. Fisheries can rapidly change their targeting characteristics due to economic conditions and/or management actions. In the case of SBT, this has been evident in the Australian fishery, as farming became established, and was evident in the Japanese longline fishery when restrictive quotas were introduced. We illustrate the assessment implications of assuming constant selectivity when it actually does change, and examine whether or not selectivity changes can be estimated reliably. We also examine the implications of attempting to estimate a change when selectivity actually is constant.

2. Fishery selectivity might change gradually in a systematic way, such as in relation to a changing age composition. Some fisheries dis-proportionately follow a strong cohort because it results in higher CPUE. Alternatively, in the case of the Australian purse seine fishery, there seems to be a preferential targeting of particular ages that are optimal for the aquaculture industry. This would tend to produce a constant catch-at-age/size composition over time, and limit the amount of information available about relative cohort strength. We simulated the former issue in an adult (longline feeding grounds) fishery, and the latter in a juvenile (purse seine) fishery.

3. Fishery selectivity is often thought to be a predominantly size-based process (e.g. net mesh sizes can allow smaller fish to pass through; hook lures appeal to fish with particular mouth size characteristics). Size selective mortality can have short and long term consequences. In the short term, the length-at-age distribution of fish can change, depending on the magnitude of the size selectivity effect and the magnitude of the fishing mortality relative to natural mortality. In the longer term, this could lead to a long lasting effect on the population genetics. For the SESAME SBT simulations, we were interested in the possible short term effects on the length-at-age distribution. Since the assessment models that are generally used tend to assume purely age-based

> selectivity, we are curious what effect size selective mortality would have for SBT stock assessment.

A major source of variability in selectivity potentially arises due to heterogeneity in the spatio-temporal distributions of fishing fleets and the fish population. Different fishing fleets usually have different selectivity characteristics, such that changes in the relative effort among fleets changes the global selectivity of the aggregated fishery. Models that dis-aggregate fishing fleets into units with relatively homogeneous selectivity should be able to provide a reasonable approximation to the global selectivity changes in this case, even if the selectivity of each individual fleet is assumed constant. Temporal variability in the distribution of the fish population will also affect the global selectivity of a fishery, even if selectivity is effectively constant within each sub-region of the fishery and for every individual fleet. MULTIFAN-CL can use spatial dis-aggregation with age-specific migration rates to potentially explain some of the heterogeneity in the fish distribution. If the model can adequately describe changes in the spatio-temporal distribution of the fish, and partition fleets into homogeneous sub-units, then the major sources of variability in global selectivity might be adequately described. However, there are also systems in which fishery selectivity changes in a manner that cannot be easily described by any practical dis-aggregation of fleets and sub-populations. The SBT longline fishery seems to be such a case, in which it seems that fishers can target substantially different age compositions without changing their fishing behaviour in a manner that is easily recognized in the spatial distribution of the data. In the SESAME SBT simulations, we did not explicitly simulate spatial selectivity effects, but we did simulate scenarios where selectivity changed for other reasons. In contrast, the SPC-OFP YFT simulations involved spatial structure and regional data dis-aggregation, so that global selectivity effects might also be recognized in an assessment model through spatial representation. This issue is discussed further under spatial representation below.

3.3.6 Objective VI - Catchability Assumptions for Relative Abundance Indices

One of the key inputs for most dynamic stock assessment models is some sort of relative abundance index, and for large highly migratory pelagic fisheries this is usually derived from commercial CPUE. There is a vast literature describing the problems of using commercial CPUE in this manner, however, there is often no alternative. Usually measures are taken to standardize effort data with the intent of making CPUE proportional to abundance (e.g. to account for spatio-temporal heterogeneity in local abundance/catchability or the relative effectiveness of different fishing gear). There are different methods for doing this, potentially yielding substantially different results, and there is usually no real indication of when standardization has succeeded. The implications of several approaches on the assessment of YFT are illustrated in Hampton and Kleiber (2003). It is possible to evaluate performance of different standardization methods relative to each other conditional on a given assessment model structure and the other data. In this manner, if the assumptions are correct, one might be able to correctly conclude that a given standardization method is preferable. But if the model is sensitive to assumptions, or if all standardization methods are subject to similar errors (e.g. unquantifiable efficiency improvements), the relative quality of fit might not be helpful for evaluating the absolute performance of standardization.

An alternative, or complementary, approach for dealing with potentially incomplete effort standardization involves allowing an integrative assessment model to estimate changing catchability over time. This can be done with various assumed functional relationships (e.g. catchability related to abundance or effort) or assuming a random walk time series model, which admits that catchability can increase or decrease randomly over time, under the influence of the other data. This latter approach is commonly used for many fisheries in a MULTIFAN-CL WCPO tuna assessment (but note that in this case, the widespread and fairly homogeneous Japanese longline fleet is generally not allowed temporal catchability variability). It is also a feature in the SCALIA assessment framework, but we have no real appreciation of whether or not time series changes in catchability can be estimated reliably, or the data requirements that would allow this to be accomplished. As part of the SESAME project, we tried to address the following objectives:

1) Illustrate the impact of changing catchability on stock assessment model inferences when it is assumed constant.

2) Investigate the ability of assessment models to estimate changing catchability using a random walk time series model, given a range of data quality (i.e. total catch, catch-at-age, catch-at-length and tag releases/recaptures).

3) Investigate the implications of allowing a stock assessment model to estimate changing catchability using a random walk time series model, when catchability is actually constant.

We defined operating models with different assumptions about the relationship between fishing mortality and effort to test these effects. As with the selectivity issue above, we note that the calculation and interpretation of catchability potentially could have a strong temporal component that might be attributed to spatial effects, and this is discussed under spatial structure below.

3.3.7 Objective VII - SCTB-MWG Assessment Model Evaluation Project and Assumptions about Fishery Spatial Structure

Around the time that SESAME began, the Standing Committee on Tuna and Billfish Methods Working Group (SCTB-MWG) encouraged participation in an independent project to evaluate the performance of MULTIFAN-CL and other assessment models using data generated by the Secretariat of the Pacific Community – Oceanic Fisheries Programme (SPC-OFP) yellowfin tuna (YFT) simulator (Labelle 2002, 2003). Other analysts were encouraged to provide assessments, and in addition to SCALIA and the production models applied as part of the SESAME project, A-SCALA (e.g. Maunder and Watters 2003) and an ADAPT-VPA (Bigelow 2002) were tested. Many of the ideas and methods of the MWG were adopted in SESAME to maintain compatibility. The YFT simulator provided a complementary extension to the SESAME SBT simulations. The independent operating model had a considerably different emphasis (as outlined in the Methods), and should give us a greater insight into our ability to make generalizations about assessment model performance. The final synthesis of SCTB-MWG YFT simulation results have not been completed at the time of writing, but we were able to present preliminary results from the application of production models, SCALIA, and a reproduction of some MULTIFAN-CL results of Labelle (2003). For the purposes of SESAME, we apply and discuss our assessment model applications primarily in the context of considering the implications of spatial structure as described below.

Fish populations and fishing fleets are never homogeneously distributed in space and this can have important implications for assessment and management. The heterogeneous distributions usually mean that different portions of the fish population are exposed to fishing gear at different times, and this affects the resultant global CPUE and/or global selectivity. If one assumes that selectivity and the relationship between effort and fishing mortality are constant over time, these spatial effects can lead to serious biases in assessment inferences, even though the assumptions might be reasonable for a given sub-population. From a management perspective, even if the global population is adequately described by an assessment model, there might be important spatial issues related to catch allocation or spatial regulations that still need to be addressed (e.g. if population mixing rates are relatively low, there is effectively a sub-population structure that could result in localized overfishing problems even though the global exploitation rate is low). Recognition of these problems encourages the inclusion of spatial structure in assessment models, with MULTIFAN-CL at the forefront for the representation of tuna populations. Spatial effects can also manifest themselves in other ways, e.g. via stock and recruitment relationships, natural mortality, and/or growth rates, but we do not consider these effects in this study.

It is fairly common for assessment models to use data dis-aggregated into fisheries with fairly homogeneous characteristics (e.g. operation in the same general area and/or with the same gear), but the global fish population is often assumed to be homogeneous and potentially vulnerable to any fishing fleet. This is the approach used in SCALIA, Butterworth et al. (2003) and A-SCALA, among others. This assumption is obviously incorrect in that no matter how much effort is applied, a localized fishing fleet cannot catch fish that are in a different area. But this might not be a substantial issue if the fishing pressure is low relative to mixing rates (or natural mortality rates). The structure of some assessment models (e.g. SCALIA) is designed so that both catchability and selectivity can potentially vary over time. In this manner, the model has sufficient freedom to potentially describe the global dynamics very well. However, in practice, it is not clear that temporal variability in global catchability and selectivity can be reliably estimated.

MULTIFAN-CL has taken the most ambitious approach to the spatial problem in that fishing fleets are dis-aggregated by gear-type and spatial area, and the fish population is spatially dis-aggregated with migration rates between areas explicitly modelled. In theory, this allows a more realistic representation of the population and potentially admits the sub-stock structure in a manner that might be useful for managers. However, it remains unclear whether the spatial representation can be appropriately defined. Statistical areas might have little relation to homogeneous population units, or interannual migration variability might overwhelm continuous mixing assumptions. And even if the model spatial structure and migration assumptions are essentially correct, we do not know what the data requirements would be for reliable parameter estimation.

Our Virtual Stock Model (VSM) operating model was designed with the capacity to simulate spatial dynamics, but we did not use this feature for the SESAME SBT simulation testing. We would argue that the spatial structure is less of an issue for SBT than the tropical tunas, and instead focused on spatial issues in the context of the SPC-OFP YFT simulator, which was explicitly designed and parameterized to operate on a fine spatial scale. We consider the spatial question a lower priority for SBT, because of the general perception that many of the stock characteristics are actually fairly homogeneous, despite the large geographic range of the population. Most SBT seem to have many common characteristics in their migration routes, and it seems plausible that a large fraction of the relevant age component of the global population are vulnerable when the major fishing fleets are active. Caton (1991) provides an overview of perceived SBT migration dynamics and fishery characteristics. Young of the year migrate along the west coast of Australia southward from the tropical spawning areas. Juveniles from about ages 1-5 feed in the Great Australia Bight (GAB) in the southern hemisphere summer where they are currently caught by the aquaculture purse seine industry (not all fish return to the GAB every year, but there is not much evidence to indicate what proportion return or whether the individuals differ). There is an annual migration of adults to the single spawning grounds in the tropical Indian Ocean near Indonesia where they are vulnerable to the spawning grounds longline fisheries. There are also substantial concentrations of SBT at reasonably consistent feeding grounds locations and migratory corridors that are targeted by the longline fishery. It is not clear what proportion of the global population is vulnerable in these areas, but given the broad coverage of the longline fishery, there does not seem to be much evidence that major portions of the stock have had a significant spatial refuge (at least since ~1970). This is an over-simplification of SBT dynamics, but the assumption of homogeneity of the population with respect to the fisheries is probably more reasonable for SBT than the tropical tunas. The tropical tunas seem to migrate and mix to a relatively limited degree given the broad range of the species (e.g. Sibert and Hampton 2003).

Furthermore, there is also the perception that SBT longline and purse seine fisheries have the capacity to substantially change their targeting practices without making large changes to the spatial region of operation, in which case dis-aggregating the data at a coarse spatial resolution would not result in units with homogeneous selectivity/catchability anyway. As a result of these factors, spatial questions in the SBT simulations were only explored implicitly at a global level, such that spatial heterogeneity in the fish or fleet distributions could arise as temporal variability in global catchability and selectivity. We attempted to examine spatial questions more explicitly in the YFT simulations, in addressing the questions:

1. How does the performance (in terms of global population inferences) of the spatially dis-aggregated MULTIFAN-CL compare with the fishery dis-aggregated SCALIA and the very simple age-aggregated and fishery-aggregated production models (Fox and Schaefer) ?

2. Can we account for the spatial issues by estimating temporal variability in catchability and/or selectivity in the SCALIA model?

We expect that more detailed results regarding the SPC-OFP YFT simulations will arise from further analyses by the SCTB-MWG.

3.3.8 Objective VIII - Uncertainty Quantification

Inevitably there are limits to the reliability with which attributes of fisheries systems can be estimated, and it is generally good practice to provide some description of this estimation uncertainty in addition to the "best" estimate. The impetus for fisheries scientists and managers to consider uncertainty about the status of fish stocks is entrenched in international agreements (FAO code of conduct for responsible fisheries, FAO 1996; UN agreement on the conservation and management of straddling stocks, UN 1994). The precautionary approach for fisheries management prescribes that management decisions need to be more cautious when uncertainty is higher (i.e. operate with a lower probability of causing permanent or long term changes to the system, which often means catching fewer fish). There are many methods for expressing uncertainty in fisheries assessment models (see review in Patterson et al. 2001), but they are generally poorly tested, and historically fisheries scientists have probably been guilty of under-estimating uncertainty about fish dynamics. However, it has also been suggested that the shifting emphasis on uncertainty quantification can be exploited to prevent management actions from disrupting the status quo (Schweder 2001). There needs to be an appropriate balance between the two types of management errors: taking disruptive action when there is no problem, vs: failing to act when there is a problem (e.g. Quinn 2003). The effective quantification of uncertainty and appropriate expression of uncertainty in a context for management decisions is an important modelling issue that remains unresolved. In the SESAME project, we attempted to examine assessment model uncertainty quantification in relation to four sub-topics:

A. Estimator Performance
B. Statistical Uncertainty Estimation (conditional on a model)
C. Model Uncertainty
D. Assessment Uncertainty and Fisheries Management

These are not mutually exclusive topics, but this partitioning forms a useful distinction for discussing the types of results presented in this report. We define the terms for our purposes in the following sub-sections.

3.3.8.A Estimator Performance

In this report, we refer to "estimator performance" as the degree of agreement between the "best" point estimates from an assessment model and the actual values. We generally refer to the parameter and state values at the Maximum Posterior Density (MPD) as the best point estimates. We use the term MPD rather loosely to mean the parameter estimates with the best global fit to the objective function, whether it comes from a strict Bayesian model with formal priors and posteriors, or a frequentist model, in which likelihood penalties might be invoked to express prior beliefs in a manner analogous to Bayesian priors. For any given assessment model and data set we would like to know:

1. How reliable are the MPD estimates (of historical stock dynamics, current stock status, management reference points, etc.) likely to be, in terms of the bias and precision that we can expect if the fishery system conforms closely with our assumptions, and how robust are the model inferences likely to be if the system exhibits plausible characteristics that are not consistent with our assumptions?

We address the question by repeatedly applying models to data sets simulated with stochastic variability. Thus, this is a form of uncertainty that usually cannot be quantified and expressed in a routine stock assessment using a single data realization. This topic is the primary underlying objective for most of the SESAME project, and is repeatedly examined in relation to the questions outlined in objectives I-VII (sections 3.3.1-3.3.7) above. We re-iterate the topic as a component of uncertainty quantification to emphasize that these results constitute one component of the uncertainty that should be admitted in the provision of advice emanating from a stock assessment model.

3.3.8.B Statistical Uncertainty estimation

The second topic that we consider under uncertainty quantification relates to the narrower issue of statistical uncertainty estimation, conditional on a single data set and a single model structure. This element of uncertainty often receives the greatest amount of attention in stock assessment, because it is usually perceived to be the most tractable problem, with a large body of supporting statistical theory. Different inference paradigms (e.g. Bayesian or frequentist; see Hilborn and Mangel 1997, Maunder 2003) lead to theoretically different measures of uncertainty, but they are generally interpreted in fundamentally the same way when it comes to making decisions. For complicated assessment models such as MULTIFAN-CL, confidence intervals for quantities of interest are most often estimated using the multi-variate normal approximation calculated from the inverse Hessian matrix at the MPD (combined with the delta method for quantities derived from the estimated parameters). There is a general recognition that these confidence intervals are probably not very good in many cases, but they are used as a rough approximation primarily for the pragmatic reason that they are computationally easy to calculate. Sampling Importance Resampling (SIR) (e.g. McAllister and Ianelli 1997) and especially Markov Chain Monte Carlo (MCMC) (e.g. Patterson 1999) methods for approximating the Bayesian posterior distributions are gaining popularity for relatively complicated models, but are not yet computationally practical for routine use on the most highly parameterized models. In the SESAME project, we only examined one aspect of statistical uncertainty estimation, in attempting to address the question:

2. Is the multi-variate normal approximation (as usually applied in models like MULTIFAN-CL and SCALIA) likely to provide a reasonable representation of confidence intervals for estimated quantities of interest?

3.3.8.C Model Uncertainty

We refer to "model uncertainty" as the general problem of model specification/selection, and the implications of making arbitrary assumptions in fisheries models. There is simply not enough information to estimate some system features reliably, and different subjective specifications often result in substantially different inferences (Schnute and Richards 2001). We define model uncertainty to encompass all of the assessment model assumptions that are imposed by the analyst using subjective judgment. This includes everything from the fundamental dynamic equations governing population dynamics, to the specification of observation error probability distributions. There are obvious cases where the issue of model uncertainty is easy to appreciate – e.g. how should one compare the inferences from a classical mass action model of population dynamics with an Individual-Based Model (IBM), when the two are structured completely differently and parameter estimates are based on different types of data? In this report, we are often interested in more subtle instances of model uncertainty, e.g. two models might be identical except for the assumed variances in process and/or observation errors.

As with Estimator Performance above, this issue is also an underlying theme throughout most of the report, within objectives I-VII, outlined in sections 3.3.1-3.3.7. The inferences that we make on this subject are derived from exactly the same results (i.e. examination of MPD bias, precision and robustness), except they are framed against a broader question:

3. Can we make some useful generalizations regarding model specification/selection for pelagic fisheries stock assessment?

We only examine this question from the point of view of MPD estimation performance, and do not consider the usefulness of model diagnostics for assessing the quality of fit between model predictions and observations for individual data realizations. Diagnostics are commonly used (in combination with prior beliefs) within an actual stock assessment during the process of model selection (to identify the "best" model) or model weighting (to average inferences across models). But, it was beyond the scope of SESAME to produce an automated expert system to simulate the model evaluation processes that generally occur during real stock assessment.

3.3.8.D Assessment Uncertainty and Fisheries Mangement

This project does suggest that there are often likely to be non-trivial limitations to the quality of inferences that we can expect from any stock assessment model. This view is getting broader recognition in the fisheries literature, and methods for dealing with the problem are emerging. The best methods for dealing with assessment uncertainty might be realized through changes in the traditional interface between science and management (particularly using formal Management Procedures), and this will likely impact on the manner in which stock assessments are conducted in the future. This heading was defined as a logical place with which we might attempt a broad synthesis of all the results from the project, discuss the results in the context of the recent literature and speculate on the implications for future stock assessment and fisheries management.

4 METHODS

4.1 SIMULATION-ESTIMATION METHODOLOGY

The simulation-estimation approach for evaluating statistical models is intuitively simple (Fig. 1). In the real world, we never know the actual state of a fish population, so we can never know exactly how well our stock assessment model has described the fishery. Instead, we use an operating model (e.g. Linhart and Zuchini 1986, Hilborn and Walters 1992) to simulate fish and fishery dynamics, and data characteristics that are known exactly. This is the general approach that has been used in many assessment model evaluation studies (e.g. NRC 1998, ICES 1993). The performance of stock assessment models for making inferences can then be evaluated by applying each model to the simulated data, as would be done in a real assessment, and comparing estimated values of interest (e.g. fishing mortality, stock biomass, etc) to the known values. Repeated application of the process gives some indication of the statistical reliability of the estimators. We should never believe that our operating model is a very accurate representation of the real world, but it should be sufficient for illustrating the relative importance of the different system characteristics of interest. At the simplest level, operating models can be designed to correspond perfectly to assessment model assumptions, and the resulting estimates should give an excellent indication of the estimator bias and variance that can be expected conditional on the assessment model being “correct”. But the more meaningful applications involve operating models that are considerably more complicated than the assessment models, and typically include plausible features that could not be quantified in the real world. In this way, it is openly recognized that the assessment model is “wrong”, and the simulations provide a measure of how reliable the model is likely to be when applied under more “realistic” conditions, including when untestable assumptions are violated.

While the overall approach is simple, many problems arise related to the specification of the operating model, the handling of prior knowledge about the underlying stock dynamics, the choice of performance indicators, and the difficulty in trying to automate the process of model selection/evaluation without a thorough examination of diagnostic output that would normally occur during a real stock assessment. We attempt to justify our decisions in the operating model descriptions, but serious unresolved problems remain and are detailed in the discussion of Methodological Limitations (5.12). The operating model specification, assessment models selected and performance indicators for evaluation are all described in the following sections.

SESAME required a complicated organizational framework for handling a large amount of simulated data and results. Several independent pieces of software were involved, and standardized file formats were required to integrate everything together. Results from ~25 different operating model scenarios are included in this report. For most SBT scenarios, there were 10 stochastic state and data realizations generated (40 for the 2003 SCTB-MWG YFT study). The number of assessment models applied varied, depending on the operating model scenario (usually ~10). Thus, the results presented consist of several thousand assessment model fittings run over the course of this project.

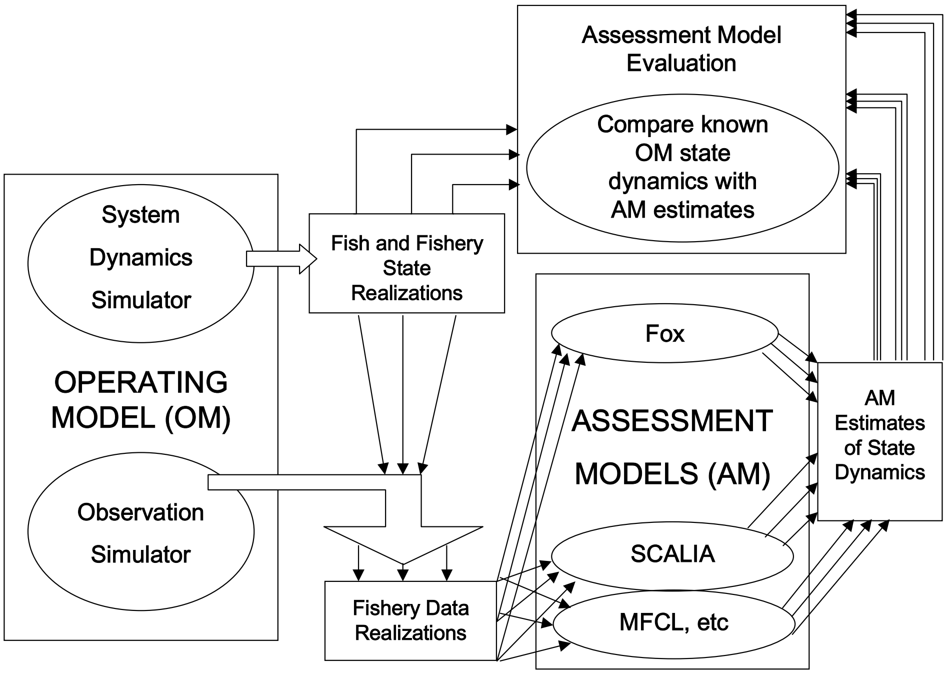

Assessment Model Evaluation
Compare known OM state dynamics with AM estimates
System Dynamics Simulator
Fish and Fishery State Realizations
OPERATING MODEL (OM)
Observation Simulator
Fishery Data Realizations
Fox
ASSESSMENT MODELS (AM)
SCALIA
MFCL, etc
AM Estimates of State Dynamics

Fig. 1. **Outline of simulation-estimation methodology for stock assessment model evaluation.**

Given the time constraints, we had to strike a balance between the number of operating model scenarios to explore, the number of state and data realizations to generate, the number of assessment models to apply, and how thoroughly to investigate the specification effects of different assessment models. We chose to examine a larger number of operating and assessment models to get a broad overview of issues that are likely to be problematic in real assessments. The downside of this decision was that we used fewer stochastic realizations than we would have liked (limiting the statistical power of the study), and we probably have less confidence that we know how to resolve the problems that were identified because each scenario was not explored in detail.

We note that the simulation-estimation approach is related to, but distinct from, the Management Procedure (MP or Management Strategy Evaluation, MSE) approach that can be used to evaluate and compare assessment models within the context of an overall management plan (e.g. Punt 1996, CCSBT 2002). The MP approach also involves hidden operating models that are used to simulate fish and fishery dynamics and data collection. But the MP approach involves a feedback-control cycle in which management decisions (e.g. TAC setting), population dynamics and fishery events are simulated iteratively within each fishery realization. In the SESAME project, each stock assessment model is fit only once to a given data realization, and not updated with additional information. In some ways, the MP approach is more attractive for evaluating assessment models, because the performance indicators should be more readily defined from management objectives, and the iterative application to a changing stock gives a better idea of performance under systematically changing conditions that are likely to be encountered as a fishery evolves. However, the iterative nature of MP applications means that any decision rule based on an assessment model requires refitting the assessment model each time a decision is required (i.e. 30 fittings for one 30 year projection), in combination with many stochastic realizations to represent uncertainty in future dynamics. In the case of CCSBT MP development, this involves tens of thousands of model fittings, and this is computationally prohibitive for complicated models. Fortunately, it seems to often be the case that very simple assessment models, or even data-based decision rules, work as well or better than sophisticated assessment models within an MP framework. However, it is also usually the case that the MP framework is predicated upon a sophisticated assessment used to specify the operating model(s), and quantify the uncertainty about the current stock status and future dynamics. Thus the effectiveness of the MP evaluation procedure is potentially limited by the quality of the operating model, and the results of studies like SESAME are directly relevant to this operating model conditioning.

4.2 Operating Models

4.2.1 VSM: a generic fishery simulation model

The Virtual Stock Model (VSM) software package was designed as part of SESAME to potentially simulate a broad range of fish populations and fishery dynamics, and was parameterized to represent several alternative representations of the SBT fishery system. VSM consists of two distinct parts: the system dynamics simulator, which describes the fishery and population dynamics, and the observation simulator, which simulates the data collection process. The underlying dynamics are based on fairly standard fisheries modelling assumptions and implemented using difference equations iterated on an arbitrary time step. VSM dynamics represent processes that are commonly found in sophisticated stock assessment models, including age structure, stock-recruitment relationship and multiple fisheries with distinct effort patterns and selectivity. Fisheries observations potentially consist of fishery-specific total catch, catch-at-length, catch-at-age, effort, tag release events including lengths at release, and tag recoveries (including release event and length-at-release). Research surveys can be implemented as special cases of fisheries if required. VSM also includes many plausible features that are not likely to be found in an assessment model, including arbitrary spatial structure with migration dynamics, and temporal variability in catchability, selectivity and size-at-age. The software was designed with future development in mind, and thus includes a number of additional features that have not been thoroughly tested to date, including multi-species predator-prey dynamics (fishing is actually an implementation of predation). Implementation details are described in Appendix 1 (VSM technical description).

4.2.2 VSM Parameterization of the fishery operating model to resemble the SBT system

The VSM simulator, was defined to qualitatively resemble the SBT fishery (from around 1950-2000). However, there was no explicit conditioning to actual data. The exploitation history of the fishery was intended to be similar in the different operating models, but the variable production dynamics in the different scenarios limited the extent to which this could be achieved. The specific details are described in Appendix 2 (VSM parameterization for a fishery resembling SBT). Two baseline operating model scenarios (E_base and D_base) were defined, and all other models were derived from these two. E_base is the easy scenario, in which the underlying dynamics and observation characteristics are all probably better than we could ever hope for in the real world (although the exploitation and data collection history are less than ideal, and the fine temporal scale of iteration ensures that no assessment model examined conforms perfectly to this structure). D_base is the difficult scenario. We would hope that the real world is not as difficult as D_base, but arguably each individual feature is probably not unreasonably perverse. The biological parameters generally conform closely with the assumptions used in SBT stock assessment (e.g. Preece et al. 2001). Qualitatively, the main features of E_base included:

- spatially-aggregated
- dynamics are iterated in monthly time-steps (fishing, growth, and mortality)
- data aggregated in annual units
- spawning (and age 0 recruitment) occurs instantaneously every 12 months, beginning on month 1
- 4 fisheries with selectivity and catch/effort characteristics roughly corresponding to: early (Japanese) longline on spawning grounds, (Japanese) longline on feeding grounds, late (Indonesian) longline on spawning grounds and (Australian) juvenile fishery; since the model is spatially-aggregated, the actual location of each fishery is irrelevant, but spatial effects are implicitly present in the selectivity (e.g. immature fish are not available to the spawning grounds longline fishery, but are harvested by the feeding grounds longline fishery).
- 50 year exploitation history resembling SBT, with largest catches on the spawning grounds in the first 15 years, followed by increasing catch in the feeding grounds and juvenile fishery, drastic cuts to these fisheries after about 40 years, and an increasing spawning ground fishery in the last 10 years
- Fishery selectivity is purely an age-dependent process and is constant over time for all fisheries.
- natural mortality vectors are unchanging over time and decreases with age (except for senescence in the oldest ages)
- knife-edged maturity at age 10 (0% of ages 0-9 spawn, 100% of ages 10+)
- spawning potential is directly proportional to mass-at-age
- mean length-at-age is constant over time and follows a von Bertalanffy growth curve
- length-at-age is normally distributed with variance slightly smaller than estimated in Polacheck et al. (2003a)
- mass-at-length (and age) is constant over time

- recruitment follows a Beverton-Holt stock recruitment relationship with a CV ~ 0.4 for recruitment deviations.
- initial population was unfished, and in a random state determined by the stock-recruitment relationship
- The relationship between effort and fishing mortality was only reliable for the longline feeding grounds fishery (annual CV ~ 0.1, no temporal trends in catchability). Effort was intentionally misleading for all other fisheries, and all assessment models explicitly recognized this.
- catch-at-length data available for all fisheries (random sample of 1000 from the whole year)
- catch-at-age data only available for the late spawning ground fishery (random sample of 1000 from the whole year)
- 6000-12000 juvenile tag releases in years 41-45. Tagged fish were released mid-year and instantaneously assumed the characteristics of the untagged population
- tag recovery reporting rates were 100%.

D_base characteristics differed from E_base in the following ways:

- feeding grounds longline fishery selectivity varies over time, with preferential targeting of relatively abundant sizes (and a minimum target age). This is intended to produce systematically varying selectivity that is broadly consistent with economic objectives of targeting valuable ages and maximizing CPUE.
- the juvenile fishery selectivity varies over time in such a way that there is a tendency to catch a constant age composition irrespective of the relative abundance of age classes. This is intended to mimic the Australian farming practice of selecting an ideal size class for farming. This reduces the ability to estimate recruitment strength for assessment models that assume constant selectivity.
- all fisheries also have a stochastic element to selectivity imposed on the underlying relationship.
- spawning occurs continuously over a 4 month period every year
- variance on length-at-age is broader than in E_base
- recruitment follows a Beverton-Holt stock recruitment relationship with a CV ~ 0.6 for recruitment deviations.
- The relationship between effort and fishing mortality was much less reliable for the longline feeding grounds fishery (annual CV ~ 0.4, annual auto-correlation ~0.5)
- catch-at-length sampling is greatly reduced from E_base (sample size 50)
- catch-at-age sampling is greatly reduced from E_base (sample size 50, still only from the late spawning grounds fishery)
- tag release numbers greatly reduced (300-600 juvenile releases in years 41-45)

The small sample sizes (and tag release numbers for D_base) were intended to reflect the fact that sampling is probably never truly random, but this does not capture sampling biases that probably occur in real life.

The other SBT operating models are defined qualitatively in Table 1. Implementation details are included in Appendix 2. Each operating model was primarily intended to address the indicated objective, but some are relevant to multiple objectives.

Table 1. Qualitative comparison of different SBT operating model scenarios used in the SESAME study. Characteristics describe differences relative to E_base (designated E_x) or D_base (designated D_x). Specific details are supplied in Appendix 2.

Relates Primarily to Objective	Operating Model Scenario	Distinguishing feature(s)
I Stock and Recruitment	E_h3 D_h3	Beverton Holt Stock Recruitment curve steepness = 0.3
	E_h9 D_h9	Beverton Holt Stock Recruitment curve steepness = 0.9
"steepness" indicates the degree of recruitment compensation as SSB decreases	E_h4_r8 D_h4_r8	Beverton Holt Stock Recruitment curve steepness = 0.4 recruitment deviation auto-correlation = 0.8
	E_h8_r8 D_h8_r8	Beverton Holt Stock Recruitment curve steepness = 0.8 recruitment deviation auto-correlation = 0.8
	E_HSSR	Stock Recruitment curve functional form is a double linear "hockey stick" (i.e. recruitment increases linearly with SSB up to a maximum and then remains constant as SSB increases further) steepness = 0.6
II Catch under-reporting	E_CU20ju E_CU20llf E_CU20lls	20% juvenile fishery total catch numbers underreported 20% longline feeding grounds fishery underreported 20% longline spawning grounds fishery underreported
(Catch-at-Age/Length Sampling)	E_CA60	catch-at-age and -length sample sizes = 60 Tag releases between 600-1200 per year
IV Changes in growth	E_DDLinf	fish growth curve changes over time, roughly consistent with a density dependent effect of intra-specific foraging competition
V Selectivity	E_H45 D_H45	longline feeding grounds selectivity shifts to younger ages at year 45 (out of 50)
	E_HL	fishery selectivity is purely length-based (size selectivity effects are maintained in the fish population)
	E_HTS	fishery selectivity characteristics are defined as in D_base, but other characteristics are from E_base
VI Catchability	E_qInc D_qInc	longline feeding grounds catchability increasing exponentially at 1% per year (42% increase in efficiency over 35 years)
	E_qI D_qI	longline feeding grounds catchability related to effort; (qualitatively consistent with fleet interference)
	E_qC D_qC	longline feeding grounds catchability related to effort; (qualitatively consistent with fleet co-operation)
	E_DRq	Stock Recruitment deviation CV = 0.6 feeding grounds longline catchability annual CV = 0.5, annual auto-correlation = 0.5

4.2.3 The SPC-OFP YFT simulator and the SCTB-MWG assessment model evaluation project

The YFT simulations were devised with a somewhat different emphasis from SESAME SBT, and provided an independent and complementary opportunity with which to test assessment models. The simulated fisheries in the two systems contained many similar features, including:

- Most simulation scenarios involved multiple distinct fishing fleets with distinct exploitation histories and data dis-aggregated by fleet

- the simulators ran at finer temporal scales (monthly) than the aggregated data that was available for analysis (YFT quarterly; SBT annual)

- total catch was available for all fleets (in mass or numbers)

- effort series were available for all fleets (for SBT it was intentionally uninformative for all fleets except the longline feeding grounds)

- catch composition mostly consisted of length frequency data, with characteristics that did not exactly meet any assessment model assumptions (SBT and YFT simulators both had time-step mismatch effects relative to the data aggregation units; YFT also had contaminated length samples)

- in some scenarios, catchability trends were present, complicating the relationship between effort and fishing mortality (or CPUE and abundance)

- tag releases by area, and recapture data by fleet were included

Notable differences between the SBT and MWG YFT simulations included:

- the SPC-OFP simulator included relatively fine-scale (5 X 5 degree) spatial dynamics with migration behaviour and recruitment linked to dynamic SST fields (with inter and intra-annual variability). The SESAME SBT simulator was only run in a spatially-aggregated mode.

- SBT simulations all included 3 Longline (LL) and 1 Purse Seine (PS) fishery. The YFT simulations explored 5 scenarios with different fishing fleets and data spatial aggregation units:
  1) 1F X 1R = 1 LL Fishery; 1 Region
  2) 2F X 1R = 2 Fisheries (1LL, 1 PS); 1 Region
  3) 4F X 2R = 4 Fisheries (2LL, 2PS); 2 Regions
  4) 7F X 7R = 7 Fisheries (7LL); 7 Regions
  5) 16F X 7R = 16 Fisheries (7LL, 6PS, 3 artisanal); 7 Regions

- very little prior information on natural mortality was provided with the YFT simulations, and the mortality-at-age vectors differed between scenarios. Some SBT assessment models used the known values of M, while others attempted to estimate M.

- YFT fisheries generally experienced lower overall depletion compared to SBT (i.e. the stock size was not reduced as much, so there was less stock size contrast). The SBT biomass trajectories were generally close to a one way trip, with some recovery near the end. YFT biomass trajectories had substantial decreases and increases.

- YFT seemingly did not attempt to test different stock recruitment relationships, while SBT scenarios explicitly tested a range of recruitment dynamics, including different levels of productivity, auto-correlation in the random errors that determine cohort strength, and alternative stock-recruitment functional relationships.

- One fishery (late longline spawning) in the SBT simulations included some direct ageing data which covered a few years near the end of the time series. All other fisheries in the YFT and SBT simulations only had Catch-at-Length data.

- In YFT scenario 4 (7F X 7R) tag reporting rates ranged between 17-56%. (We assumed 100% reporting in SCALIA applications)

- In YFT scenario 5 (16F X 7R) the simulated population had sex-specific growth and mortality characteristics. (We assumed sexes were identical in the SCALIA and production model applications)

We present some results from the 2003 YFT project in this report, but it is expected that intersessional work under the auspices of the SCTB-MWG will lead to a more comprehensive synthesis of the 2003 results for discussion at SCTB-17 in 2004.

4.3 Assessment Models

Table 2 compares the different assessment model specifications that we have included for the SBT objectives and briefly describes the intention behind each. General descriptions of the various models follows, and technical implementation details are included in the appendices.

A qualitative description of models applied to the SPC-OFP YFT simulated data for the 2003 SCTB MWG are listed in Table 3. Specific details are supplied under the individual assessment model sections and in the appendices. We do not describe much of the 2002 YFT study in this report (see SCTB-MWG 2002, Labelle 2002, Kolody 2002, Ricard and Kolody 2002), because the 2003 study was conducted more thoroughly and was more comprehensive in scope.

Table 2. Qualitative description of assessment model specifications applied to the simulated SBT data sets. SCALIA models are defined relative to the reference model SC_base.

Assessment Model	Specific Implementation	Defining features
Age-Aggregated Production Models (AAPMs)	f_calc	Fox surplus production model
	s_calc	Schaefer surplus production model
Age-Structured Production Models (ASPMs)	ASPM_d2g	Age-Structured Production Model using selectivity from E_base operating model; deterministic recruitment from an estimated Beverton-Holt SR
	ASPM_d6g	Age-Structured Production Model with selectivity calculated using simple approximation from Catch-Length distribution and equilibrium age structure assumptions; deterministic recruitment from an estimated Beverton-Holt SR
	ASPM_sto	Age-Structured Production Model using selectivity from E_base operating model; annual stochastic recruitment deviations estimated from Beverton-Holt SR
SCALIA	SC_base	SCALIA reference model
	SC_Mest	natural mortality estimated
	SC_noHTS	selectivity constant over time
	SC_qTS1	longline catchability temporal variability estimated (random walk CV ~ 0.01)
	SC_qTS5	longline catchability temporal variability estimated (random walk CV ~ 0.05)
	SC_EL	natural mortality estimated; catchability temporal variability estimated; length-at-age mean and variance estimated
	SC_noTags	tagging data not used
	SC_1ideal	specifications closely resemble E_base operating model
	SC_2ideal	specifications closely resemble D_base operating model
	SC_CA60	uses cohort-sliced age data (in addition to direct aged data for late spawning grounds)
	SC_189	small CA/CL effective sample sizes
Integrated Analysis using cohort-sliced CA data	BIH_2	Polacheck and Preece (2001) model (resembling Butterworth et al. (2003)) and parameterized to superficially resemble SCALIA version SC_BIH
MULTIFAN-CL	mf_yft	MULTIFAN-CL implementation as taken from the website example application for simulated YFT; adapted for SBT
	mf_scan	MULTIFAN-CL implementation modified to resemble SC_base
	mf_qTS	as mf_scan, except feeding grounds longline catchability temporal variation estimated

Table 3. Assessment models applied to the SPC-OFP YFT simulated data sets.

	Assessment Model	Defining features
Age-Aggregated Production Models	Fox	Fox surplus production model using CPUE from one of the largest longline fisheries
	Fox_agg	Fox surplus production model using the nominal CPUE (total catch of all longline fisheries / total effort of all longline fisheries)
	Schaefer	Schaefer surplus production model using CPUE from one of the largest longline fisheries
	Schaefer_agg	Schaefer surplus production model using the nominal CPUE (total catch of all longline fisheries / total effort of all longline fisheries)
Age-Structured Production Models	ASPM_d6g	Age-Structured Production Model with selectivity calculated using simple approximation from the Catch-at-Length distributions and equilibrium age structure assumptions
(Results from the ASPMs were withdrawn from the SCTB study because of numerical problems)	ASPM_sto	Age-Structured Production Model with selectivity calculated using simple approximation from the Catch-at-Length distributions and equilibrium age structure assumptions; stochastic quarterly recruitment deviations estimated from a Beverton-Holt SR
SCALIA	SCALIA	11 different specifications for the 2003 study are detailed under the SCALIA section. There was not a systematic comparison of models; differences between specifications related to: effort deviation CVs (0.1 – 0.4) effective sample sizes of CL data downweight by (0.1-0.01) max(effective sample sizes of CL data) (200 - 1000) effective tag release co-efficients (0.01 - 1.0) recruitment deviation CV (0.2 – 0.8) length-at-age (mean and variance fixed or estimated) different methods of ageing tags constraints on mortality estimation selectivity and catchability (constant or temporally variable)
other	MULTIFAN-CL A-SCALA ADAPT	these models were applied by other analysts in the SCTB MWG and are not detailed as part of this report; however we do reproduce some of the MULTIFAN-CL results for comparison.

4.3.1 Age-aggregated and age-structured Production Models

Production models are very simple to implement and interpret, and have a long history in fisheries stock assessment. For SBT, they are currently being considered within the context of Management Procedure development (e.g. Butterworth and Mori 2003, Polacheck et al. 2003b), but have also been interpreted as alternative assessments in parallel with more sophisticated models (Butterworth and Plaganyi 2001). In attempting to determine whether the complicated integrative models are actually adding anything new to our understanding of fishery systems and quality of advice to managers, these simple models provide a benchmark that we can compare with. The models that we included are briefly described in Appendix 3.

The age-aggregated models (Fox and Schaefer) that we examined are particularly simple to implement, requiring only a time series of total catch, and a relative abundance index (e.g. CPUE). They require estimation of only 2 free parameters (carrying capacity and intrinsic growth rate), while other parameters (e.g. CPUE catchability) are calculated analytically. It would have been worth considering a more generalized production model (e.g. Pella-Tomlinson), except that this would probably require additional constraining assumptions on the “shape” parameter, and we were interested in the simplest options possible for testing in an automated context.

Age-structured production models (ASPMs) are more sophisticated, in that they represent the age structure of a population and potentially the associated time lags observed in the dynamics of relatively long-lived populations. It can be argued that these models are very simple, in that only stock recruitment relationship parameters need to be estimated with an objective function minimizer, and the only data requirements are total catch by fishery and a relative abundance index. However, in this form, these models also require fixed input of natural mortality and selectivity by fishery, which must be derived by separate and not necessarily simple analyses. We tested two forms of ASPM, one with recruitment as a deterministic function of the estimated stock recruitment relationship, the other with annual recruitment deviations estimated around the mean stock recruitment relationship. For the stochastic case, this required estimation of 50 extra parameters for the SBT simulations and 148 for the YFT simulations. In the stochastic case, the ASPMs do not really represent a simple function minimization problem. In the SBT simulations, we also compared performance between ASPMs with the correct selectivity taken as fixed input, and an ASPM with the selectivity derived from a simple empirical calculation based on the catch-at-length distributions and equilibrium age structure assumptions.

The Fox, Schaefer and ASPMs were also applied to the SPC-OFP YFT simulated data. However, in the YFT case, the ASPM analyses were withdrawn because of numerical problems (discussed in 5.1.2).

4.3.2 SCALIA: a generic fisheries stock assessment model

SCALIA (Statistical Catch-at-Age/Length Integrated Analysis) is a flexible stock assessment framework that was initially developed for Southern Bluefin Tuna stock assessment (Kolody and Polacheck 2001). The majority of SCALIA features have been implemented in other stock assessment models in the past and the acronym is thus probably most meaningful as an indicator of the group involved with development, rather than any specific features. SCALIA arose out of the recognition that a number of features in MULTIFAN-CL might address some of the outstanding issues in the existing methods for SBT assessment, building on the approach of Butterworth et al. (2003) and Polacheck and Preece (2001). As a result of the SESAME and SCTB-MWG model evaluation projects, SCALIA evolved many additional features and currently allows the analyst substantial control over what is fixed input and what is estimated and how over-parameterization is constrained.

Technical details of most SCALIA features are included in Appendix 5. Key points include:

- spatially-aggregated, age-structured population
- dis-aggregation of fisheries into an arbitrary number of distinct fleets
- data include some or all of the following for each fishery: effort time series, total catch in numbers or mass, catch-at-age or catch-at-length, tag recoveries (including age or length at release)
- estimated parameters potentially include: effective effort deviations (errors in the relationship between effort and fishing mortality), fishery selectivity (including temporal variability), catchability (including temporal variability), natural mortality by age, estimation of length-at-age relationship, stock recruitment relationship (including deviations from the mean relationship), and tag reporting rates (assumed constant over time but different for each fishery).
- fixed input requirements include (in addition to the parameters above if they are not estimated): the functional form of the stock-recruitment relationship and all variance-related parameters (including standard deviations or penalty weightings for error distributions, effective sample sizes for catch-at-length and catch-at-age data, and effective release/recapture weighting factor for tag data).

For each SBT simulation, the number of estimated parameters was typically ~450, most of which were related to effort deviations and selectivity. However, the flexible nested structure means that SCALIA can also represent rather simple models such as an Age-Structured Production Model (although it would not be the most efficient implementation). Similarly, in the manner illustrated in Butterworth et al. (2003), SCALIA can approximate the main features of an ADAPT-VPA, by specifying annual changes in selectivity and large effective sample sizes for the catch-at-age/length frequency distributions.

SCALIA is implemented with the AD Model Builder software (Otter Research, Victoria, Canada, http://otter-rsch.com/), which provides computationally efficient function minimization, and different methods for statistical uncertainty estimation. We did not generally include uncertainty in most of the SESAME simulation testing, except in relation to Objective VIII, where the Inverse Hessian matrix – delta method was used to calculate confidence intervals.

To address the SESAME SBT objectives, we defined a variety of SCALIA models with different assumptions. The reference case specification (SC_base) is defined in

Table 4, and a range of alternatives are defined relative to the reference case in Table 5. The assumptions ranged from highly constrained models specified with reliable prior knowledge (depending on the OM scenarios) to rather unconstrained models with large variances and considerable structural freedom.

SCALIA specifications for the SPC-OFP YFT simulations were primarily intended to examine the treatment of alternative methods for admitting spatial structure (i.e. by allowing varying degrees of variability in selectivity and catchability, including temporal trends). A baseline specification is defined in Table 6 and deviations from the baseline are defined in Table 7.

Table 4. The specification for SCALIA model SC_base, a compromise of features resembling applications to real SBT data, and from which a number of others are derived in Table 5. Terms are defined in Appendix 5.

Description	Value
total catch observation error app. CV	0.01
stock-recruitment relationship log-scale CV (t < -5; -4 <= t <= 50)	0.01; 0.6
stock-recruitment auto-correlation	0
catch-at-length effective sample size (proportion of observed)	1
catch-at-length maximum effective sample size	200
effective tag release co-efficient	1
tag mixing time (timesteps)	1
tag reporting rates (all fisheries)	1.0
tag age estimation	Cohort-slicing
maximum effective effort deviation app. CV	0.2
effort deviation prior scaling exponent	0
temporal change in selectivity app.CV	0.05
temporal change in selectivity – number of timesteps between changes for longline feeding grounds and juvenile fishery	5
selectivity curvature penalty (pseudo-length-based parameterization used)	2.0
length-at-age mean	Correct values used as fixed input
length-at-age variance	Estimated
number of length-based selectivity parameters	8
mortality-at-age	Correct values used as fixed input
mortality-at-age curvature penalty and CV on deviations from mean	n.a.
Beverton-Holt Stock Recruitment relationship steepness	Estimated
catchability temporal variability app. CV	0
temporal change in catchability – number of timesteps between changes for longline feeding grounds and juvenile fishery	n.a.

Table 5. Assessment model specifications for the SBT simulations. SCALIA models are defined relative to the reference model SC_base.

Assessment Model	Defining features
SC_base	SCALIA reference model
SC_Mest	natural mortality estimated mortality-at-age deviation from mean and second difference curvature penalty weighting = 0.2
SC_noHTS	selectivity constant over time
SC_qTS1	longline catchability temporal variability estimated (5 y blocks); random walk CV ~ 0.01
SC_qTS5	longline catchability temporal variability estimated (5 y blocks); random walk CV ~ 0.05
SC_EL	natural mortality estimated mortality-at-age deviation from mean and second difference curvature penalty weighting = 0.2 catchability temporal variability estimated (5 y blocks); random walk CV=0.01 length-at-age mean and variance estimated tag ageing with fractional fish, and weighted by N(a)
SC_noTags	tagging data not used
SC_1ideal	specifications closely resemble E_base: • growth specifications match E_base scenario • effort deviation CV = 0.1 • recruitment deviation CV ~ 0.4 • CA/CL effective sample sizes = 1000 • selectivity constant
SC_2ideal	specifications closely resemble D_base: • growth specifications match D_base scenario • effort deviation CV = 0.4 • CA/CL effective sample sizes = 60 • catchability variable (10 y blocks)
SC_BIH	cohort-sliced age data used (in addition to direct aged data for late spawning grounds) CA effective sample sizes = 60
SC_CA60	CA and CL effective sample sizes = 60

Table 6. The specification for SCALIA model SM_1921 is presented as a reference case, with other models defined relative to this case in Table 7. Terms are defined in Appendix 5.

Description	Value
total catch observation error app. CV	0.01
stock-recruitment relationship log-scale CV (t < -5; -4 <= t <= 148)	0.01; 0.6
stock-recruitment auto-correlation	0
catch-at-length effective sample size (proportion of observed)	0.1
catch-at-length maximum effective sample size	200
effective tag release co-efficient	0.1
tag mixing time (quarters)	1
tag reporting rates (all fisheries)	1.0
tag age estimation	fractional ages weighted by N(t,a)
maximum effective effort deviation app. CV	0.2
effort deviation prior scaling exponent	1
temporal change in catchability app. CV	na
selectivity curvature penalty (pseudo-length-based parameterization used)	2.0
length-at-age mean	fixed input estimated from auxiliary data
length-at-age variance	estimated
number of length-based selectivity parameters	12
mortality-at-age curvature penalty and CV on deviations from mean	0.2
Beverton-Holt Stock Recruitment relationship steepness	estimated
catchability temporal variability	none
selectivity temporal variability	none

Table 7. SCALIA specifications for the SPC-OFP YFT simulations. Definitions indicate deviations from SM_1921 specifications given in Table 6. The YFT scenario (number of Fisheries X number of Regions) is listed under the assessment model name.

Assessment Model	Defining features
SM_1921* (1F X 1R)	reference specification
SM_2914 (2F X 1R)	max effort dev CVs by fishery = 0.4, 0.4 catch-at-length effective sample size = min(0.1 of observed, 1000) effective tag release co-efficient = 1.0 stock recruitment relationship log-scale CV = 0.8
SM_2918* (2F X 1R)	max effort dev CVs by fishery = 0.2, 0.2 stock recruitment relationship log-scale CV = 0.8 mean length-at-age estimated
SM_2930 (2F X 1R)	stock recruitment relationship steepness fixed input = 0.999 Natural mortality estimated, but the same for all ages effective tag release co-efficient = 0.01
SM_3915* (4F X 2R)	max effort dev CVs by fishery = 0.2, 0.2, 0.2, 0.2 catch-at-length effective sample size = min(0.01 of observed, 1000) stock recruitment relationship log-scale CV = 0.8
SM_3931** (4F X 2R)	max effort dev CVs by fishery = 0.1, 0.1, 0.1, 0.1 temporal variability in catchability and selectivity estimated in 4 blocks of 37 timesteps
SM_4918* (7F X 7R)	max effort dev CVs by fishery = 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2 effective tag release co-efficient = 1.0 stock recruitment relationship log-scale CV = 0.2 length-at-age not estimated tags aged by cohort-slicing
SM_4950 (7F X 7R)	max effort dev CVs by fishery = 0.3, 0.3, 0.2, 0.1, 0.2, 0.3, 0.3 effective tag release co-efficient = 0.5 stock recruitment relationship log-scale CV = 0.8 length-at-age not estimated tags aged by cohort-slicing
SM_4930 (7F X 7R)	max effort dev CVs by fishery = 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2 stock recruitment relationship steepness fixed input = 0.999 M estimated, but constant over ages effective tag release co-efficient = 0.01 length-at-age not estimated tags aged by cohort-slicing
SM_5915 (16F X 7R)	max effort dev CVs by fishery = 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4 catch-at-length effective sample size = min(0.01 of observed, 1000) effective tag release co-efficient = 1.0 stock recruitment relationship log-scale CV = 0.8 length-at-age not estimated tags aged by cohort-slicing

Assessment Model	Defining features
SM_5950 (16F X 7R)	considers large longline effort series to be highly informative max effort dev CVs by fishery = 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.2, 0.2, 0.2, 0.2, 0.1, 0.2, 0.2 effective tag release co-efficient = 0.5 stock recruitment relationship log-scale CV = 0.6 mortality-at-age curvature penalty and CV on deviations from mean = 0.1 effort deviation prior exponent = 0.5 length-at-age not estimated tags aged by cohort-slicing

* results submitted to SCTB-16 MWG

** indicates specification that failed badly and was aborted before all runs completed

4.3.3 BIH_2: an independent implementation of a SCALIA-like model

BIH_2 is an independently coded and extended implementation of Butterworth et al. (2003), described in Polacheck and Preece (2001). This assessment model provided another independent performance test, but was only applied to a restricted subset of operating models. BIH_2 and SC_BIH were both intended to examine the effects of cohort-slicing and were parameterized to be similar. However, the models have a number of different features. When performance differences are evident between the two, we expect that BIH_2 is a better indicator of the performance to be expected when cohort-slicing is used. BIH_2 was devised explicitly for use in this manner and performance was extensively reviewed as part of actual SBT assessment; SCALIA was not explicitly intended for use with cohort-slicing data, and plus-group assumptions were not critically examined.

4.3.4 MULTIFAN-CL

MULTIFAN-CL is a flexible modelling framework that was developed initially for the assessment of tuna fisheries in the WCPO (e.g. Hampton and Fournier 2001). It combines a formal treatment of growth curve estimation via the modal decomposition of catch length frequency distributions (Fournier et al. 1990), and the statistical modelling of population dynamics (Fournier and Archibald 1982). Many of the features have been adopted in other modelling approaches (e.g. SCALIA, A-SCALA); and the software continues to evolve. The main feature that MULTIFAN-CL includes that is lacking from most single species age-structured stock assessment models is the ability to dis-aggregate into arbitrary spatial units. Since the compiled software has recently become publicly available (http://www.multifan-cl.org), we decided to compare it alongside the other models in the SESAME SBT simulations, as an additional independent implementation of a complicated stock assessment model. This was a last minute addition to the SESAME project, and we recognize that an experienced MULTIFAN-CL user might have opted for different specifications. The three specifications (Table 1) that we applied to the simulated SBT data are defined in detail in Appendix 6. Fixed inputs (e.g. maturity-at-age, length-at-age, mortality-at-age) and initial parameter values were the same as the SCALIA models.

Labelle (2002, 2003) describe the application of MULTIFAN-CL to the SPC-OFP YFT simulated data. We had no part in the specification in the YFT case, but we reproduce some of the results from the 2003 study for comparison with SCALIA and the production models in this report.

4.4 CRITERIA FOR EVALUATING ASSESSMENT MODEL PERFORMANCE

Stock assessment models typically make a lot of inferences about population dynamics and inevitably some are better than others. Given that the goal of stock assessment is generally recognized to be the provision of advice to fisheries managers, the relative value of the different performance indicators should be related to management objectives. Unfortunately, management objectives are often poorly defined, and different analysts have somewhat different views about what the most important performance indicators are. We report on a range of indicators, and they vary depending on what we are trying to illustrate in the different sections. Table 8 indicates a range of quantities that we calculated as potential assessment model performance indicators. Some were adopted to maintain compatability with the results requested by the SCTB-MWG. They can be roughly broken into various (overlapping) categories related to:

- Biomass – describes the amount of some component of the fish population (e.g. spawning biomass or exploitable biomass). In stock assessment, this is often more useful (and more reliably estimated) if expressed relative to some reference point (e.g. unfished equilibrium, biomass at MSY, biomass 5 years ago, etc); this is discussed under management reference points below. Biomass relative to the biomass that would have been observed in the absence of fishing can be a useful measure for quantifying the impact of fishing when some element of the system is non-stationary (e.g. the stock recruitment relationship).

- Exploitation rates – (Catch/Biomass) an aggregated measure of fishing mortality that does not need to be interpreted relative to age structure and selectivity.

- Recruitment – useful for examining short-medium term population projections, and exploring stock recruitment relationships

- Management reference points – attempt to quantify stock characteristics that measure the state of the stock or fishing mortality relative to management objectives (e.g. B_MSY-related, F(0.1), F(rep), etc). Providing advice to managers about the status of the population and impact of the fishery are usually the main goals for stock assessment, and this is often expressed in two dimensions:1) current stock biomass relative to management objectives (i.e. is the stock over-exploited?), and 2) current harvest rates relative to management objectives (i.e. will the current fishing pressure cause the stock biomass to change in a direction that is compatible with management objectives ?).

- Management Procedures – propose a TAC according to the stock assessment model inferences and a decision rule (e.g. F(MSY)*B(current)).

Table 8. Candidate Performance Indicators used to evaluate stock assessment model inferential performance. *t* indicates an arbitrary time-step (in years for SBT simulations); *T* indicates the final time-step for which data are available (current time).

Indicator	Description
Biomass Indicators	
B(t)	Time series of total exploitable (age 1+) biomass
SSB(t)	Time series of Spawning Stock Biomass
B(T-4:T)/B(T-9:T-5) (or B_trend)	Recent Biomass trend
B(t) / B(t = 1)	Time series of Biomass at time t relative to initial Biomass
B(t) / B_NF(t)	Time series of Biomass relative to the biomass that would have occurred in the absence of fishing

*B(t = 1)
*B(t=0.2T)
*B(t=0.4T)
*B(t=0.8T)
*B(t=0.8T)
*B(t=T)
*B(0.2T)/B(t=1)
*B(0.4T)/B(t=1)
*B(0.6T)/B(t=1)
*B(0.8T)/B(t=1)
*B(T)/B(t=1)
*B_NF(T)
*B/B_NF(T)
	Absolute and relative biomass indicators at specific points in time

Fishing Mortality Indicators	
F(t)	Time series of exploitation rate = C(t)/B(t)

*F(0.2T)
*F(0.4T)
*F(0.6T)
*F(0.8T)
*F(T)
	Exploitation rate indicators at different points in time

*F(T-2:T)
*F(T-5)
*F(T-10)
	average exploitation rate over the last 3 time-steps

Recruitment Indicators	
R(t)	Time series of age 0 recruitment
*R(T-9:T)/R(1:10)	Ratio of recent recruitment over initial recruitment
Management-related indicators	
*MSY	Maximum Sustainable Yield
*B_MSY	exploitable biomass at MSY
*F_MSY	exploitation rate at MSY
*B(T)/B_MSY	indicator describing whether the stock is currently overfished relative to

Indicator	Description
	B_MSY reference point
*F(T)/F_MSY)	indicator describing whether the current exploitation rate will lead to an over-fished state relative to B_MSY
*(F_MSY) X (B(T))	A Management Procedure for TAC setting that should move stock size toward B_MSY regardless of current stock size

* included in the aggregate performance indicator (see text).

Not all of these indicators could be calculated for all models. The AAPMs do not distinguish between B and SSB. Estimation errors in the two are often highly correlated, so we only report B (SSB(t) is included for some models in Appendix 6). The AAPMs also confound mortality, growth and recruitment, so recruitment is not reported for these models.

Additional performance indicators are used in examining stock recruitment related issues. "Steepness" defines the amount of compensation in the stock recruitment relationship (specifically, it refers to the ratio of expected recruitment when SSB is 20% of unfished levels, over expected recruitment when SSB is at unfished equilibrium). We use recruitment Root Mean Squared Error (RMSE) to describe the "empirical variability" about the stock recruitment relationship. This is calculated from the MPD estimates of recruitment. Similarly, empirical SR_rho is the lag(1) auto-correlation of MPD recruitment deviations about the stock recruitment relationship. Both of these latter values might be substantially different from the assumptions used in model fitting, and may provide evidence of systematic lack of fit to the assumed functional form of the stock recruitment relationship.

There were some ambiguities in the working definitions that we used to calculate some of these performance indicators (e.g. biomass at the beginning of the time-step vs: the average or middle), and this presumably contributes to a small component of the performance biases. There are also different definitions for MSY-related calculations. The most common assumption is that global selectivity remains constant at current levels (i.e. using the standard catch equations, the effort and fishing mortality of all fleets is scaled proportionately). In the case of fisheries with proportional catch allocations (e.g. SBT), it makes more sense to do the MSY calculations assuming that the relative catch proportions remain constant (in this case the global selectivity changes in the equilibrium yield calculations, because with different age structures, the different fleets must exert different relative fishing pressure to obtain the same catch ratios). The SBT operating models assumed constant catch ratios among fleets, while the YFT operating models assumed constant exploitation rate ratios among fleets. VSM and the production models used constant catch ratios. SCALIA calculated MSY in both ways, and for SBT the results were generally very similar. MULTIFAN-CL, and BIH_2 used constant fishing mortality ratios. The two approaches yield the same result for the production models.

In most cases the assessment model performance is summarized graphically for a range of performance indicators, using the ratio of the estimated value (from the assessment) over the "true" value from the operating model (e.g. B(*t*, AM) / B(*t*, OM)). These ratios are presented as frequency distributions (boxplots or time series of quantiles), in a manner consistent with the SCTB-MWG YFT analysis. These allow one to make general statements like "*on average the terminal depletion was over-estimated by 20%*", or "*the absolute biomass estimates had an over-estimation bias that increased in magnitude over time*". In some cases (e.g. stock recruitment curve steepness), it can be more informative to actually show the operating model and assessment model values, rather than the ratios. We found the time series plots of several indicators (B(*t*), F(*t*), etc) to be more informative than a restricted number of point estimates, because temporal trends in the estimation bias were often the most interesting feature. We note that the use of ratios as performance indicators are potentially deceptive in some ways. When biases are large, over-estimation appears more extreme than under-estimation. If the AM and OM values differ by a factor of 2, this appears as an over-estimation of 100%, but an under-estimation of only 50%. A more subtle deception occurs with respect to the scale of comparison. If one is examining relative biomass (the ratio of B(*t*)/B(0)), a performance indicator ratio of (estimated = 0.2) / (true = 0.1) appears to be a very large error (+100%) relative to (estimated = 0.5) / (true = 0.4) an error of +20%. However, if the same quantities are examined in terms of depletion instead (i.e. where depletion = 1- B(*t*)/B(0)), the magnitude of the relative error between the two is reversed to –11% and –16% respectively.

We had hoped that there would be well-defined relationships among the performance indicators, such that a very restricted subset could be used to describe the essence of the assessment model performance. There were a number of reasonably strong relationships as might be expected (e.g. generally B(*t*) is highly correlated with SSB(*t*); B(*t*) and B(*t*-*x*) are correlated to a decreasing degree as *x* increases, errors in B(*t*) and F(*t*) are inversely related, as are B(MSY) and F(MSY)). Unfortunately, in the majority of cases, we could not interpret the performance of one indicator as very representative of the others. The relationships among some of the indicators that we chose to focus on are illustrated in Fig. 2.

We deliberately avoided the use of multivariate statistics for the synthesis, because we did not want to transform the results in a manner that obscures the nature of the estimation problems. But we did reluctantly include a simple aggregate performance indicator, because there is an irresistible desire to have complicated results reduced to a single dimension. Even the aggregate index has two dimensions of interest (bias and variance). The variance of the aggregate tended to be the focus of discussion, but even this is not always straightforward since robustness to outliers (e.g. range) might be more important than the actual variance. The aggregate is simply an arithmetic mean of an arbitrary mix of 28 performance indicators (indicated in Table 8), selected on the basis that all of the assessment models were generating estimates for these quantities. The aggregate index for each assessment model can also be calculated across multiple operating models. This latter comparison potentially allows one to rapidly compare the robustness of assessment models to a range of conditions, to identify gross performance differences. However, we did find the aggregate potentially deceptive in some cases, and examination of the individual performance indicators usually provided a more satisfactory comparison of assessment model performance.

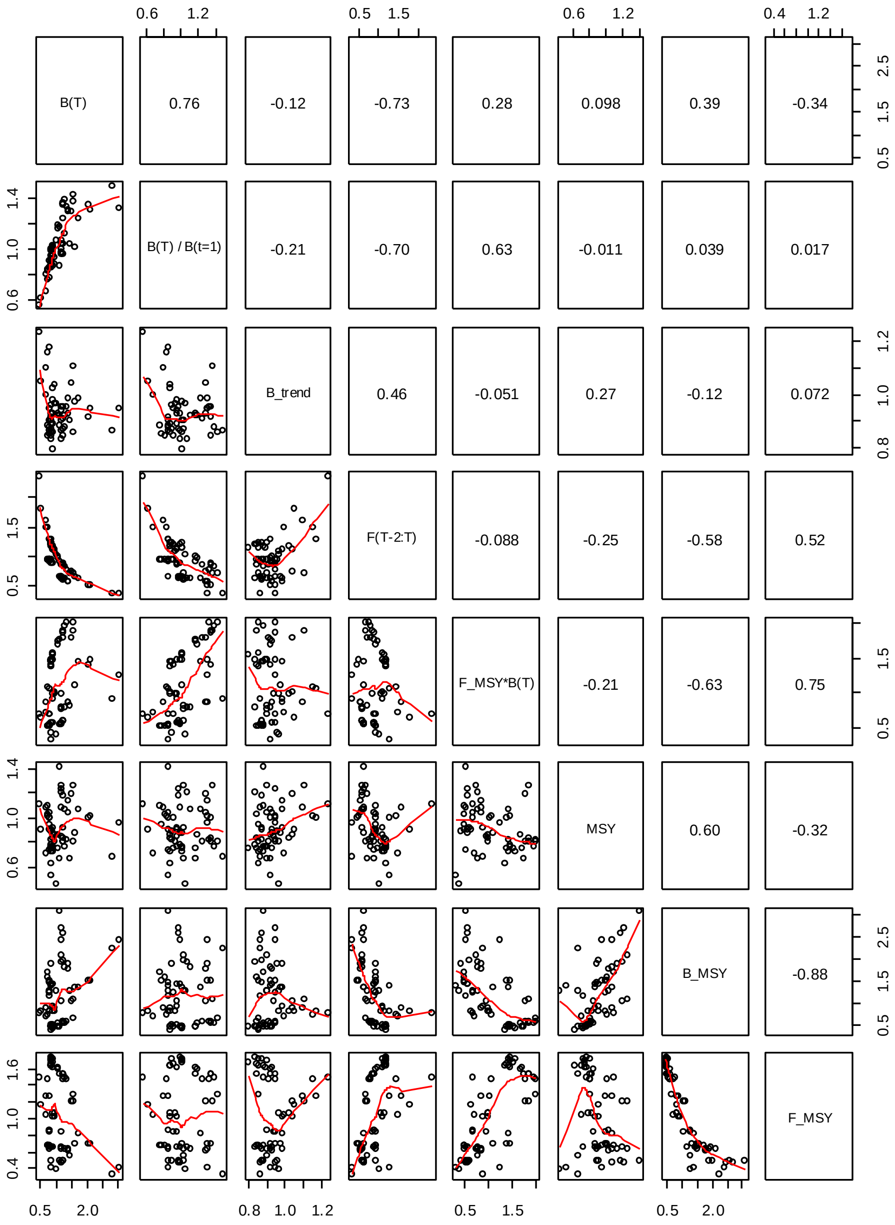


Fig. 2. Scatterplots (with LOWESS smoothing lines) and Pearson correlation coefficients illustrating typical relationships between assessment model performance indicators. Each point represents a ratio of (AM estimate)/(OM actual). The 60 points in each comparison are taken from 6 different assessment models (f_calc, ASPM_d2g, SC_base, SC_Mest, MF_YFT and MF_scan) each applied to 10 stochastic realizations from the E_base (highly informative) operating model scenario.

4.5 Database Overview

We found a relational database to be an extremely powerful tool for organizing the operating model state realization summary statistics and assessment model estimates. Every assessment model fitting to each data realization produced a uniquely identifiable file of results that were uploaded to an ODBC-compliant database. In addition to stock assessment model estimates, these files potentially contained flags that were useful for diagnosing function minimization problems (e.g. final gradients of the objective function with respect to parameters). In this manner we were able to query the database to look for suspicious results that otherwise might not be identified due to the highly automated manner that the assessment models had to be applied. The linked structure of the database facilitated an easy comparison of assessment model results with operating model "true" values across all data realizations for the operating models of interest. Data extractions were made through a simple command line argument to R software functions, such that it was easy to compare a range of assessment models for a given operating model, or a range of operating models for a given assessment model. For brevity we do not include further details about database implementation, but we recommend that any similar study should use a similar approach.

4.6 Quality Control

We endeavoured to test that operating and assessment models were implemented and documented correctly, but inevitably, coding errors occur in complicated software. The complexity of the simulations also led to mis-specifications and interpretation errors among participants, that sometimes were not recognized until well into the project when the technical documentation was being completed. When the errors were large, all affected results were re-run; relatively minor errors are documented and perhaps apparent only as peculiar model specifications. While we cannot be sure that all the errors were identified, the multitude of comparisons among independently coded population dynamics models gives us a reasonable degree of confidence that the gross features of most models behaved roughly as intended in the majority of cases.

5 RESULTS AND DISCUSSION

The SESAME SBT and SPC-OFP YFT simulation-estimation testing has given us insight into many facets of assessment model performance. The use of two independently implemented operating models provide a good illustration of how the results are potentially very specific to the simulation conditions. In general, we are left with the impression that most of the assessment models provided reasonable inferences about many aspects of stock dynamics when the data were good and key assumptions adequately satisfied. But there were important limitations to what could be estimated reliably even with excellent data and good assumptions. As the magnitude of assumption violations increases within the realm of what we would consider plausible, the potential for misleading inferences increases to an extent that we had not fully appreciated. The estimation errors were not always what we would have expected, presumably due to the complicated non-linear interactions within the model.

It proved exceedingly difficult to comprehensively discuss the diverse array of results in a coherent fashion verbally. We draw attention to the most obvious and interesting results in the discussions below, and provide some speculation on the relevant mechanisms and likely implications. We also include a substantial archive of the assessment model results in Appendix 6. These can be used to examine additional details that were not specifically addressed in the text because there were simply too many operating and assessment model combinations to examine individually.

The Results and Discussion is organized in several distinct sections, but many of them are interdependent. The first section relates to assessment model implementation issues, including automation and minimization problems, but does not comment on any specific results. Sections 5.2-5.9 present fairly detailed results in attempting to address Objectives I-VII as defined in the Introduction. Section 5.10 is organized under the heading of Objective VIII - Uncertainty Quantification, and is an ambitious attempt to provide general commentary on the limitations that we are likely to encounter in our assessment modelling endeavors, and speculates on promising methods for improving the provision of scientific advice for fisheries managers. In section 5.11, we attempt to make comments about the general performance of different implementations and specifications of assessment models. In section 5.12, we outline a number of methodological problems inherent in this type of study and our attempts to resolve the issues. Finally, the conclusions and recommendations attempt to summarize the key findings in relation to the original objectives defined in the introduction.

5.1 GENERAL COMMENTS ON ASSESSMENT MODEL IMPLEMENTATION

This section describes our general impressions of the different assessment models from the perspective of the reliability of implementation, particularly in the automated setting required for simulation-estimation testing. Comments on the actual estimation performance of the models is addressed in subsequent sections.

We note that in all cases, we cannot be sure that global minima were always identified during model fitting. There was usually an inspection of the distribution of gradients at the function minimum (included in Appendix 6), and some superficial examination of results to check if estimates of dynamics from a given assessment model were generally similar for all realizations from an individual operating model scenario. As a rule of thumb, we were generally concerned when the maximum gradient of the objective function with respect to the estimated parameters exceeded 0.1 at the function minimum (but recognize that this is an arbitrary decision that is dependent on the model parameterization). In the production models, some attempts were made to repeat model fittings to remove obviously bizarre behaviour. The more complicated integrated models generally seemed to be more robust in their minimization behaviour, but failures were evident in some of the less constrained SCALIA models, particularly under the difficult assessment scenarios. A detailed examination of every realization result was not possible. We expect that some of the outlier behaviour evident in the results can be attributed to minimization problems, but this did not seem to be a big problem.

5.1.1 Age-Aggregated Production Models

Parameter estimation in the Fox and Schaefer models often required more user interaction than the complicated models. Minimization failures characteristically resulted in parameter bounds being hit and failure to attain convergence criteria. The failed estimates were sometimes associated with chaotic behaviour, in which the population dynamics experience large amplitude, high frequency oscillations (e.g. as in Adkison 1992). In the majority of cases with obvious problems, credible behaviour was obtained by changing starting values and/or transforming the parameters used by the minimizing function. As indicated in the maximum gradient boxplots from Appendix 6, a few results from the Fox model failed to converge in the SESAME SBT results, but these should be identifiable as outliers in three of the D_x scenarios. For future automated applications, we would recommend implementing a systematic search of the parameter space to find reasonable starting values. However, we also note that in applications to real SBT assessments, the inferences from these models have demonstrated a surprising sensitivity to the reliability of the function minimizer (e.g. Ricard et al. (2002) demonstrate considerably different inferences as parameter estimates differ in the 4th-5th significant figure).

5.1.2 Age-Structured Production Models

Of all the models, we had the least success implementing the ASPMs, in part due to the particular implementations that we were testing. As with the Fox and Schaefer models, minimizations were sensitive to the initial parameter values. An automated systematic search of the parameter space prior to the objective function minimization seemed to eliminate this as a major problem. However we note that several of the ASPM results include realizations with convergence failures (Appendix 6) for the majority of the D_x scenarios and a couple of the E_x scenarios.

We had problems getting the function minimizer to converge reliably in the majority of the stochastic recruitment ASPM applications. Given the automated nature of the application, this unfortunately was not recognized until late in the SESAME project, and we do not present any of the stochastic ASPM results for serious consideration. The problem might have been overcome reasonably easily by adding more constraints to the minimization (e.g. reducing the recruitment deviation CV). However, these models were intended to be simple alternatives to the fully integrated models. In practice they do not represent a trivial parameter estimation problem, and are more of a transitional step between production models and the fully integrated models.

The ASPMs with deterministic recruitment converged much more reliably, but were prone to a numerical problem in some OM scenarios. As implemented, these models can attempt to remove too much catch in some circumstances, potentially resulting in negative fish for some age classes. The error is more likely to occur with higher fishing mortality and greater selectivity errors. Our attempts to temporarily guide the function minimizer away from the problem using objective function penalties were not very satisfactory. The function minimizer often converged to the point that was arbitrarily close to a result of 0 fish for one of the age classes, such that this was the dominant term in the objective function at the minimum. We withdrew all of the ASPM results from the SCTB-MWG YFT simulations, because 4 out of 5 OM scenarios were adversely affected. The problem was more sporadic in the SBT simulations, and the number of afflicted results are flagged in the figures with the label “Penalty Activation Count” (e.g. in Fig. 3a - "Penalty Activation Count: E_base(0)" means that 0 of the aspm_x realizations were affected by the problem, and the flag is irrelevant for all other models). The problem could be resolved in different ways, but by the time it was identified, we were not interested in investing more time in the ASPMs.

5.1.3 SCALIA

In general, we were pleased with the SCALIA implementation and minimization reliability. In both the SBT and YFT studies, once a reliable minimization procedure was established, it seemed to generally be robust across the different operating models and realizations. We do note however that the convergence of the least constrained SCALIA models (e.g. SC_EL) was often marginal (as indicated by the maximum gradient distributions in Appendix 6), and the convergence was not completely reliable for any of the SCALIA models in the D_x operating model scenarios (despite a reasonable maximum gradient, the inverse Hessian matrix could not always be calculated for estimating confidence limits).

The speed of the SCALIA implementation could undoubtedly be improved. In the SBT studies, we found the similarly parameterized MULTIFAN-CL to minimize in about half the time (efficiency in terms of the number of function calls and the function evaluation time was not explicitly compared). We observed a similar factor of 2 difference in the most complicated YFT scenario (16 F X 7 R), despite the fact that MULTIFAN-CL was also simulating migration dynamics. This latter procedure took about 24 hours, including inverse Hessian calculation, on a 2.6 GHz Pentium 4 PC (RAM was not the limiting factor in either case).

5.1.4 MULTIFAN-CL

MULTIFAN-CL converged reliably in all the SESAME SBT simulations to which it was applied (although this was a small subset relative to SCALIA and the production models). MULTIFAN-CL has more options and different specification protocol than SCALIA, and none of our analysts had any prior experience working directly with the software. We seemed to get it working and producing plausible results (qualitatively good agreement between predictions and observations) without much problem, but we could have overlooked something important.

5.2 BASELINE ASSESSMENT MODEL PERFORMANCE

The results from application of a range of assessment models to the baseline operating model scenario (E_base) are illustrated in Fig. 3. A number of points are evident from looking at these figures:

- Bias and variance in the MPD estimates for most performance indicators is substantial for many of the assessment models. The time series plots often have disturbing temporal trends in the biases, and these differ substantially among models.

- For any individual performance indicator, the differences among assessment models are usually substantial, i.e. different models have different biases (but there are similarities in that, in general the SC_x models are more similar to each other than the MF_x models).

- The Aggregate PI does not generally appear to be very biased. Presumably this reflects the fact that it is a compilation across indices with biases in opposing directions.

- The models that were specified with the closest agreement to E_base (SC_1Ideal, SC_noHTS) seemed to provide the best assessment inferences in terms of the time series of biomass and recruitment; but other models (notably aspm_d2g, f_calc and SC_2Ideal) often performed better on the management-related estimates.

- The variance of recruitment estimates was high for the ASPM_x, MF_x and BIH_2 models relative to SC_x. Since aspm_x relies on deterministic recruitment, these estimates will always be a poor approximation to a stochastic time series. In the case of BIH_2, this presumably reflects the limitation of cohort-slicing. MF_x recruitment estimates were more variable than BIH_2; presumably this is because MF_x did not use any of the spawning ground direct ageing data.

- The large recruitment estimation errors in BIH_2 and MF_x did not have corresponding effects on the relative biomass estimates for these models. Presumably there is a short term negative auto-correlation in recruitment errors that averages out over several cohorts in the biomass calculations.

- A number of the SCALIA models had substantial and similar (trends in) recruitment estimate biases over the last ~15 years (SC_base, SC_noHTS, SC_qTS, SC_2Ideal), but this was not really evident in the other SC_x models or BIH_2. The MF_x models also seemed to have some minor trends in recruitment biases over the last ~10 years.

- The models that estimated natural mortality (SC_Mest, SC_EL, MF_YFT) generally had worse absolute and relative biomass biases than the most similar models that used the true M from the operating model.

- Stock recruitment curve steepness was estimated rather poorly by the majority of models. The SCALIA models tended toward under-estimation, while MULTIFAN-CL and BIH_2 tended toward over-estimation. Corresponding biases were evident in most of the other management-related indicators as well. Although MSY was generally reasonably well estimated. ASPM_d2g seemed to have the best performance on several of the management-related indicators (and steepness), but performed rather poorly on the biomass time series estimates.

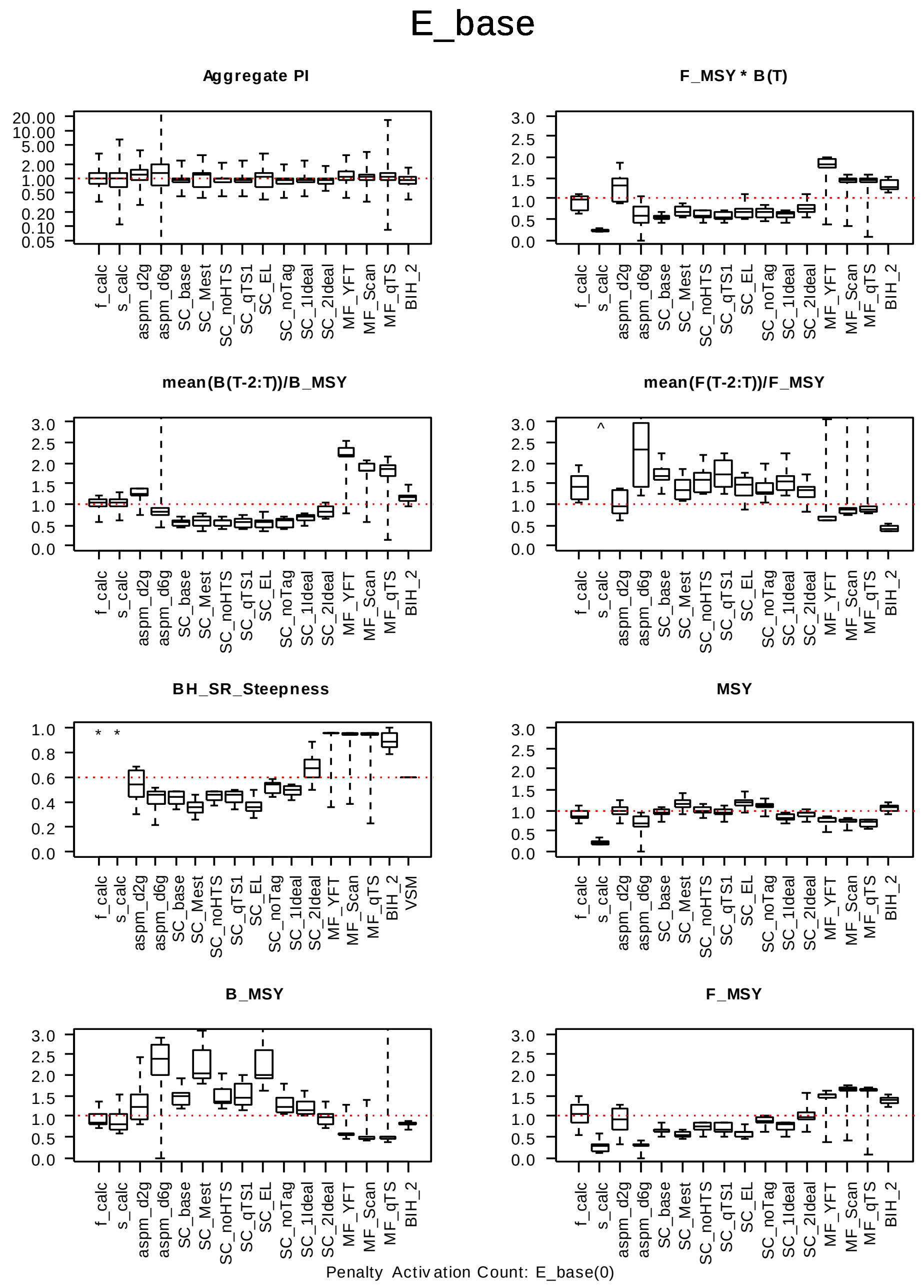


Fig. 3a. Boxplots of MPD performance indicators resulting from the application of a range of assessment models to 10 data realizations simulated from the E_base SBT operating model. Individual values represent the ratios of (AM estimated)/(OM actual) in all cases except BH_SR_Steepness which shows the actual values (the Aggregate PI is an arithmetic mean of 28 ratios). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

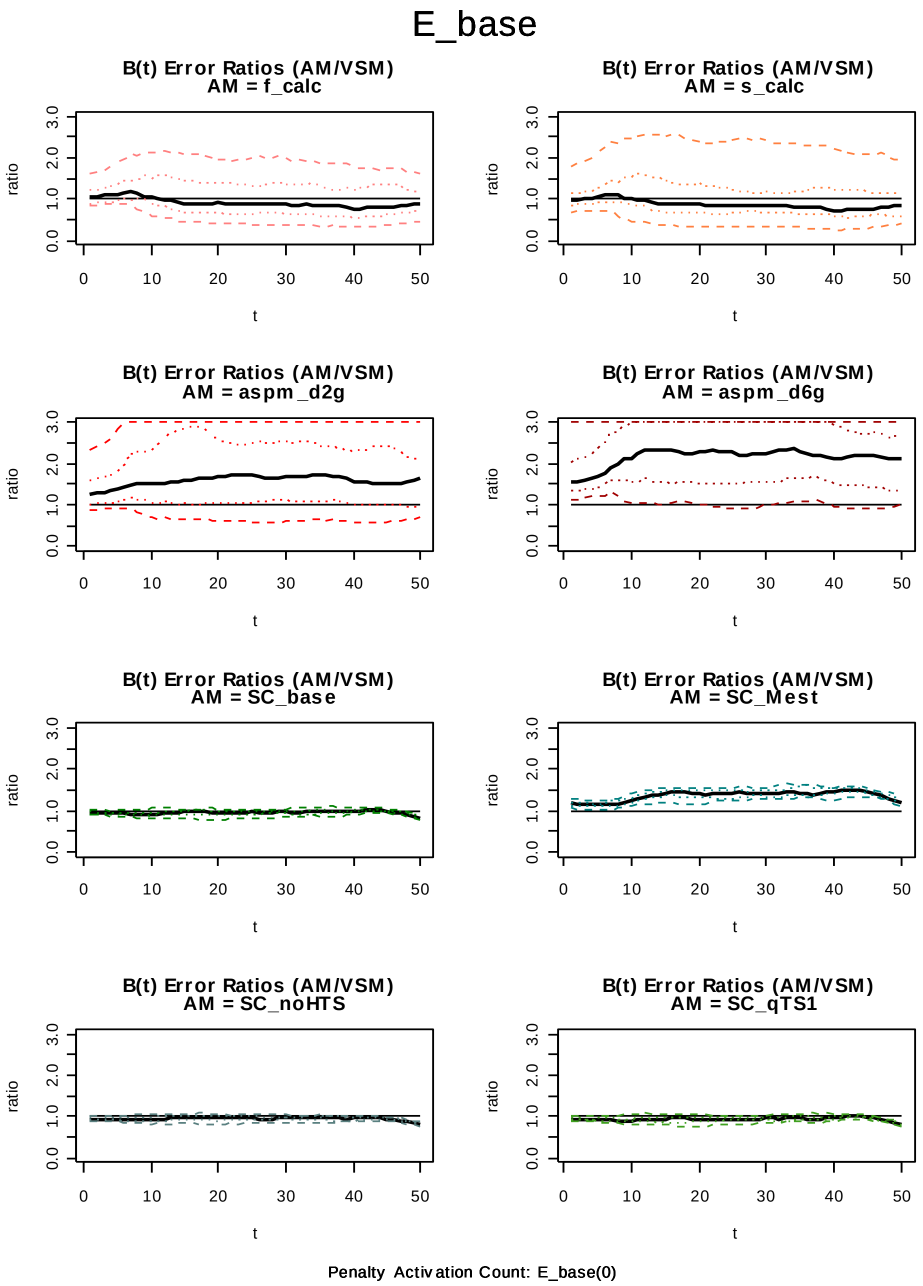


Fig. 3b. **Time series of performance indicators (AM estimated)/(OM actual) resulting from the application of AMs to 10 data realizations from the E_base OM scenario. Lines indicate rankings 1, 3, (5+6)/2, 8 and 10. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

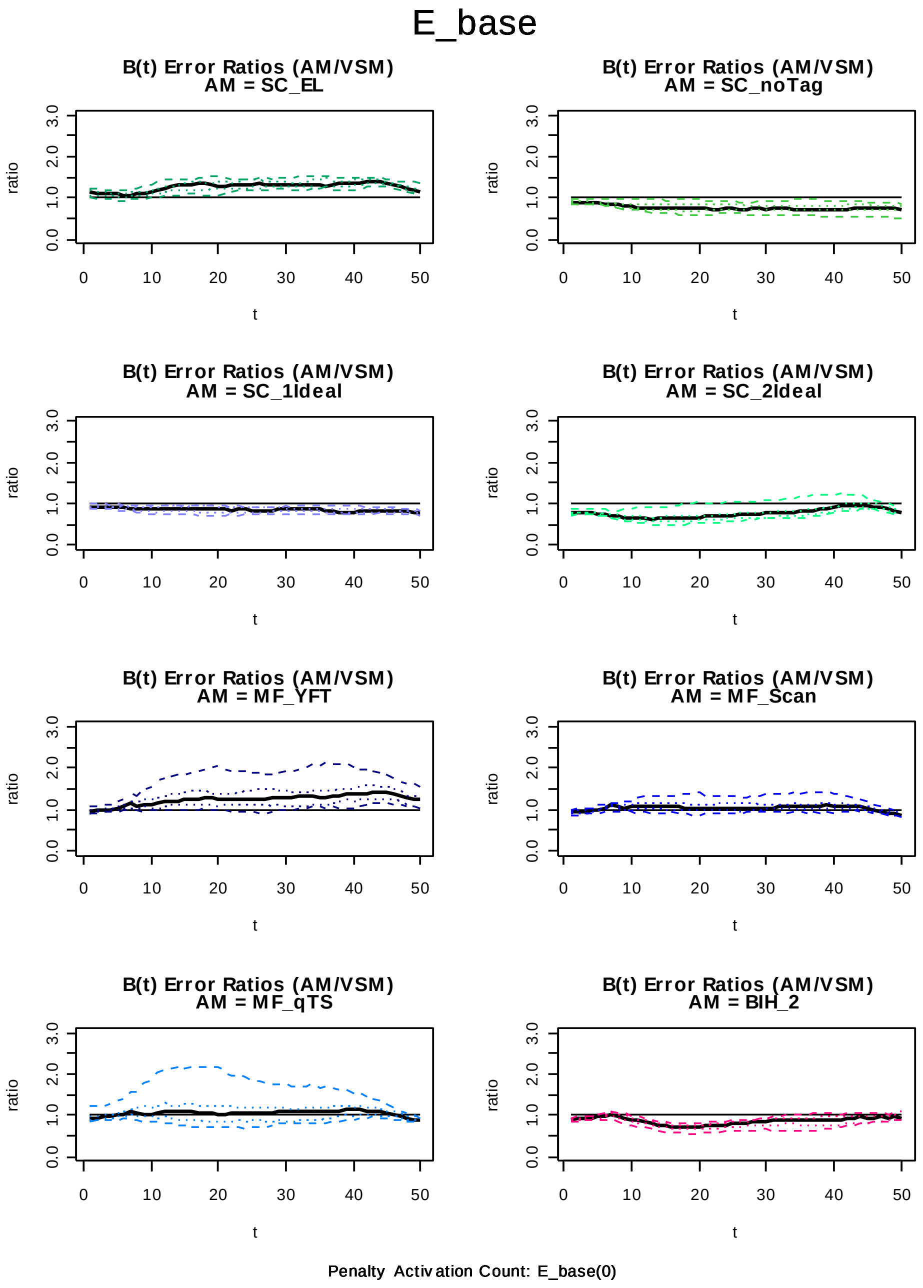

E_base
B(t) Error Ratios (AM/VSM)
AM = SC_EL
B(t) Error Ratios (AM/VSM)
AM = SC_noTag
B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
B(t) Error Ratios (AM/VSM)
AM = MF_YFT
B(t) Error Ratios (AM/VSM)
AM = MF_Scan
B(t) Error Ratios (AM/VSM)
AM = MF_qTS
B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0)

Fig. 3b (cont.)

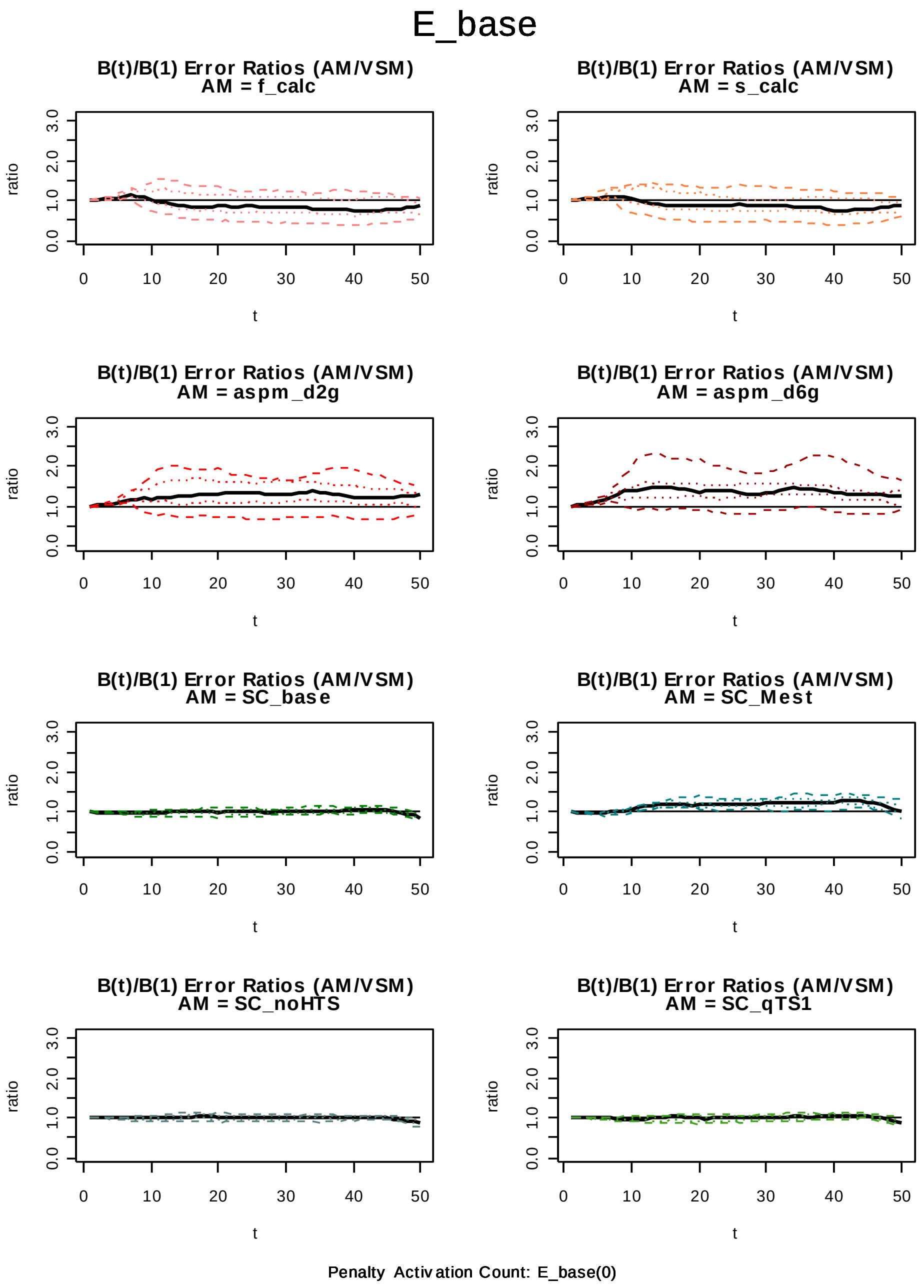
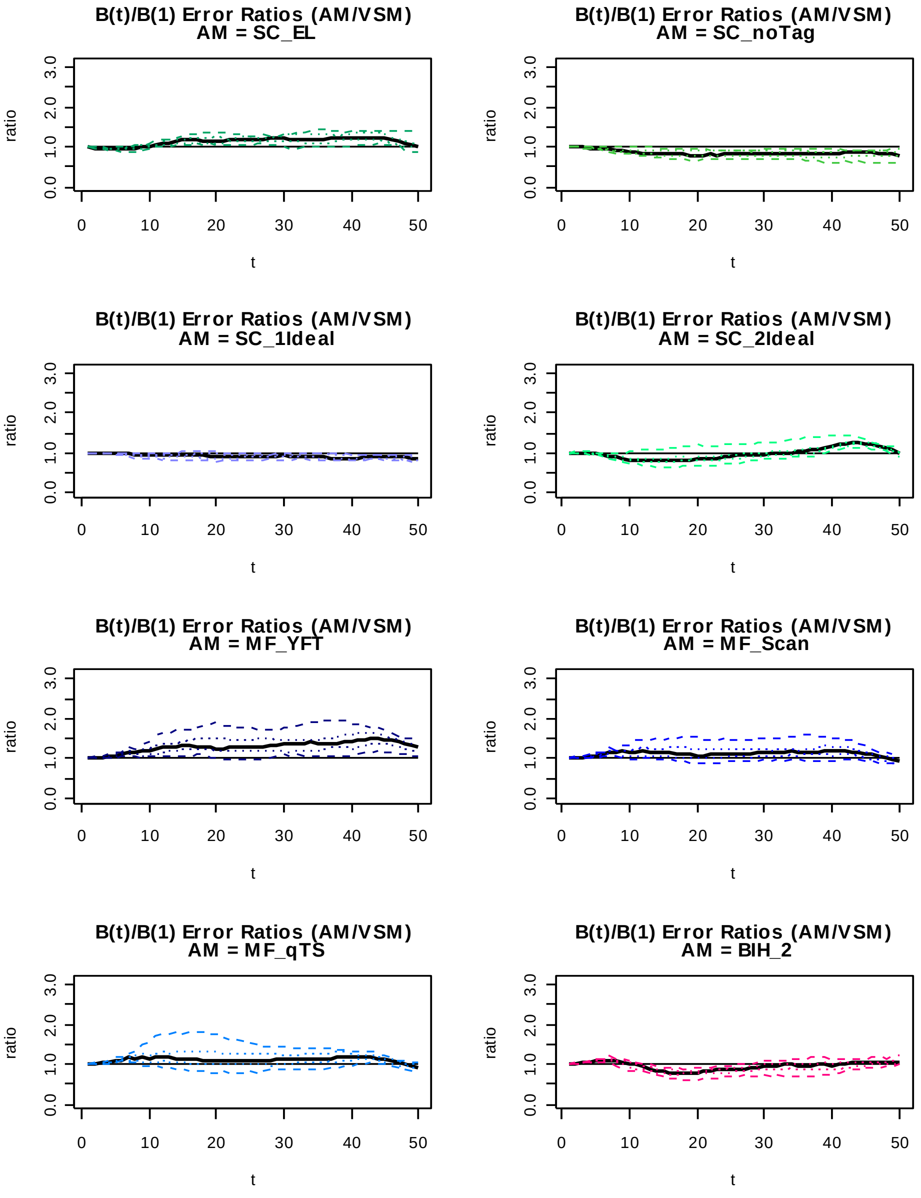

E_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = f_calc
B(t)/B(1) Error Ratios (AM/VSM)
AM = s_calc
B(t)/B(1) Error Ratios (AM/VSM)
AM = aspm_d2g
B(t)/B(1) Error Ratios (AM/VSM)
AM = aspm_d6g
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_Mest
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_noHTS
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0)

Fig. 3b (cont.)

E_base



B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_EL

B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_noTag

B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_1Ideal

B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_2Ideal

B(t)/B(1) Error Ratios (AM/VSM)
AM = MF_YFT

B(t)/B(1) Error Ratios (AM/VSM)
AM = MF_Scan

B(t)/B(1) Error Ratios (AM/VSM)
AM = MF_qTS

B(t)/B(1) Error Ratios (AM/VSM)
AM = BIH_2

Penalty Activation Count: E_base(0)

Fig. 3b (cont.)

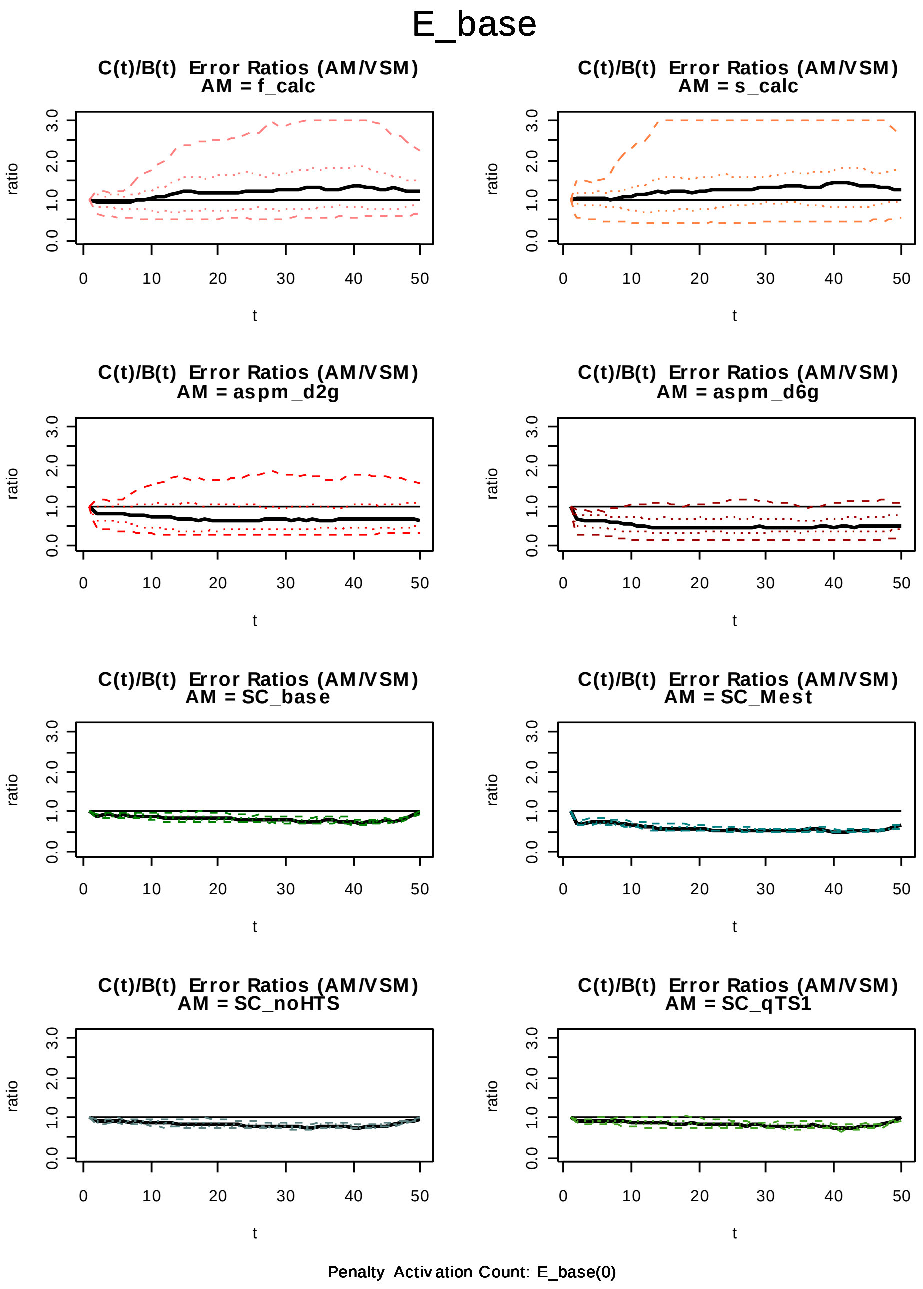

E_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = f_calc
C(t)/B(t) Error Ratios (AM/VSM)
AM = s_calc
C(t)/B(t) Error Ratios (AM/VSM)
AM = aspm_d2g
C(t)/B(t) Error Ratios (AM/VSM)
AM = aspm_d6g
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_Mest
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_noHTS
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0)

Fig. 3b (cont.)

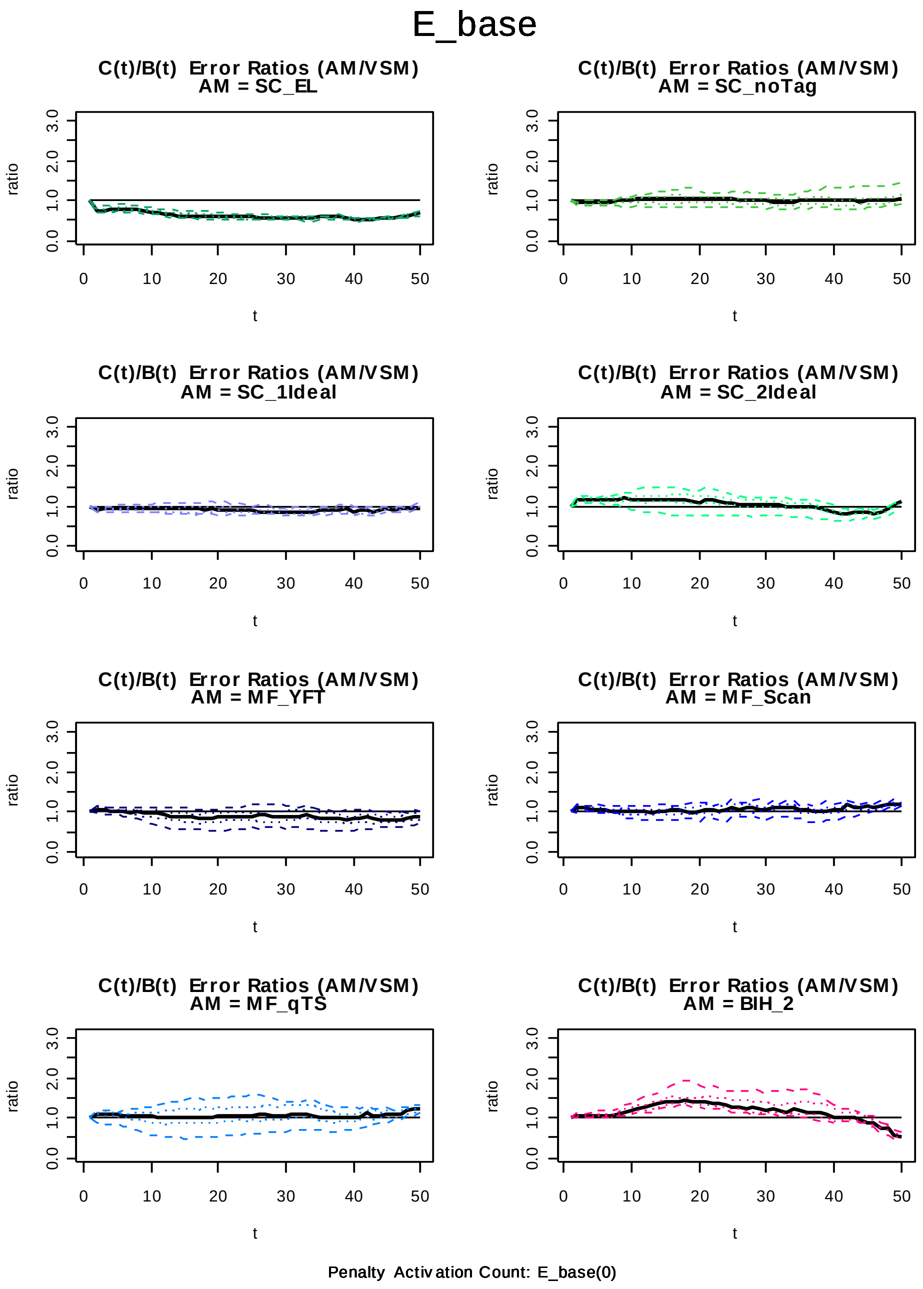

E_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_EL
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_noTag
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
C(t)/B(t) Error Ratios (AM/VSM)
AM = MF_YFT
C(t)/B(t) Error Ratios (AM/VSM)
AM = MF_Scan
C(t)/B(t) Error Ratios (AM/VSM)
AM = MF_qTS
C(t)/B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0)

Fig. 3b (cont.)

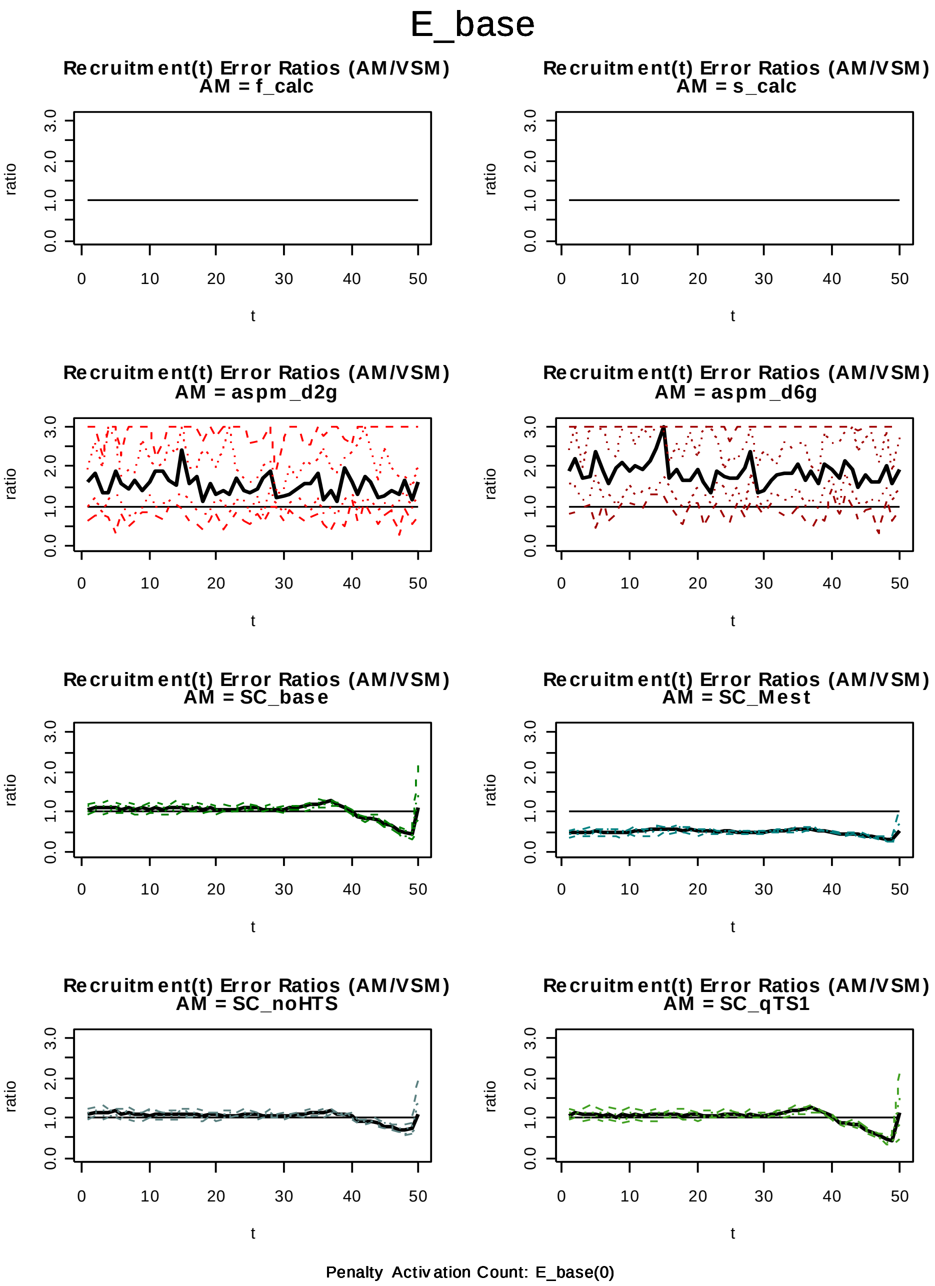
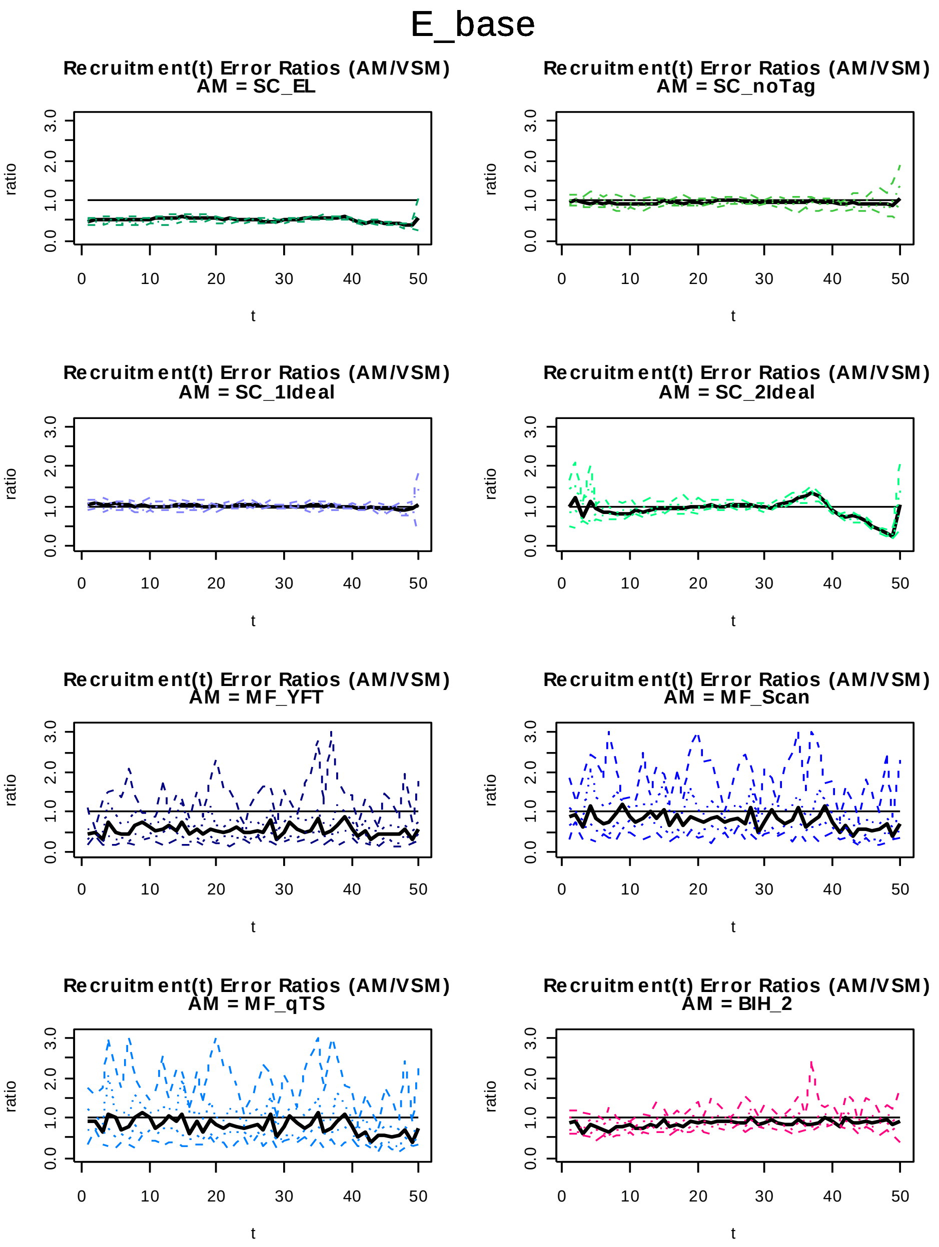

E_base
Recruitment(t) Error Ratios (AM/VSM)
AM = f_calc
Recruitment(t) Error Ratios (AM/VSM)
AM = s_calc
Recruitment(t) Error Ratios (AM/VSM)
AM = aspm_d2g
Recruitment(t) Error Ratios (AM/VSM)
AM = aspm_d6g
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_Mest
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_noHTS
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0)

Fig. 3b (cont.)



Penalty Activation Count: E_base(0)

Fig. 3b (cont.)

The E_base operating model probably approaches the upper limits of assessment performance that we could expect in the real world for a fishery with a data history like SBT. The D_base scenarios probably represent a more reasonable indication of the performance that we might expect. All assessment models had substantial difficulties making reliable inferences for the D_base scenario (Fig. 4). We would not consider that any individual component of the D_x scenarios was greatly outside of the range that could be considered plausible for the SBT fishery. We note the following points from (Fig. 4):

- All of the performance indicators demonstrated substantial bias and/or high variance in comparison with the inferences from E_base.

- Strong temporal trends in biases were evident in most of the time series estimates, although median performance on the relative biomass estimators was not bad in some cases.

- The aggregate performance indicator suggests that SC_2ideal (specified to most closely resemble the D_base OM characteristics) has the best overall performance, but this is not obvious from inspecting the individual performance indicators that are all highly variable.

- On the basis of the aggregate PI, it seems as though the SCALIA models with estimated variability in catchability (SC_qTS1, SC2Ideal, but not SC_EL) performed slightly better than the other SCALIA models. The same is true of MF_qTS relative to the other MF_x models. The production models and BIH_2 generally performed more poorly than the others.

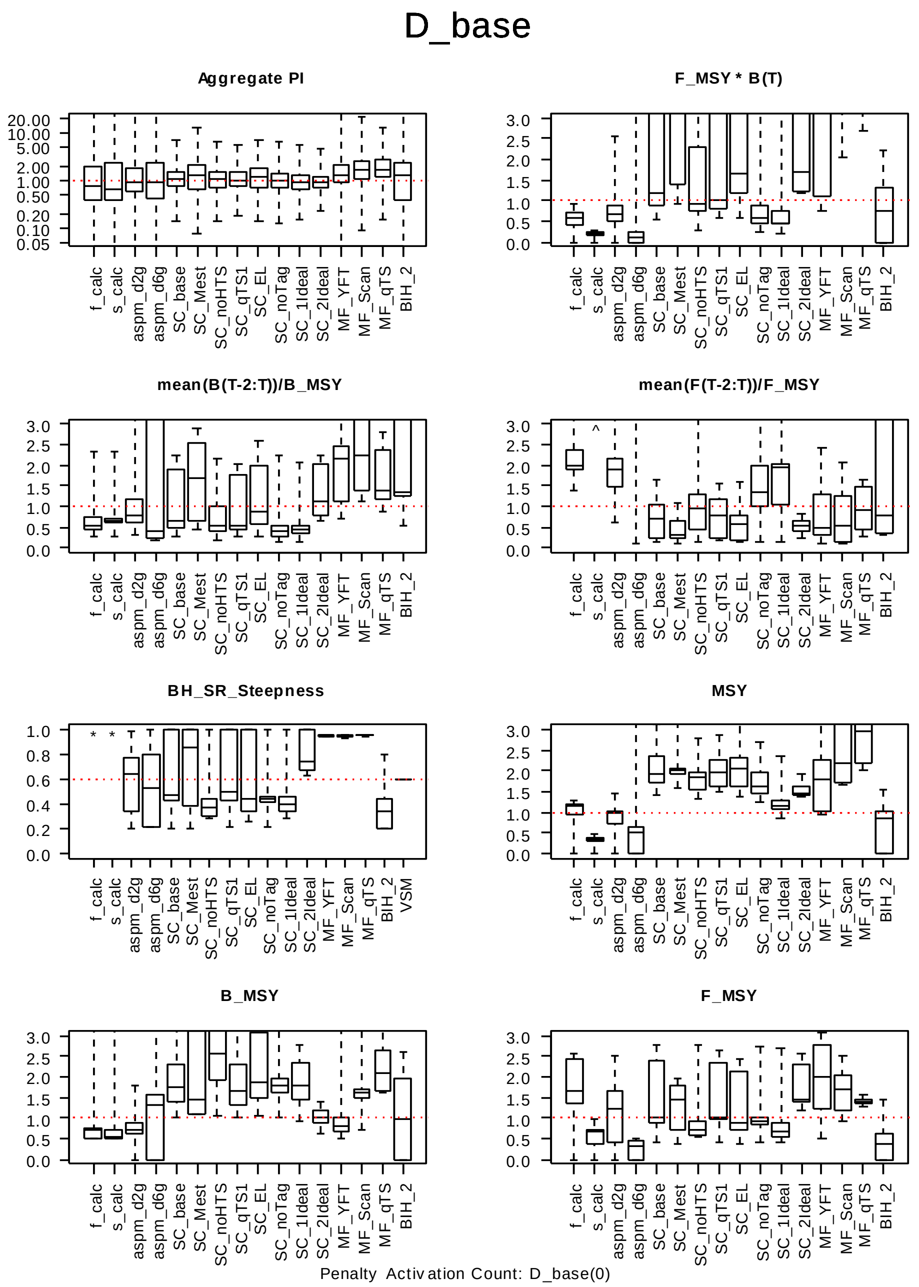


Fig. 4a. Boxplots of MPD performance indicators resulting from the application of a range of assessment models to 10 data realizations simulated from the D_base SBT operating model. Individual values represent the ratios of (AM estimated)/(OM actual) in all cases except BH_SR_Steepness which shows the actual values (the Aggregate PI is an arithmetic mean of 28 ratios). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

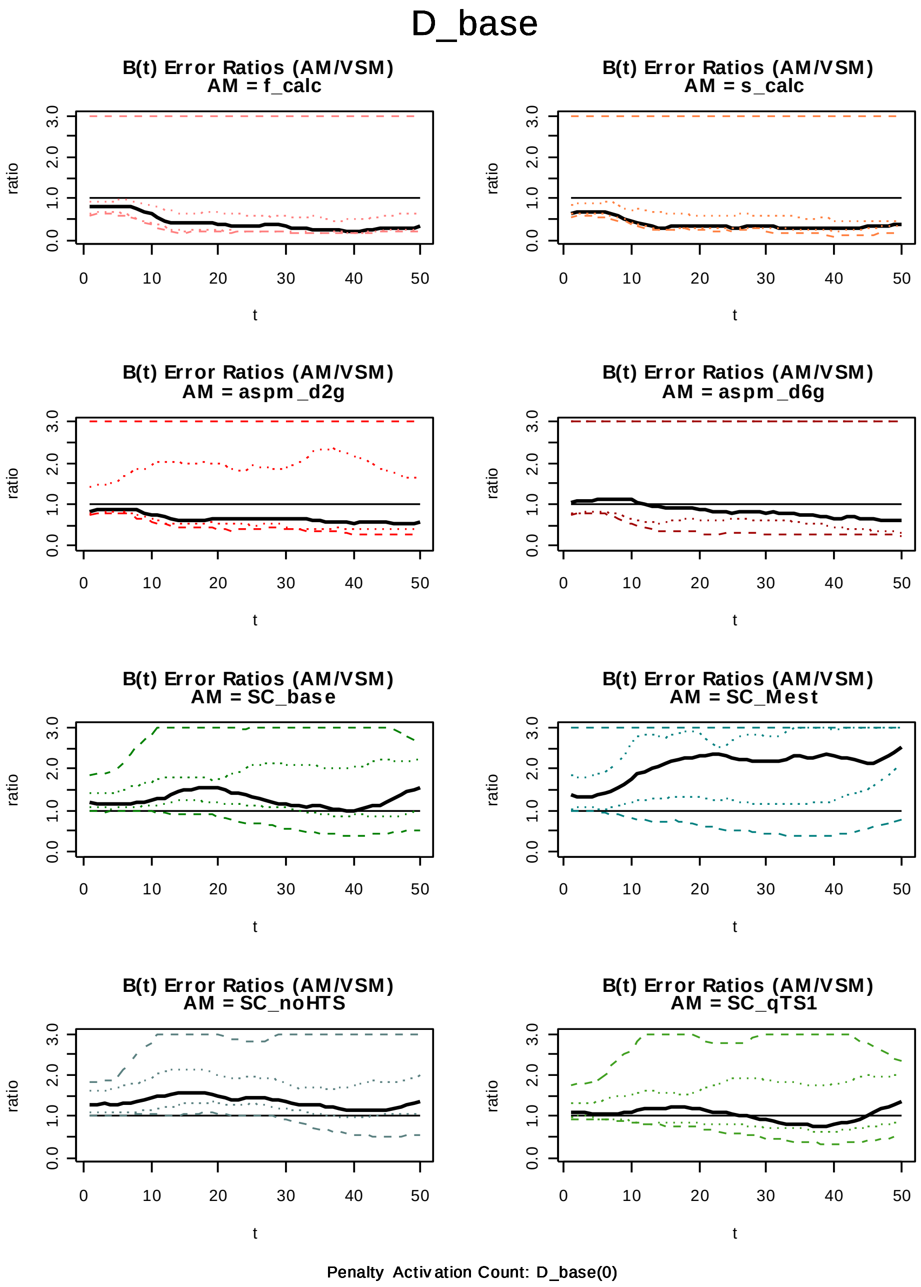


Fig. 4b. **Time series of performance indicators (AM estimated)/(OM actual) resulting from the application of AMs to 10 data realizations from the D_base OM scenario. Lines indicate rankings 1, 3, (5+6)/2, 8 and 10. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

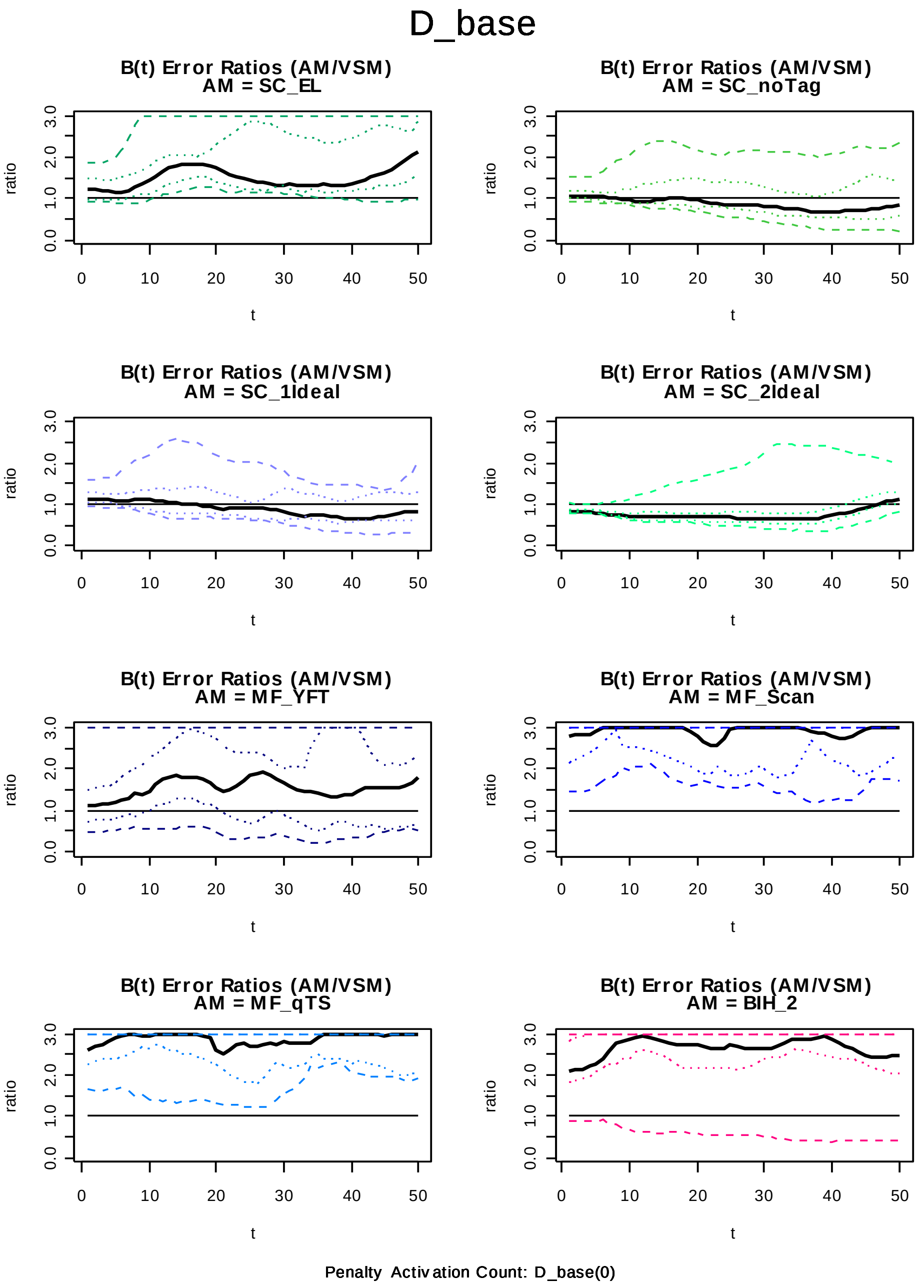

D_base
B(t) Error Ratios (AM/VSM)
AM = SC_EL
B(t) Error Ratios (AM/VSM)
AM = SC_noTag
B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
B(t) Error Ratios (AM/VSM)
AM = MF_YFT
B(t) Error Ratios (AM/VSM)
AM = MF_Scan
B(t) Error Ratios (AM/VSM)
AM = MF_qTS
B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: D_base(0)

Fig. 4b (cont.)

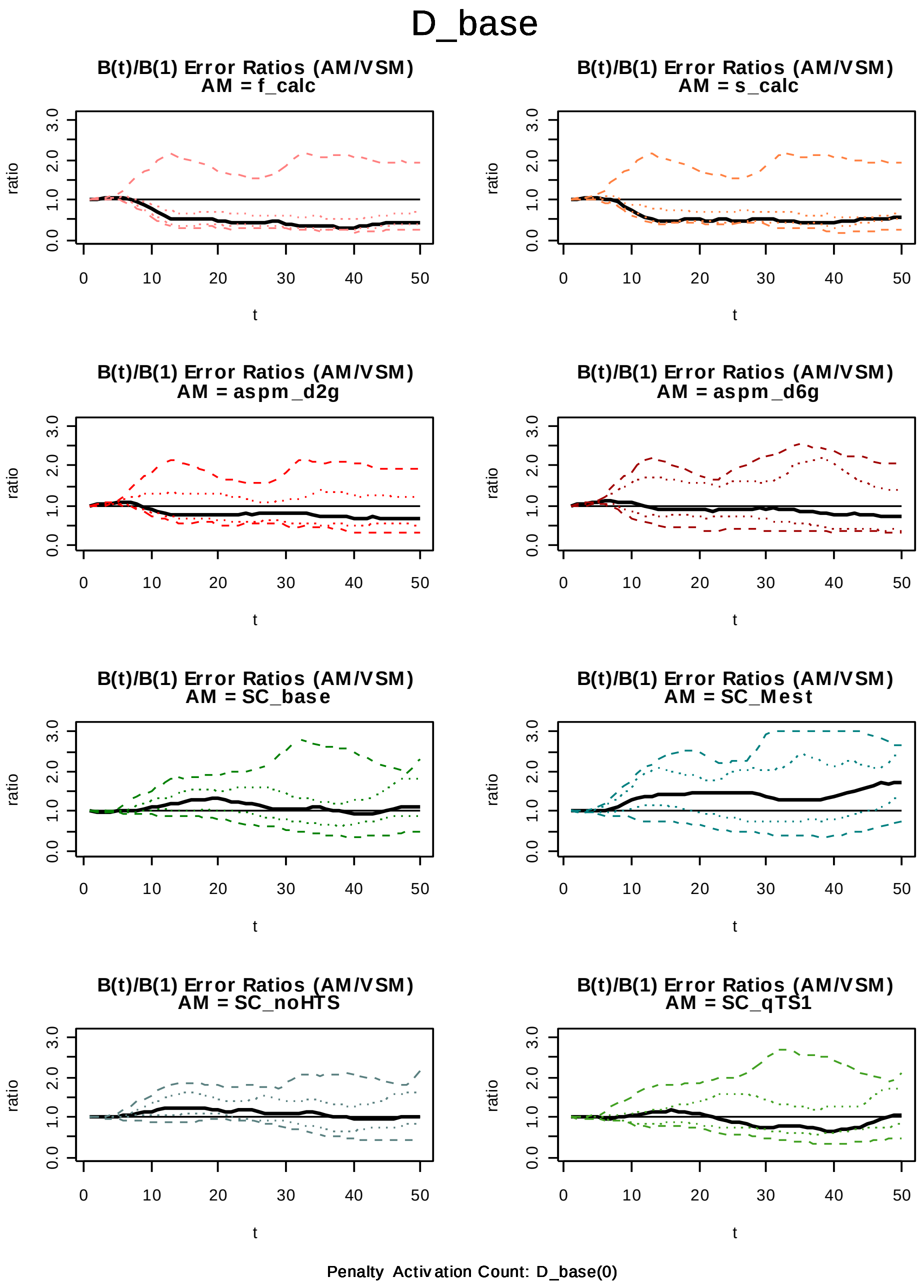

D_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = f_calc
B(t)/B(1) Error Ratios (AM/VSM)
AM = s_calc
B(t)/B(1) Error Ratios (AM/VSM)
AM = aspm_d2g
B(t)/B(1) Error Ratios (AM/VSM)
AM = aspm_d6g
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_Mest
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_noHTS
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: D_base(0)

Fig. 4b (cont.)

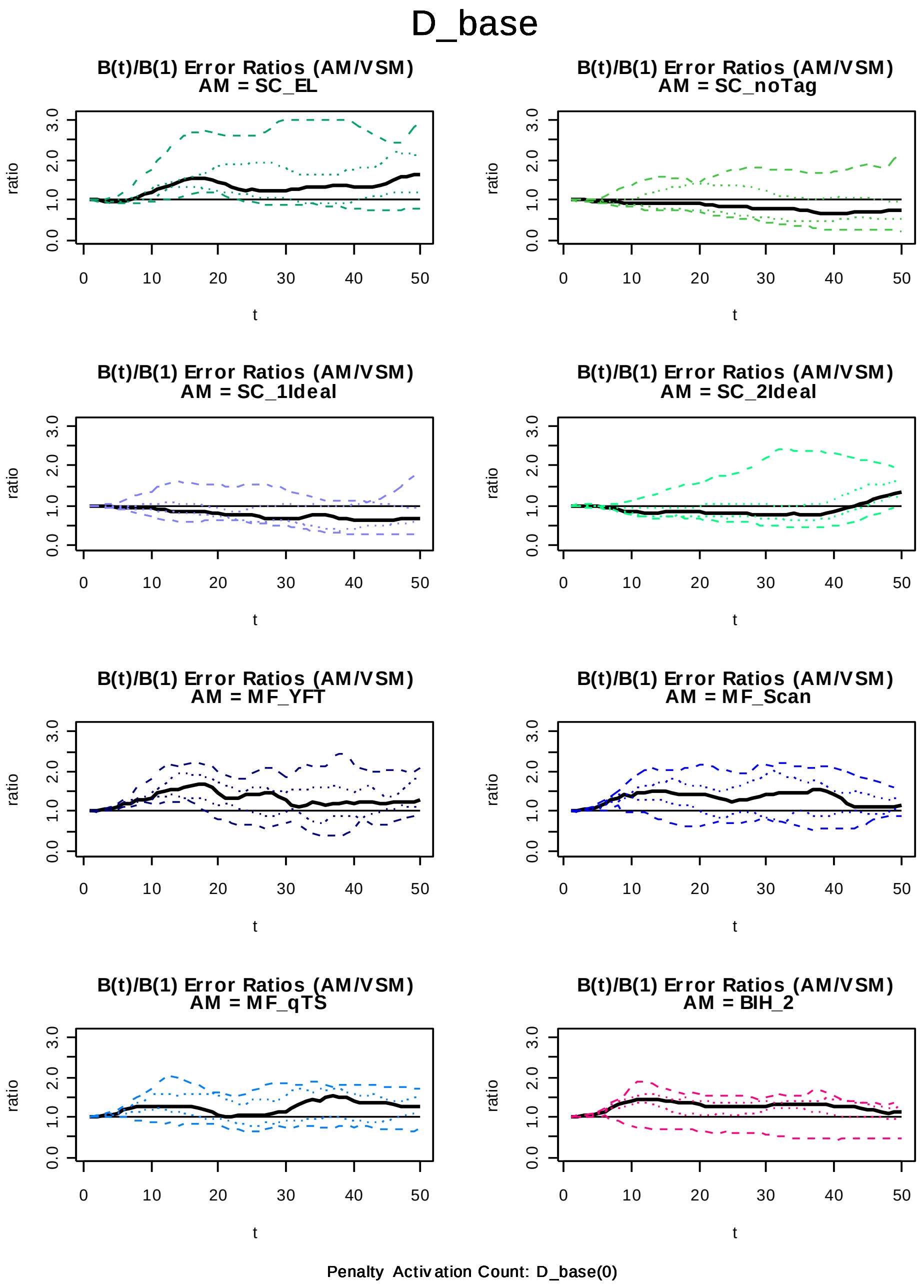

D_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_EL
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_noTag
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_2Ideal
B(t)/B(1) Error Ratios (AM/VSM)
AM = MF_YFT
B(t)/B(1) Error Ratios (AM/VSM)
AM = MF_Scan
B(t)/B(1) Error Ratios (AM/VSM)
AM = MF_qTS
B(t)/B(1) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: D_base(0)

Fig. 4b (cont.)

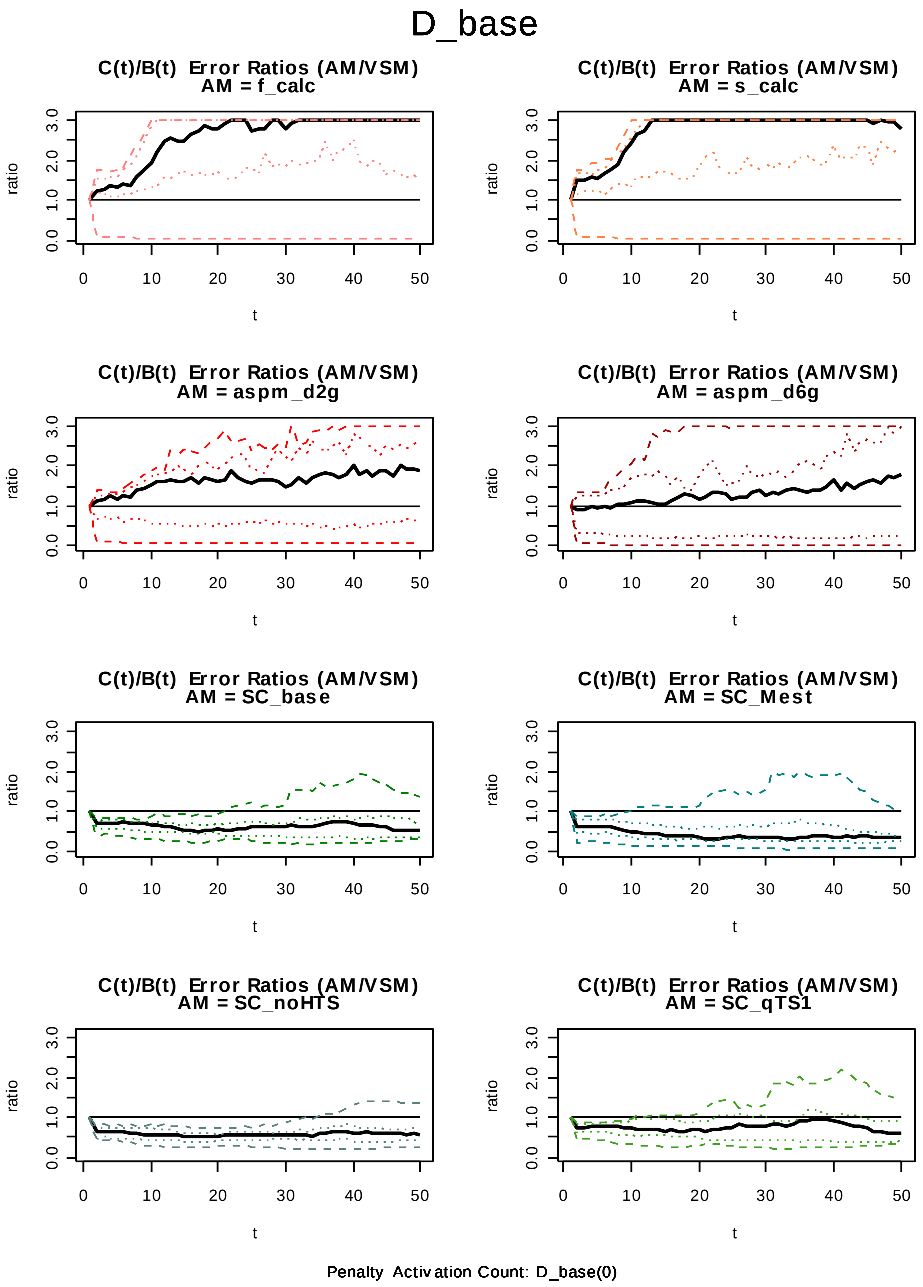

D_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = f_calc
C(t)/B(t) Error Ratios (AM/VSM)
AM = s_calc
C(t)/B(t) Error Ratios (AM/VSM)
AM = aspm_d2g
C(t)/B(t) Error Ratios (AM/VSM)
AM = aspm_d6g
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_Mest
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_noHTS
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: D_base(0)

Fig. 4b (cont.)

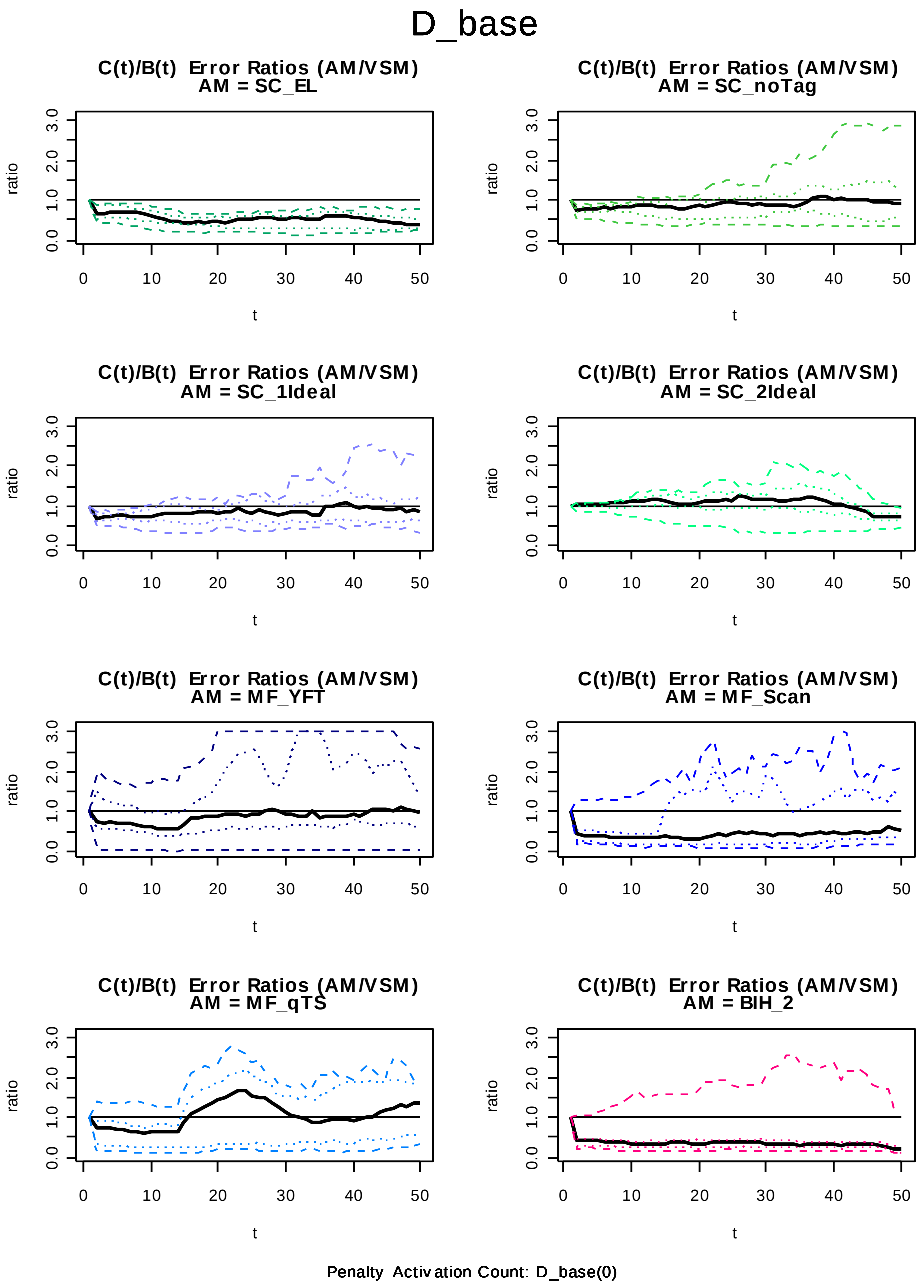

D_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_EL
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_noTag
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
C(t)/B(t) Error Ratios (AM/VSM)
AM = MF_YFT
C(t)/B(t) Error Ratios (AM/VSM)
AM = MF_Scan
C(t)/B(t) Error Ratios (AM/VSM)
AM = MF_qTS
C(t)/B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
0
10
20
30
40
50
t
Penalty Activation Count: D_base(0)

Fig. 4b (cont.)

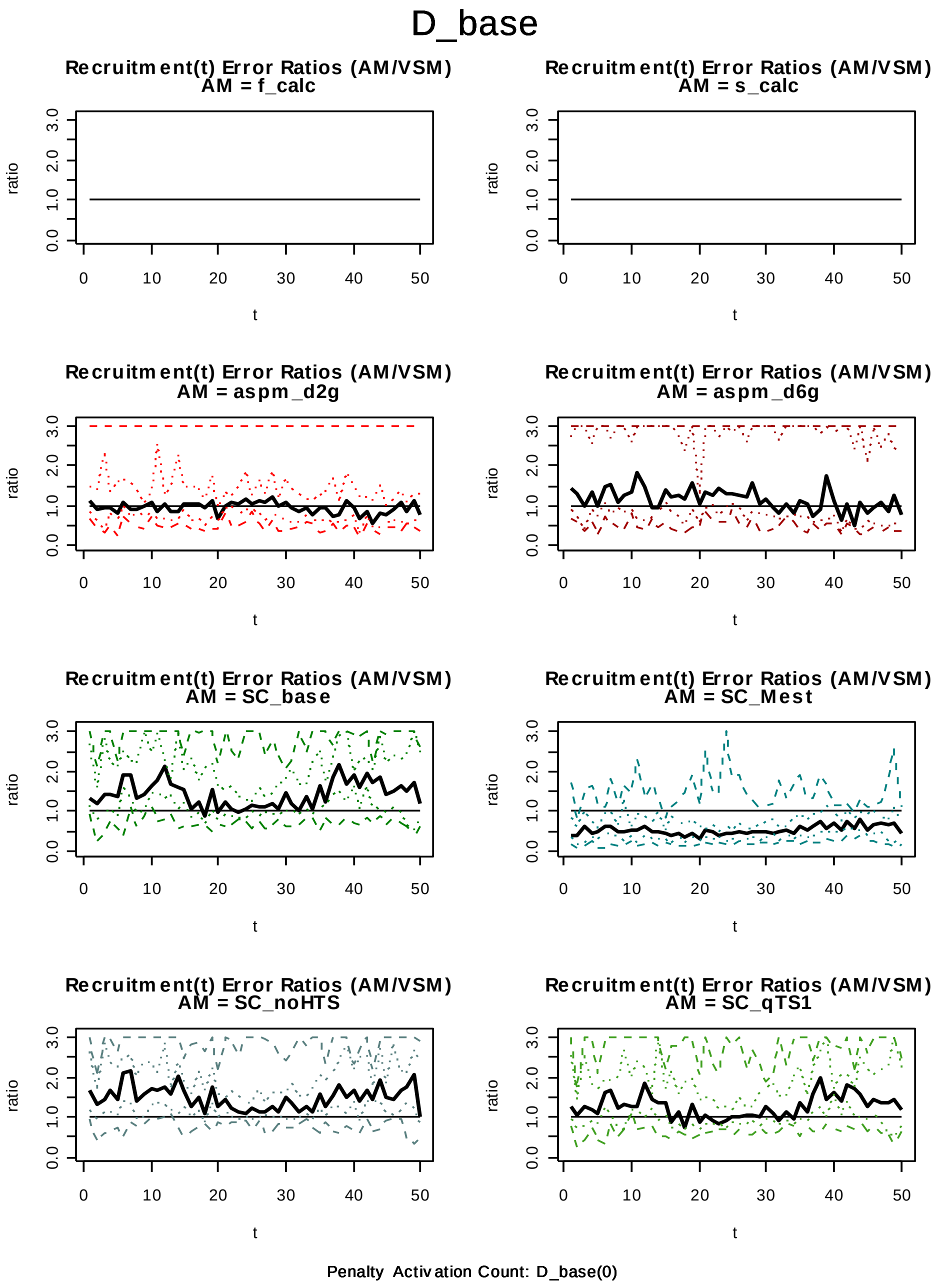

D_base
Recruitment(t) Error Ratios (AM/VSM)
AM = f_calc
Recruitment(t) Error Ratios (AM/VSM)
AM = s_calc
Recruitment(t) Error Ratios (AM/VSM)
AM = aspm_d2g
Recruitment(t) Error Ratios (AM/VSM)
AM = aspm_d6g
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_Mest
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_noHTS
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: D_base(0)

Fig. 4b (cont.)

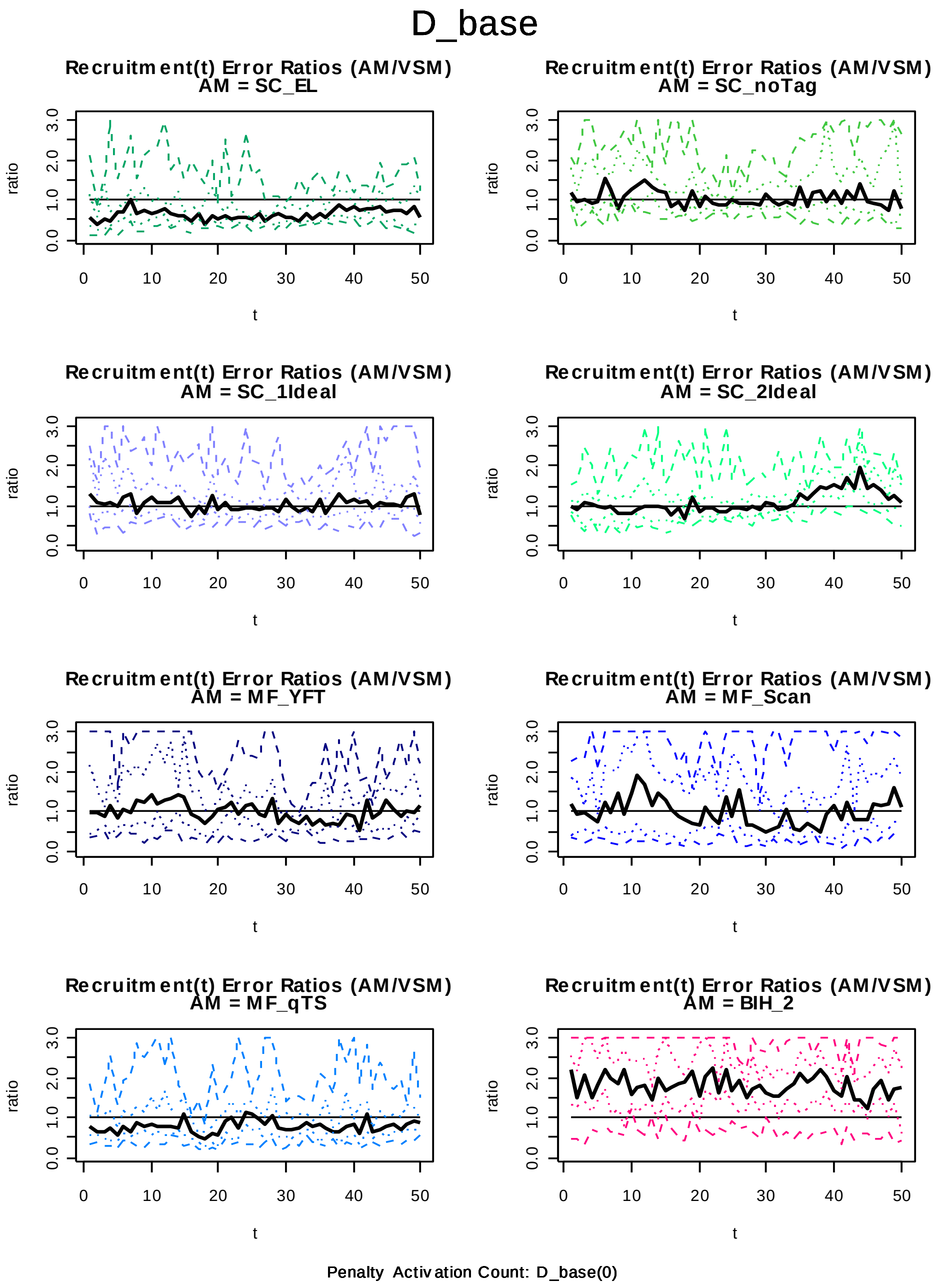

D_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_EL
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_noTag
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
Recruitment(t) Error Ratios (AM/VSM)
AM = MF_YFT
Recruitment(t) Error Ratios (AM/VSM)
AM = MF_Scan
Recruitment(t) Error Ratios (AM/VSM)
AM = MF_qTS
Recruitment(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: D_base(0)

Fig. 4b (cont.)

Four additional operating model scenarios were specified to test if the degradation in assessment performance between E_base and D_base could be disproportionately attributed to a particular specification change. Fig. 5 compares the performance of a restricted set of AMs using only the aggregate PI. These plots suggest that:

- performance of the different assessment models is affected differently by the different intermediate OM scenarios.

- The reduction in the CL and CA sample sizes (E_CL60) had a negligible effect on the AMs examined. This possibly reflects the fact that operating model length-at-age characteristics are never entirely consistent with AM assumptions (e.g. due to within year growth, natural mortality and variability in monthly catch rates), such that beyond a certain sample size, nothing more is really gained.

- Stochastic selectivity variation (E_stoH) had a negligible effect on the AMs examined.

- Systematic selectivity variation in the OM (E_HTS) had a negligible effect on the SC_x AMs examined. However, the production models appear to be adversely affected for reasons that are unclear; perhaps due to the effect of changing selectivity on the interpretation of CPUE.

- Inflated variability in effort deviations and recruitment variability (E_DRq) had a large impact on SC_base and the production models. We would expect that the auto-correlated catchability errors would be the the largest single factor contributing to the performance degradation between E_base and D_base, but presumably the interaction with the other complicating factors is not additive.

- there was surprisingly little difference in performance when AM SC_EL was applied to the various OMs. This probably reflects the fact that SC_EL is highly over-parameterized, prone to minimization failures and provides rather poor performance under a large range of conditions.

The general impressions provided by the aggregate performance indicators are further supported by examining SC_base time series estimation performance. The relative biomass estimates shown in Fig. 6 indicate a negligible effect for E_HTS, minor effects for E_CL60 and E_stoH and a major effect for E_DRq (though not as large as D_base). E_DRq is the operating model with substantial and auto-correlated errors in the relationship between fishing mortality and effort, and the estimation errors observed for this scenario emphasizes the key role of the relative abundance index in most stock assessment models. The corresponding series of recruitment estimates suggest some slightly different bias characteristics. E_CL60 actually results in better recruitment estimates than E_base, because results for E_base have strong bias trends in the last ~15 years (as do results for E_stoH and E_HTS). There is presumably some structural incompatability between E_base and SC_base that causes a recruitment bias when SCALIA gives too much weight to the CL data. The small CL sample size presumably results in a noisy signal that SC_base cannot track too closely, because the modal progression signal is weak among consecutive years. However, we note that this speculation is not consistent with the fact that SC_1ideal (with a large CL effective sample size) did not have a recruitment bias problem for E_base (Fig. 3b), so it seems likely that some other interaction is at work. Recruitment estimates also show high variability in the early part of E_stoH. This presumably reflects the fact that the stochastic selectivity in E_stoH is dependent on the amount of effort, and all fisheries have low effort in the early years except for the spawning ground fleet which is not informative for recruitment.

We hope that the magnitude of the baseline operating model error characteristics that we have defined in E_base and D_base envelop the real SBT situation (ignoring the other major assumptions that we treat separately). In the following sections, we would tend to interpret E_x scenarios as the upper limits of how well we could expect to do in a real assessment. If the simulated assessments are failing in the E_x scenarios, we have serious concerns about performance in a real assessment. In contrast, we believe that many of the characteristics in D_x might be more challenging than the real world data. Consistently reliable inferences from the D_x series would give us reasonable confidence that the estimates would probably be okay in many real world applications.

The use of only 10 realizations for each assessment model/operating model combination is rather minimal. Certainly we expect the estimation behaviour in the tails to be poorly described, but the 10 replicates appear sufficient to illustrate the large qualitative differences in estimation bias and variance characteristics among assessment models. We note that the 2003 MWG used 40 simulated data sets for each OM scenario, but the general character of the results was not obviously improved.

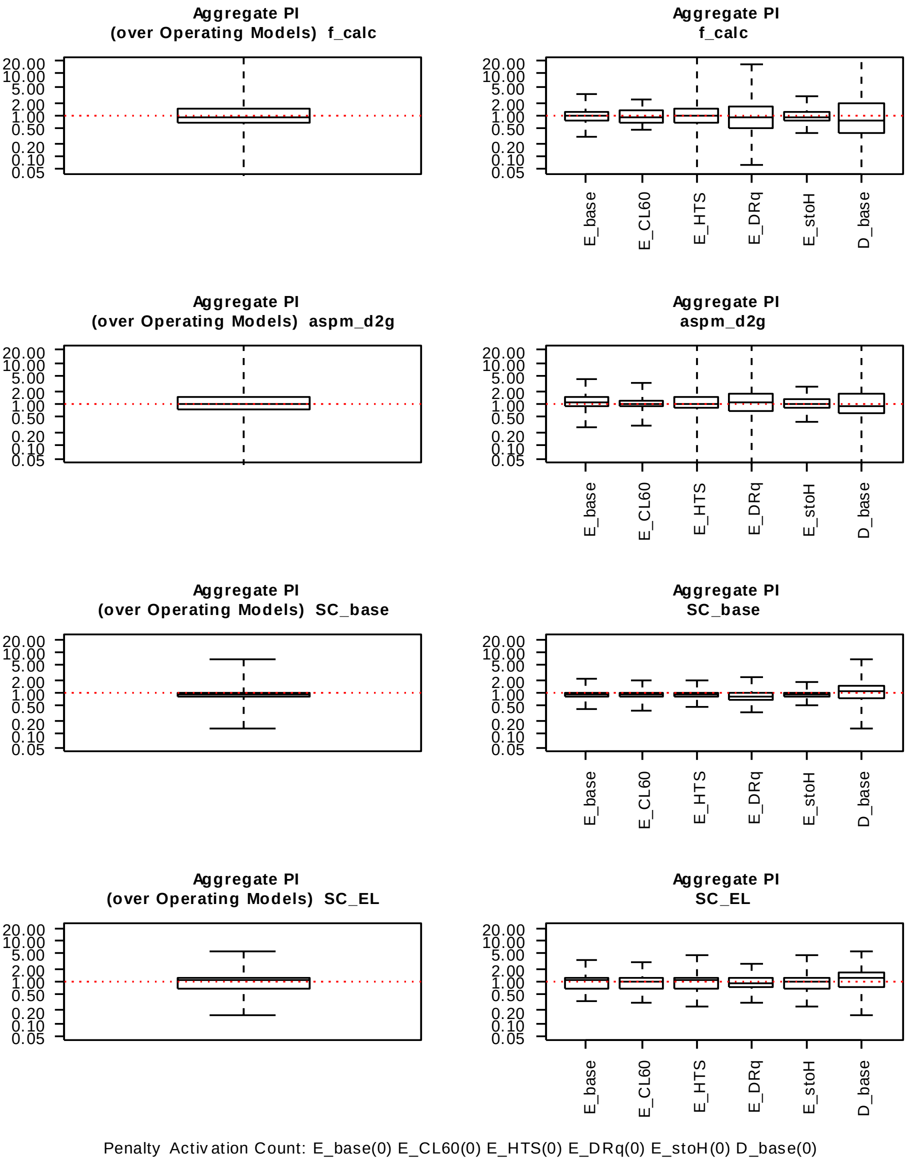


Fig. 5. **Aggregate performance of assessment models when fit to a range of OMs including the relatively easy E_base, difficult D_base and 4 intermediate scenarios. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

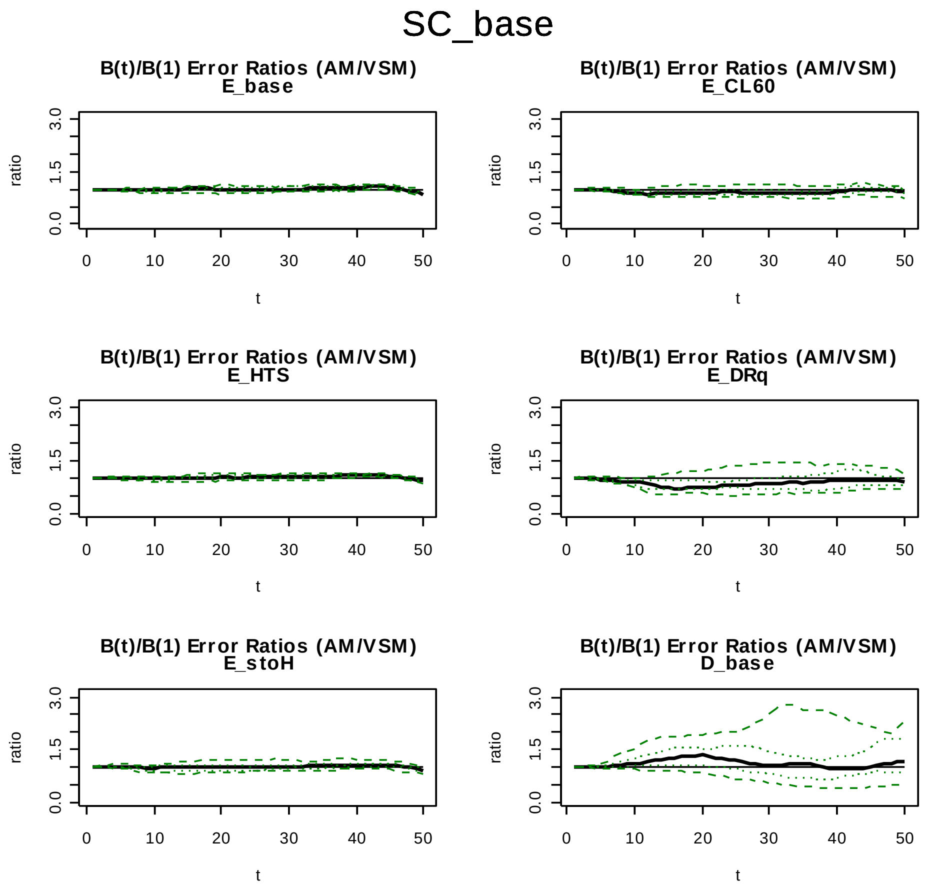


Fig. 6. Time series of performance indicators (AM estimated)/(OM actual) resulting from the application of AM SC_base to 10 data realizations from a range of OMs including E_base, D_base and 4 intermediate scenarios. Lines indicate rankings 1, 3, (5+6)/2, 8 and 10. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

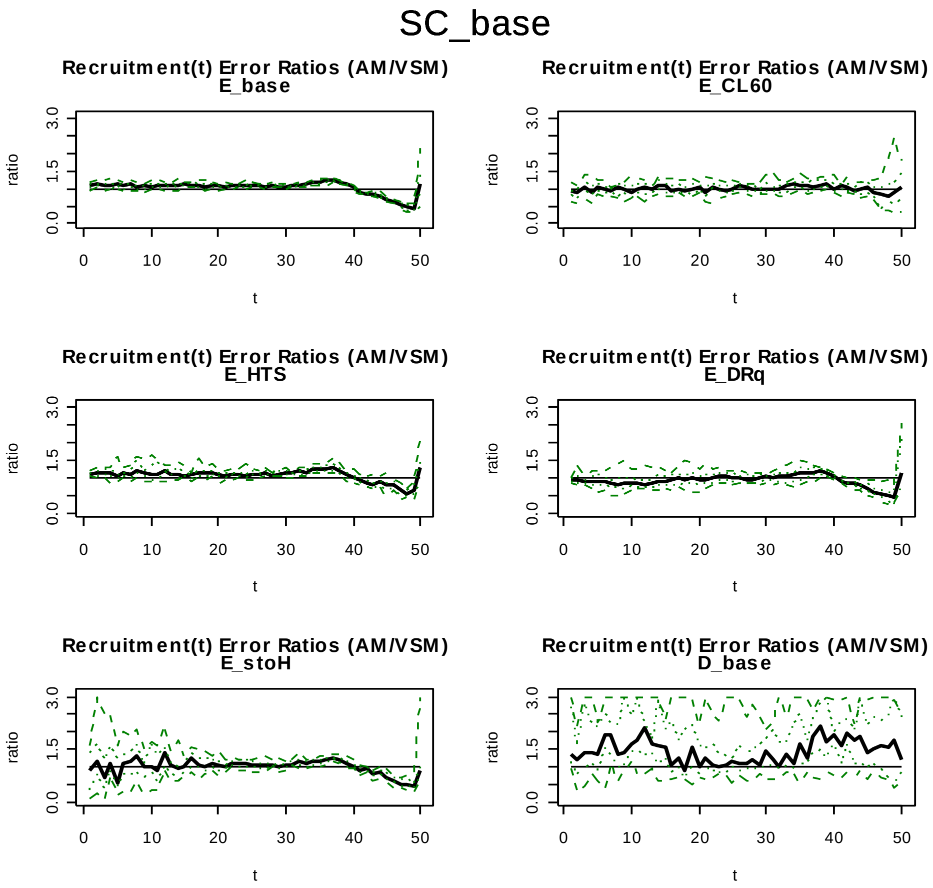

SC_base
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_CL60
Recruitment(t) Error Ratios (AM/VSM)
E_HTS
Recruitment(t) Error Ratios (AM/VSM)
E_DRq
Recruitment(t) Error Ratios (AM/VSM)
E_stoH
Recruitment(t) Error Ratios (AM/VSM)
D_base
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 6 (cont).

5.3 Objective I - Stock Recruitment Relationship Estimation

Estimation of stock recruitment relationships is important for quantifying how the productivity of the stock is likely to change as spawning biomass changes, and has important implications for estimating sustainable catches. In the case of the SBT fishery, the spawning stock is near the lowest ever levels, and the stock recruitment assumptions are key factors driving the future productivity scenarios explored in the operating models used to evaluate candidate MPs (see Operating Model exploration in CCSBT 2003). Three factors related to future recruitment potentially have a large influence on the management actions that are likely to be taken for SBT in the near future: the degree of compensation in the stock recruitment relationship (steepness), the magnitude of the year to year variability in recruitment, and the auto-correlation in recruitment variability (the degree to which recruitment can deviate from the long term expected value for a sustained time period). In this section our discussions emphasize these points over the other stock assessment inference issues.

Fig. 7-Fig. 11 illustrate our exploration into the estimability of stock recruitment relationships using a range of assessment models. These figures include the actual steepness estimates, empirical auto-correlation among recruitment deviations and the emperically calculated variability (RMSE) of the estimated recruitment deviations (but note that not all values were available or even relevant for some assessment models). The aggregate PIs and management status indicators are included to provide a general indication of overall model performance. Refer to Appendix 6 for more detailed description of model performance for specific indicators (including time series estimates of recruitment).

We note the following points regarding the estimation of stock recruitment relationships and model performance under different production scenarios given excellent data and prior knowledge about the functional form of the Beverton-Holt stock recruitment relationship (Fig. 7):

- SCALIA and ASPM models had reasonable capacity to distinguish between high (E_h6) and low (E_h3) productivity curves (Fig. 7a-c), but there were considerable estimation errors evident. Performance in the E_h3 scenario is difficult to compare because all the SCALIA models included a lower steepness bound of 0.3 for calculating MSY-related quantities (and hence convergence to the lower bound results in a perfect steepness estimate for purposes of MSY calculations).

- The majority of assessment models seemed to have a steepness under-estimation bias (MPD estimates often converged to the lower bound which corresponds to an absence of surplus production). aspm_d2g and SC_2Ideal were better than most, and SC_BIH was clearly the worst. But we note that the MF_x models tended to over-estimate steepness in the baseline scenarios (these models were not run against most OM scenarios and MF_x applications to E_base are not repeated here).

- aspm_d2g generally provided the best estimates of management-related quantities (at least among those examined here). However, this model had greater variability and biases in biomass estimates than many of the SCALIA models (evident in the terminal relative biomass estimates, aggregate performance indicators and Appendix 6).

- The SCALIA models produced empirical estimates of recruitment variability (Rec RMSE) that were very similar to the true values, despite the different input specifications (CV = 0.4 or 0.6).

- The empirical auto-correlation was fairly consistent among SCALIA models and usually 0 < (Rec lag(1) rho) < 0.4 (compared with the actual OM value of 0); SC_2Ideal and SC_BIH were substantially higher. In the case of SC_BIH, we would expect this to be due to age estimation via cohort-slicing. Presumably the auto-correlation > 0 in most cases is related to the systematic lack of fit that arises in part because of the errors in steepness.

The effect of adding recruitment auto-correlation (rho = 0.8) to two operating models with high (E_h8_r8) and low (E_h4_r8) productivity is illustrated in Fig. 8, from which we note:

- the variance in the steepness estimates increased relative to the OM scenarios with no auto-correlation, and the under-estimation bias remained.

- The SCALIA recruitment deviation RMSE was generally substantially lower than the real CV (irrespective of the different input values of 0.4 or 0.6). This illustrates the manner in which strong auto-correlation makes it more difficult to distinguish between different stock recruitment curves, particularly with a relatively small time series (e.g. data might be equally consistent with a low steepness curve and small independent deviations, or a high steepness curve with large correlated deviations (positive at high SSB and/or negative at low SSB). It also indicates that estimation of the variance around the stock recruitment relationship might be difficult if there truly is auto-correlation (or a systematic lack of fit to the SR function).

- The empirical auto-correlation from all the AMs examined was fairly consistent with the actual OM values.

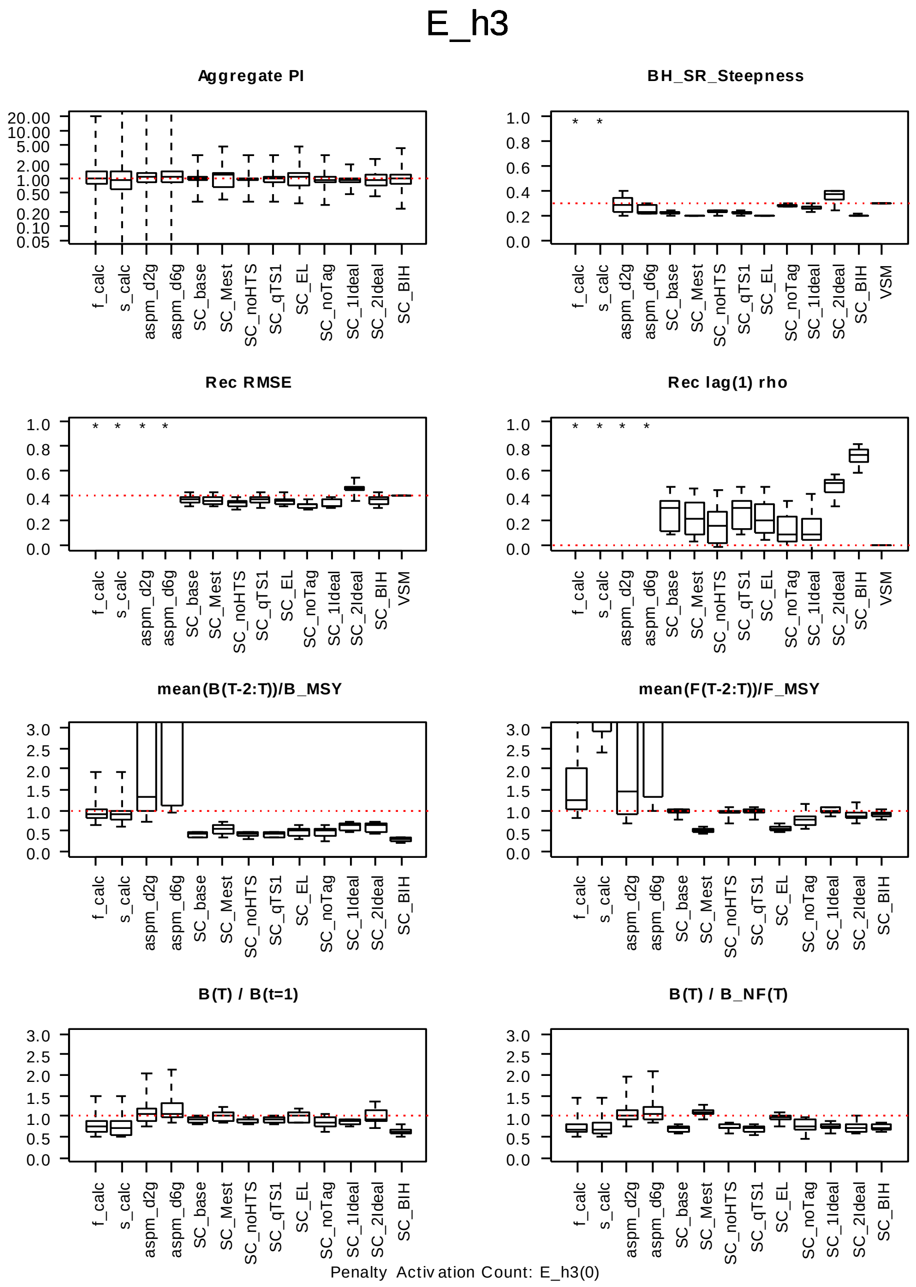


Fig. 7a. **Boxplots of MPD performance indicators resulting from the application of a range of assessment models to operating models with excellent data characteristics and a range of stock recruitment relationships and no recruitment deviation auto-correlation. Individual values represent the ratio of (AM estimated)/(OM actual) in all cases except BH_SR_Steepness, Rec RMSE and Rec lag(1) rho which show the actual values. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8**

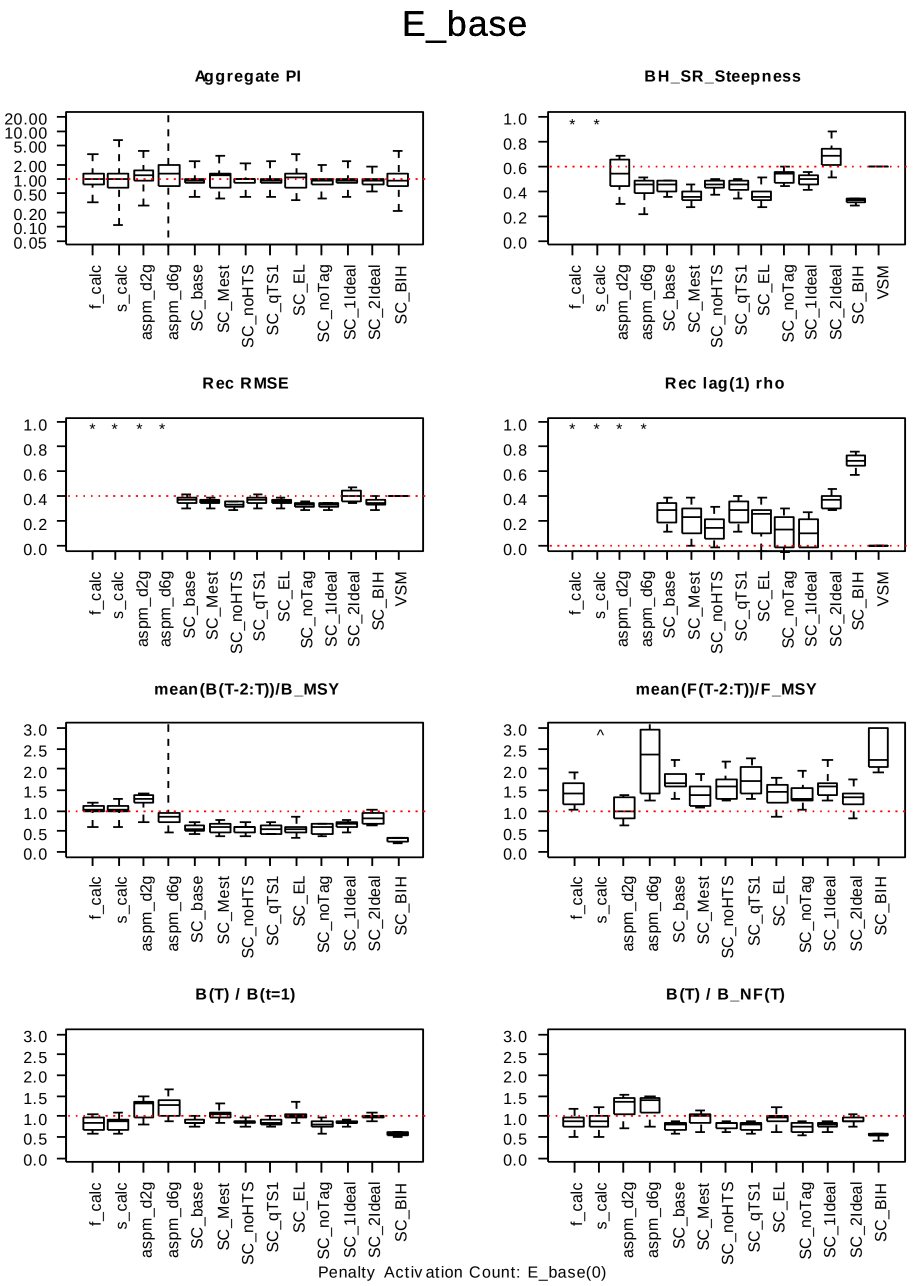
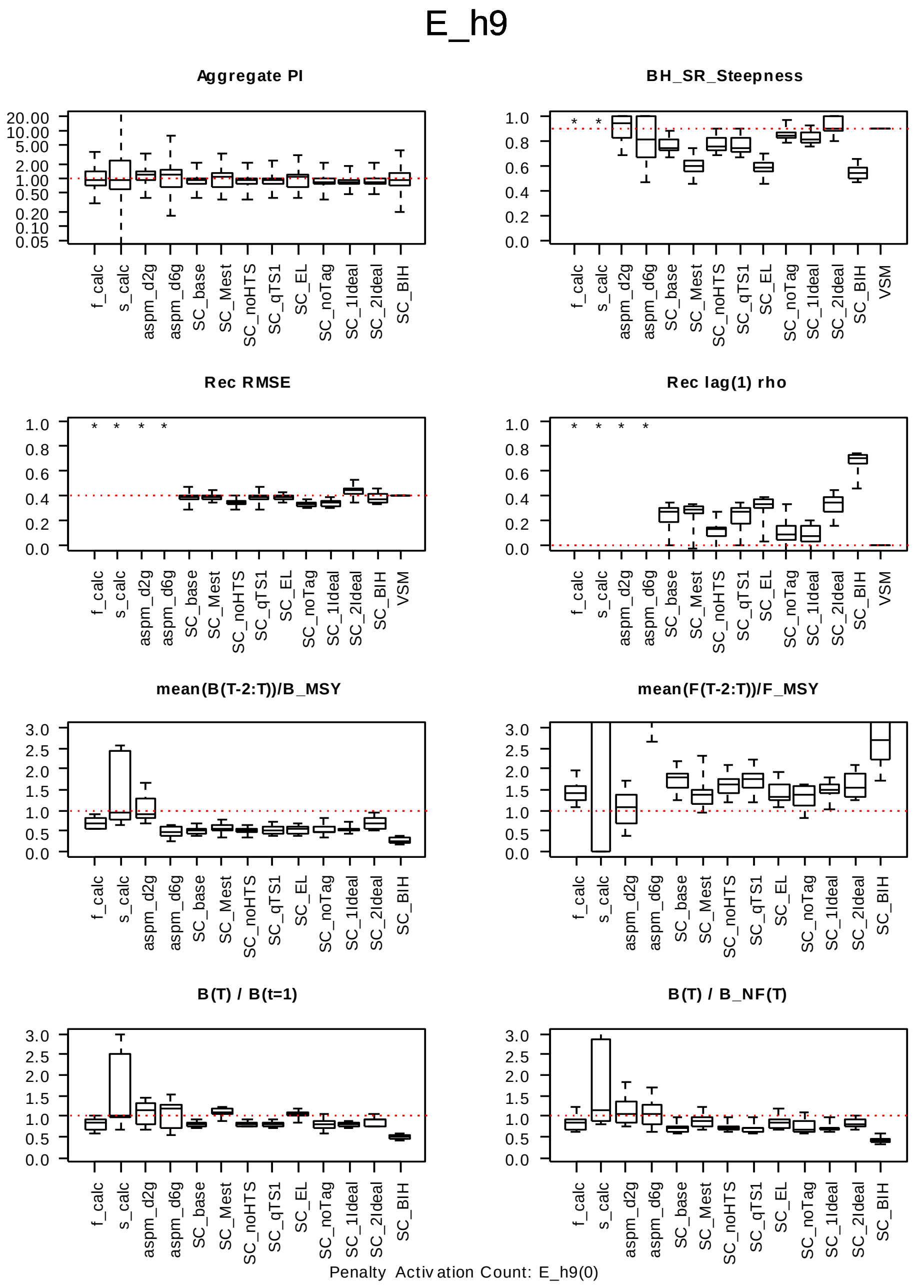

E_base
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
f_calc
s_calc
aspm_d2g
aspm_d6g
SC_base
SC_Mest
SC_noHTS
SC_qTS1
SC_EL
SC_noTag
SC_1Ideal
SC_2Ideal
SC_BIH
BH_SR_Steepness
1.0
0.8
0.6
0.4
0.2
0.0
VSM
Rec RMSE
Rec lag(1) rho
mean(B(T-2:T))/B_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
mean(F(T-2:T))/F_MSY
B(T) / B(t=1)
B(T) / B_NF(T)
Penalty Activation Count: E_base(0)

Fig. 7b.



E_h9
Aggregate PI
BH_SR_Steepness
Rec RMSE
Rec lag(1) rho
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
B(T) / B(t=1)
B(T) / B_NF(T)
f_calc
s_calc
aspm_d2g
aspm_d6g
SC_base
SC_Mest
SC_noHTS
SC_qTS1
SC_EL
SC_noTag
SC_1Ideal
SC_2Ideal
SC_BIH
VSM
Penalty Activation Count: E_h9(0)

Fig. 7c.

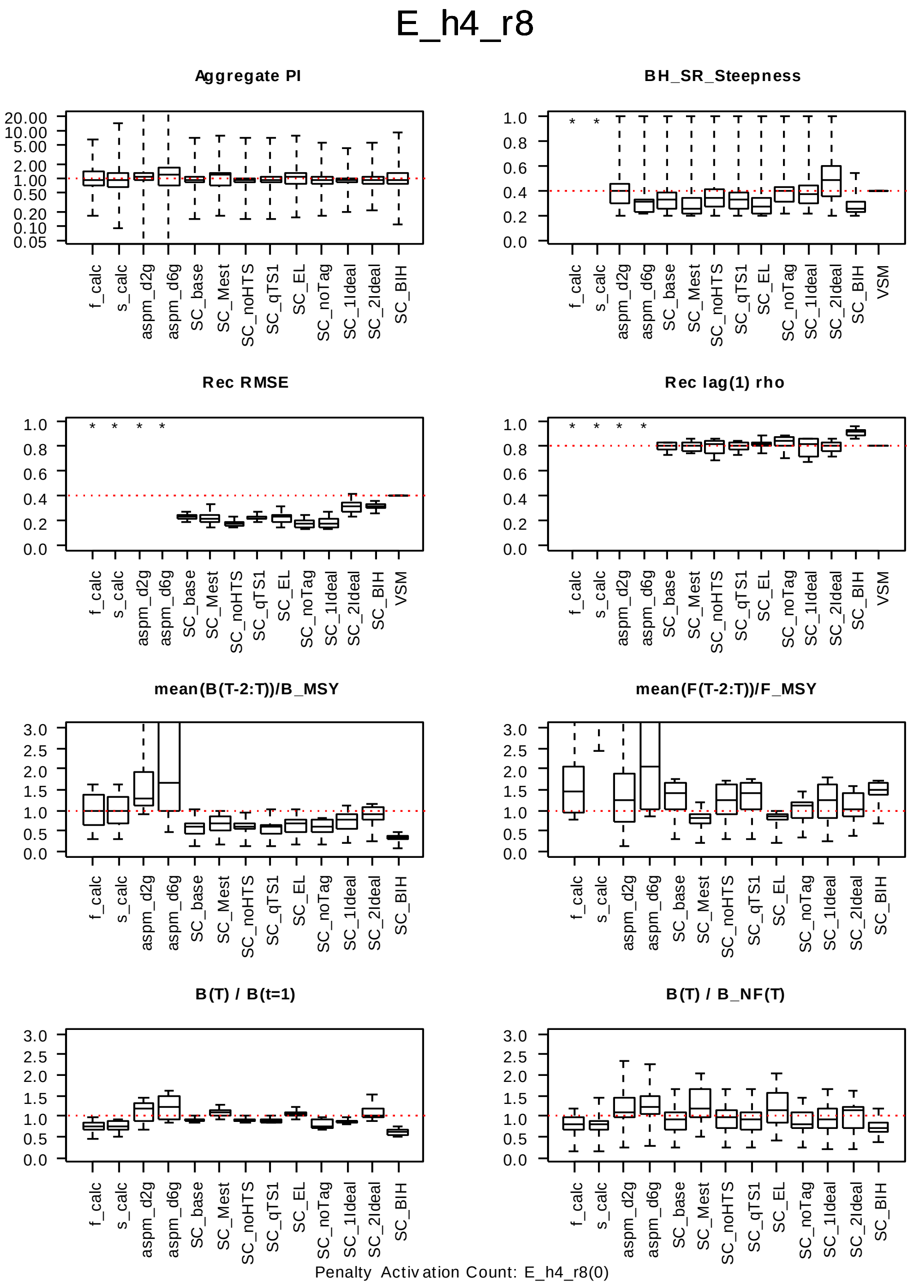


Fig. 8a. Boxplots of MPD performance indicators resulting from the application of a range of assessment models to operating models with excellent data characteristics, a range of stock recruitment relationships and high recruitment deviation auto-correlation. Individual values represent the ratio of (AM estimated)/(OM actual) in all cases except BH_SR_Steepness, Rec RMSE and Rec lag(1) rho which show the actual values. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8

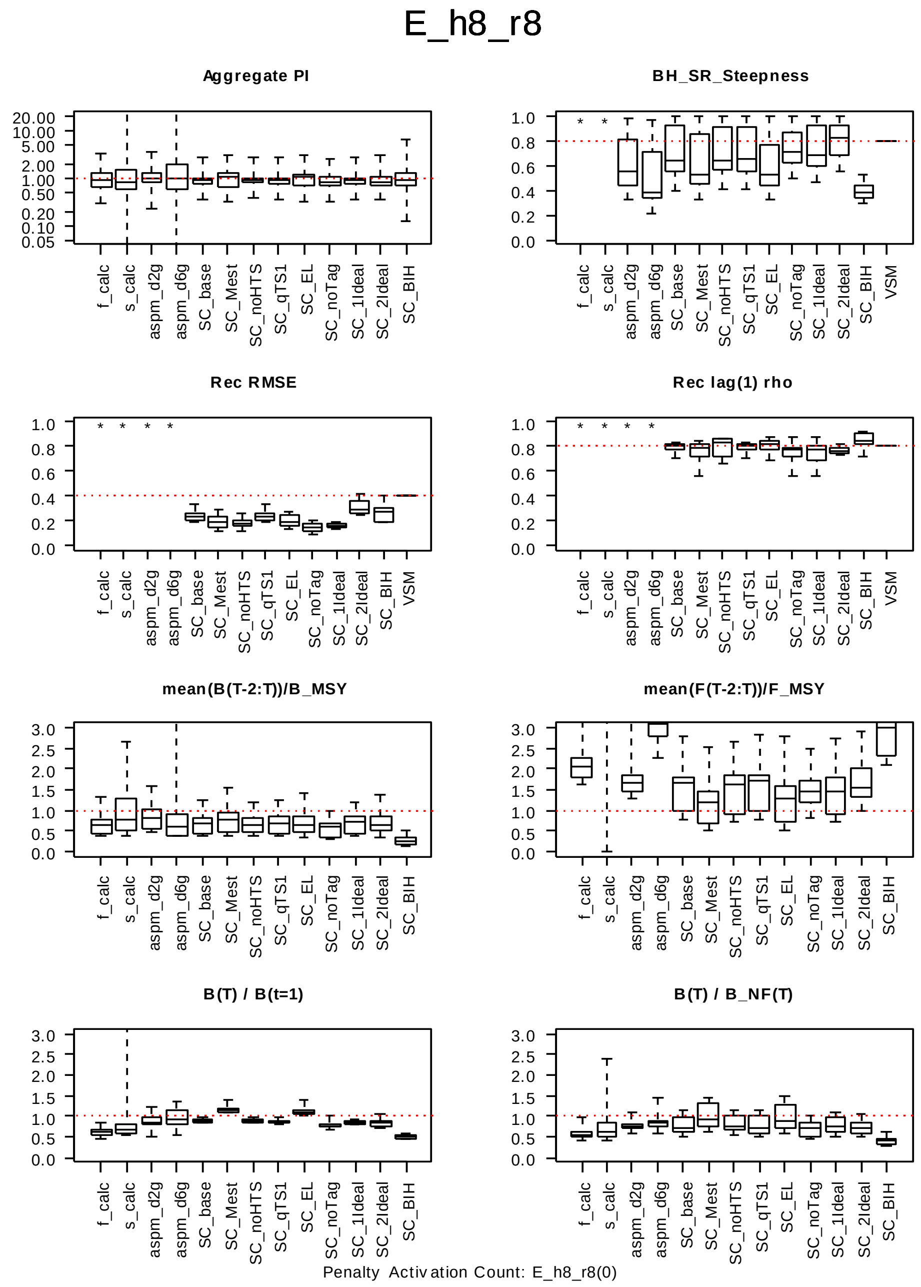

E_h8_r8
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
BH_SR_Steepness
Rec RMSE
Rec lag(1) rho
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
B(T) / B(t=1)
B(T) / B_NF(T)
1.0
0.8
0.6
0.4
0.2
0.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
f_calc
s_calc
aspm_d2g
aspm_d6g
SC_base
SC_Mest
SC_noHTS
SC_qTS1
SC_EL
SC_noTag
SC_1Ideal
SC_2Ideal
SC_BIH
VSM
Penalty Activation Count: E_h8_r8(0)

Fig. 8b.

Given the difficult D_x OM scenarios, and again prior knowledge of the functional form of the stock recruitment relationship, we make the following comments about the assessment model performance (from Fig. 9):

- bias and variance in the SR steepness estimates increased relative to the E_x scenarios, but most AMs did seem to have some capacity to distinguish high from low steepness on average. The consistently better performers included aspm_d2g, SC_base, SC_noHTS, SC_qTS1 and SC_2ideal. But all AMs got the steepness badly wrong in at least some realizations.

- Recruitment deviation RMSE remained fairly consistent among assessment model specifications. There was a slight under-estimation bias when auto-correlation was low, but there was no obvious dis-agreement when auto-correlation was high (unlike the E_x scenarios).

- Relative to the E_x scenarios without auto-correlation, the empirical recruitment auto-correlation output from the SCALIA models was much higher and more variable (median values often >0.5). When high auto-correlation was present in the D_x OMs, the empirical AM estimates of auto-correlation were much more variable in comparison with the E_x scenarios, and were generally lower than the true values.

- All AMs examined had much worse overall performance for the D_x OM scenarios than the E_x scenarios. We would generally conclude that the age-aggregated production models were worse than the rest. It is difficult to generalize among the SCALIA and aspm_x models, but on the basis of the aggregate performance indicators, we would probably conclude that SC_2Ideal performed the best against the D_h3, D_base and D_h9 OMs.

When auto-correlation was added to the operating models (D_h4_r4, D_h8_r8), the capacity to estimate steepness was further decreased (Fig. 10), with most models having a substantial probability of getting the steepness estimate badly wrong. It also becomes increasingly difficult to make meaningful comments about relative performance of AMs when all models are performing this poorly, but we can probably conclude in the basis of the aggregate PI that the AAPMs performed worse than the others.

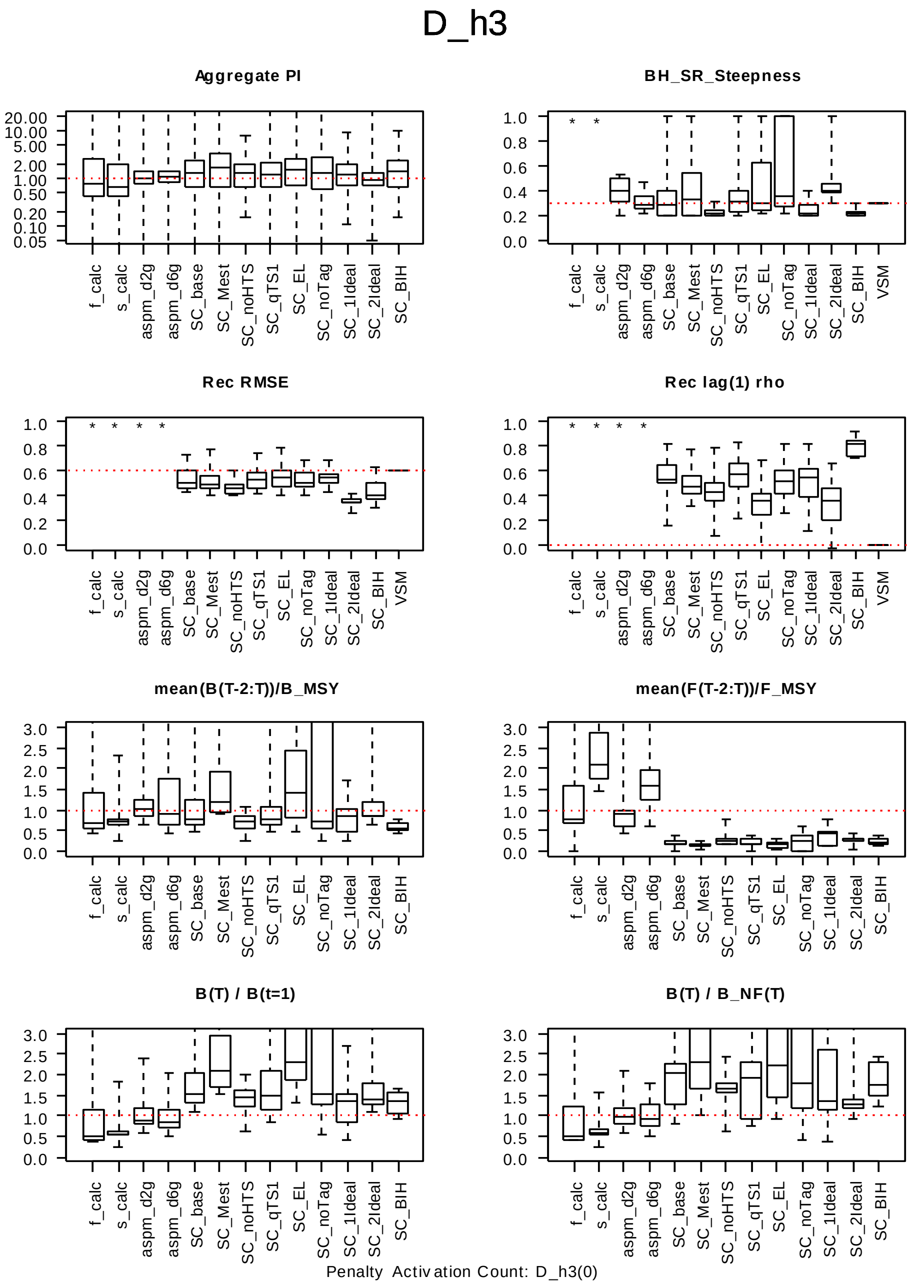


Fig. 9a. Boxplots of MPD performance indicators resulting from the application of a range of assessment models to operating models with difficult data characteristics, a range of stock recruitment relationships and no recruitment deviation auto-correlation. Individual values represent the ratio of (AM estimated)/(OM actual) in all cases except BH_SR_Steepness, Rec RMSE and Rec lag(1) rho which show the actual values. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

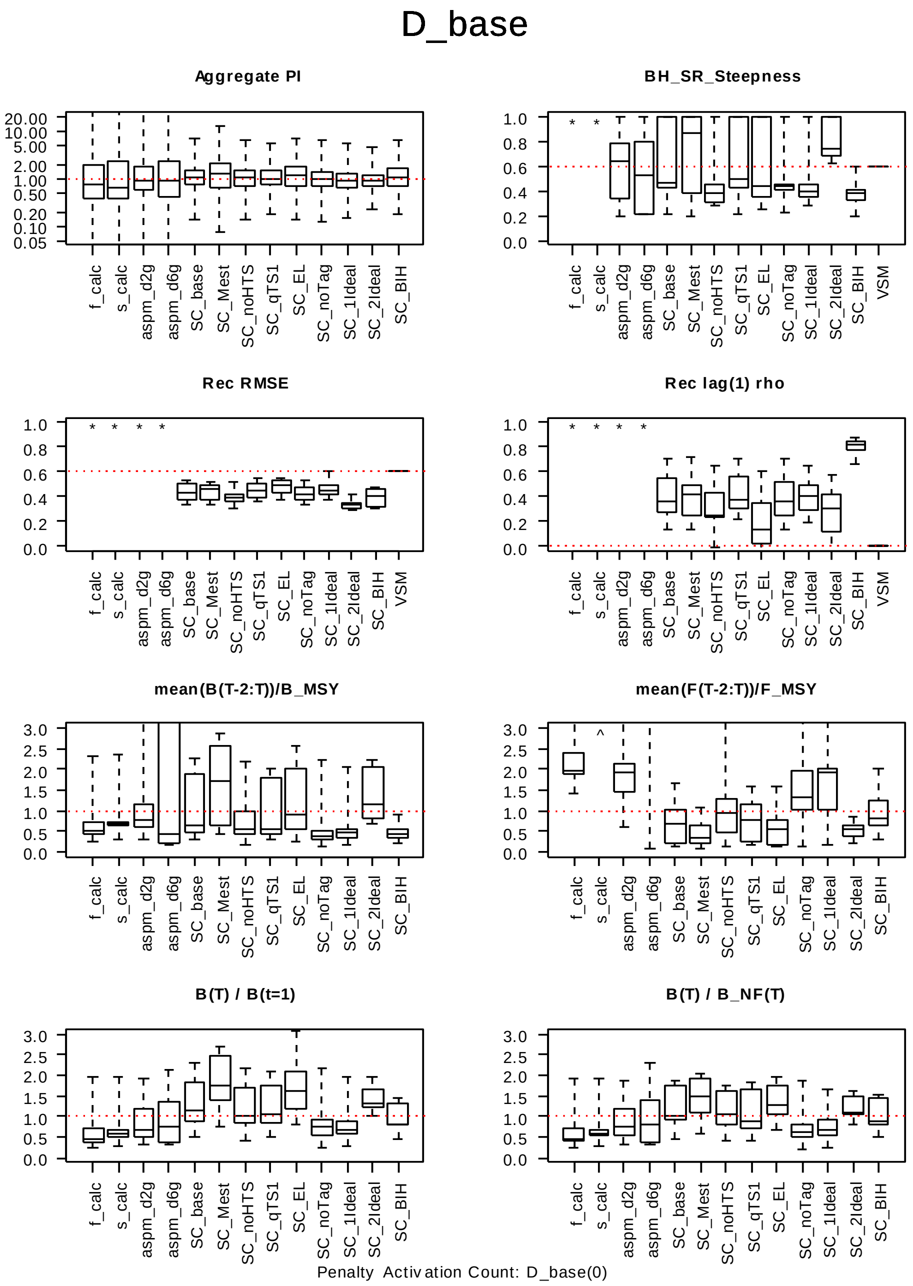

D_base
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
f_calc
s_calc
aspm_d2g
aspm_d6g
SC_base
SC_Mest
SC_noHTS
SC_qTS1
SC_EL
SC_noTag
SC_1Ideal
SC_2Ideal
SC_BIH
BH_SR_Steepness
1.0
0.8
0.6
0.4
0.2
0.0
VSM
Rec RMSE
Rec lag(1) rho
mean(B(T-2:T))/B_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
mean(F(T-2:T))/F_MSY
B(T) / B(t=1)
B(T) / B_NF(T)
Penalty Activation Count: D_base(0)

Fig. 9b.

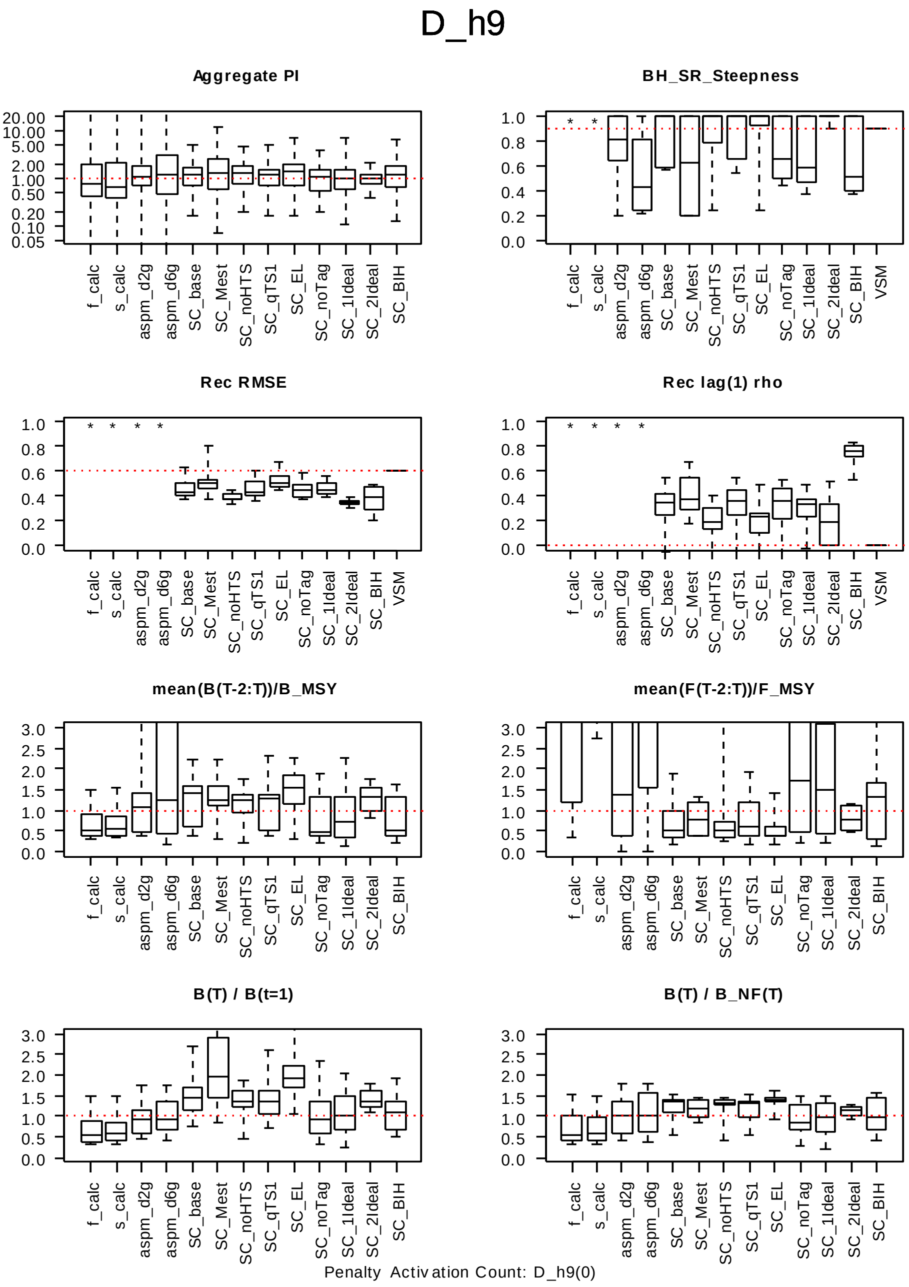

D_h9
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
f_calc
s_calc
aspm_d2g
aspm_d6g
SC_base
SC_Mest
SC_noHTS
SC_qTS1
SC_EL
SC_noTag
SC_1Ideal
SC_2Ideal
SC_BIH
BH_SR_Steepness
1.0
0.8
0.6
0.4
0.2
0.0
*
*
VSM
Rec RMSE
* * * *
Rec lag(1) rho
* * * *
mean(B(T-2:T))/B_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
mean(F(T-2:T))/F_MSY
B(T) / B(t=1)
B(T) / B_NF(T)
Penalty Activation Count: D_h9(0)

Fig. 9c.

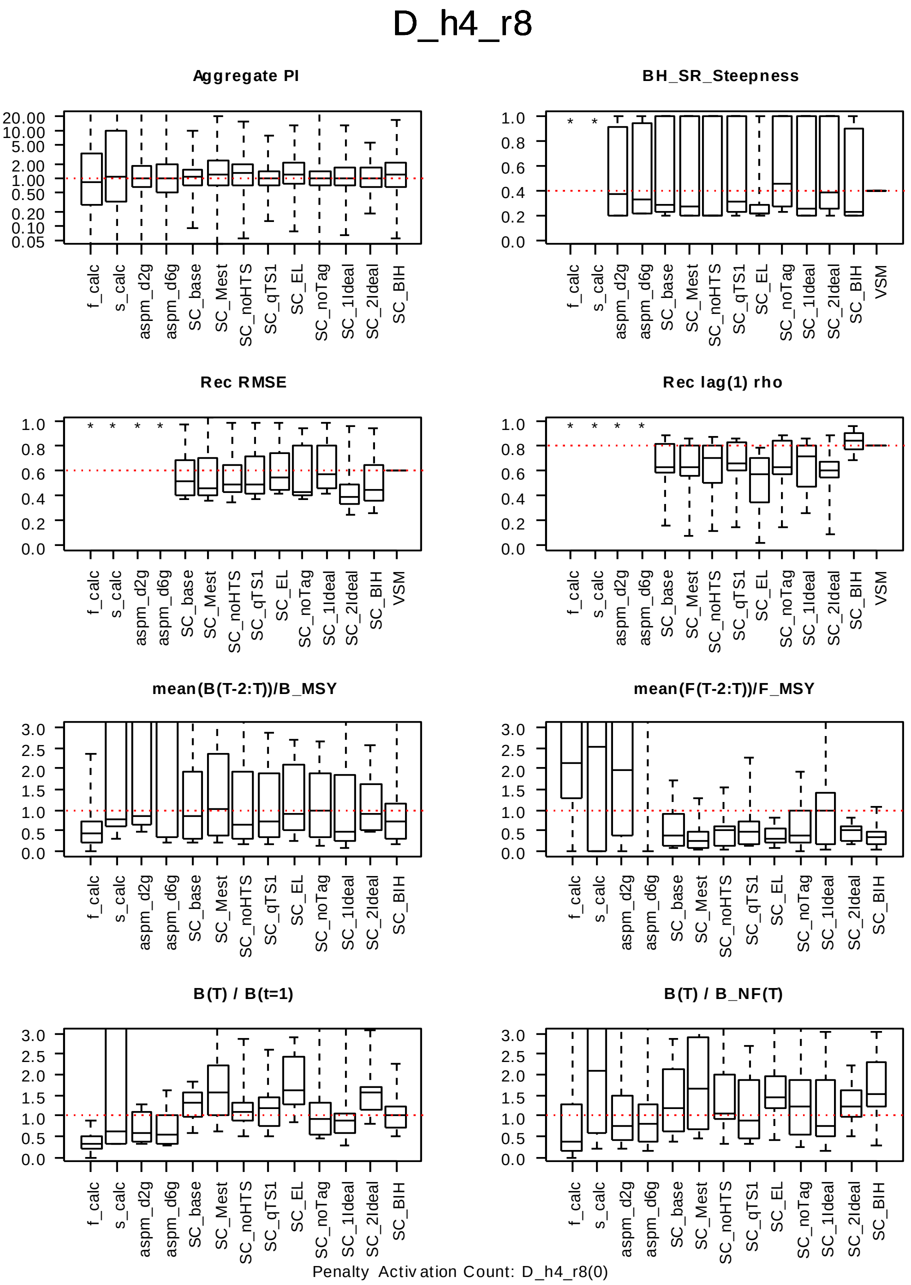


Fig. 10a. **Boxplots of MPD performance indicators resulting from the application of a range of assessment models to operating models with difficult data characteristics, a range of stock recruitment relationships and no recruitment deviation auto-correlation. Individual values represent the ratio of (AM estimated)/(OM actual) in all cases except BH_SR_Steepness, Rec RMSE and Rec lag(1) rho which show the actual values. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

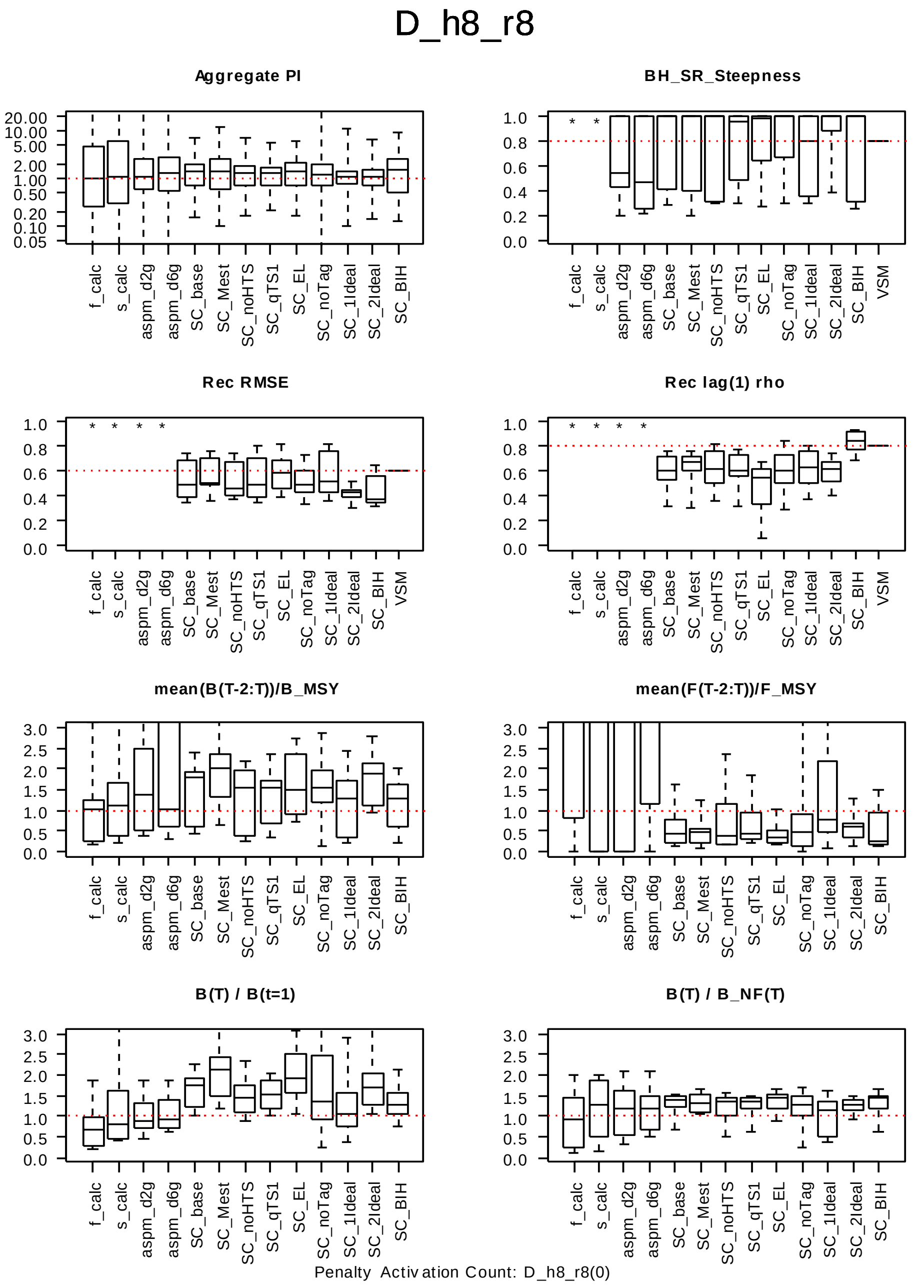

D_h8_r8
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
f_calc
s_calc
aspm_d2g
aspm_d6g
SC_base
SC_Mest
SC_noHTS
SC_qTS1
SC_EL
SC_noTag
SC_1Ideal
SC_2Ideal
SC_BIH
BH_SR_Steepness
1.0
0.8
0.6
0.4
0.2
0.0
VSM
Rec RMSE
Rec lag(1) rho
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
B(T) / B(t=1)
B(T) / B_NF(T)
Penalty Activation Count: D_h8_r8(0)

Fig. 10b.

The preceding results suggest that, on average, most of the assessment models have some capacity to distinguish between high and low steepness, and this does provide justification for attempting to estimate it. However, all of the preceding trials were conditional on the correct, and rather strong, assumption that the functional form of the stock recruitment relationship is a Beverton-Holt function. We also considered one case (E_HSSR) in which the functional form of the SR violated this assumption. In E_HSSR, the steepness was nominally the same as E_base (0.6). From this scenario, we could not conclude that the assumption violation had a particularly strong effect on assessment performance (Fig. 11):

- Steepness estimates for E_HSSR were generally not biased low, as observed for E_base, but this is not really an informative comparison, unless we are specifically interested in recruitment precisely at SSB = 0.2*SSB(unfished)

- The recruitment deviation RMSE and auto-correlation were very consistent across AMs and similar to the E_Base results (e.g. slightly low for RMSE and somewhat high for the auto-correlation). We would have expected the auto-correlation to be higher than E_base, reflecting the systematic lack of fit caused by assuming the wrong functional form for the SR relationship. Perhaps the difference in functional form was not large enough to be an issue, in this case, as the systematic recruitment errors are no worse than for models that get the steepness estimate wrong for E_base.

- The incorrect SR had very little effect on the time series estimates of relative biomass or recruitment. Qualitatively, these estimates appear to be as good as in applications to E_base.

We do not provide any illustration of the MSY-related estimates for E_HSSR (including the Aggregate PIs), because the operating model encountered numerical problems in this scenario.

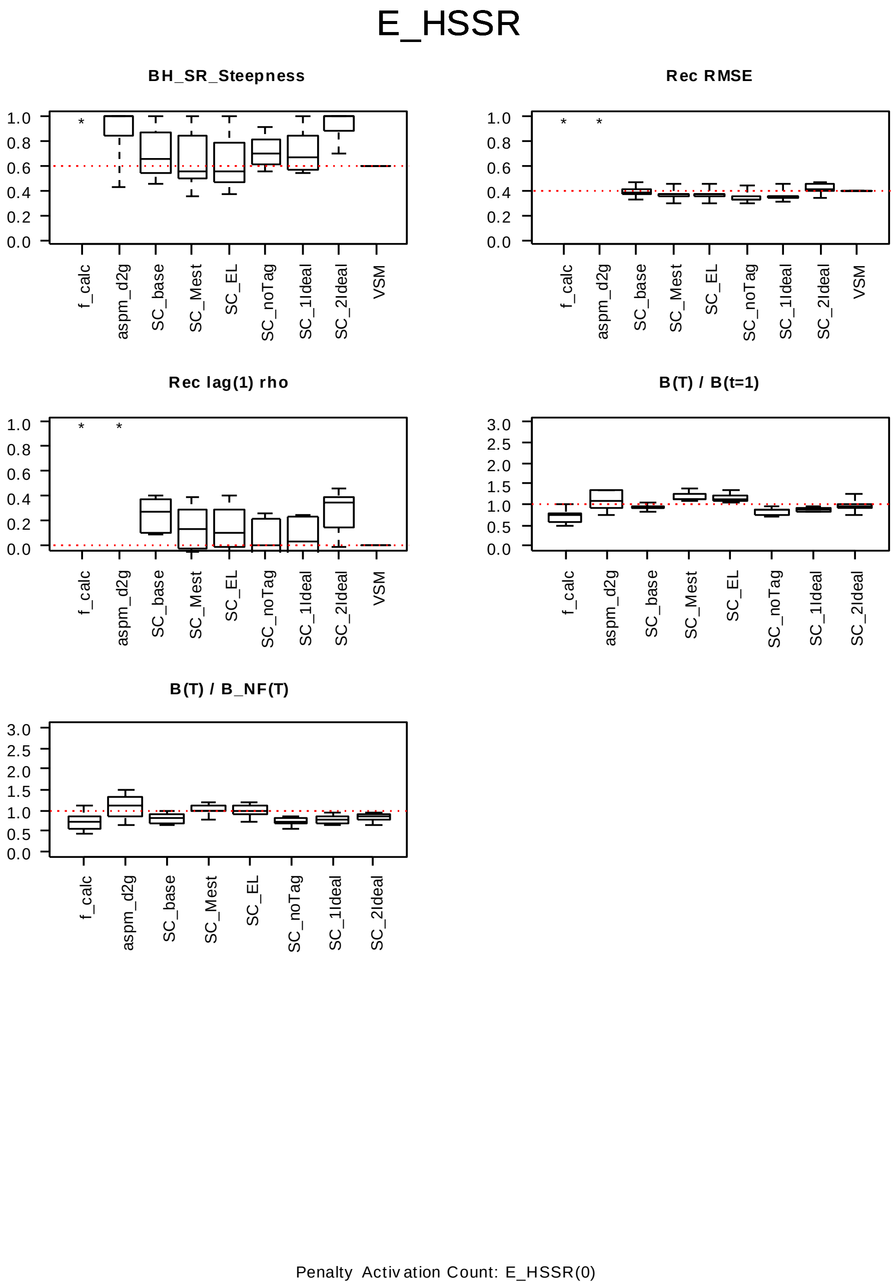


Fig. 11a. Boxplots of MPD performance indicators resulting from the application of a range of assessment models to an operating model with excellent data characteristics, but a stock recruitment relationship functional form that does not conform to the assessment assumptions (E_HSSR). Individual values represent the ratio of (AM estimated)/(OM actual) in all cases except BH_SR_Steepness, Rec RMSE and Rec lag(1) rho which show the actual values. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

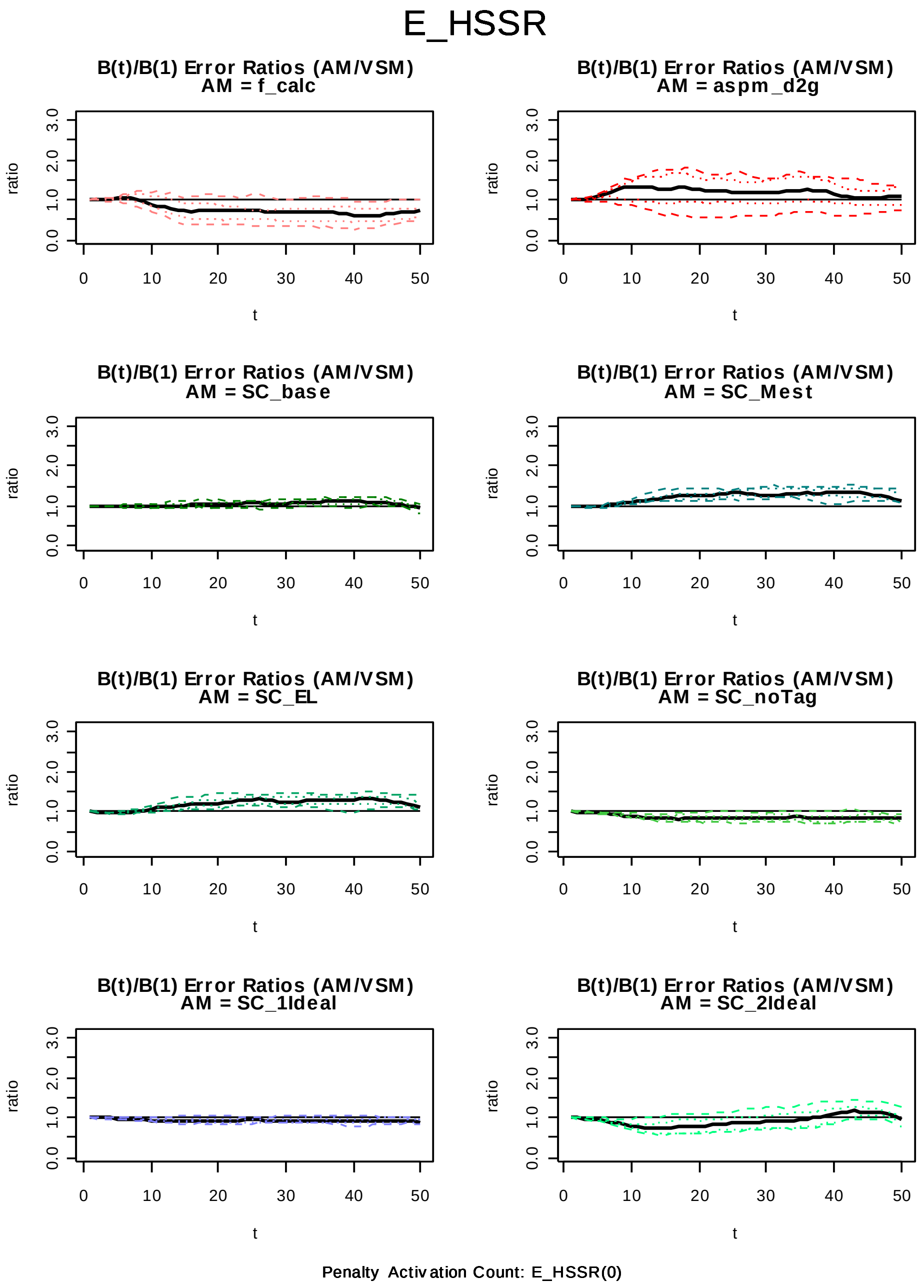


Fig. 11b. Time series of performance indicators (AM estimated)/(OM actual) resulting from the application of several AMs to 10 data realizations from the SBT operating model E_HSSR. Lines indicate rankings 1, 3, (5+6)/2, 8 and 10. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

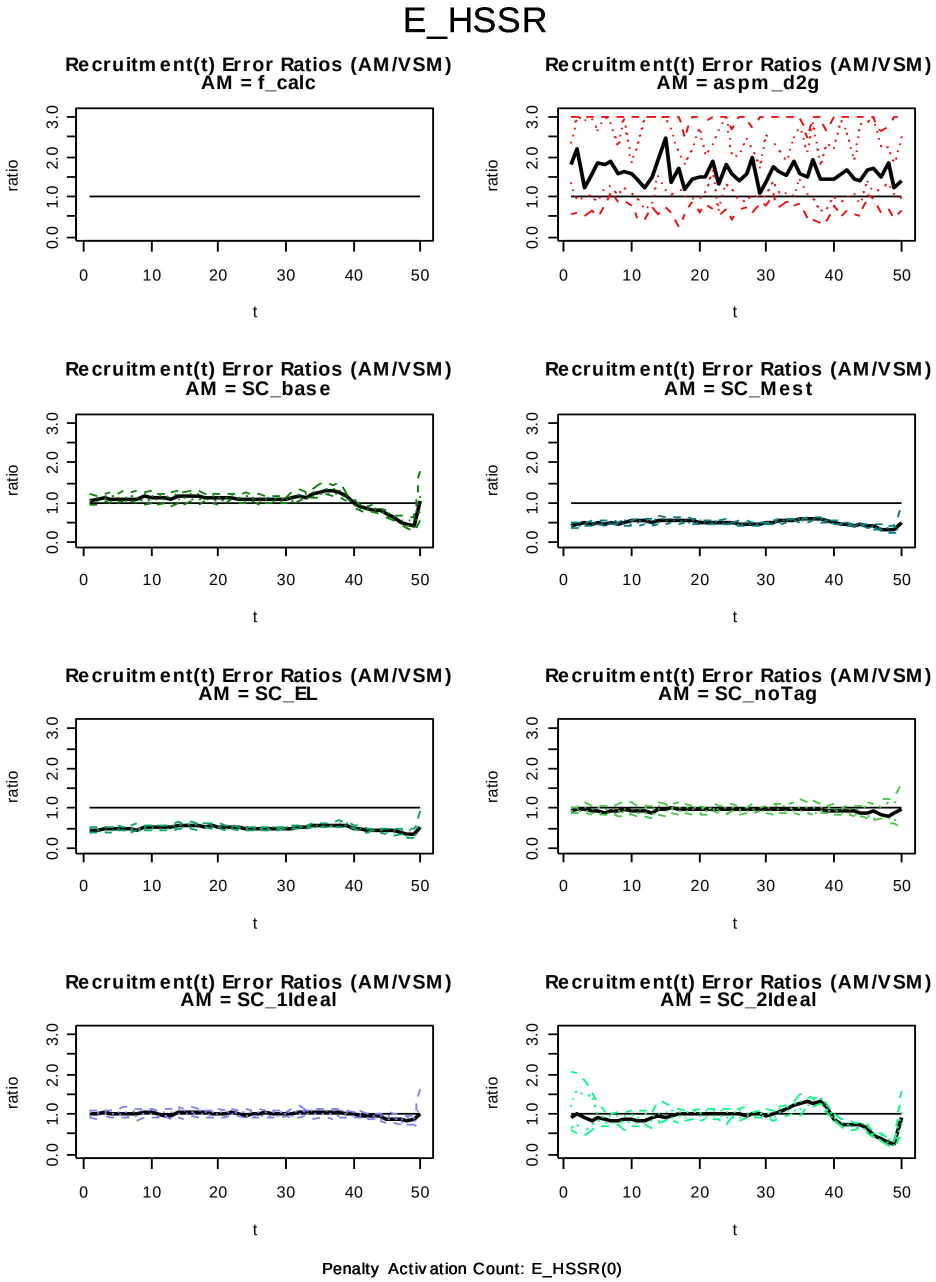

E_HSSR
Recruitment(t) Error Ratios (AM/VSM)
AM = f_calc
Recruitment(t) Error Ratios (AM/VSM)
AM = aspm_d2g
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_Mest
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_EL
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_noTag
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_HSSR(0)

Fig. 11b(cont.)

Overall, these SR-related simulations do broadly support the decisions that have been taken by the CCSBT-SC with respect to the formulation of operating models for MP development. Specifically:

- it was recognized that steepness is difficult to estimate reliably, and a range of alternatives were considered as plausible.
- a lower bound was placed on the recruitment variability to avoid under-estimating the true value
- auto-correlation in the recruitment deviations was imposed in forward projections, to minimize the probability of predicting sudden changes in recruitment that are not consistent with recent trends (this provides some short-term mitigation for using a poor stock recruitment curve).

We had expected that the imposition of an incorrect stock recruitment functional form would cause elevated recruitment auto-correlation, and perhaps it did, but it was not obvious relative to the cases with the correct curves, because the relationship is usually estimated rather erroneously anyway. Similarly, the time series of recruitment did not seem to be adversely affected by assuming the wrong SR. But this could be a co-incidental result. The worst recruitment estimates would be expected in the most recent years if the SR form was wrong (possibly also the initial age structure), because the data are poorest for discriminating these cohorts. However, as long as the spawning biomass in the most recent years corresponds to two points in which the actual and predicted recruitment values are similar (e.g. SSB ~ 0.2*SSB(unfished) in this case), we would expect good recruitment prediction. We expect that incorporating recruitment auto-correlation into the CCSBT MP operating model can largely compensate for the effects of an incorrect stock recruitment curve over the critical short-term period in which drastic management actions might be required for SBT.

It is currently unclear what we have learned about stock recruitment curve estimation from the SCTB-MWG. The 2002 study was uninformative in that all analysts were given prior knowledge that there was no relationship between recruitment and stock size in the YFT simulations. In the 2003 study, analysts were provided with information that there was some sort of a relationship active on a relatively fine spatio-temporal scale (and linked to dynamic SST fields). At this time, we do not know how this stock-recruitment relationship would have scaled up to the global population, so we are not sure if the steepness estimates even provide a meaningful basis for evaluation. Most of the assessment models applied in 2003 assumed a Beverton-Holt curve. The Multifan-CL and A-SCALA analysts used priors to constrain the steepness estimate (with a fairly high mode). There were no constraints on the SCALIA steepness. The SCALIA estimates were highly variable among operating model scenarios and assessment model specifications (Fig. 12). Presumably the operating model productivity did not actually vary much between OM scenarios, so this variability is an indication of the sensitivity to the SCALIA specifications (and interactions with the exploitation history and/or data aggregation units). Hopefully more insight from the 2003 study will become available as the SCTB-MWG analysis progresses.

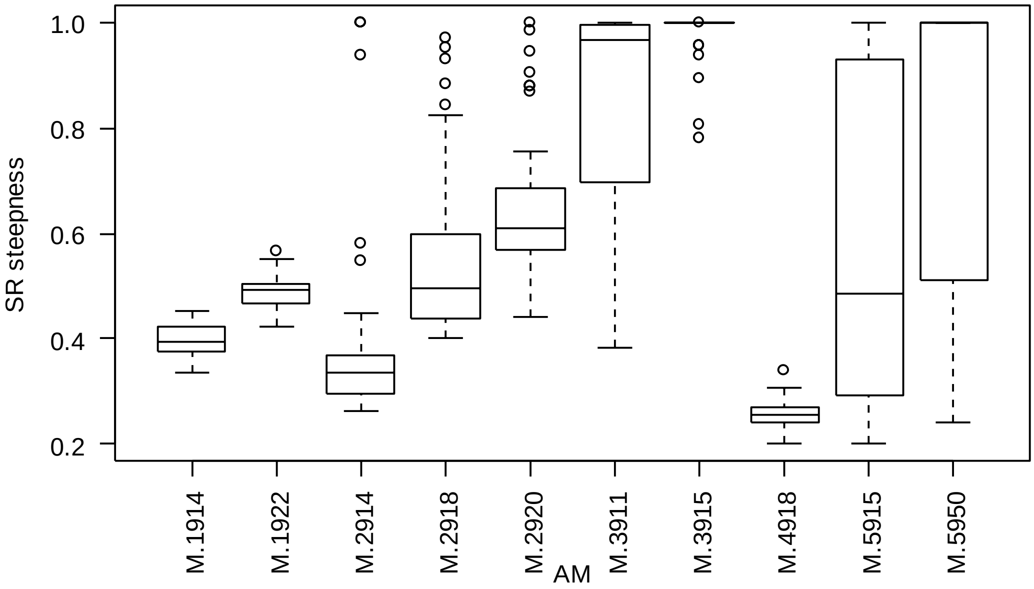


Fig. 12. **Boxplots of the steepness estimates from different SCALIA applications to the SCTB MWG YFT simulations. The first digit in the model number identifies the operating model scenario to which it was applied (e.g. M.1914 corresponds to 1F X 1R, M.2914 corresponds to 2F X 1R, etc). AMs defined in Table 6 -Table 7.**

5.4 Objective II - Assessment Implications of Catch Under-Reporting Biases

Total catch by fishery (in mass or numbers) is a key input to most stock assessment models, but it is usually very difficult to estimate catches taken outside of formal monitoring programs (e.g. IUU fishing). We tried to get some appreciation of the likely effects of these errors on stock assessment inferences by simulating under-reporting biases in the major fisheries. Fig. 13 shows representative results for the SCALIA model SC_base comparing the ideal operating model scenario E_base, with three other scenarios in which one of the fishing fleets under-reported total catch by ~20% (E_C20j = juvenile fishery, E_C20f = longline feeding grounds and E_C20s = both early and late spawning grounds). For all assessment models tested, the estimation performance seemed to be largely unaffected by the reporting biases. The reporting bias in the juvenile fishery had the smallest effect. The feeding grounds and spawning grounds under-reporting generally produced biases in the management-related quantities that were in opposite directions (or perhaps it is more correct to say that they affected the pre-existing biases in opposite directions). The time series plots indicate that the biomass, exploitation rate and recruitment time series were all very similar in all cases. Presumably, the magnitude of the estimation biases resulting from each scenario depend on the selectivity of the fishery and the magnitude of the catch over time for the biased fishery (both in absolute terms and relative to the other, unbiased fisheries). However, given that the effects were generally smaller than expected, we did not try to further explain the mechanisms by which the estimation biases are introduced.

We expected a larger effect from the under-reporting because some of our exploratory trials (not shown) seemed to show a surprising sensitivity to an (unintended) ~5% global catch over-estimation bias; but we did not pursue the over-estimation scenarios further, because they seemed less plausible than the under-reporting scenarios. The inclusion of catch under-reporting scenarios in actual SBT assessments seemed to have a more dramatic effect than we observed here (Polacheck and Preece 2001). We also note that the WCPO bigeye tuna assessment (Hampton et al. 2003) seemed to have peculiar recruitment trend estimates roughly co-inciding with increases in the large, but poorly quantified, fisheries in Indonesia and the Phillipines. This was identified as a potentially misleading issue for the assessment (at SCTB-16). Presumably the simulation scenarios would have been more challenging (and realistic in most cases) if there was a temporal trend in the reporting bias, but this was not explicitly explored.

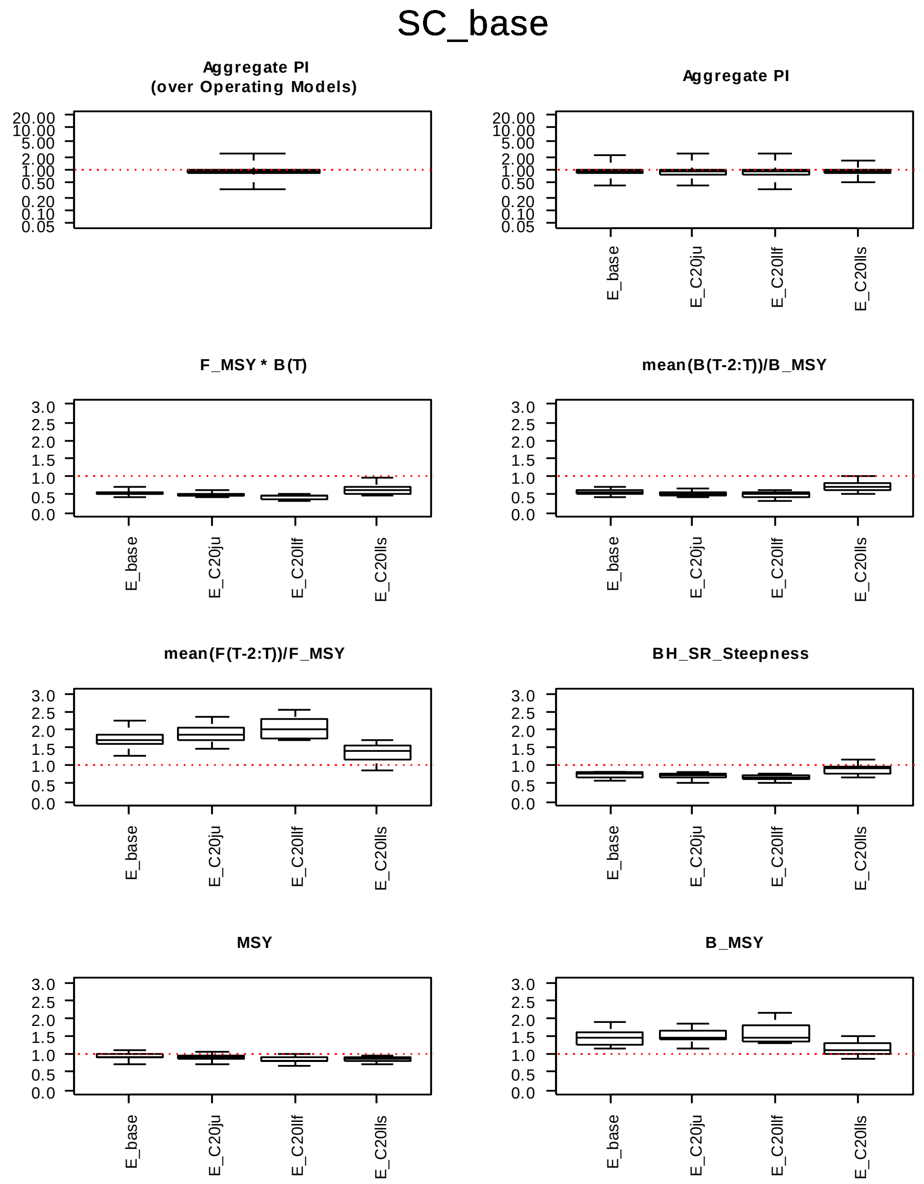


Fig. 13a. MPD performance indicators resulting from the application of SCALIA assessment model SC_base to the baseline SBT operating model (E_base) and three scenarios with 20% catch under-reporting biases in one of the fisheries (E_C20j = juvenile fishery, E_C20f = longline feeding grounds and E_C20s = both early and late spawning grounds). Each assessment model was applied to 10 data realizations from each OM. Individual values represent the ratio of (AMestimated)/(OM actual) in all cases except BH_SR_Steepness which shows the actual values. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

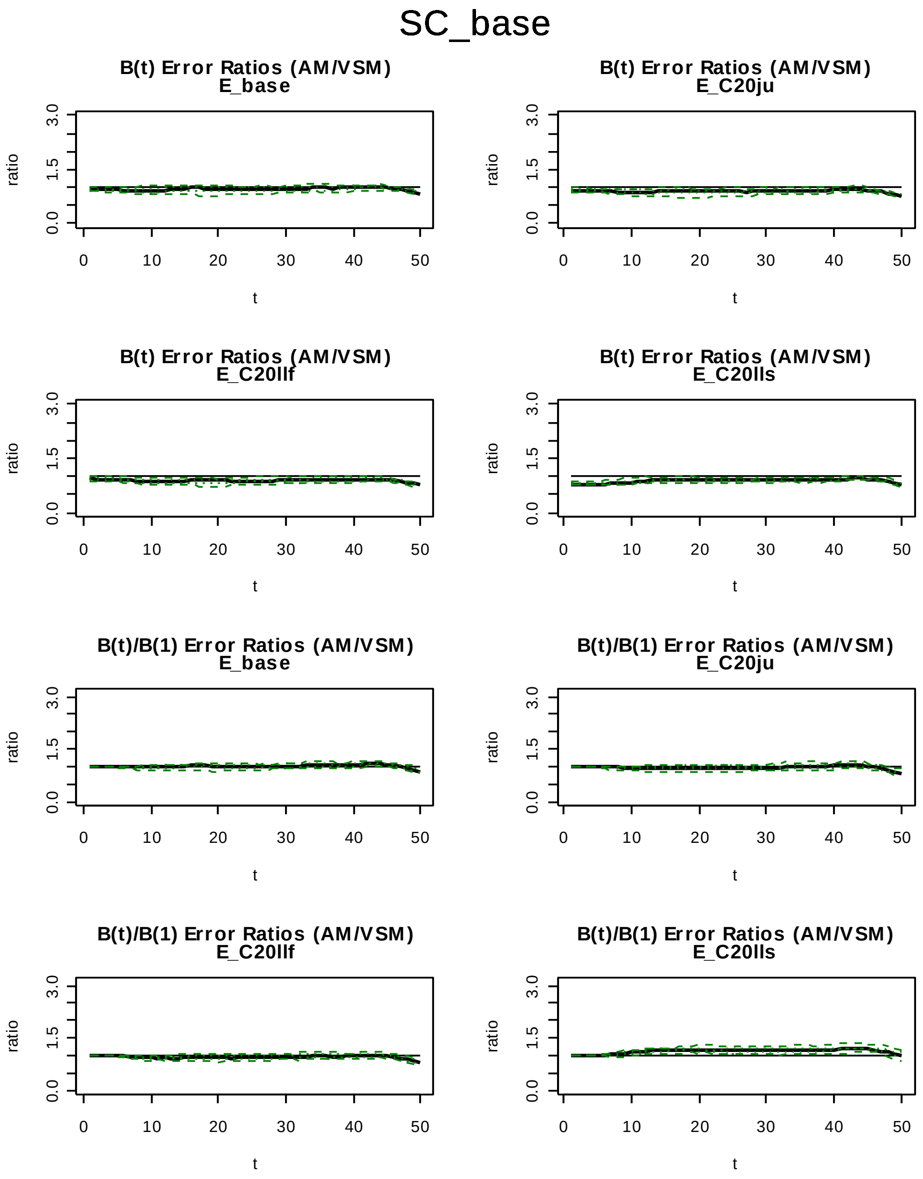


Fig. 13b. **Time series of performance indicators (AM estimated)/(OM actual) resulting from the application of AM SC_base to 10 data realizations from the baseline SBT operating model (E_base) and three scenarios with 20% catch under-reporting biases in one of the fisheries (E_C20j = juvenile fishery, E_C20f = longline feeding grounds and E_C20s = both early and late spawning grounds). Lines indicate rankings 1, 3, (5+6)/2, 8 and 10. OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

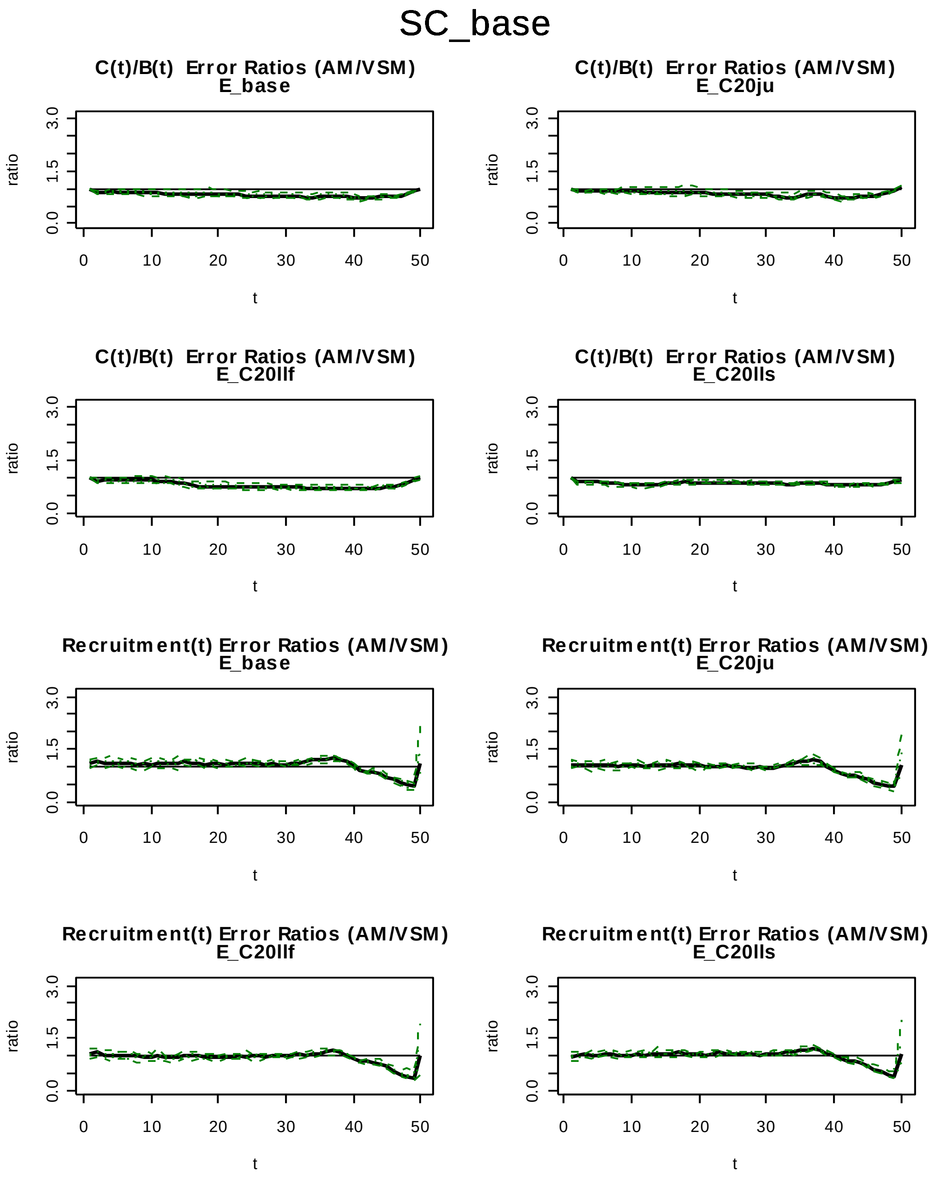

SC_base
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_C20ju
C(t)/B(t) Error Ratios (AM/VSM)
E_C20llf
C(t)/B(t) Error Ratios (AM/VSM)
E_C20lls
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_C20ju
Recruitment(t) Error Ratios (AM/VSM)
E_C20llf
Recruitment(t) Error Ratios (AM/VSM)
E_C20lls
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 13b (cont.)

5.5 Objective III Age Estimation from cohort-slicing vs: Catch-at-Length

Age-structured stock assessment models are ideally suited to use age composition data and it is only relatively recently that methods for directly using length composition data have become popular. There is a long history in SBT assessment of estimating the age composition of the catch by cohort-slicing the catch length frequency distributions. This is known to be an unreliable method for ageing older fish, but it is not clear what the implications are in the context of an integrated assessment.

Our simulation results suggest that the use of cohort-sliced age data in an assessment does result in some characteristic estimation errors, but the overall assessment performance might not be any worse than models that use catch-at-length, depending on the inferences that one is interested in. The performance of two different cohort-sliced, catch-at-age models (SC_BIH and BIH_2) are presented alongside two similar catch-at-age/length models SC_base and SC_1ideal) in Fig. 14 for the baseline operating model (E_base). Aside from the data used, the main differences between SC_BIH/BIH_2 and SC_base/SC_1ideal are the catch-at-age and catch-at-length effective sample size assumptions (and constant selectivity in the case of SC_1ideal). From Fig. 14 we observe:

- The aggregate performance indicator suggests that the catch-at-length models and BIH_2 perform similarly, while SC_BIH has clearly worse performance. We did not explicitly examine why this is the case, but we recognize that SCALIA was never thoroughly tested to use the cohort-sliced data, and hence would expect BIH_2 to provide a better representation of what can be achieved with cohort-slicing.

- SC_BIH and BIH_2 tend to have larger biomass biases than the CL models, and there are some similarities in the temporal pattern of the biases. However, BIH_2 exhibits excellent biomass estimation towards the end of the time series. Presumably cohort-slicing tends to produce particular estimation errors for the initial population, while cohorts that are observed repeatedly in a series of fisheries over time are probably estimated much better.

- SC_BIH and BIH_2 both have stronger temporal trends in exploitation rate biases than the CL models. It is curious that BIH_2 actually has the worst terminal exploitation rate bias (part of an alarming downward trend), given that it had excellent terminal biomass estimates. This probably indicates an error (definition inconsistency) in either the biomass or exploitation rate calculations.

- Relative to the CL models, SC_BIH and BIH_2 had recruitment estimates that were much more variable, and highly auto-correlated (0.6-0.7 for SC_BIH; presumably similar for BIH_2). These are the differences that one would expect because cohort-slicing consistently mis-allocates a certain proportion of a particular age class into adjacent age-classes. The mis-allocation causes an estimation error for any given cohort, and inflates the auto-correlation because recruitment anomalies are spread among consecutive cohorts (e.g. a single very large recruitment event of age *a* is interpreted as a large recruitment event at age *a* and above average for ages *a* - 1 and *a* + 1). SC_base and SC_BIH also had disturbing trends in the recruitment bias at the end of the time series as in section 5.5. These were not evident for SC_1ideal (not shown) or BIH_2.

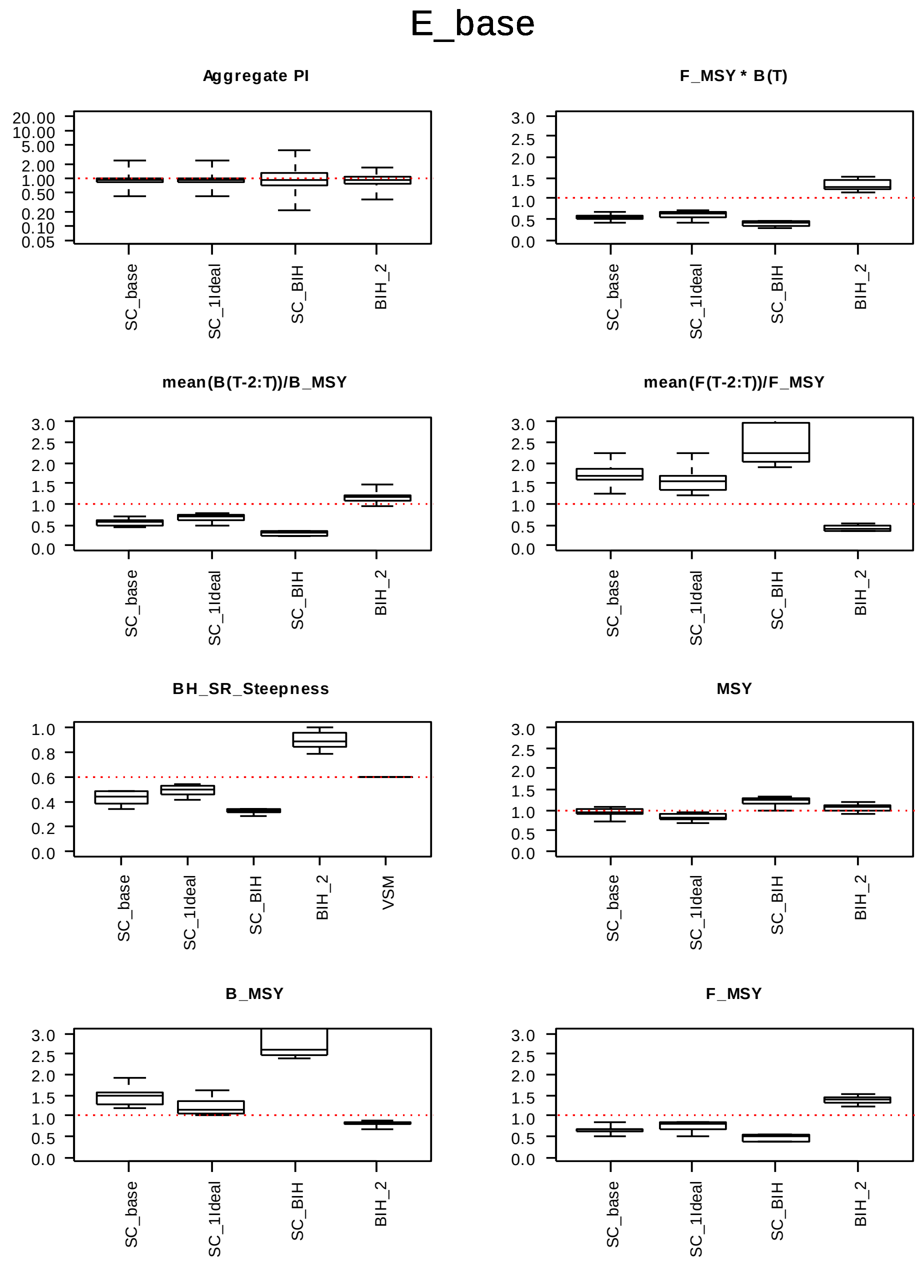


Fig. 14. Comparison of assessment models that use cohort-slicing (SC_BIH, BIH_2) with models that use catch-at-length (SC_base, SC_1ideal), when applied to 10 simulated data realizations for a fishery resembling SBT (E_base). Each PI is a ratio of (AM estimated)/(OM actual), except stock-recruitment related quantities which show actual values. Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

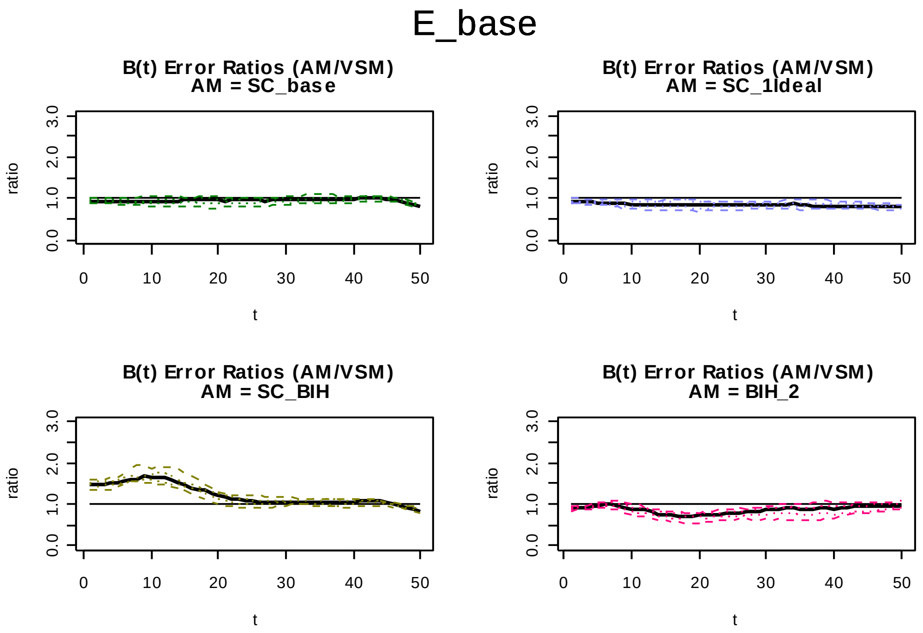

E_base
B(t) Error Ratios (AM/VSM)
AM = SC_base
B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t) Error Ratios (AM/VSM)
AM = SC_BIH
B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50

Fig. 14. (cont.)

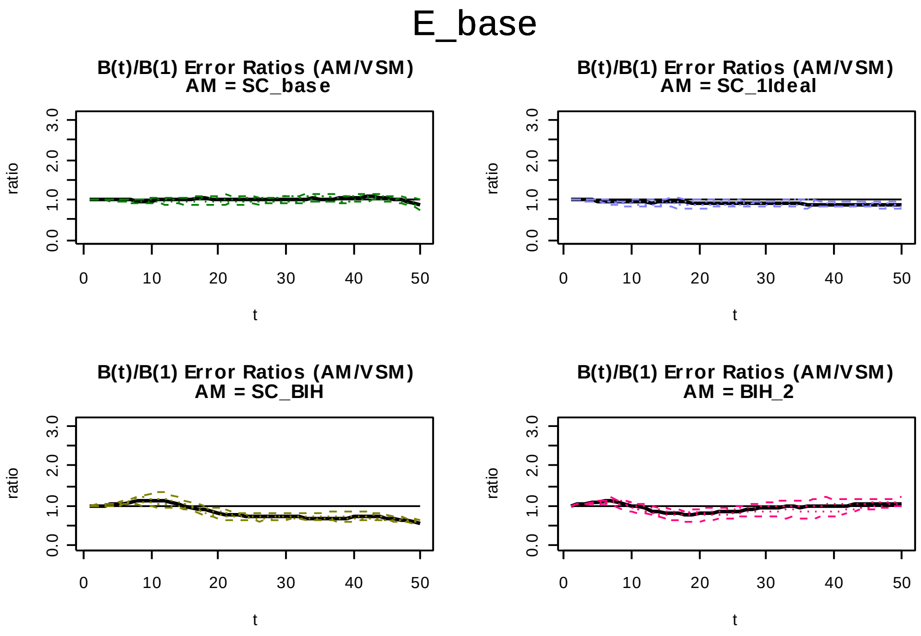

E_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_BIH
B(t)/B(1) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50

Fig. 14. (cont.)

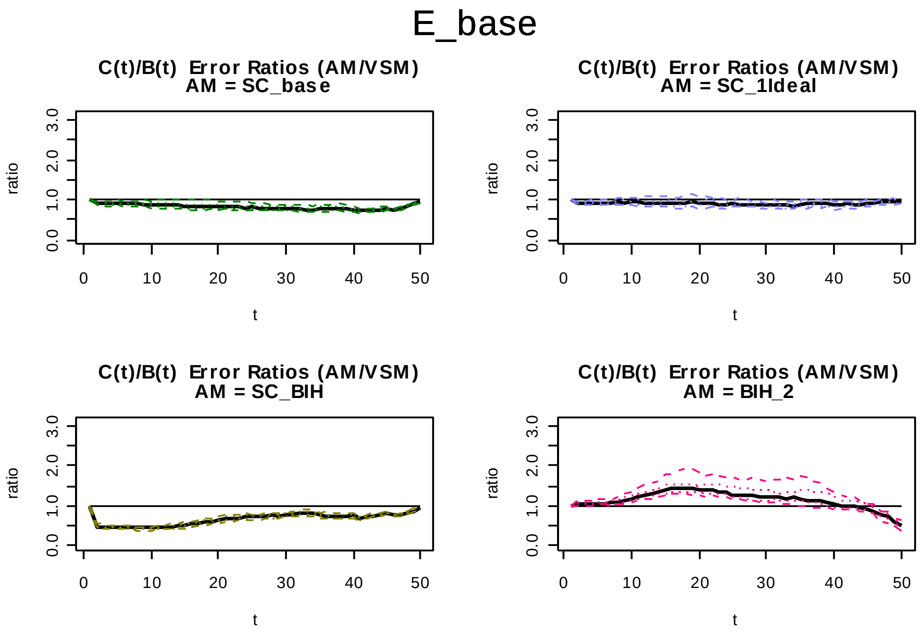

E_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_BIH
C(t)/B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50

Fig. 14. (cont.)

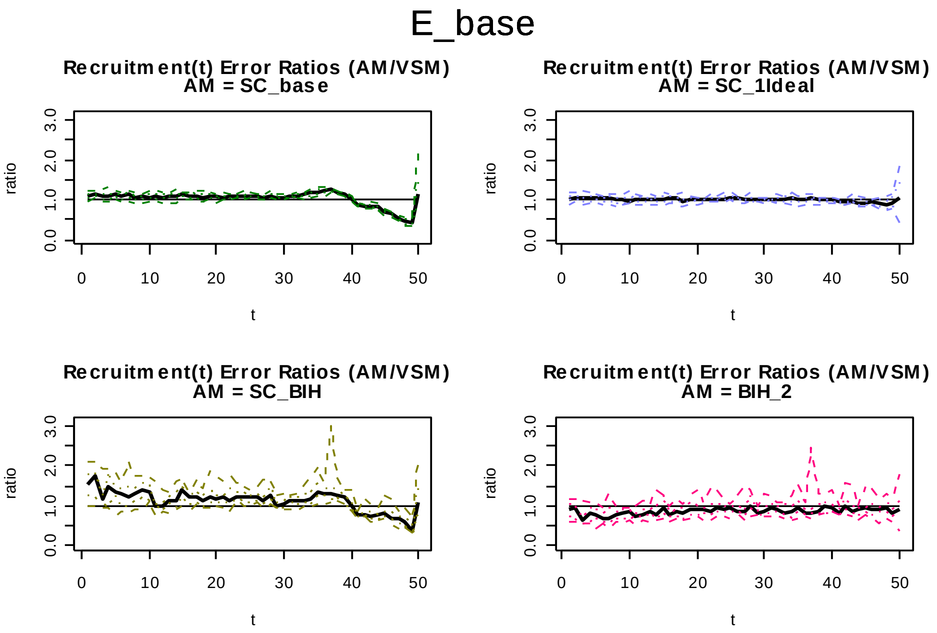

E_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_BIH
Recruitment(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
0
10
20
30
40
50
t

Fig. 14. (cont.)

These comparisons were repeated with the challenging D_base scenario (Fig. 15). All AMs performed substantially worse in this case, and the relative performance among models differs from the E_base case. We note the following:

- BIH_2 had the worst performance with respect to the aggregate performance indicator; the management-related indicators were highly variable among models and difficult to generalize.

- The cohort-sliced models seemed to have worse absolute biomass estimates than the CL models, but they are all highly variable. It is not clear which models performed better in terms of relative biomass estimates (SC_base is arguably slightly better).

- SC_1ideal seemed to have the best exploitation rate estimates.

- All models had substantial variability in recruitment estimates; SC_1ideal is slightly better than the others, but it is not obvious whether SC_base is any better than the cohort-sliced CA models.

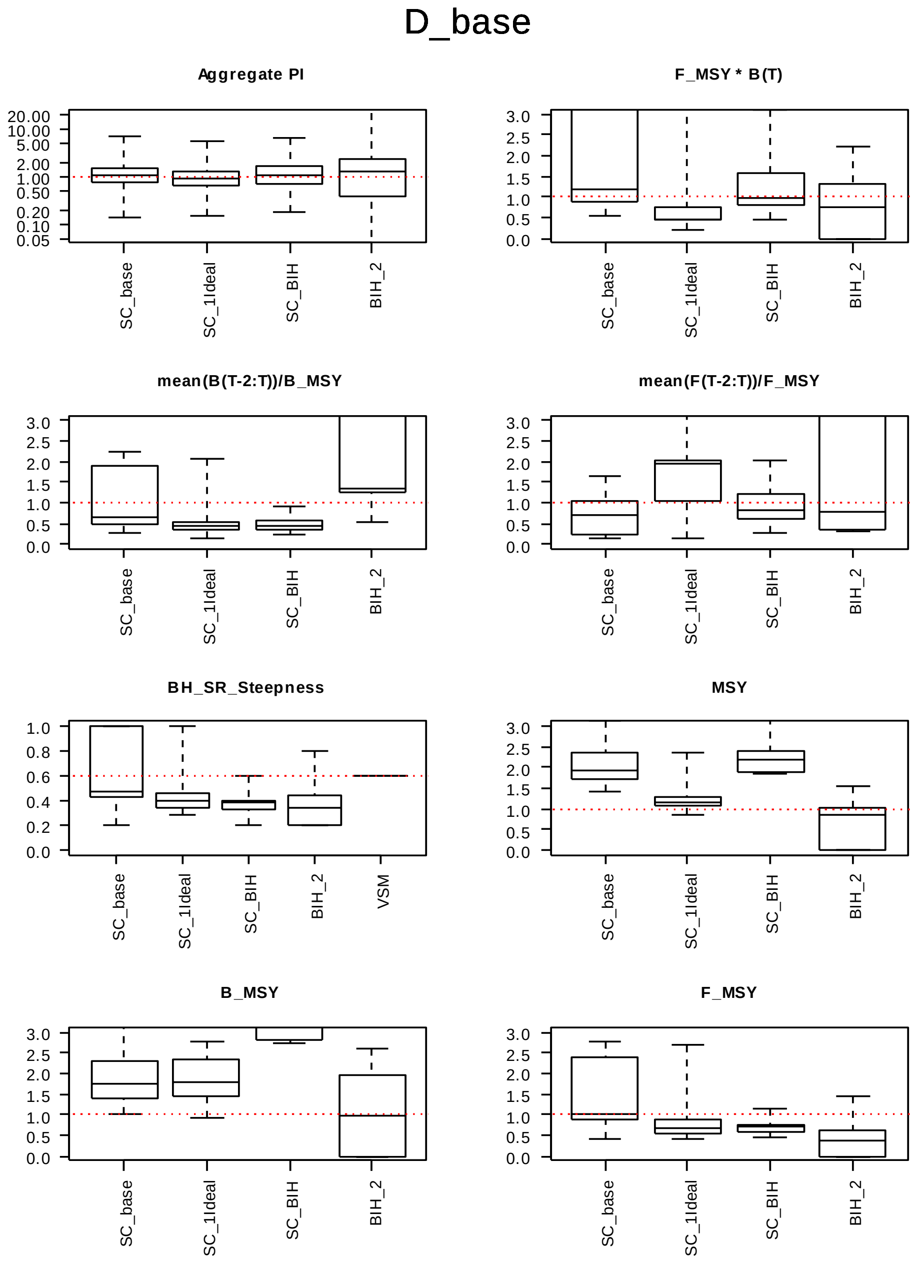


Fig. 15. **Comparison of assessment models that use cohort-slicing (SC_BIH, BIH_2) with models that use catch-at-length (SC_base, SC_1ideal), when applied to 10 "difficult" simulated data realizations for a fishery resembling SBT (D_base). Each PI is a ratio of (AM estimated)/(OM actual), except stock-recruitment related quantities which show actual values. Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

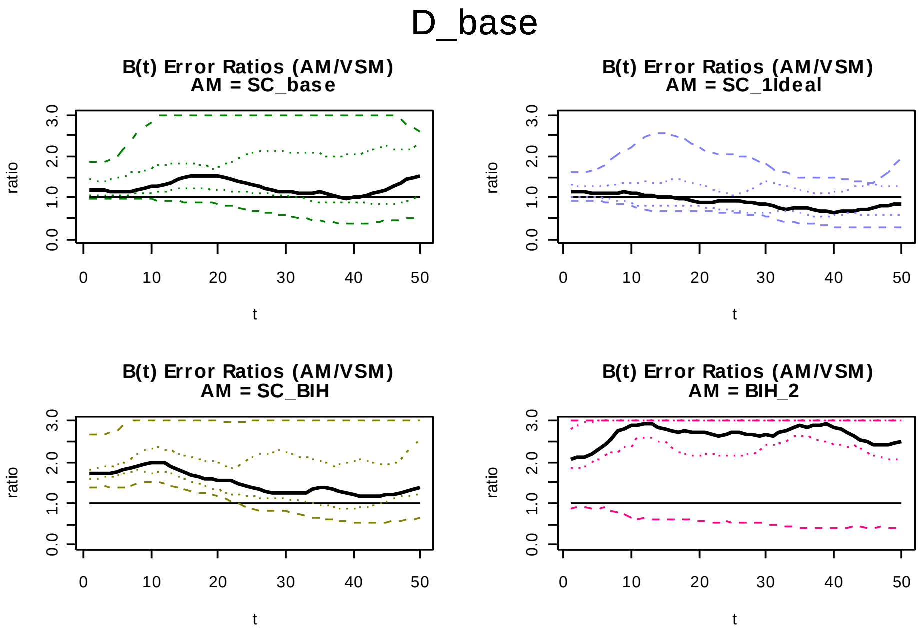

D_base
B(t) Error Ratios (AM/VSM)
AM = SC_base
B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t) Error Ratios (AM/VSM)
AM = SC_BIH
B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50

Fig. 15. (cont.)

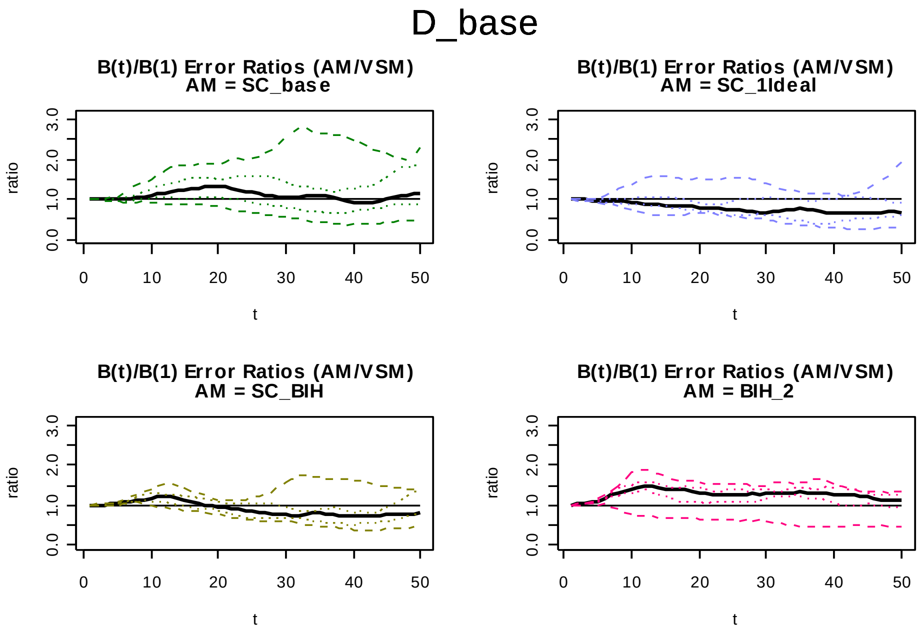
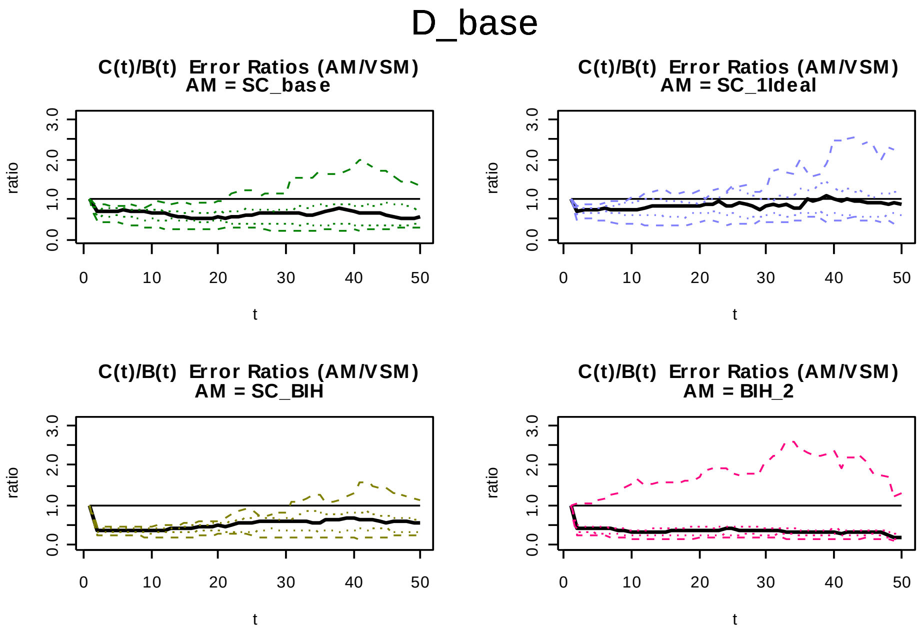

D_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_BIH
B(t)/B(1) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50

Fig. 15. (cont.)



D_base

C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_base

C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal

C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_BIH

C(t)/B(t) Error Ratios (AM/VSM)
AM = BIH_2

Fig. 15. (cont.)

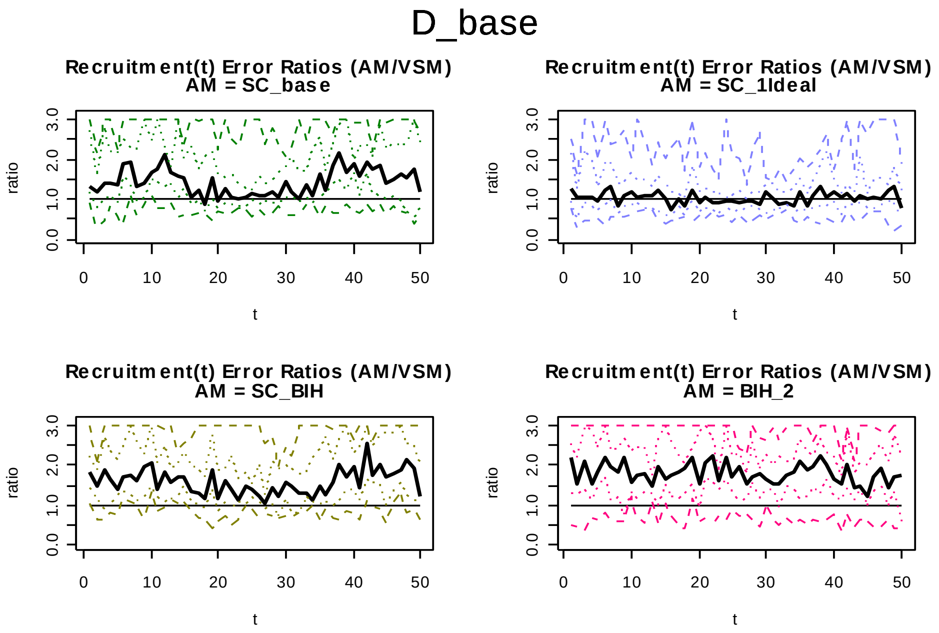

D_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_BIH
Recruitment(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
0
10
20
30
40
50
t

Fig. 15. (cont.)

5.6 OBJECTIVE IV - ASSESSMENT IMPLICATIONS OF UNRECOGNIZED CHANGES IN SBT LENGTH-AT-AGE

Regardless of whether one uses cohort-slicing or catch-at-length prediction, there remains a potential problem in the interpretation of the SBT catch length frequency distribution on the spawning grounds. Very large SBT are observed in the modern Indonesian SBT fishery, but not the historical Japanese fishery in the 1950s-1960s. This can be interpreted in a number of ways, but there are not sufficient data to directly distinguish which theory is correct. To date, nobody has examined how assessment model inferences would change if the length-at-age distribution was assumed to have changed over time (e.g. potentially due to density dependent intra-specific competition). We simulate this effect here with the operating model E_DDLinf, and compare it with the baseline model (E_base) that uses constant length-at-age.

The simulations suggest that if the SBT length-at-age decreased around the 1950s-1960s, this might have substantial implications for stock assessment inferences, depending on which models are applied. Representative results are illustrated in Fig. 16**,** from which we note the following points:

- f_calc performance was largely the same irrespective of the change in growth. This is not surprising given that the catch-at-length and length-at-age information is not directly used. s_calc (not shown) was similarly unaffected.

- ASPM_d6g was adversely affected by the growth change, while ASPM_d2g was much less affected (not shown). This difference is presumably related to the fact that aspm_d6g used an analytical selectivity calculation based on the CL data, while ASPM_d2g uses the correct selectivity and is only indirectly affected by the CL data via total catch and age-length-mass relationships.

- SC_base typifies the problems demonstrated by several SCALIA models (including SC_qTS and SC_2ideal which are not shown). The management performance indicators diverged greatly between E_base and E_DDLinf (except MSY was estimated well in both cases presumably because opposing biases in B_MSY and F_MSY cancelled out). The absolute and relative biomass estimates have a serious time series trend in bias, with under-estimation in the intermediate years, increasing to serious over-estimation in the last 10 years. Exploitation rate biases are opposite to the biomass estimates. Recruitment has a mild under-estimation bias in the early years, and substantial over-estimation for many of the last 10 years (but in this case recruitment is not particualrly well estimated in E_base either). Presumably these models make a bad estimate of the initial age structure, and the bad biases in the most recent years are caused by resultant errors in the estimated stock recruitment curve.

- Other SCALIA models had rather different performance patterns. SC_noHTS (selectivity constant over time) had large biases in biomass, exploitation rate and recruitment over most of the time series. SC_noTag and SC_1ideal were qualitatively similar (not shown).

- The assessment models that estimated natural mortality were not as adversely affected by the shift in growth as the other SCALIA models. SC_Mest performance on management-related estimates is mixed between E_base and DD_Linf; biomass and exploitation rate estimates are similar or better, and recruitment estimates are strongly biased in both cases (but in opposite directions). SC_EL performance was similar (not shown). Neither SC_Mest or SC_EL attempts to estimate a change in length-at-age, so we would assume that the difference in performance between E_base and DD_Linf reflects a change in the trade-off of the estimation biases and results in an improvement in some cases for largely spurious reasons.

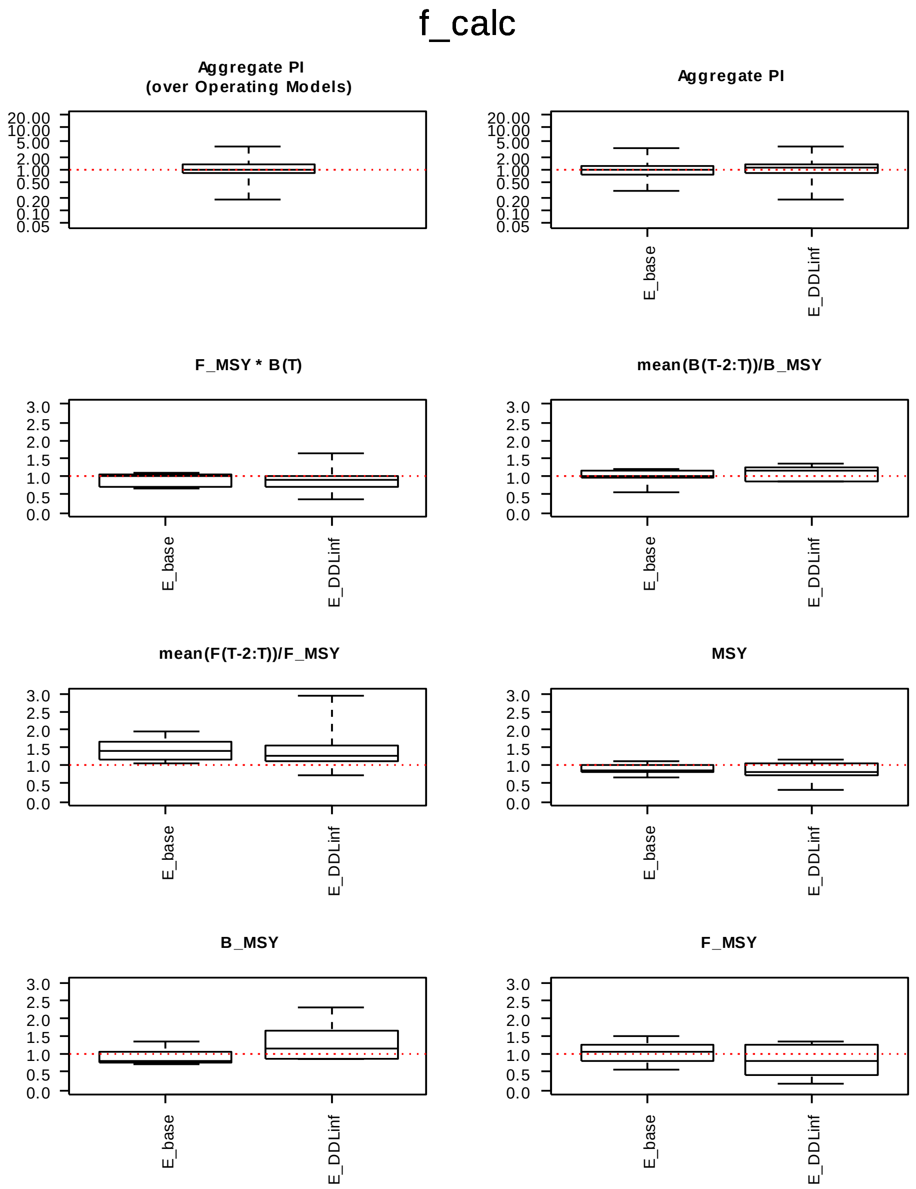


Fig. 16. Stock assessment modelling implications of an unrecognized shift in the length-at-age distribution for a simulated fishery system resembling SBT. Each AM was applied to 10 simulated data realizations from OM E_base (growth curve constant) and OM E_DDLinf (growth curve changes). Each PI is a ratio of (AM estimated)/(OM actual). Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

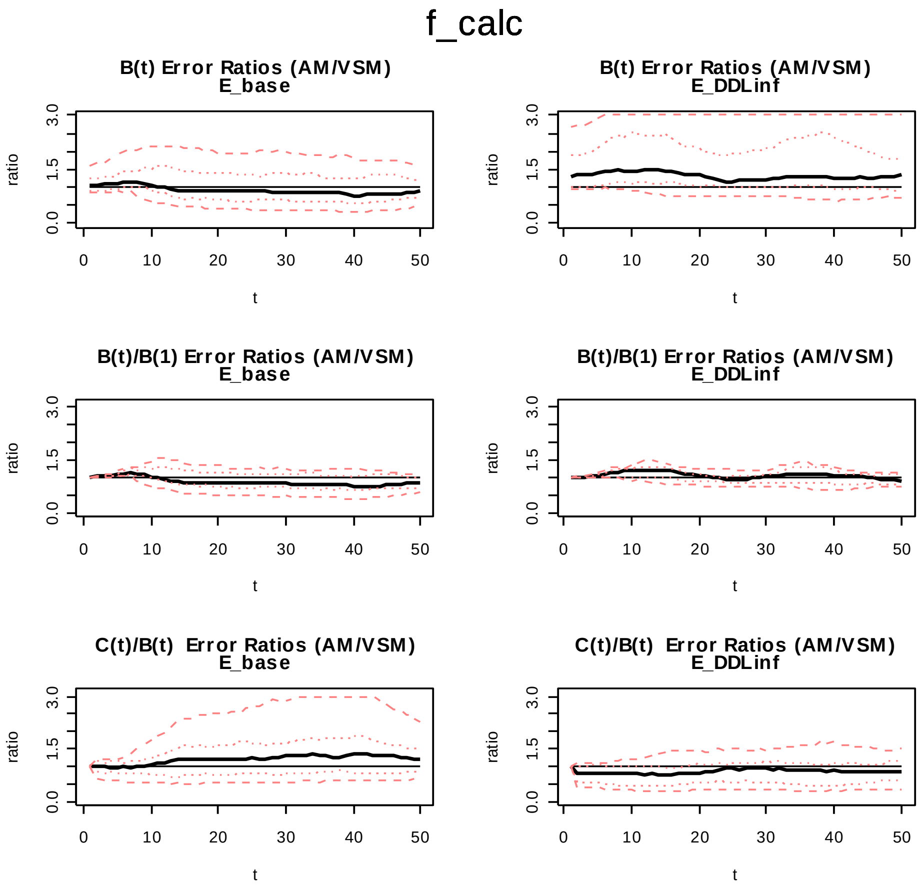

f_calc
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 16. (cont.)

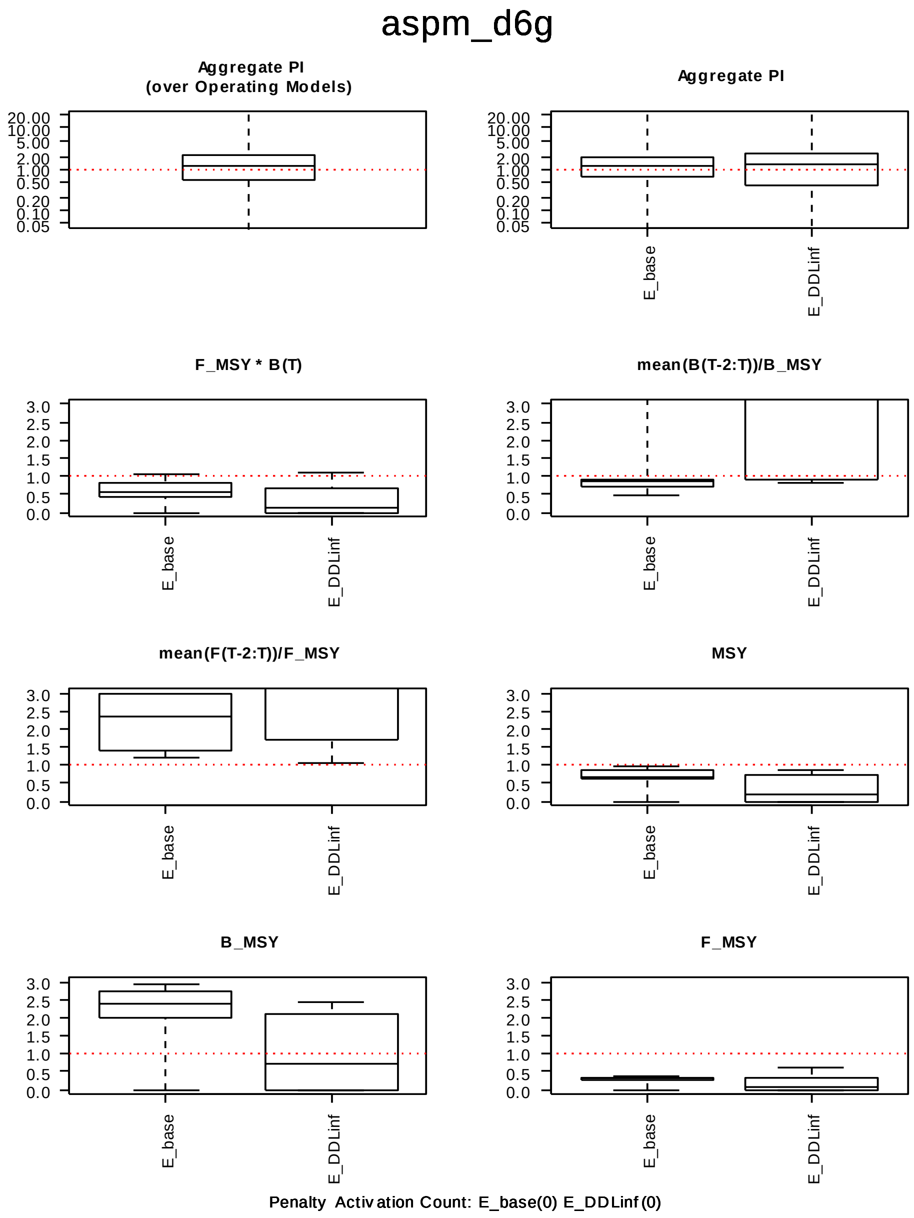

aspm_d6g
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_base
E_DDLinf
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Penalty Activation Count: E_base(0) E_DDLinf(0)

Fig. 16. (cont.)

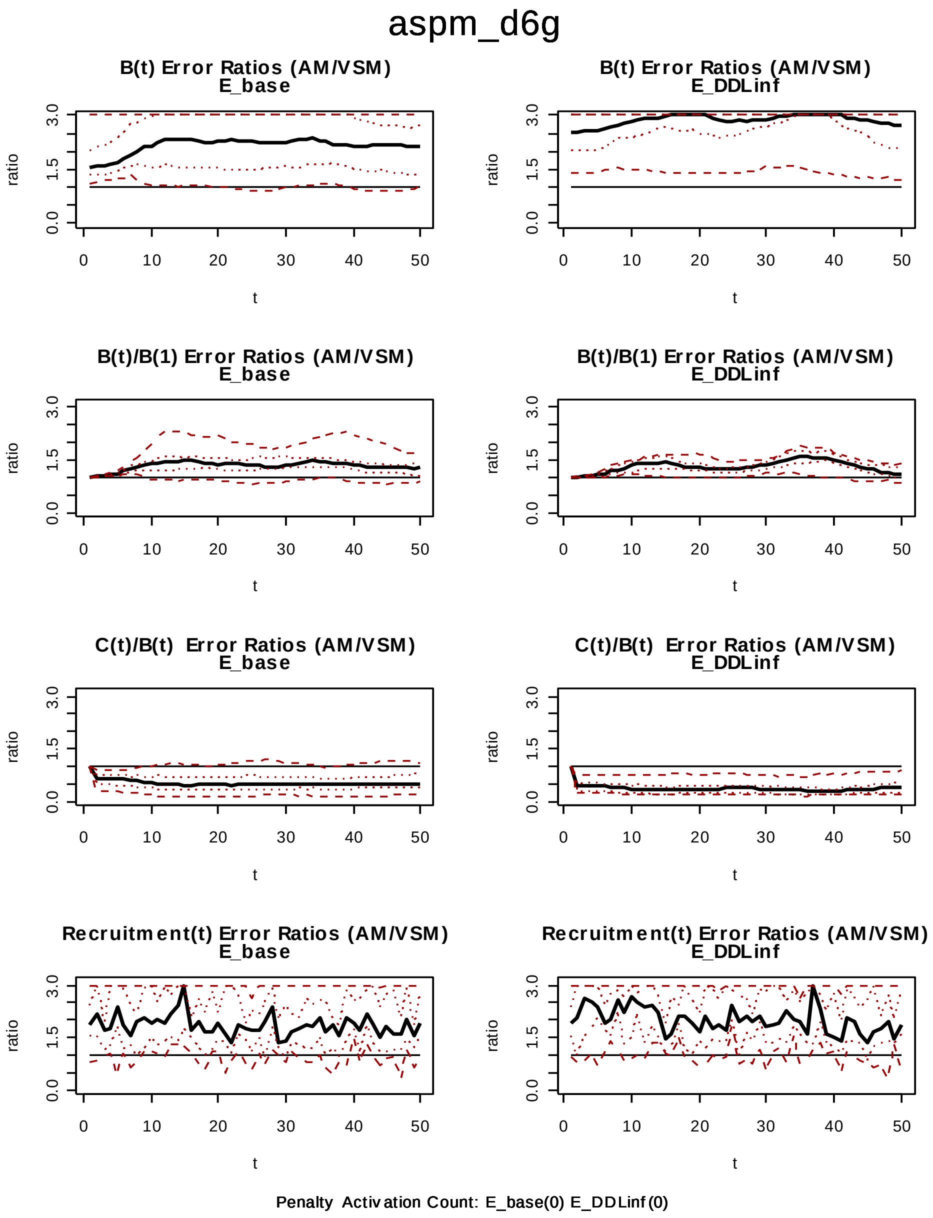

aspm_d6g
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0) E_DDLinf(0)

Fig. 16. (cont.)

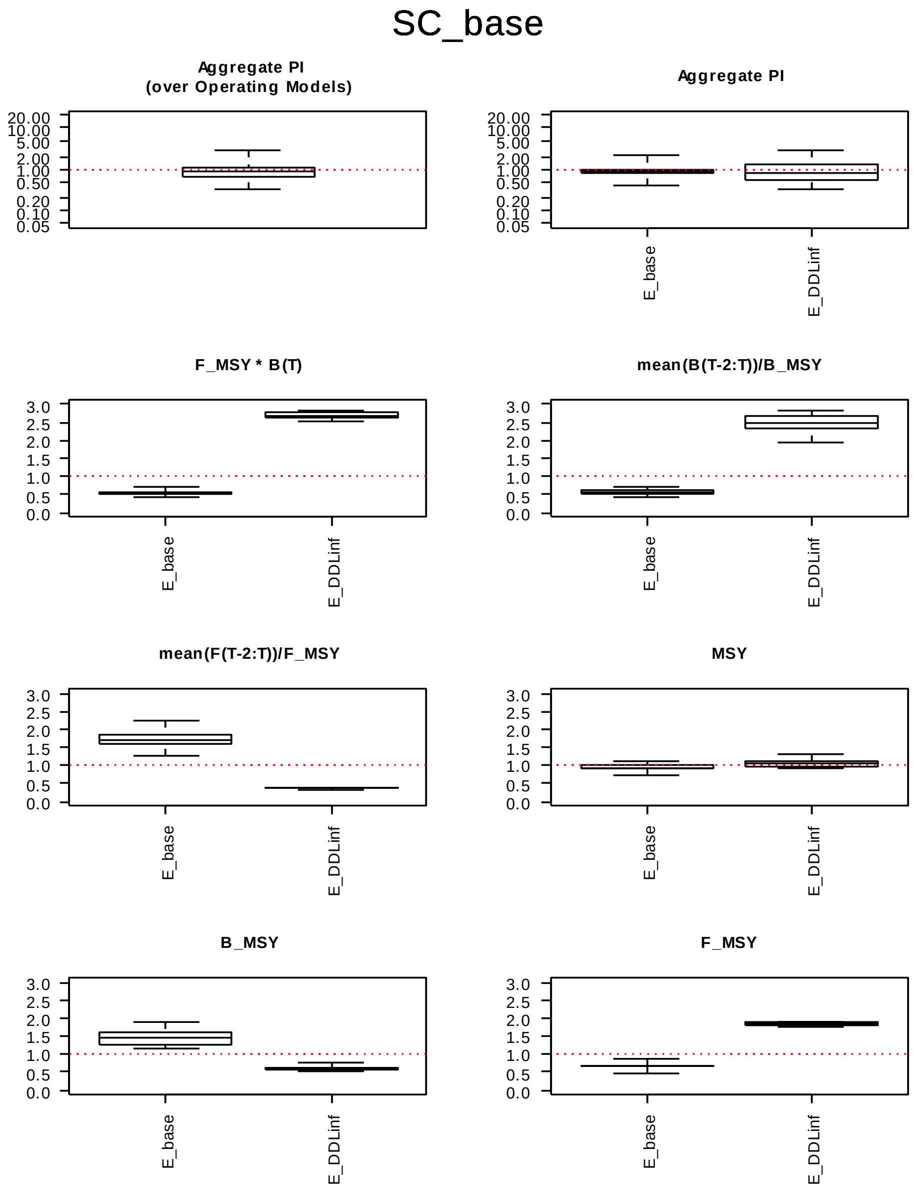

SC_base
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_base
E_DDLinf
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Fig. 16. (cont.)

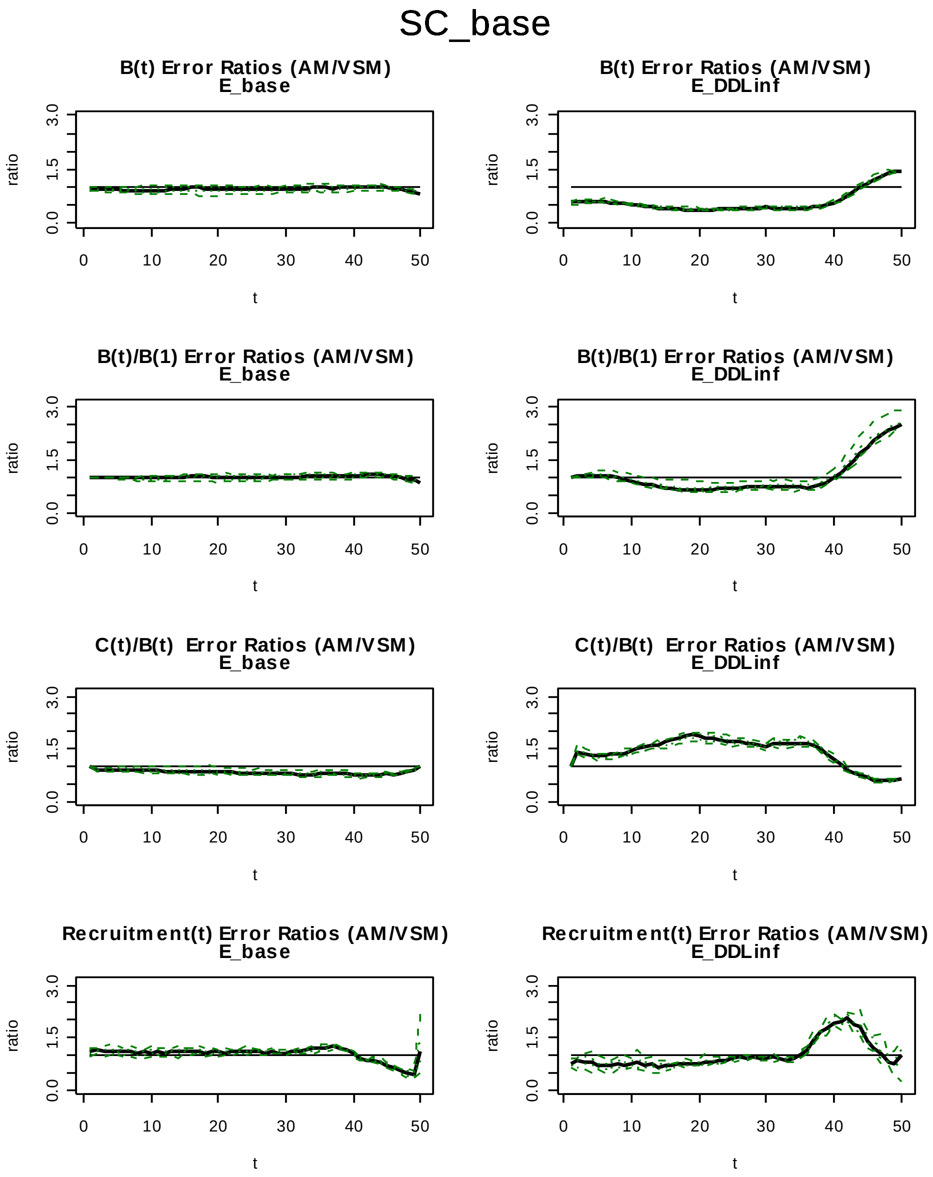

SC_base
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 16. (cont.)

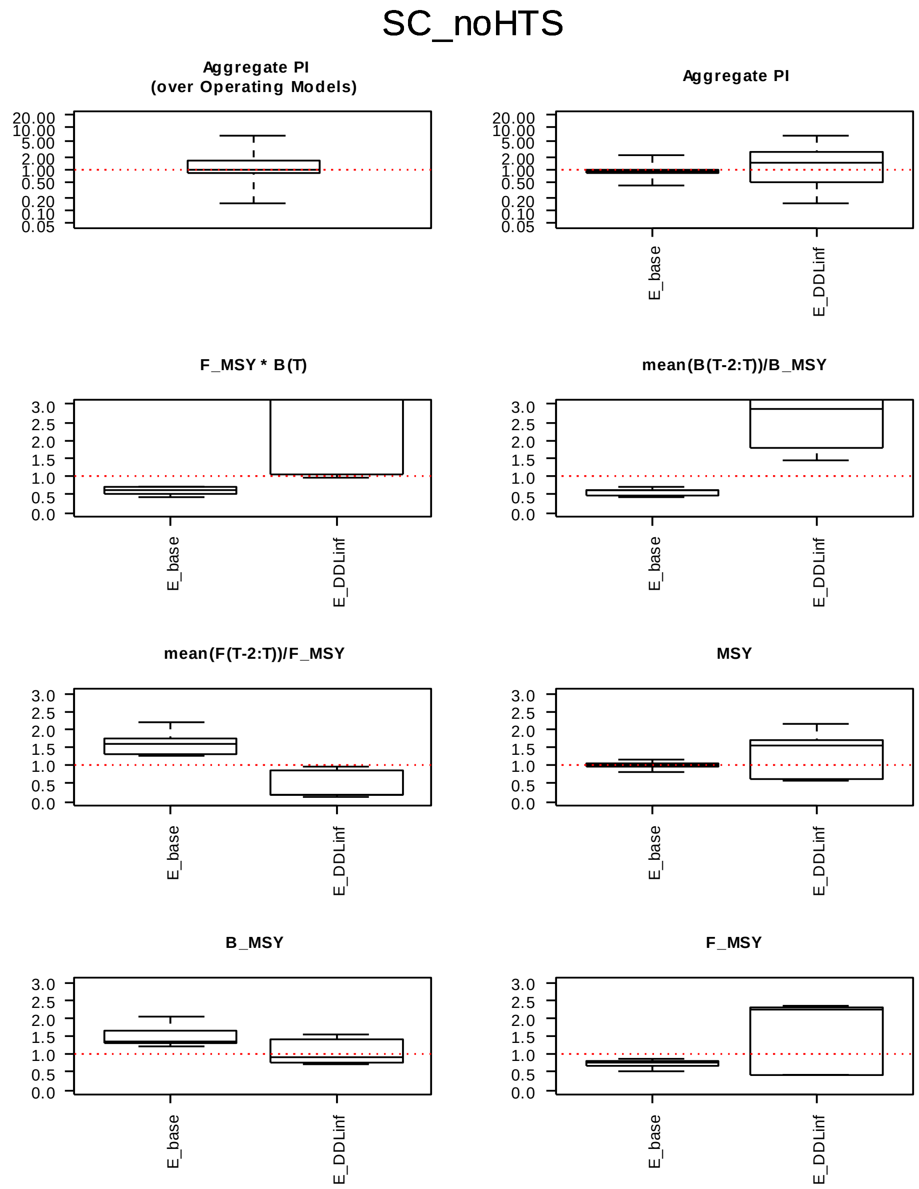

SC_noHTS
Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
Aggregate PI
E_base
E_DDLinf
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Fig. 16. (cont.)

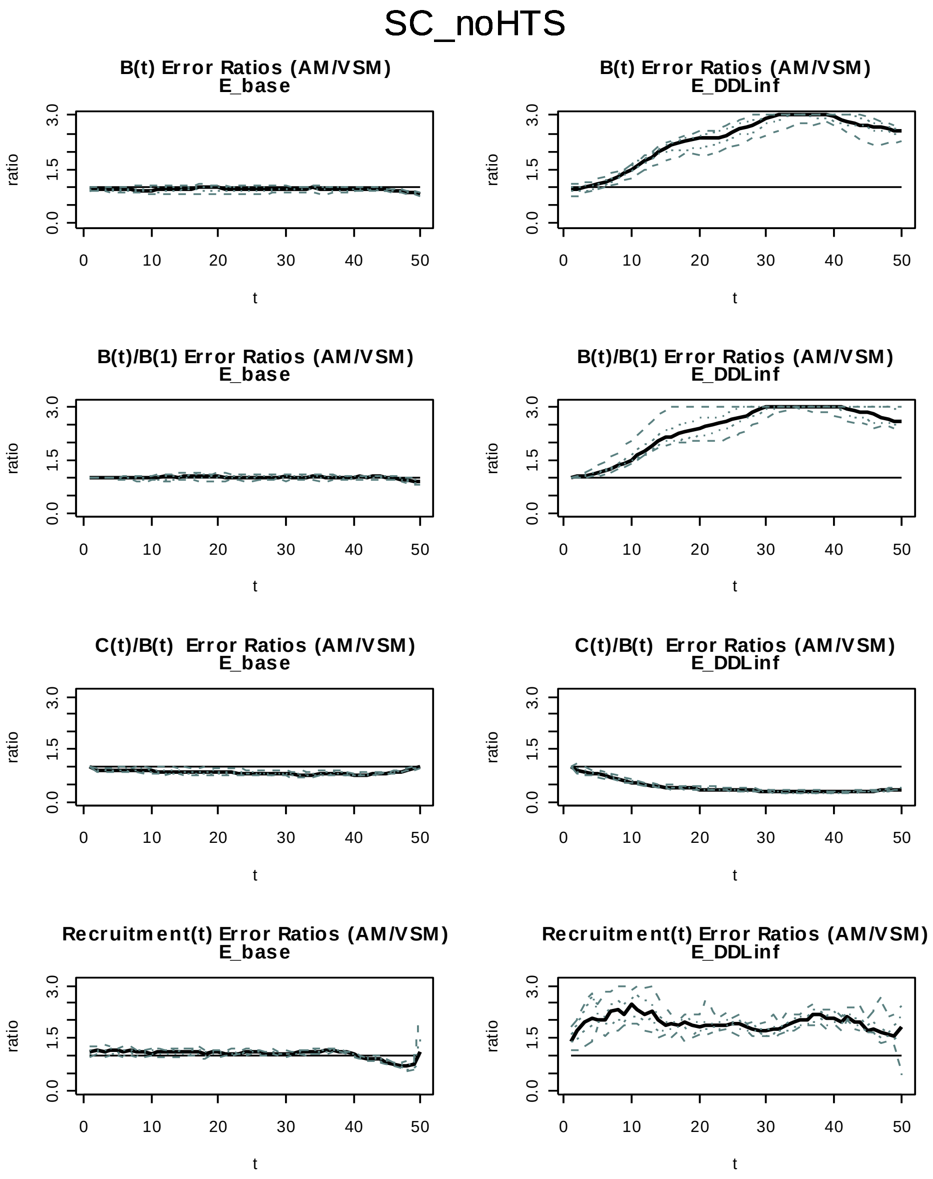

SC_noHTS
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 16. (cont.)

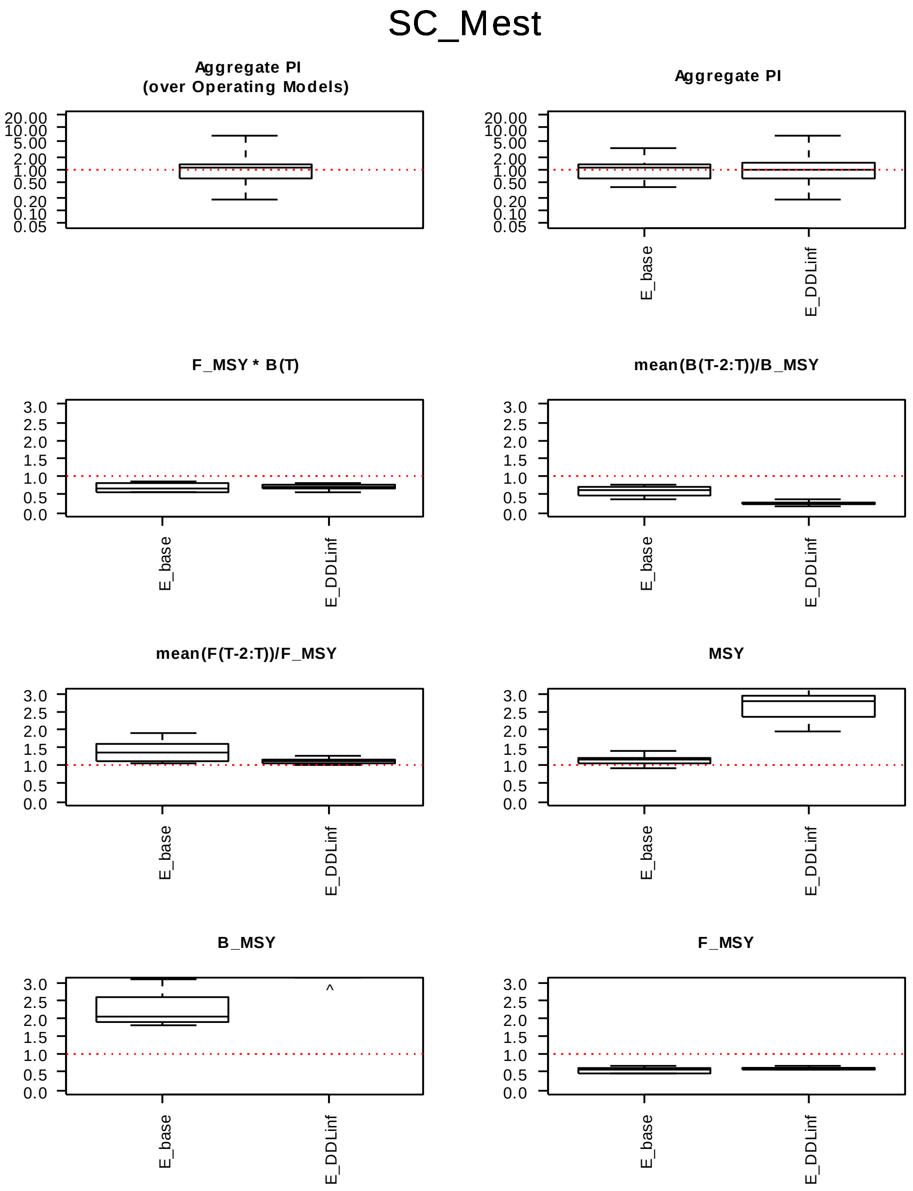

SC_Mest
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_base
E_DDLinf
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
^

Fig. 16. (cont.)

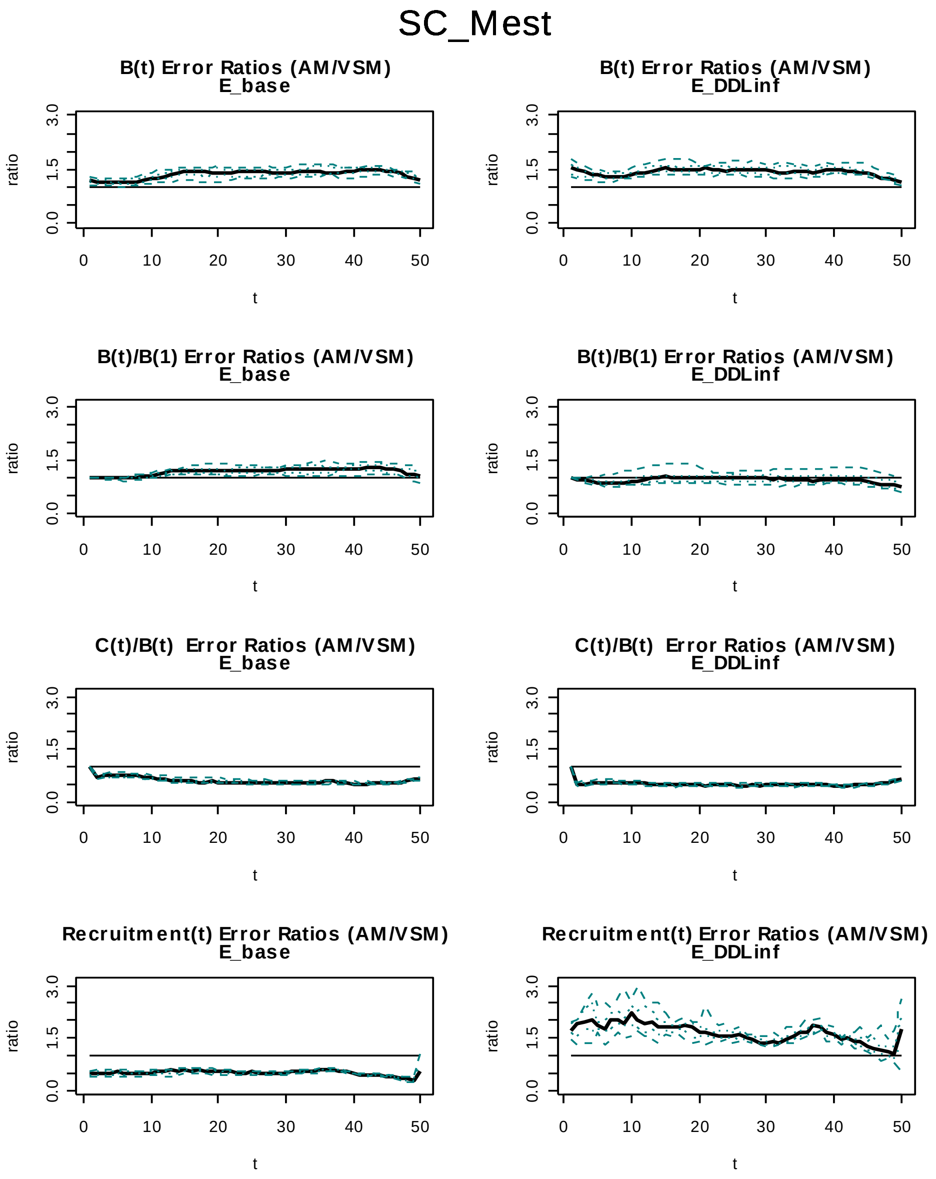

SC_Mest
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 16. (cont.)

5.7 OBJECTIVE V - ASSESSMENT IMPLICATIONS OF FISHERY SELECTIVITY ASSUMPTIONS

Fishery selectivity describes the relative rates of fishing mortality among different age classes in a population, such that if selectivity remains constant over time, then the relative abundance of age-classes in the catch provide good information about relative abundance of age-classes in the population. Most age-structured stock assessment models attempt to exploit this idea to some extent. However, if selectivity changes over time, the information content from the constant selectivity assumption is reduced, and can potentially be very misleading in some cases. In this section, we look at some of the effects of selectivity variability on the performance of different assessment models.

5.7.1 TEMPORAL VARIABILITY IN SELECTIVITY

The results of AM applications to the E_base, D_base and E_HTS scenarios described in 5.2 suggested that gradual systematic changes in selectivity linked to the age structure of the population did not have a major effect on stock assessment results. This is consistent with CCSBT Management Procedure robustness test results (Polacheck et al. 2003b, Hiramatsu et al. 2003), which examined the effects of temporal variability in selectivity on the performance of candidate MP performance. However, we would expect sudden sustained shifts in selectivity to have a substantial effect, particularly on recruitment estimates, when strong separable assumptions are applied in the assessment models. Fig. 17 shows the performance of several assessment models (indicated in the figure header) when a shift in the main longline fishery occurred 5 years before the end of the time series (E_H45), compared with the baseline OM (E_base). From these representative results, we note the following:

- Performance of the AAPMs was largely unaffected by the selectivity shift. This is not very surprising, given that these models do not attempt to represent age structure.

- the ASPM performance was highly dependent on the input selectivity vector. aspm_d2g (selectivity constant and correct up to the point of the change) was minimally affected; whereas aspm_d6g (selectivity constant and calculated using a simple estimate derived from length frequencies and equilibrium age structure assumptions) performance was substantially worse.

- SCALIA performance was variable in a manner that is consistent with our understanding of the interaction between CL sample sizes and variable selectivity. SC_base (moderate CL effective sample size and temporally variable selectivity) was only modestly affected by the selectivity shift, presumably indicating that the shift was reasonably well estimated. In contrast, the constant selectivity models SC_1ideal and SC_noHTS (the latter not shown), interpreted the strong change in CL signal as a shift in recruitment, with corresponding implications for the biomass and management-related estimates.

- despite the constant selectivity assumption, the aspm_d2g model was not sensitive to the selectivity shift as was SC_1ideal. The absence of CL data in the likelihood limit the extent to which numbers-at-age could deviate to explain the change in CL signal. Furthermore, the deterministic recruitment would not allow the model to estimate large recruitment anomalies in the most recent years even if the CL data was used.

- Estimation implications of a shift in selectivity 10 years before the final year (not shown) were qualitatively very similar to the shift at T-5.

The effect of the selectivity shift is similar in the difficult scenario (D_H45), except that the problems are superimposed on top of the other estimation errors characteristic of the D_x scenarios (Fig. 18). There is a modest change in performance between D_base and D_H45 for f_calc, aspm_d2g and SC_base, but it is minor compared with the effect on SC_1ideal. Clearly there is potential to reliably estimate selectivity changes, and this is something that should be attempted if there are suspicious changes in recruitment, particularly if there is auxilliary information about changes in fishing industry practices. But we note that if multiple fisheries were to change their targeting practices simultaneously (e.g. in response to a global management action), we do not know if the effect could be estimated.

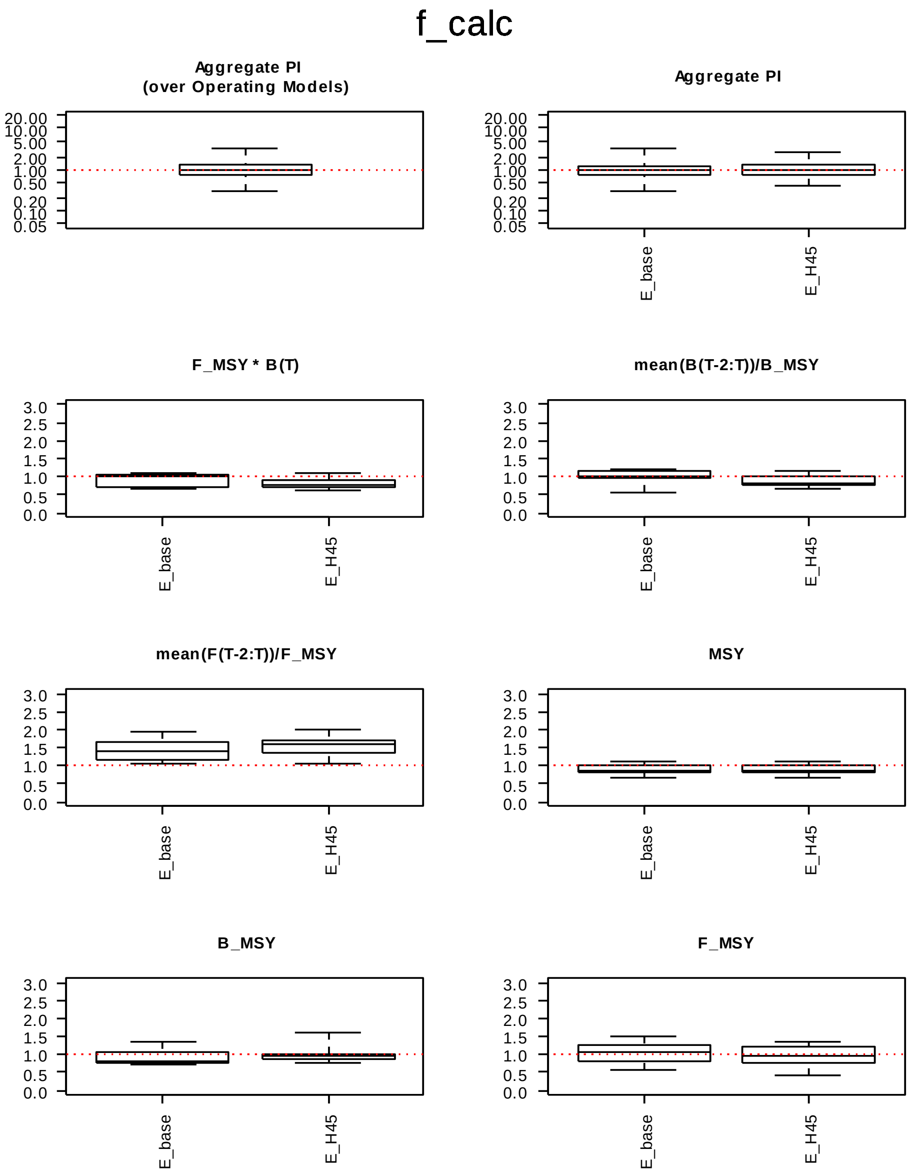


Fig. 17. **Performance indicators resulting from the application of a range of AMs to 10 simulated data realizations from OMs E_base, and E_H45 (shift in longline fishery selectivity 5 years before the end of the time series). Each PI is a ratio of (AM estimated)/(OM actual). Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs (indicated at top of page) are defined in Table 2 and PIs in Table 8.**

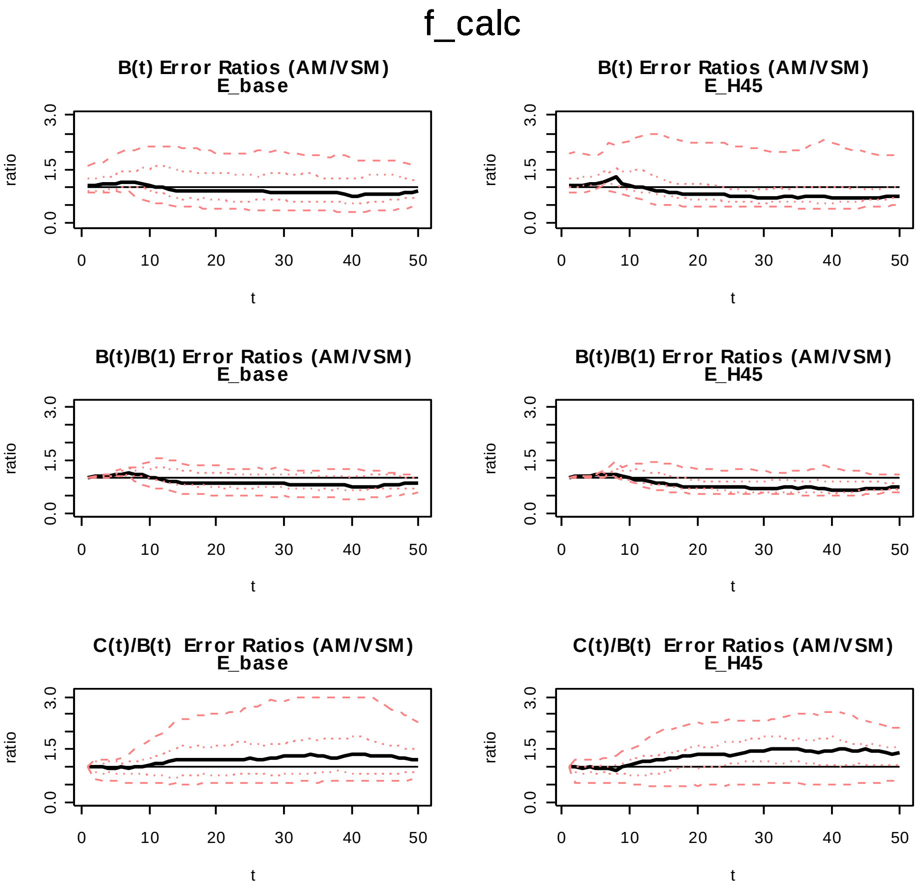

f_calc
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 17. (cont.)

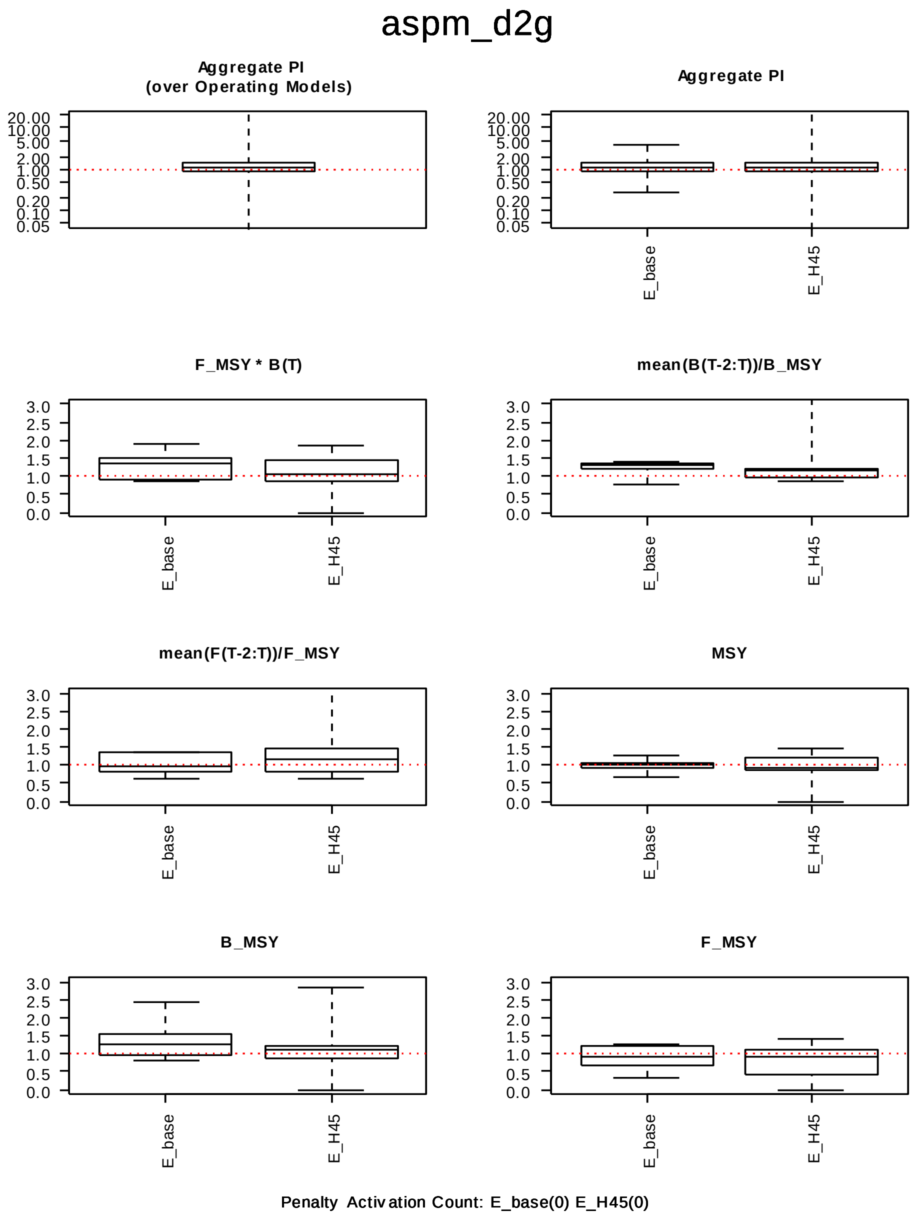

aspm_d2g
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
E_base
E_H45
Penalty Activation Count: E_base(0) E_H45(0)

Fig. 17. (cont.)

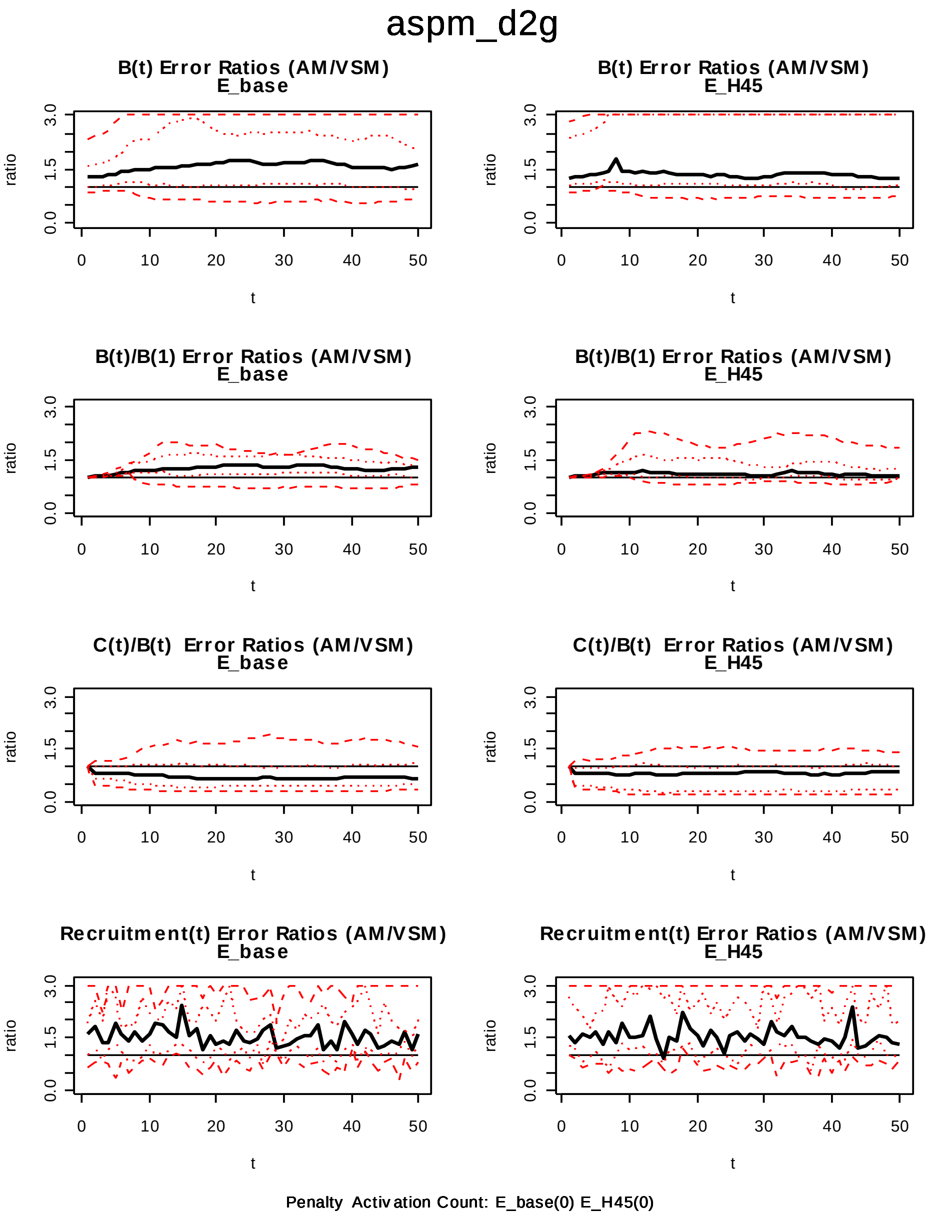

aspm_d2g
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_H45
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0) E_H45(0)

Fig. 17. (cont.)

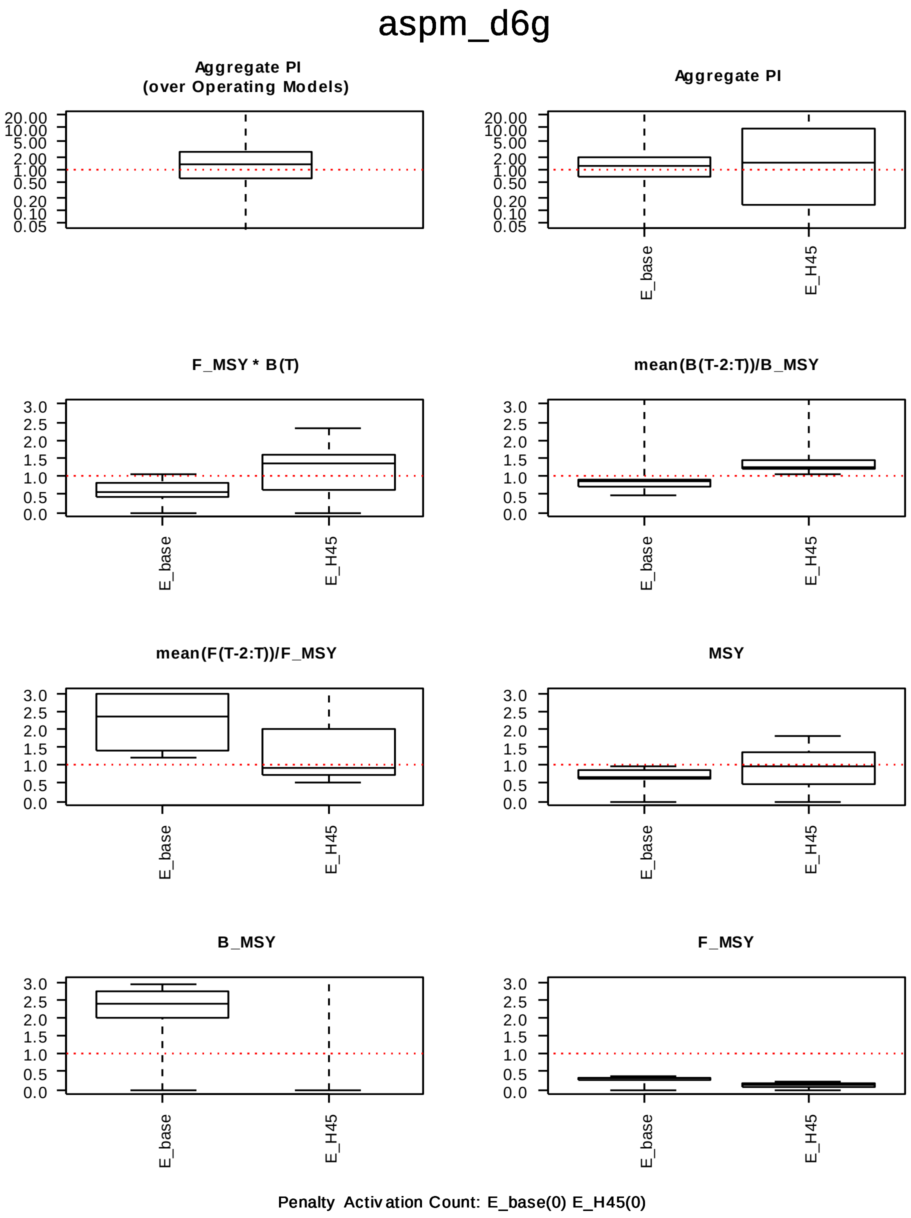

aspm_d6g
Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
Aggregate PI
E_base
E_H45
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
Penalty Activation Count: E_base(0) E_H45(0)

Fig. 17. (cont.)

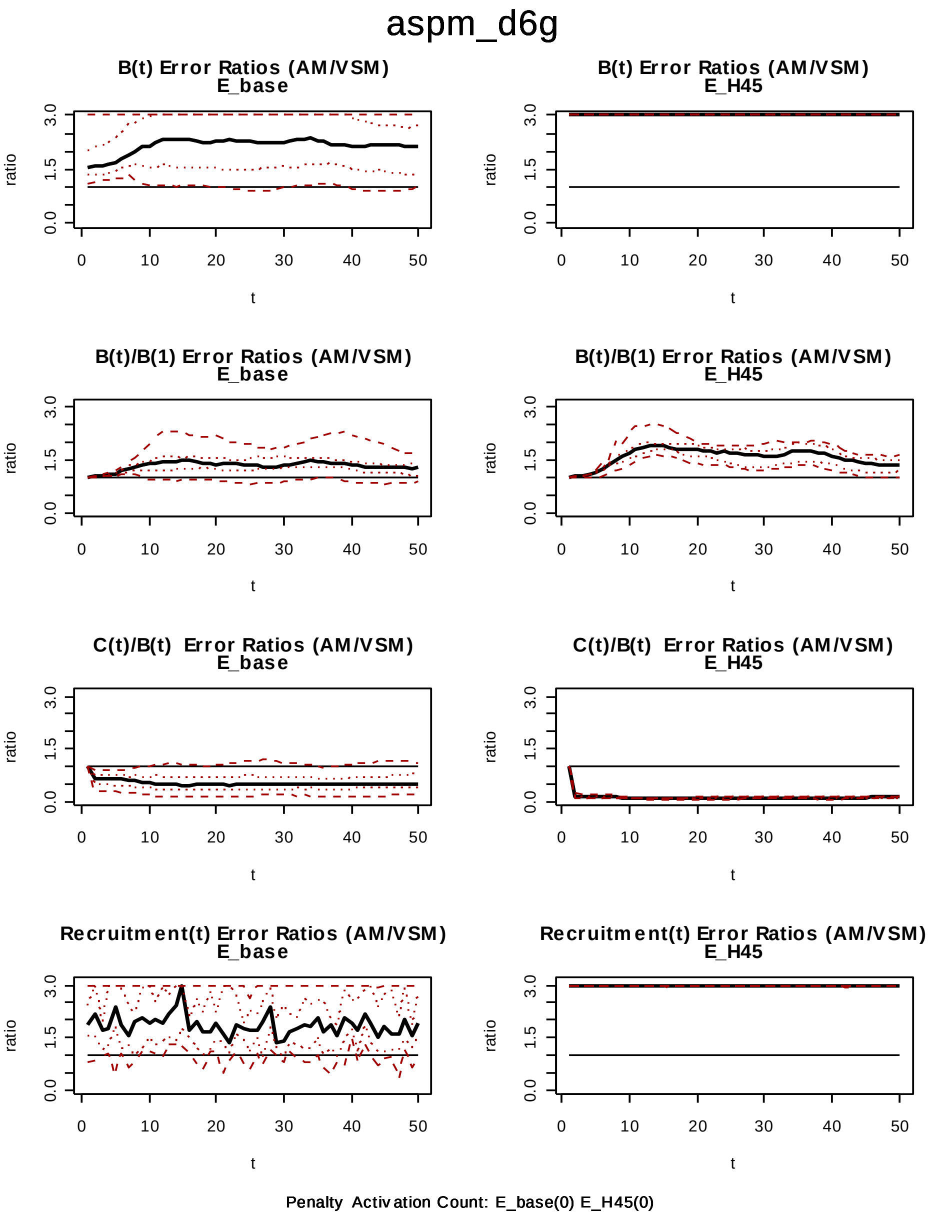

aspm_d6g
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_H45
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0) E_H45(0)

Fig. 17. (cont.)

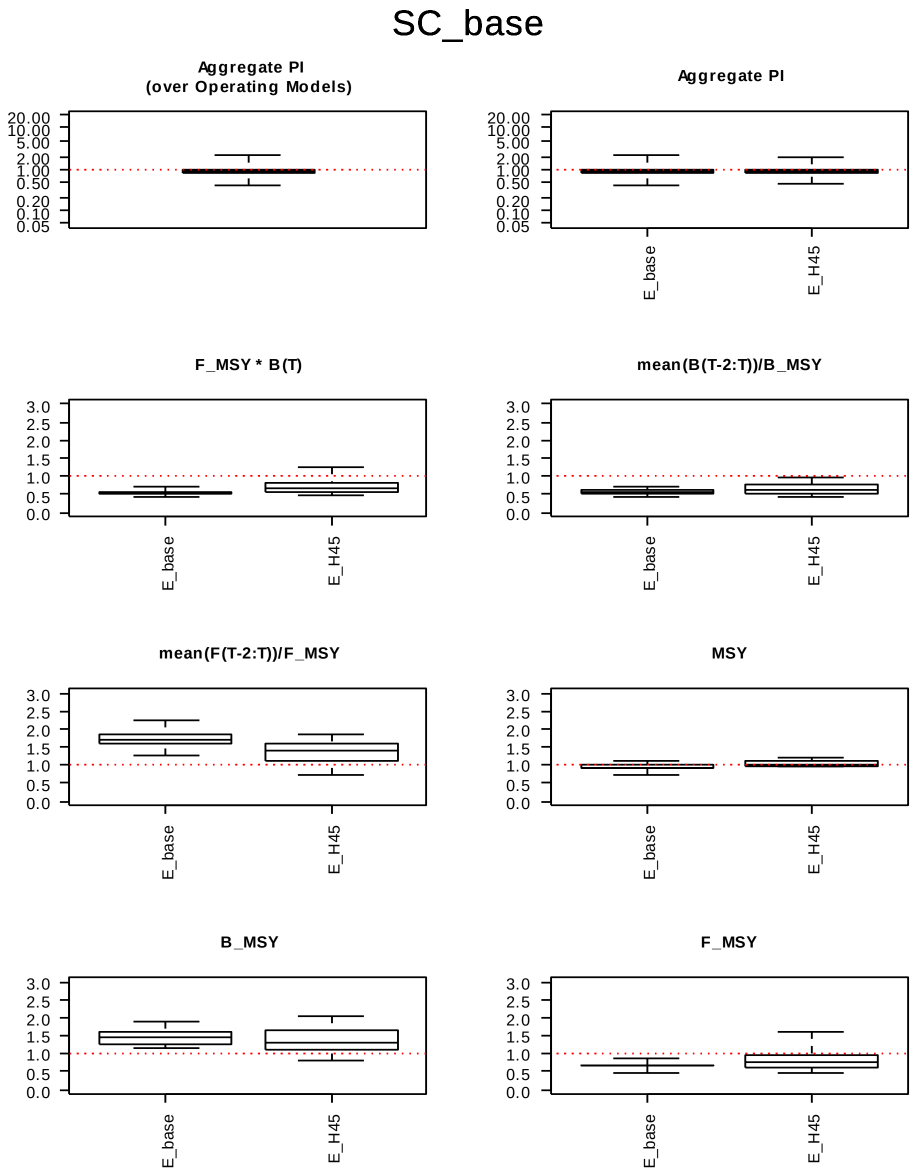

SC_base
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_base
E_H45
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Fig. 17. (cont.)

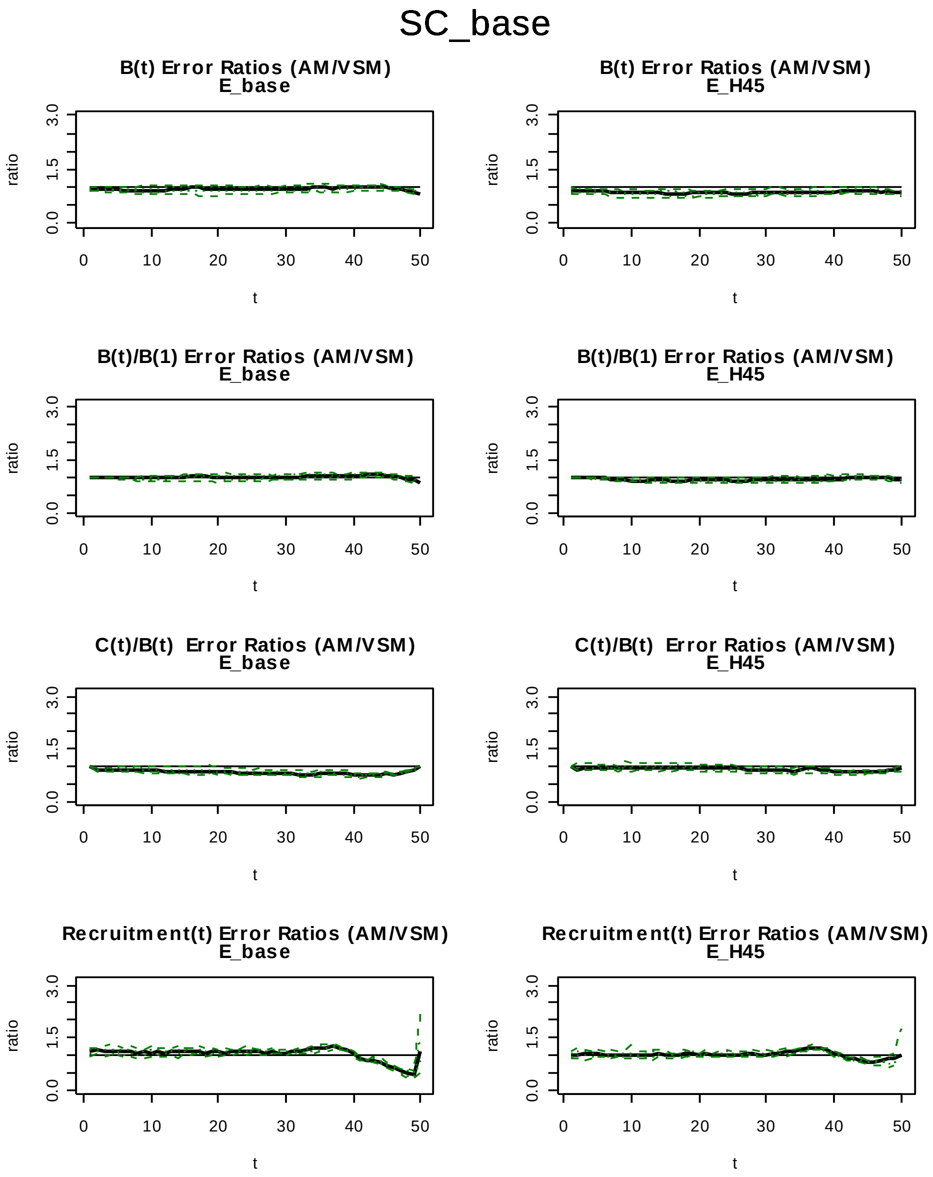

SC_base
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_H45
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 17. (cont.)

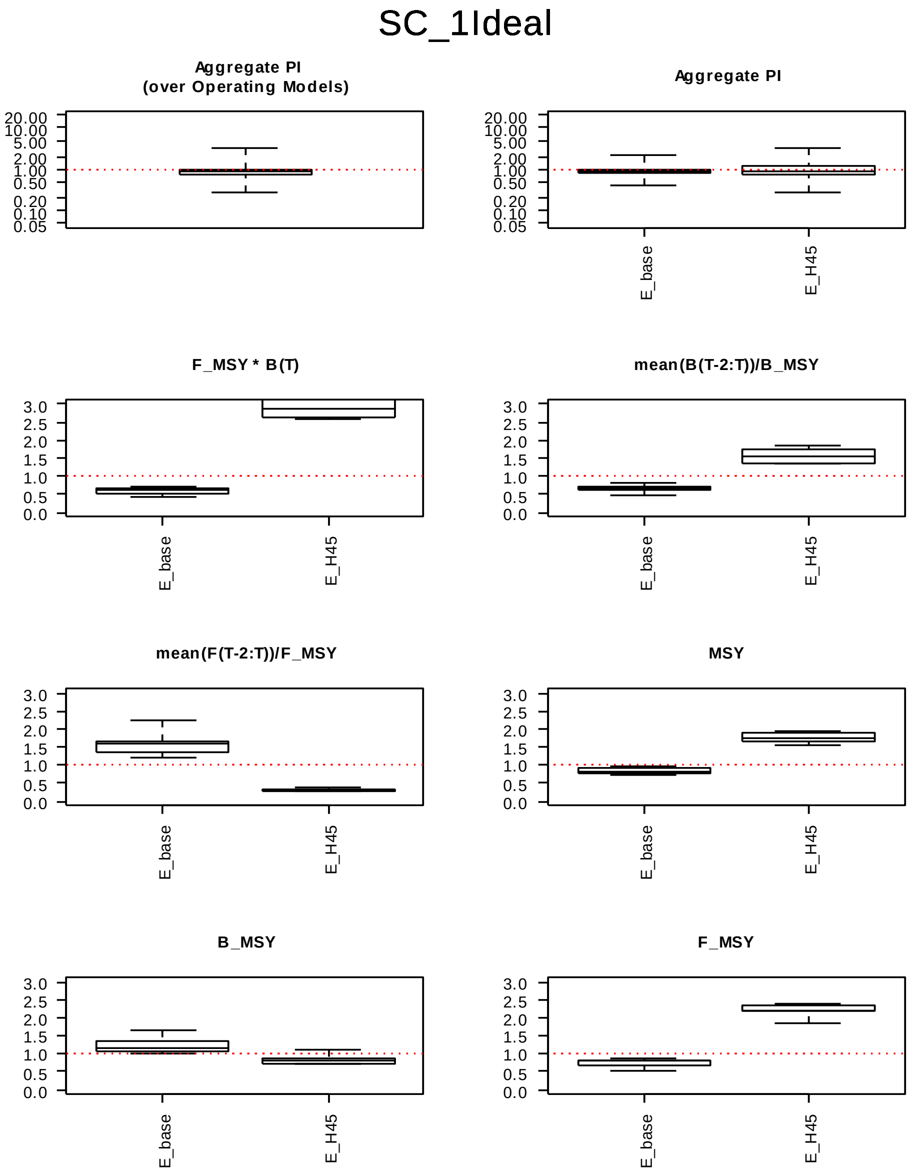
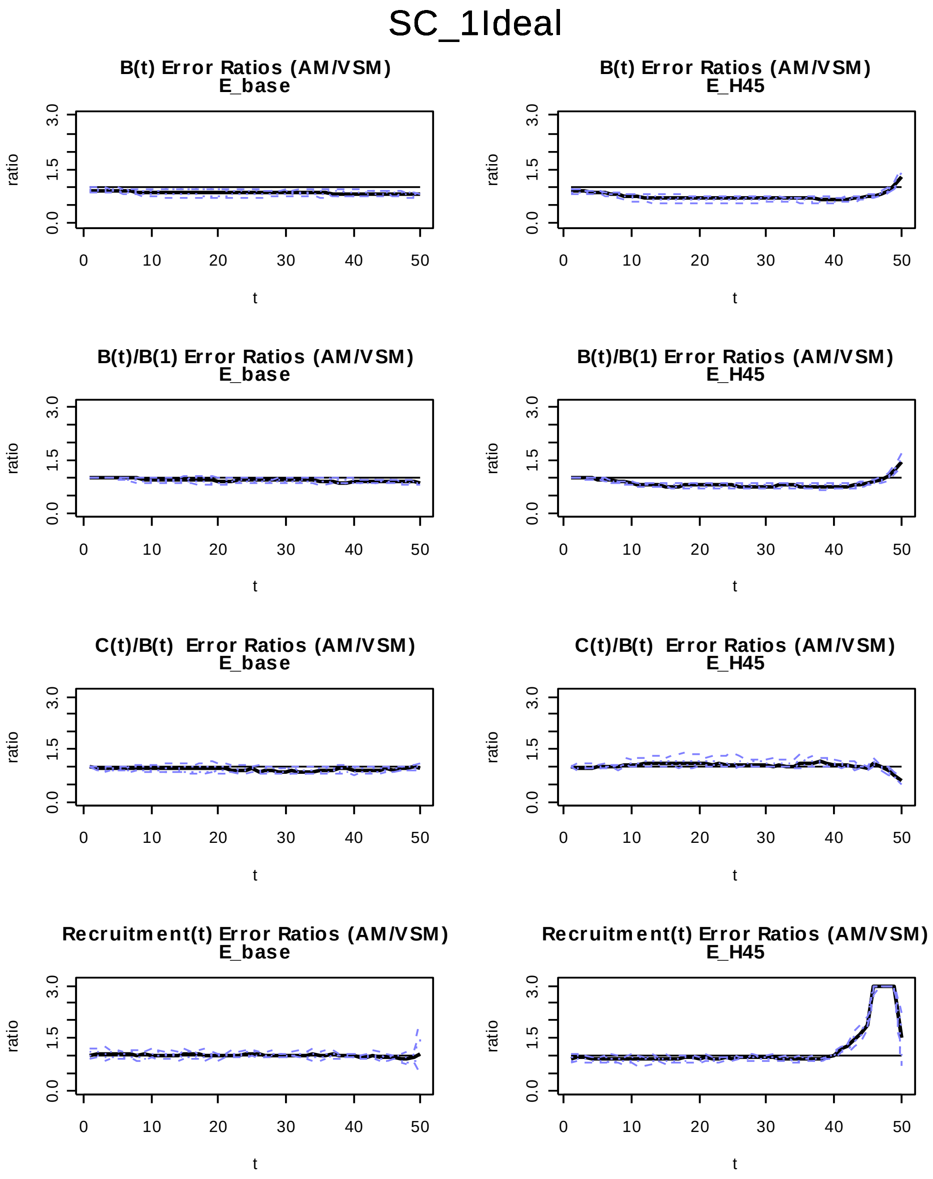

SC_1Ideal
Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
Aggregate PI
E_base
E_H45
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Fig. 17. (cont.)


SC_1Ideal

B(t) Error Ratios (AM/VSM) E_base

B(t) Error Ratios (AM/VSM) E_H45

B(t)/B(1) Error Ratios (AM/VSM) E_base

B(t)/B(1) Error Ratios (AM/VSM) E_H45

C(t)/B(t) Error Ratios (AM/VSM) E_base

C(t)/B(t) Error Ratios (AM/VSM) E_H45

Recruitment(t) Error Ratios (AM/VSM) E_base

Recruitment(t) Error Ratios (AM/VSM) E_H45

Fig. 17. (cont.)

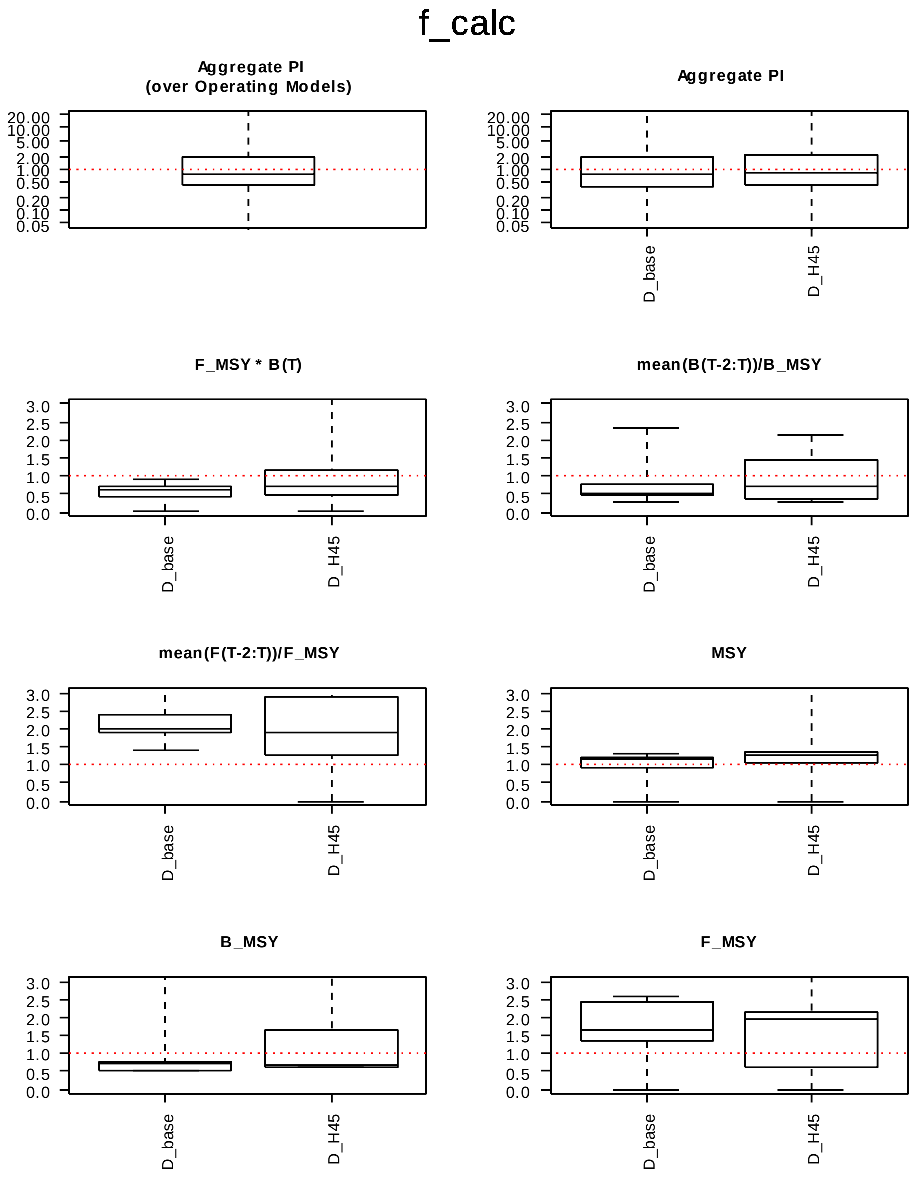


Fig. 18 **Performance indicators resulting from the application of a range of AMs to 10 simulated data realizations from the difficult OM scenario D_H45 (shift in longline fishery selectivity 5 years before the end of the time series). Each PI is a ratio of (AM estimated)/(OM actual). Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

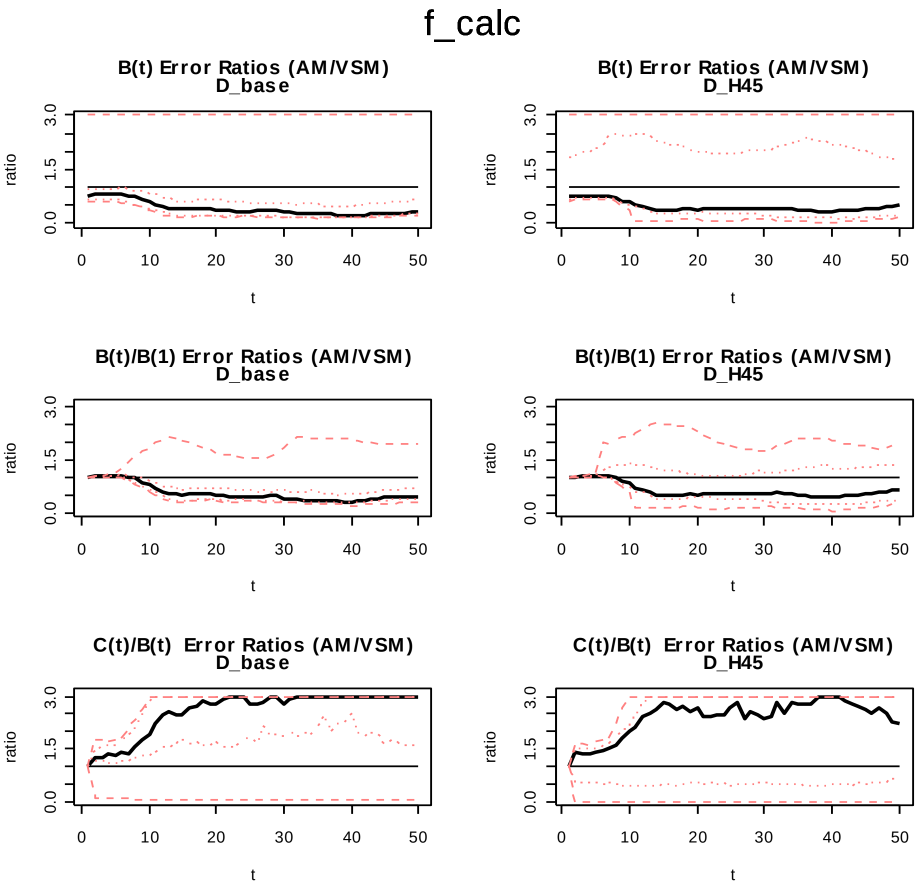

f_calc
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_H45
B(t)/B(1) Error Ratios (AM/VSM)
D_base
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
C(t)/B(t) Error Ratios (AM/VSM)
D_base
C(t)/B(t) Error Ratios (AM/VSM)
D_H45
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Fig. 18. (cont.)

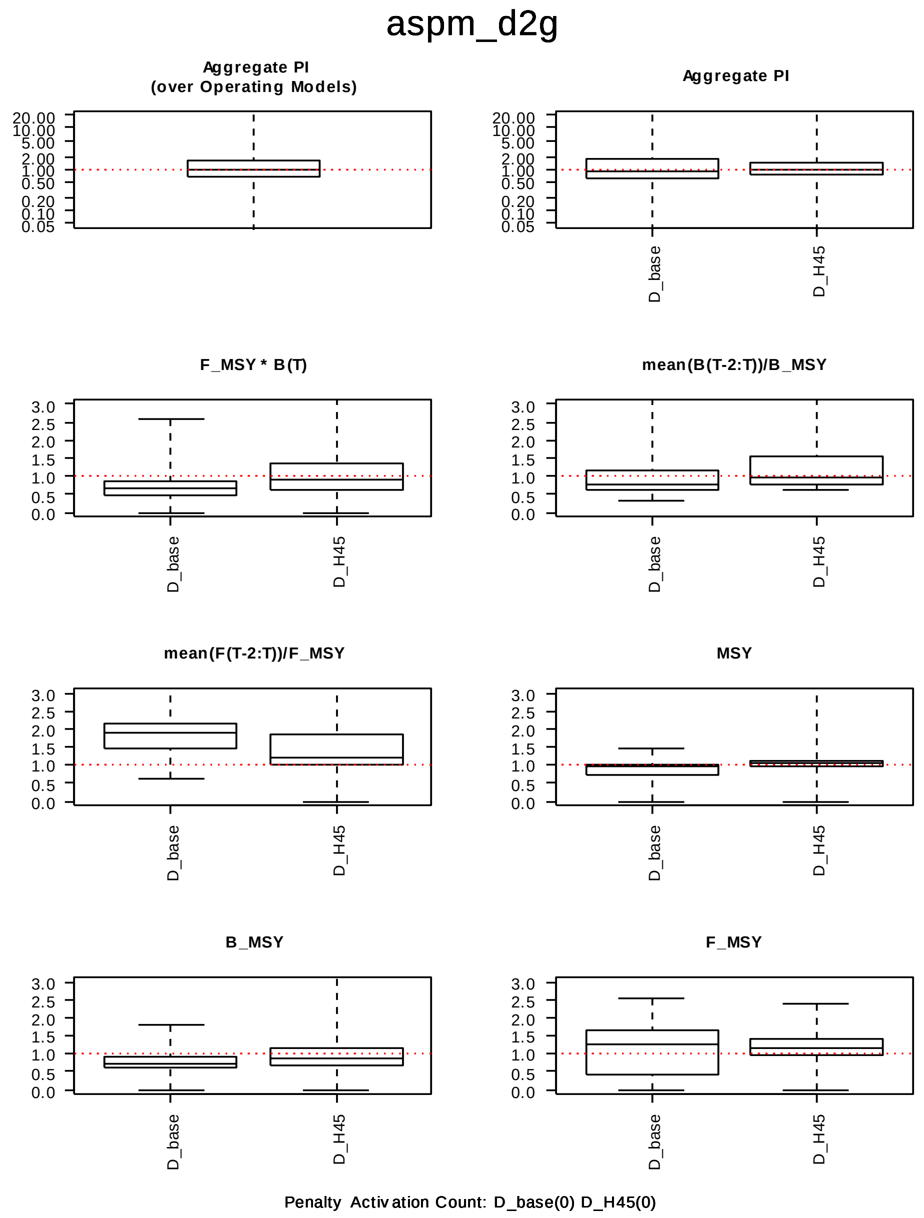

aspm_d2g
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
D_base
D_H45
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Penalty Activation Count: D_base(0) D_H45(0)

Fig. 18. (cont.)

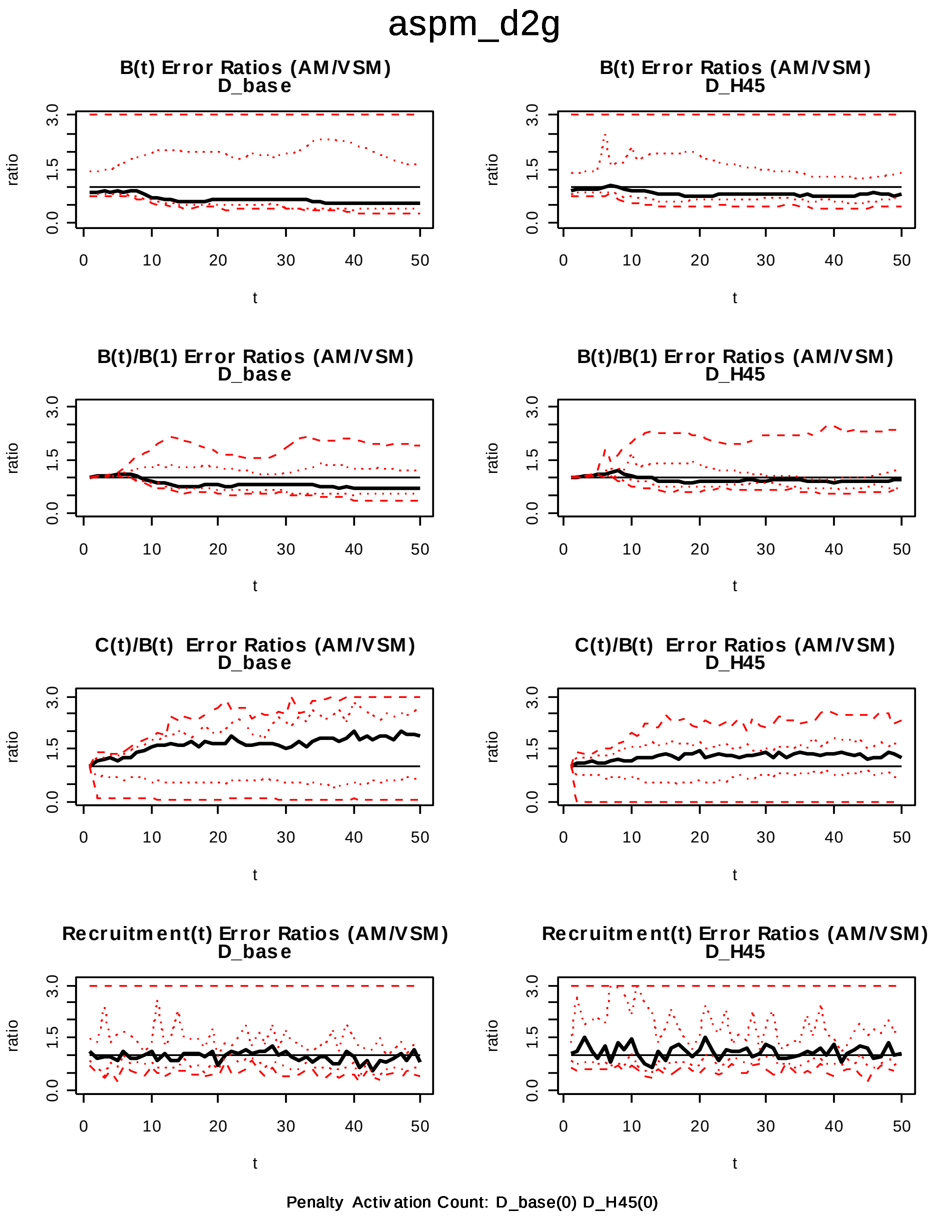

aspm_d2g
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_H45
B(t)/B(1) Error Ratios (AM/VSM)
D_base
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
C(t)/B(t) Error Ratios (AM/VSM)
D_base
C(t)/B(t) Error Ratios (AM/VSM)
D_H45
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_H45
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: D_base(0) D_H45(0)

Fig. 18. (cont.)

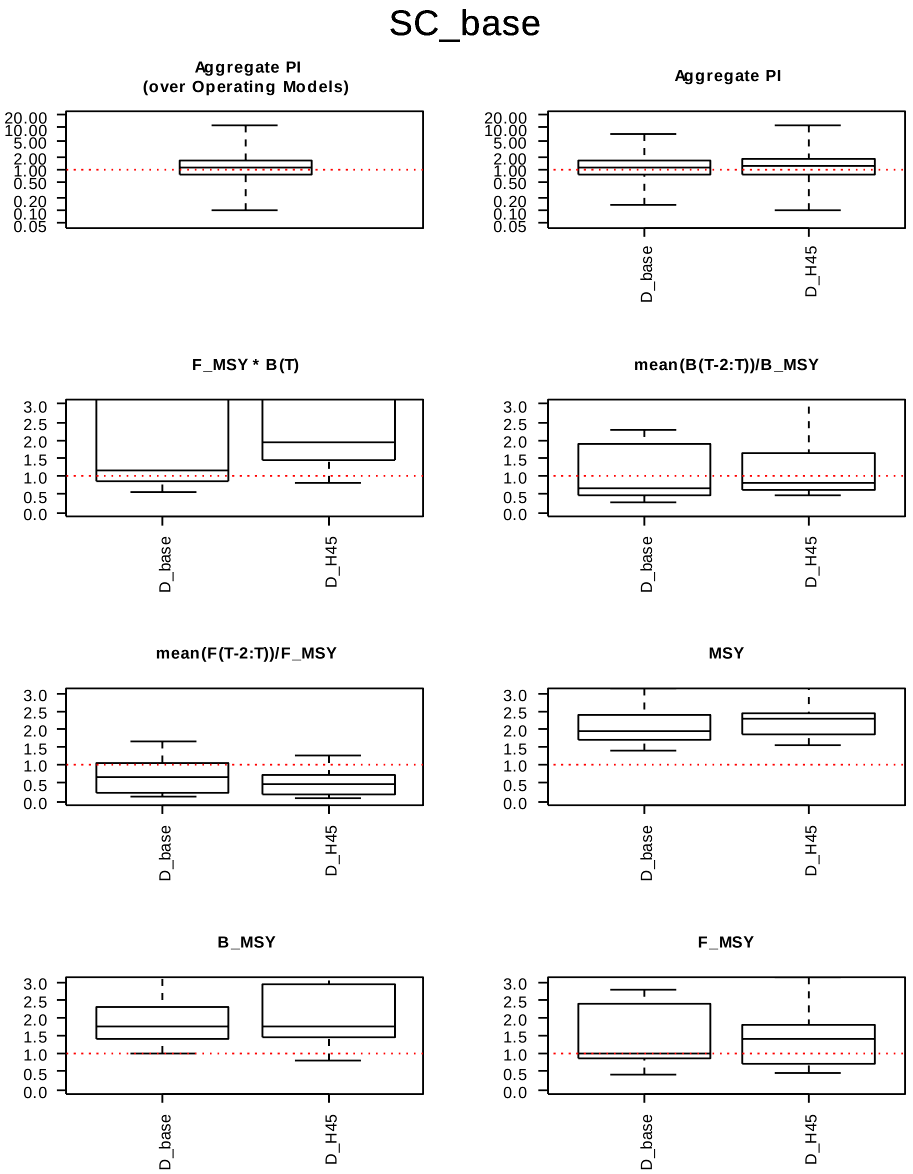

SC_base
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0
D_base
D_H45

Fig. 18. (cont.)

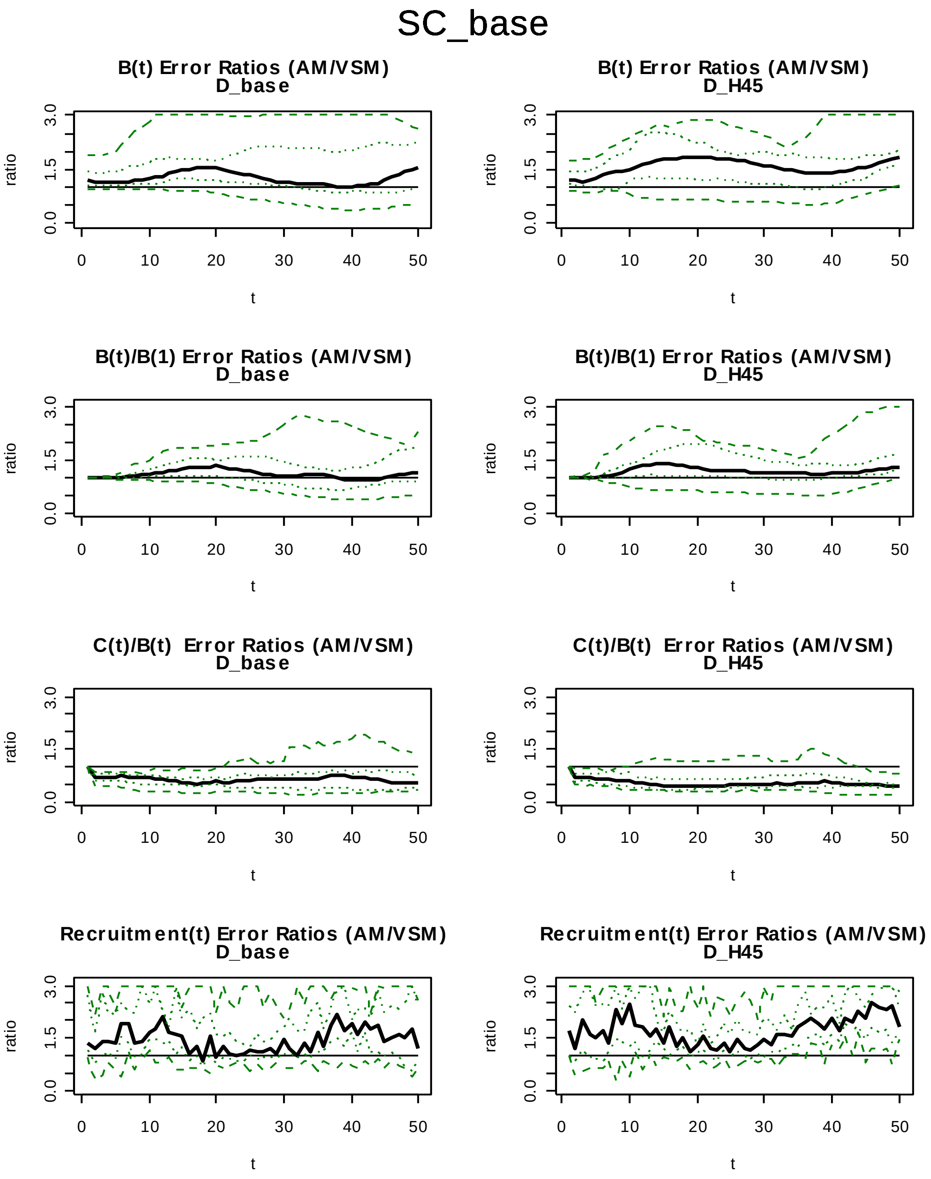

SC_base
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_H45
B(t)/B(1) Error Ratios (AM/VSM)
D_base
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
C(t)/B(t) Error Ratios (AM/VSM)
D_base
C(t)/B(t) Error Ratios (AM/VSM)
D_H45
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_H45
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Fig. 18. (cont.)

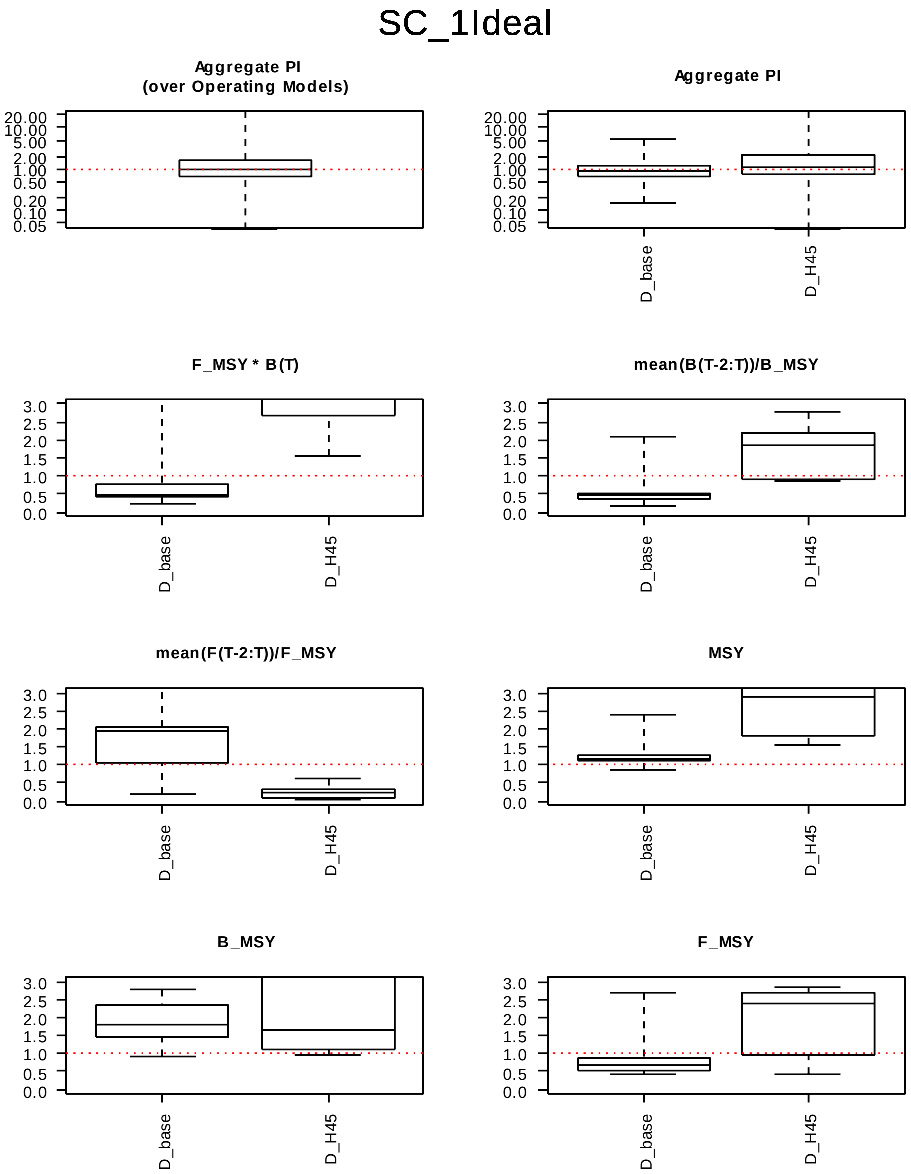

SC_1Ideal
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
D_base
D_H45
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B_MSY
F_MSY
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Fig. 18. (cont.)

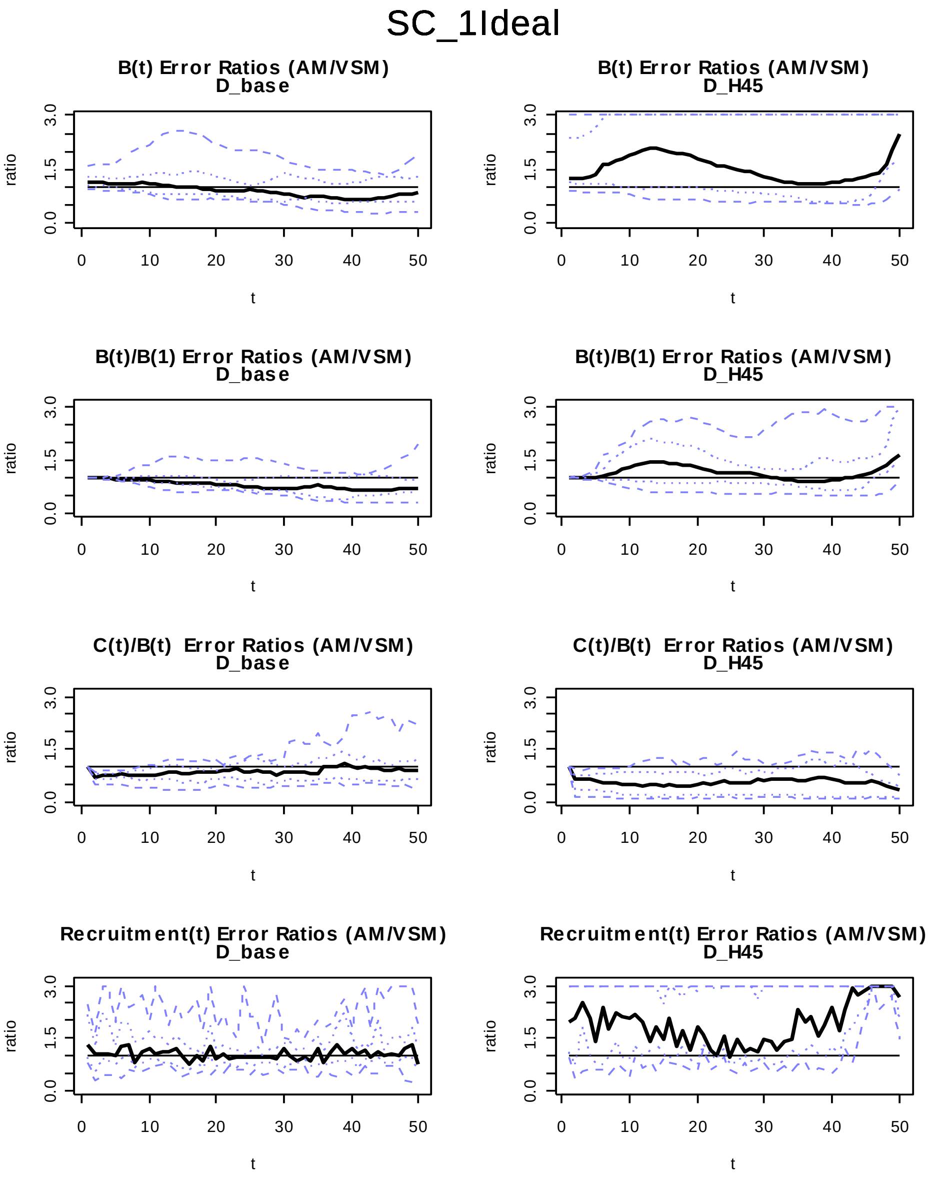

SC_1Ideal
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_H45
B(t)/B(1) Error Ratios (AM/VSM)
D_base
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
C(t)/B(t) Error Ratios (AM/VSM)
D_base
C(t)/B(t) Error Ratios (AM/VSM)
D_H45
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_H45
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Fig. 18. (cont.)

5.7.2 LENGTH-BASED FISHERY SELECTIVITY

We tested one possible implementation of purely size-based selectivity in VSM operating model specification E_HL and found that performance of all models was only slightly different from E_base (Fig. 19; only results from SC_base are shown). One could probably contrive size selectivity scenarios that would be troublesome for the models (e.g. by inflating the variance on the length-at-age, implementing a different form of growth, and/or exaggerating the selectivity effect with knife-edged functions), but we do not see this avenue of exploration as a high priority for SBT, given the other assessment issues that clearly do cause problems.

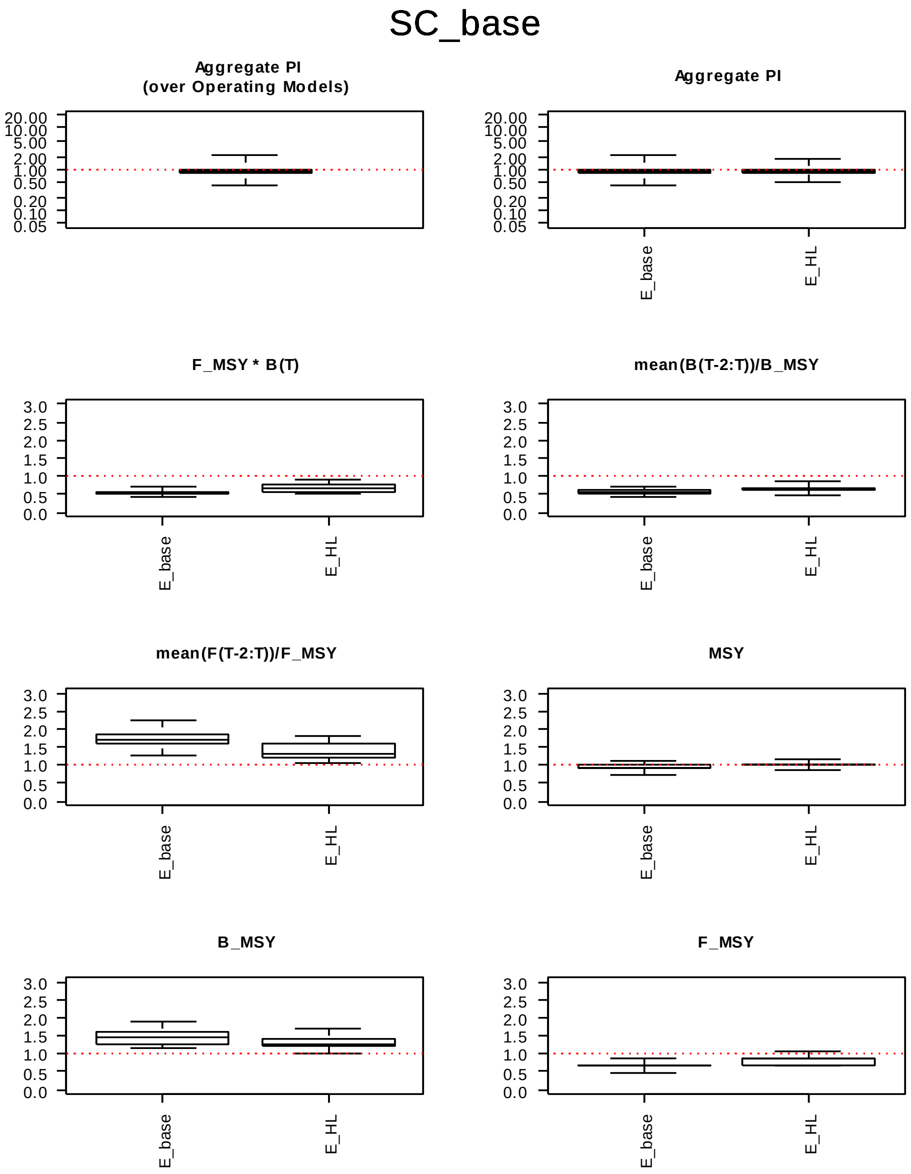


Fig. 19. **MPD performance indicators resulting from the application of a range of assessment models to 10 simulated data realizations from operating model E_HL, which has purely length-based fishery selectivity. Each PI is a ratio of (AM estimated)/(OM actual). Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

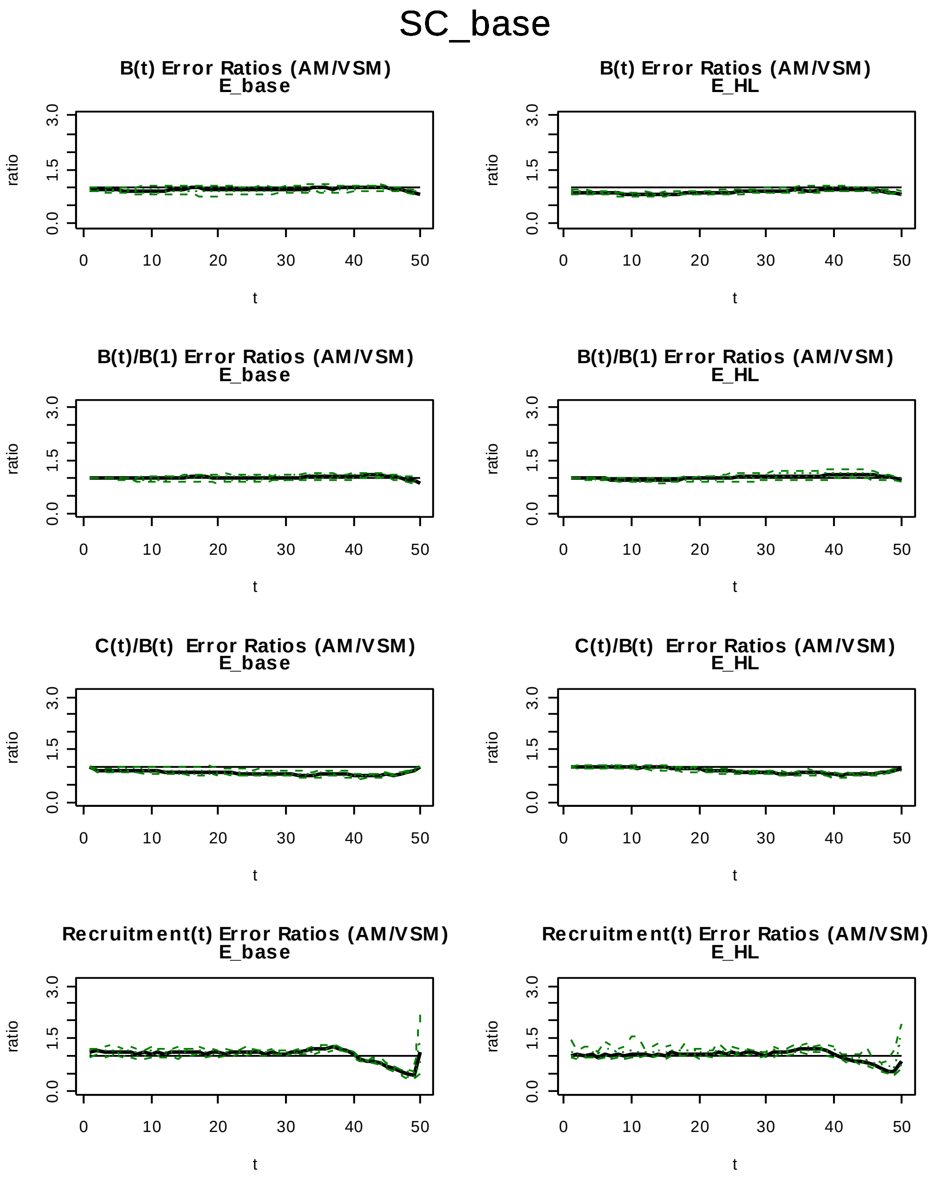

SC_base
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_HL
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_HL
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_HL
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_HL
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 19 **(cont.)**

5.8 OBJECTIVE VI - ASSESSMENT IMPLICATIONS OF CATCHABILITY TEMPORAL VARIABILITY IN RELATIVE ABUNDANCE INDICES

Most dynamic stock assessment models use some form of relative abundance index (usually derived from CPUE for oceanic pelagic fisheries) as one of the primary data inputs. This section is intended to illustrate how assessments can be badly misleading if the abundance index is mistakenly assumed to be directly proportional to abundance, and explores how effectively integrative models can estimate temporal variability in catchability, given the other sources of data that are available for SBT (i.e. total catch, catch-at-length, catch-at-age, tag releases and recoveries). The operating models that we examine in relation to this section include the ideal case, E_base (catchability constant), E_qInc (catchability increasing over time at 1% per year), E_qI (catchability increases and decreases over time in relation to effort, in a manner that is qualitatively consistent with co-operation among fishers), and E_qC (catchability increases and decreases over time in relation to effort, in a manner that is qualitatively consistent with interference among fishers; the catchability pattern is opposite to E_qI). The D_x scenarios are analogous to the E_x scenarios, but the catchability problems are overlaid on all the other data issues that make D_x difficult (including an auto-correlated pattern to the effort deviations).

Assessment model performance was clearly dependent on the reliability of CPUE as a relative abundance index, but the estimation errors were not always consistent with what we would have anticipated. Fig. 20 provides an overview of the performance of several representative models, which illustrate the following:

- The difficult OM scenarios (D_x) were generally more problematic than the easy scenarios (E_x), with a couple of conspicuous exceptions.

- Within the E_x and D_x scenarios, the production models generally had the worst performance for scenarios E_qC and D_qC respectively (based on the aggregate performance index). E_qC and E_qI had biomass estimation biases in opposite directions as might have been expected (because the temporal patterns in catchability are opposite in the two scenarios). We expected that the production model estimates would result an increasing biomass error trend over time for the E_qInc and D_qInc scenarios, but this was not obvious. aspm_d2g did demonstrate an increasing biomass error trend for E_qInc, but there were many differences in the estimation characteristics between E_base and E_qInc, such that it was not even obvious that overall performance was worse for E_qInc than E_base.

- Among the E_x scenarios, all the SCALIA models (with the exception of SC_noTag) demonstrated the worst performance against scenario E_qInc. All SCALIA models had similar problems with E_qC as did the production models, but these were not as serious as with E_qInc. It was not obvious which scenario was the most problematic in the D_x series for the SCALIA models (except D_qC seemed somewhat worse for SC_noTag). All the SCALIA models produced an increasing trend in the biomass estimation error for E_qInc as would be expected for an increasing catchability trend. However in all cases (except SC_noTag) the trend in biomass error greatly exceeded the catchability trend in the operating model.

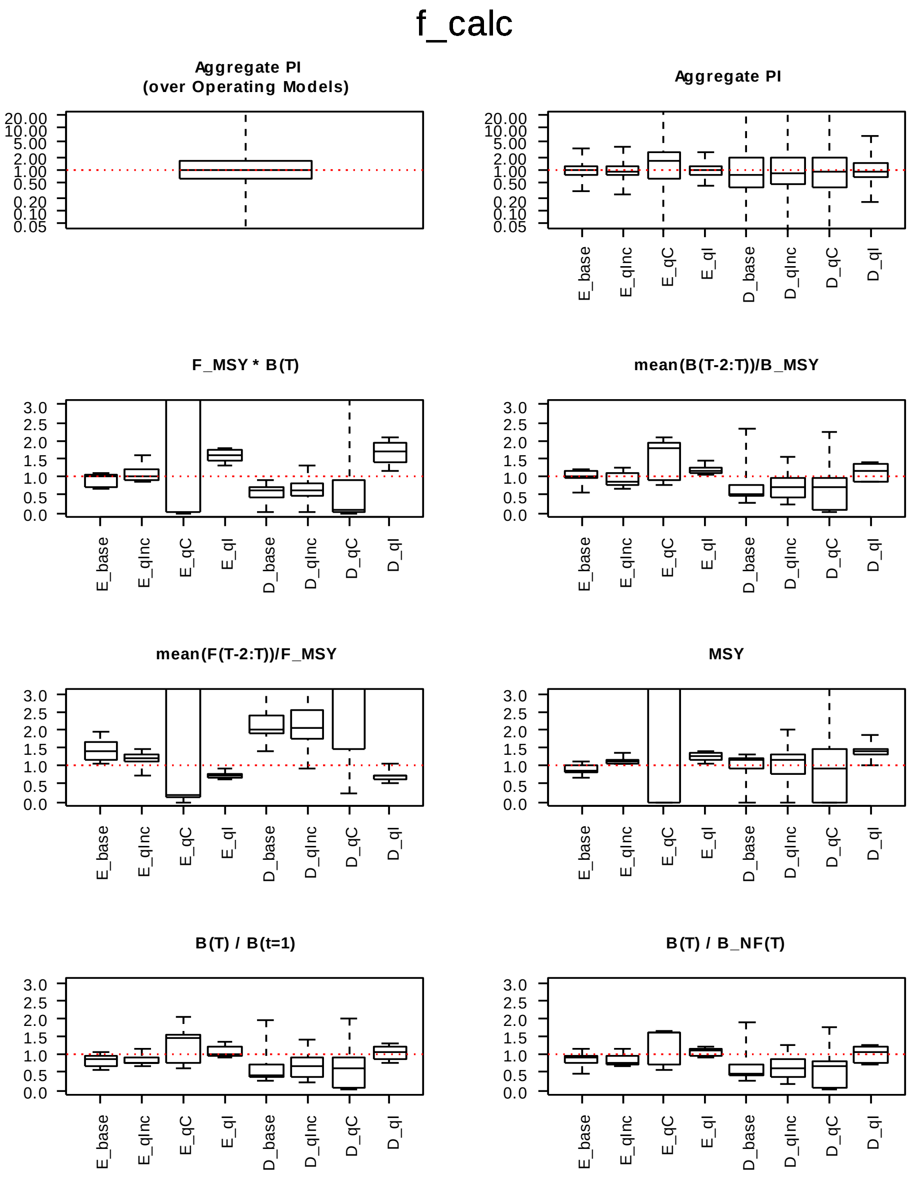


Fig. 20. Overview of the performance of assessment models when challenged by operating models with a range of assumptions about the effort – fishing mortality relationship. Each AM was applied to 10 simulated data realizations from each OM. Each PI is a ratio of (AM estimated)/(OM actual). Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.

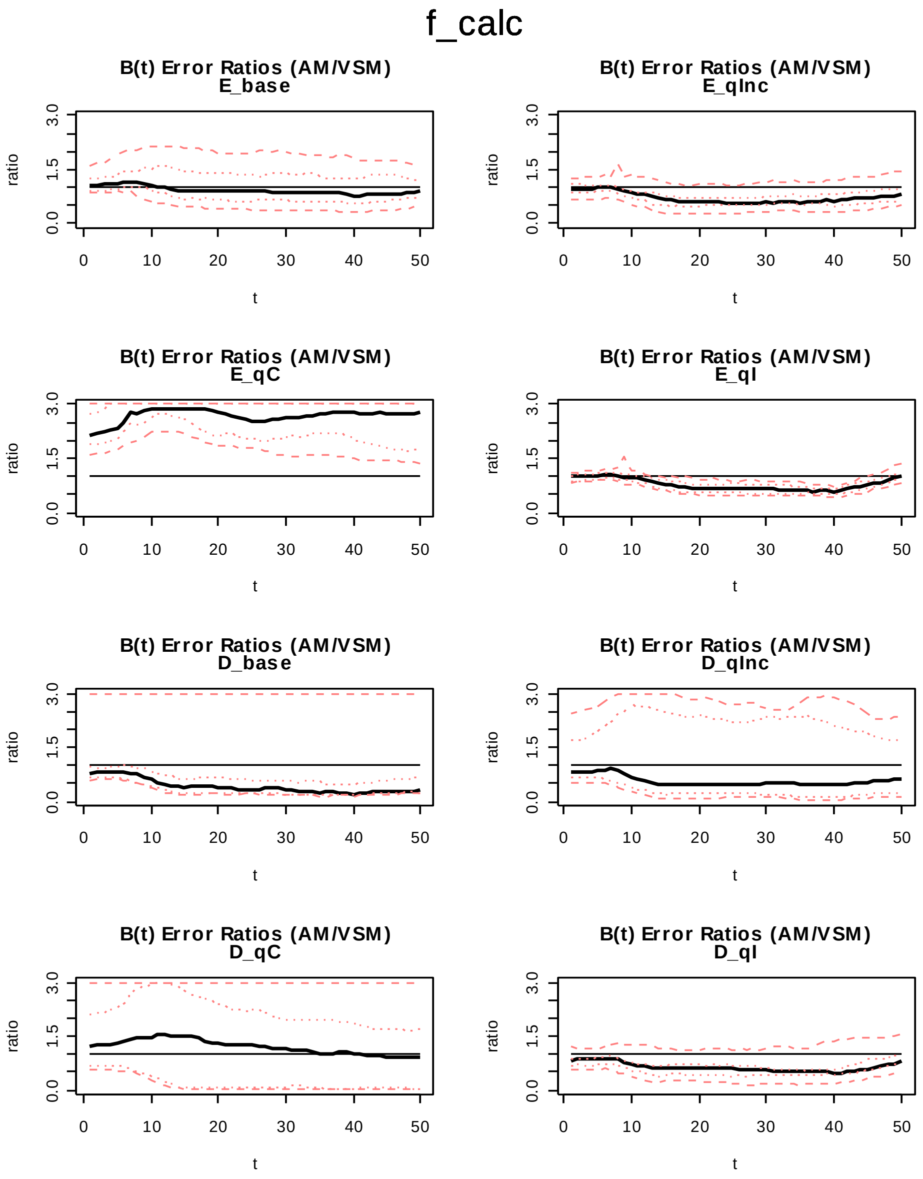

f_calc
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Fig. 20(cont.)

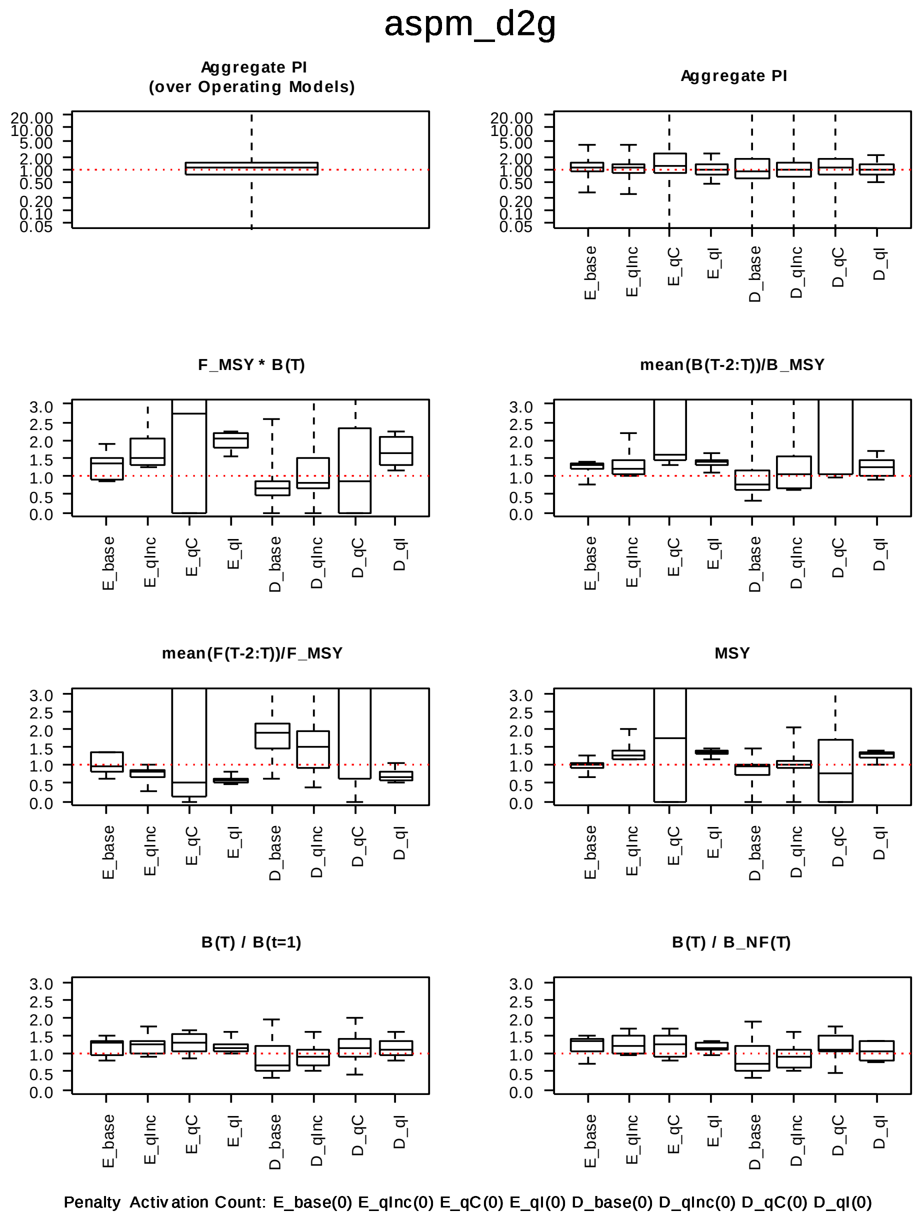

aspm_d2g
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B(T) / B(t=1)
B(T) / B_NF(T)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
E_base
E_qInc
E_qC
E_ql
D_base
D_qInc
D_qC
D_ql
Penalty Activation Count: E_base(0) E_qInc(0) E_qC(0) E_ql(0) D_base(0) D_qInc(0) D_qC(0) D_ql(0)

Fig. 20(cont.)

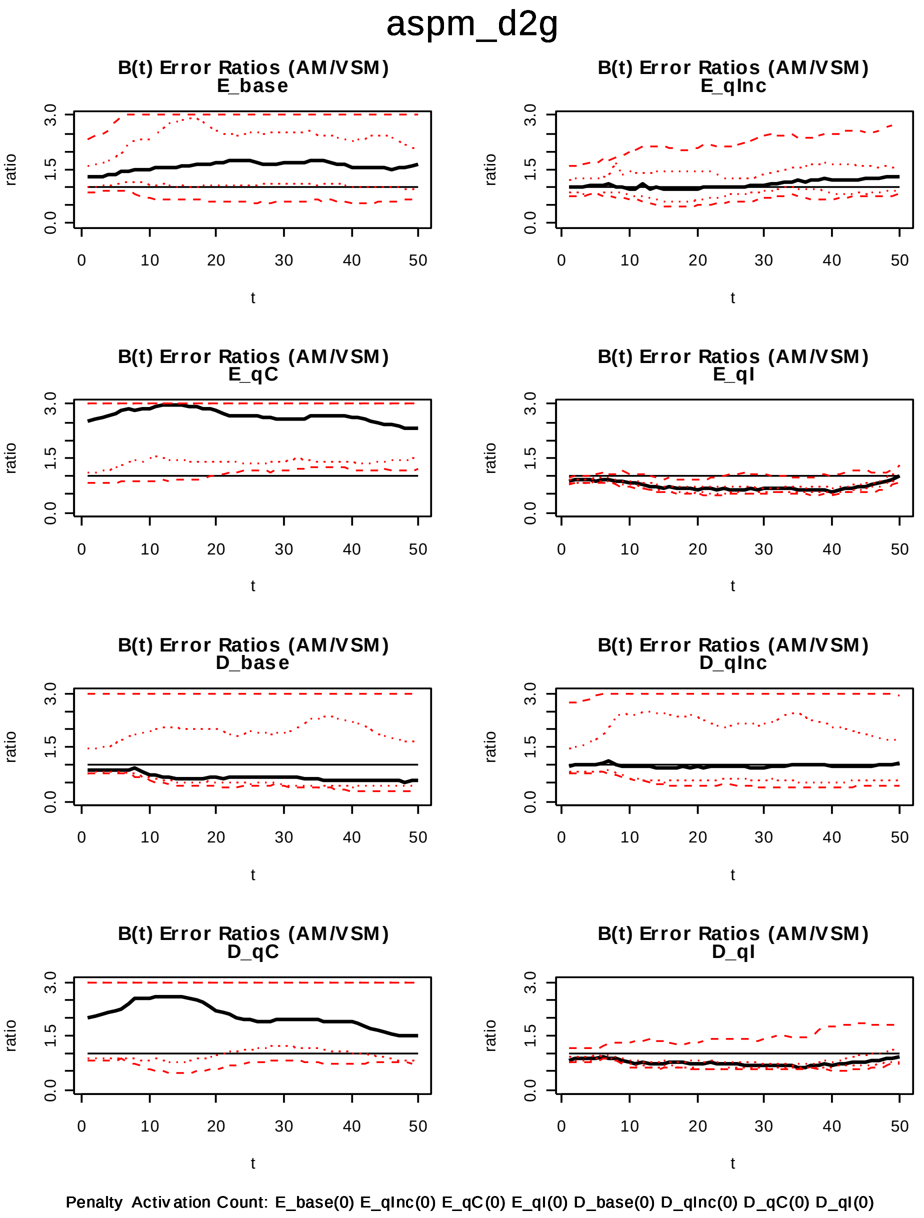

aspm_d2g
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_base(0) E_qInc(0) E_qC(0) E_qI(0) D_base(0) D_qInc(0) D_qC(0) D_qI(0)

Fig. 20(cont.)

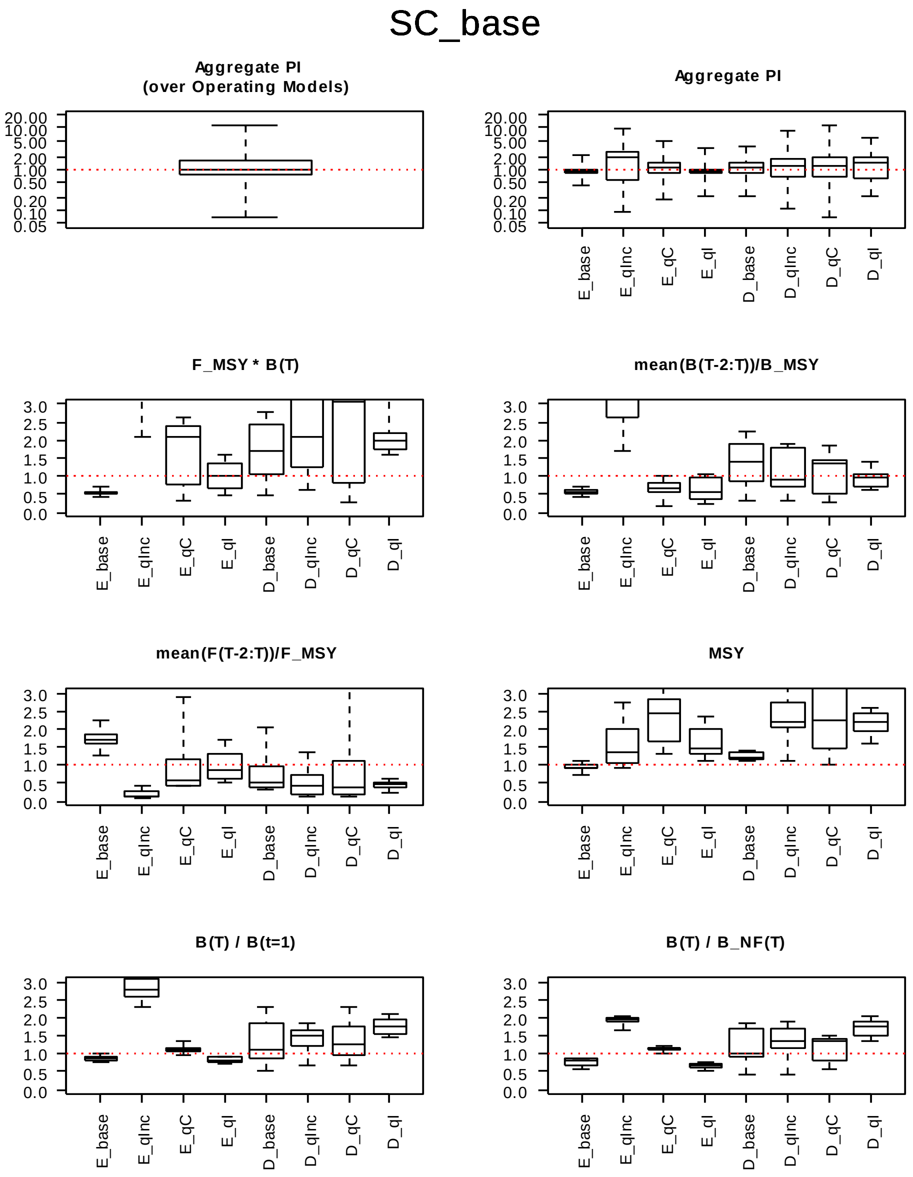

SC_base
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B(T) / B(t=1)
B(T) / B_NF(T)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
E_base
E_qInc
E_qC
E_ql
D_base
D_qInc
D_qC
D_ql

Fig. 20(cont.)

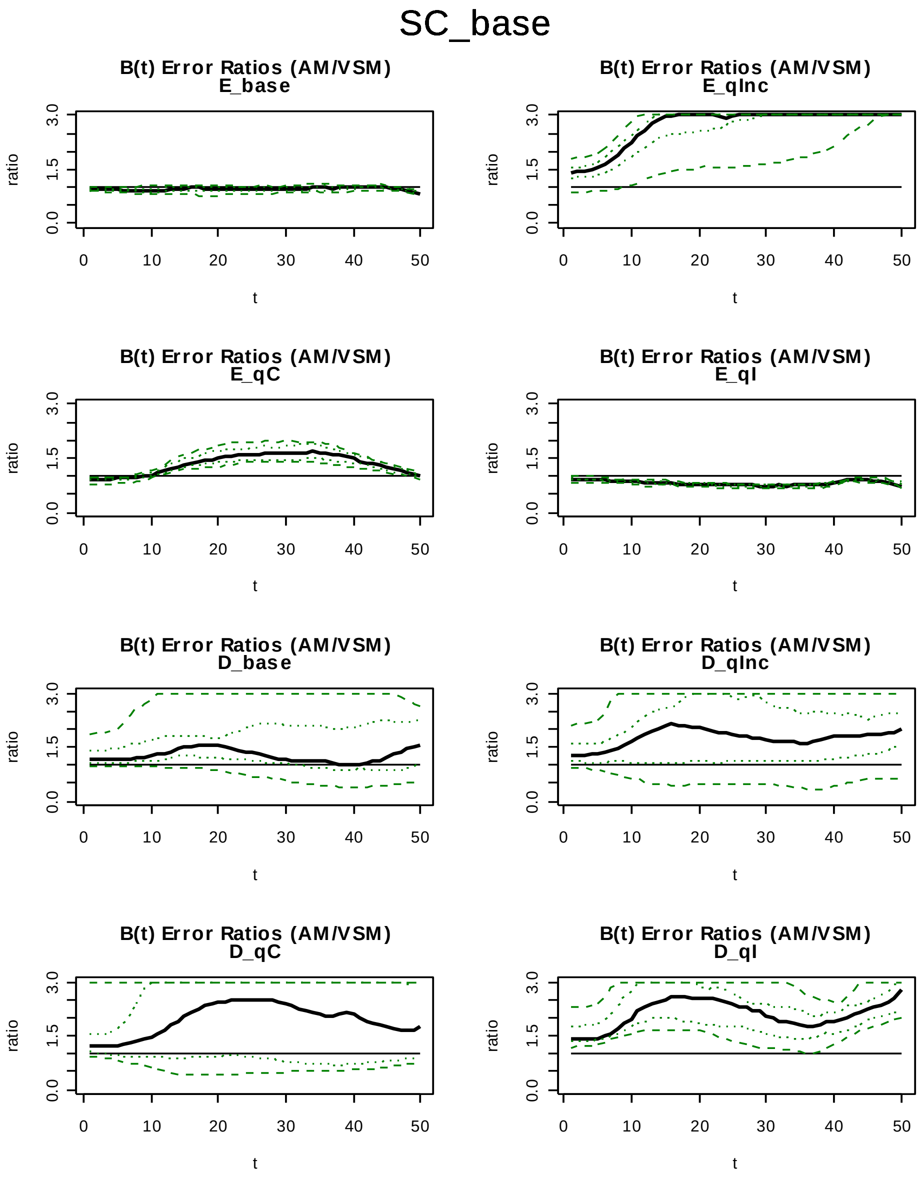

SC_base
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 20(cont.)

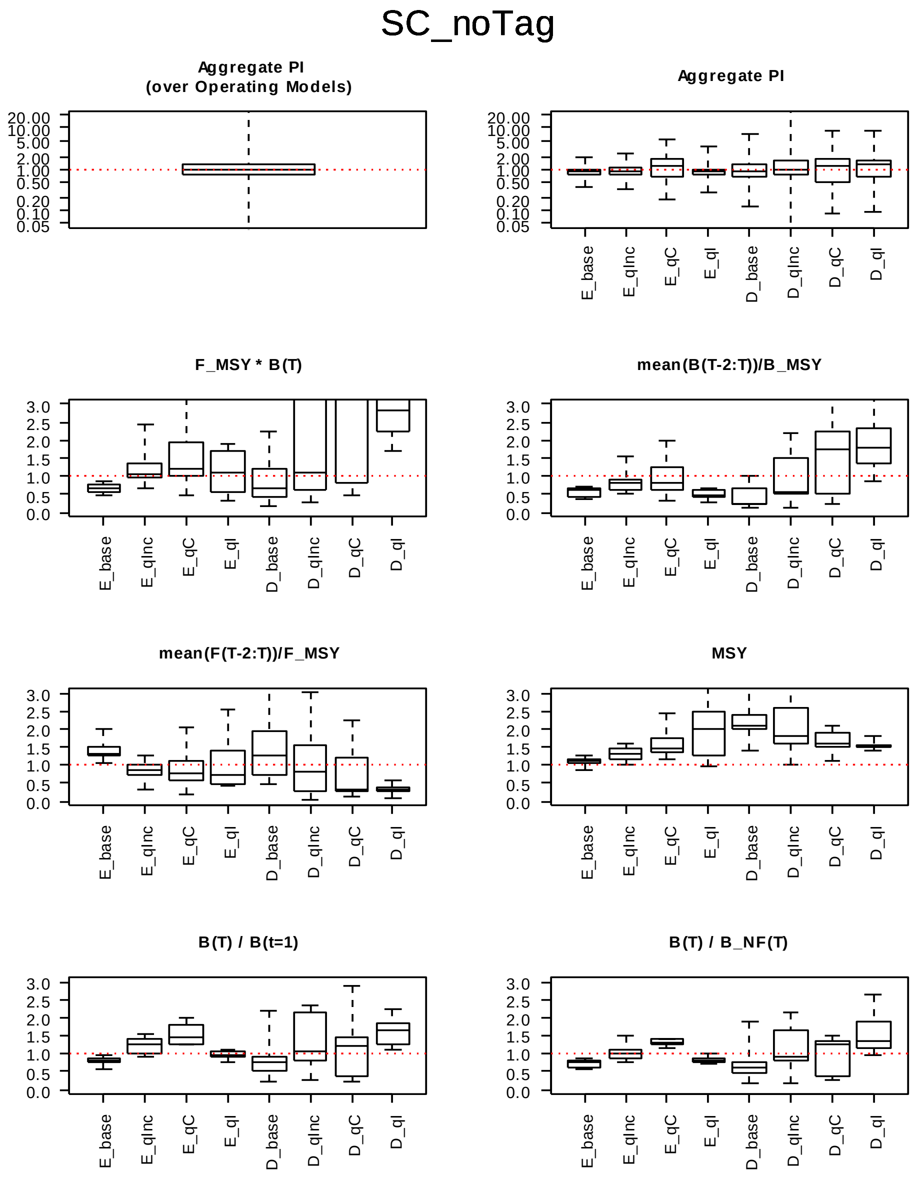

SC_noTag
Aggregate PI
(over Operating Models)
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
F_MSY * B(T)
mean(B(T-2:T))/B_MSY
mean(F(T-2:T))/F_MSY
MSY
B(T) / B(t=1)
B(T) / B_NF(T)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
E_base
E_qInc
E_qC
E_qI
D_base
D_qInc
D_qC
D_qI

Fig. 20(cont.)

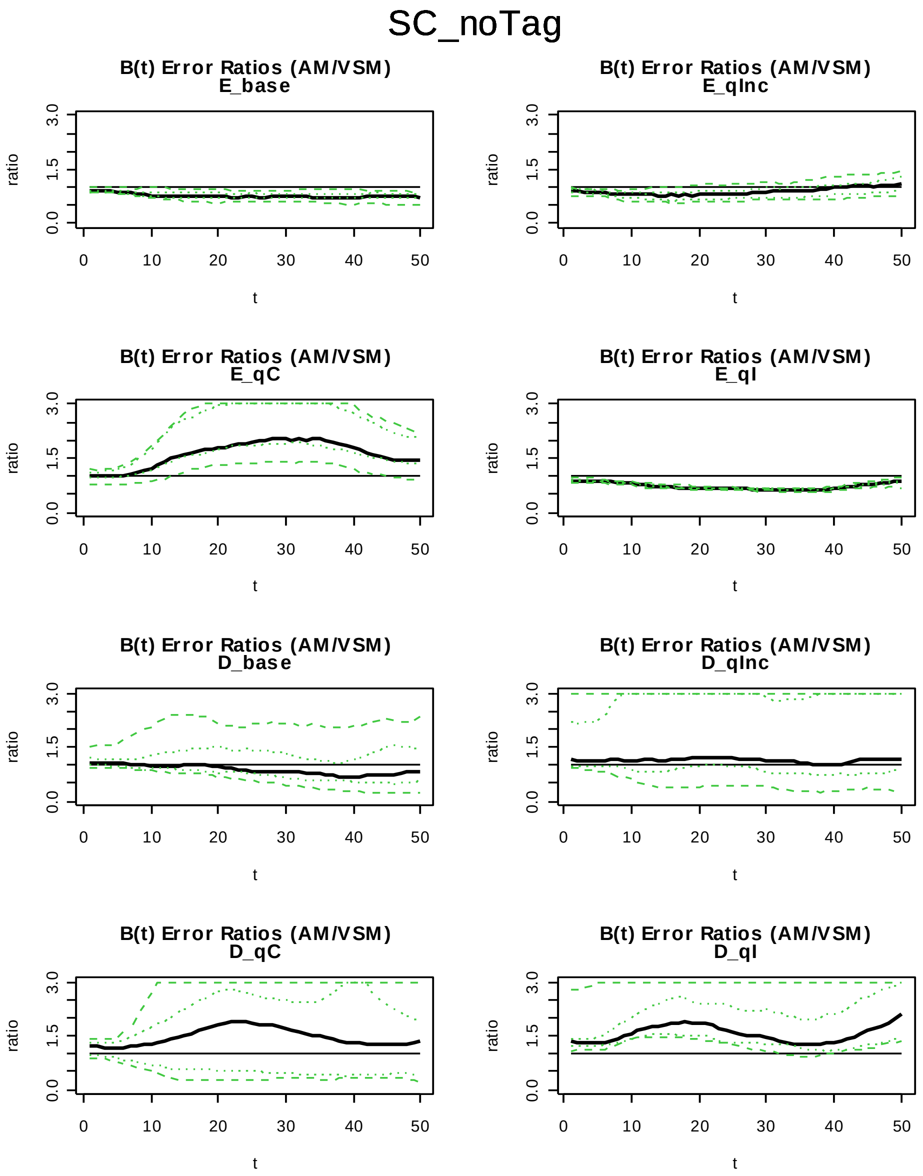

SC_noTag
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Fig. 20(cont.)

Results from a broad range of assessment models are presented for the problematic E_qInc scenario in Fig. 21. From these plots we note:

- The aggregate performance indicator suggests that the SCALIA model SC_noTag seemed to have the best overall performance, followed by the production models. The remainder of the SC_x, MF_x and BIH_2 models were substantially worse.

- Aside from the production models and SC_x, all assessment models demonstrated serious estimation errors that are consistent with an unrecognized trend in catchability, but the magnitude of the bias seems to be greater than the catchability trend. Absolute and relative biomass estimates had an increasing over-estimation bias over time, while exploitation rates had a corresponding increasing under-estimation bias over time. Recruitment estimation errors showed highly variable patterns depending on the model.

- AMs that attempted to estimate temporal variability in catchability were not very successful. On the basis of the aggregate PI, it might be argued that SC_qTS1 and MF_qTS were slightly better than the comparable models that did not estimate catchability variability; but they were clearly worse than SC_noTag and the production models. SC_qTS1 and MF_qTS do have qualitatively different errors than the other SC_x and MF_x models, as evident in the stronger biomass error trends toward the end of the time series (such that it is not at all clear that these models perform better).

It is curious that the production models and SC_noTag generally did not seem to be as badly affected by the catchability trend as the other complicated integrative models. It suggests that there is some unexpected interaction with the tagging data causing this behaviour. The magnitude of the problem and consistency across different assessment models could also indicate a bug (or specification error) in the operating model. But we would have expected that a serious inconsistency in the tag dynamics would have been evident in many other OM scenarios as well. This does suggest a potential problem for the integrated models.

This type of result does provide justification for examining different sources of data independently to identify conflicting trends. We note that Schnute and Hilborn (1993) describe an approach for admitting that each data source has a certain possibility of being incorrect in the context of a particular model. This might be an effective method for admitting more uncertainty within a single model framework, but the polymodal likelihood surfaces that result might preclude analyses of the type undertaken here, in which the MPD estimates are the main focus.

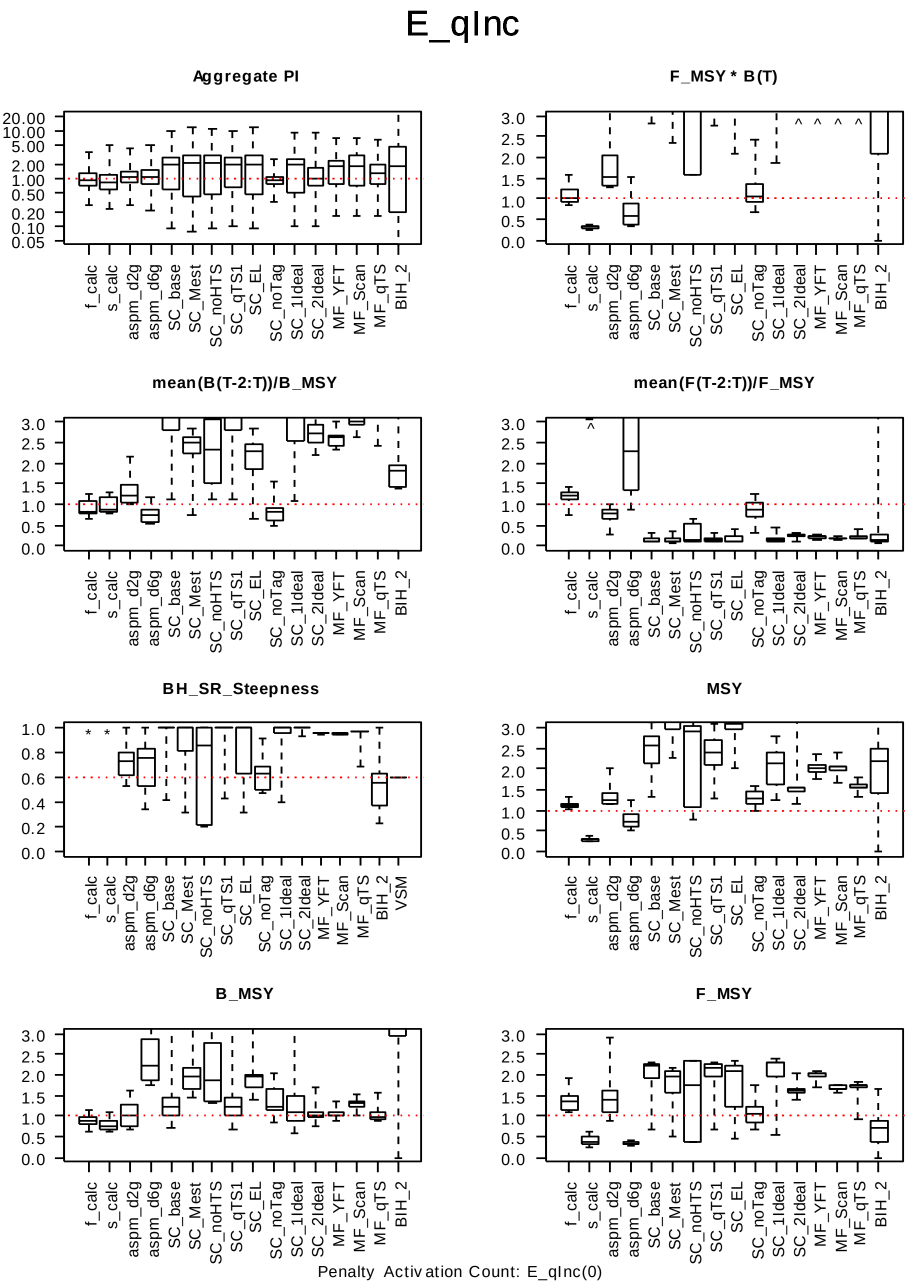


Fig. 21. **Comparison of assessment model performance for the OM scenario with an increasing trend in longline fishery catchability (E_qInc). All AMs were applied to 10 simulated data realizations. Each PI is a ratio of (AM estimated)/(OM actual). Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

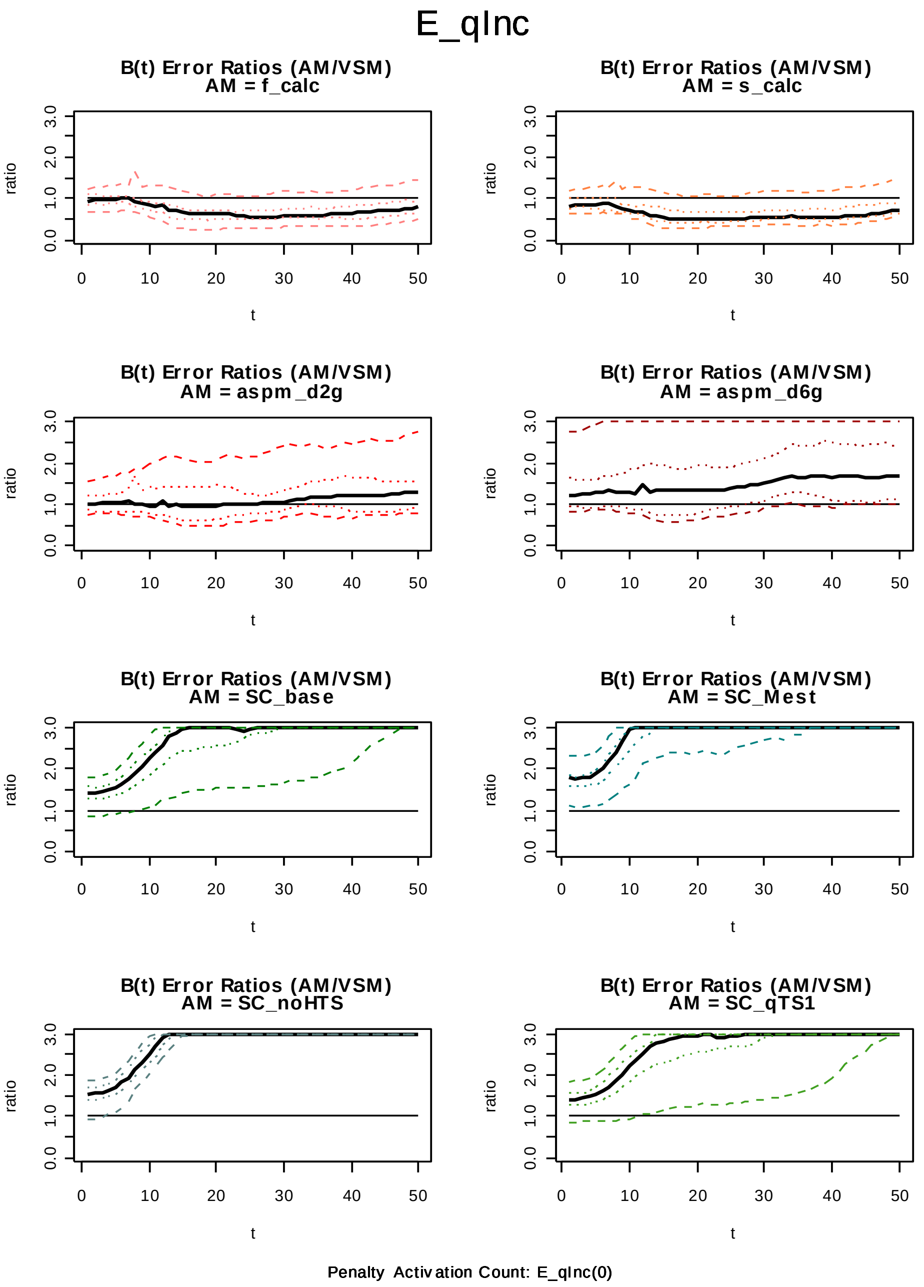

E_qInc
B(t) Error Ratios (AM/VSM)
AM = f_calc
B(t) Error Ratios (AM/VSM)
AM = s_calc
B(t) Error Ratios (AM/VSM)
AM = aspm_d2g
B(t) Error Ratios (AM/VSM)
AM = aspm_d6g
B(t) Error Ratios (AM/VSM)
AM = SC_base
B(t) Error Ratios (AM/VSM)
AM = SC_Mest
B(t) Error Ratios (AM/VSM)
AM = SC_noHTS
B(t) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)

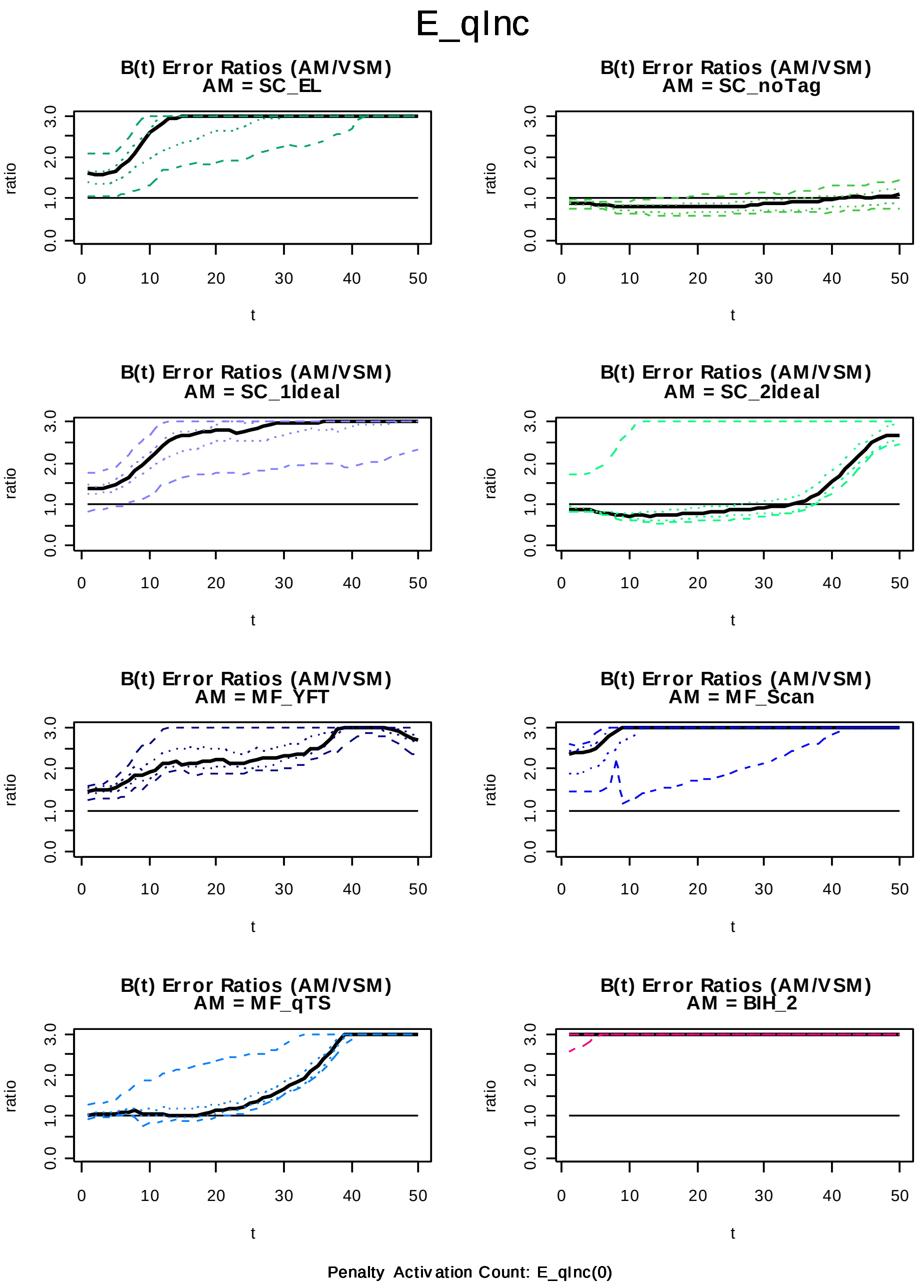

E_qInc
B(t) Error Ratios (AM/VSM)
AM = SC_EL
B(t) Error Ratios (AM/VSM)
AM = SC_noTag
B(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
B(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
B(t) Error Ratios (AM/VSM)
AM = MF_YFT
B(t) Error Ratios (AM/VSM)
AM = MF_Scan
B(t) Error Ratios (AM/VSM)
AM = MF_qTS
B(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)

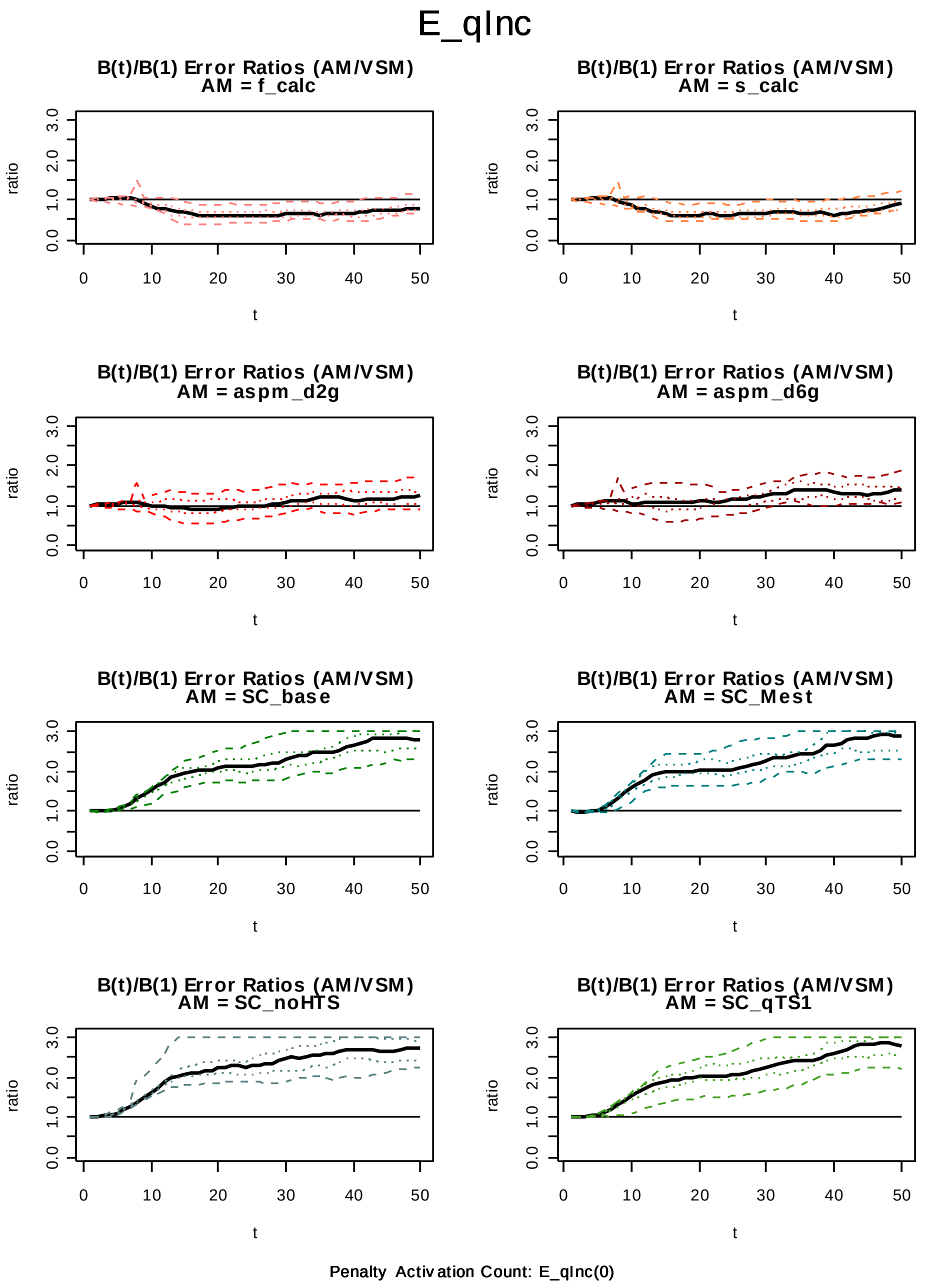
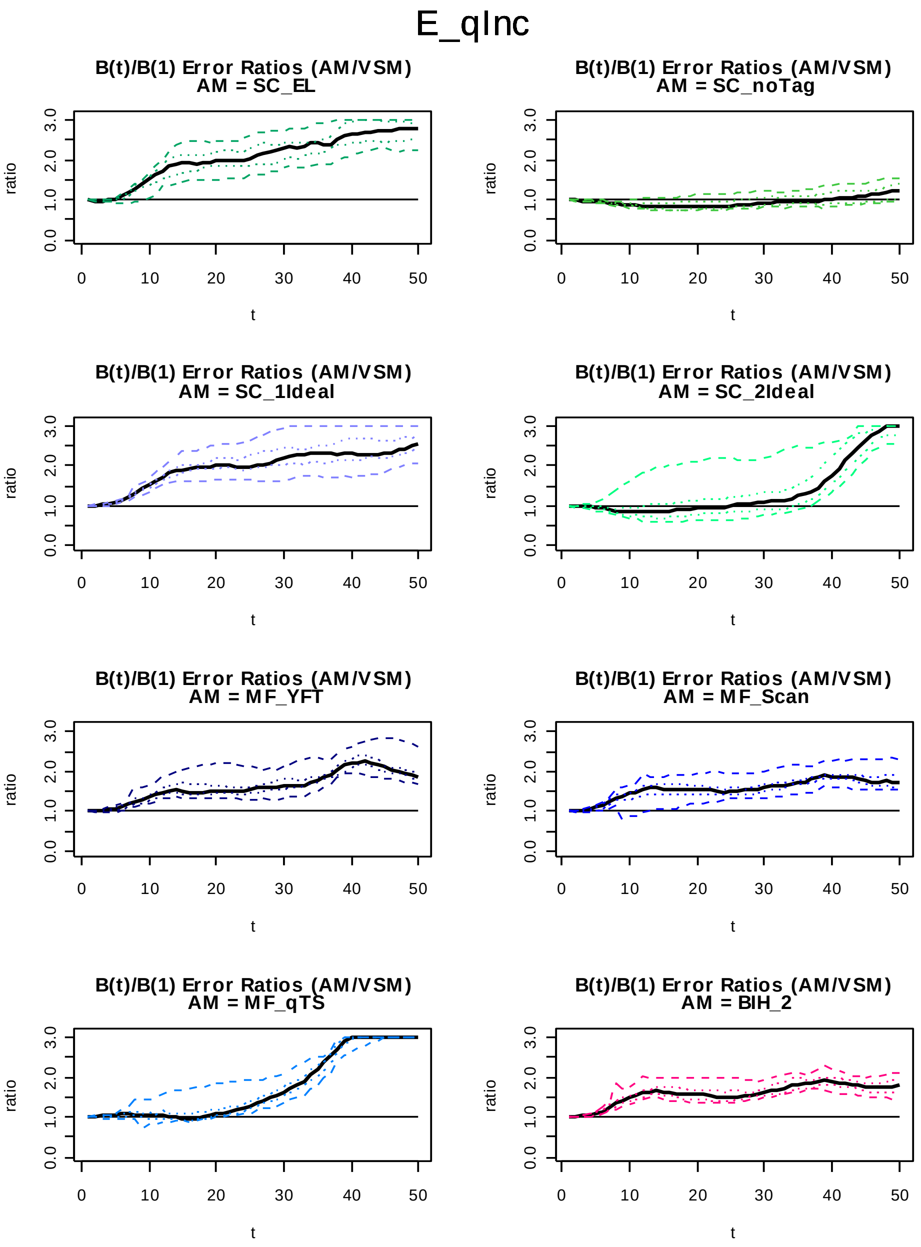

E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
AM = f_calc
B(t)/B(1) Error Ratios (AM/VSM)
AM = s_calc
B(t)/B(1) Error Ratios (AM/VSM)
AM = aspm_d2g
B(t)/B(1) Error Ratios (AM/VSM)
AM = aspm_d6g
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_Mest
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_noHTS
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)



Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)

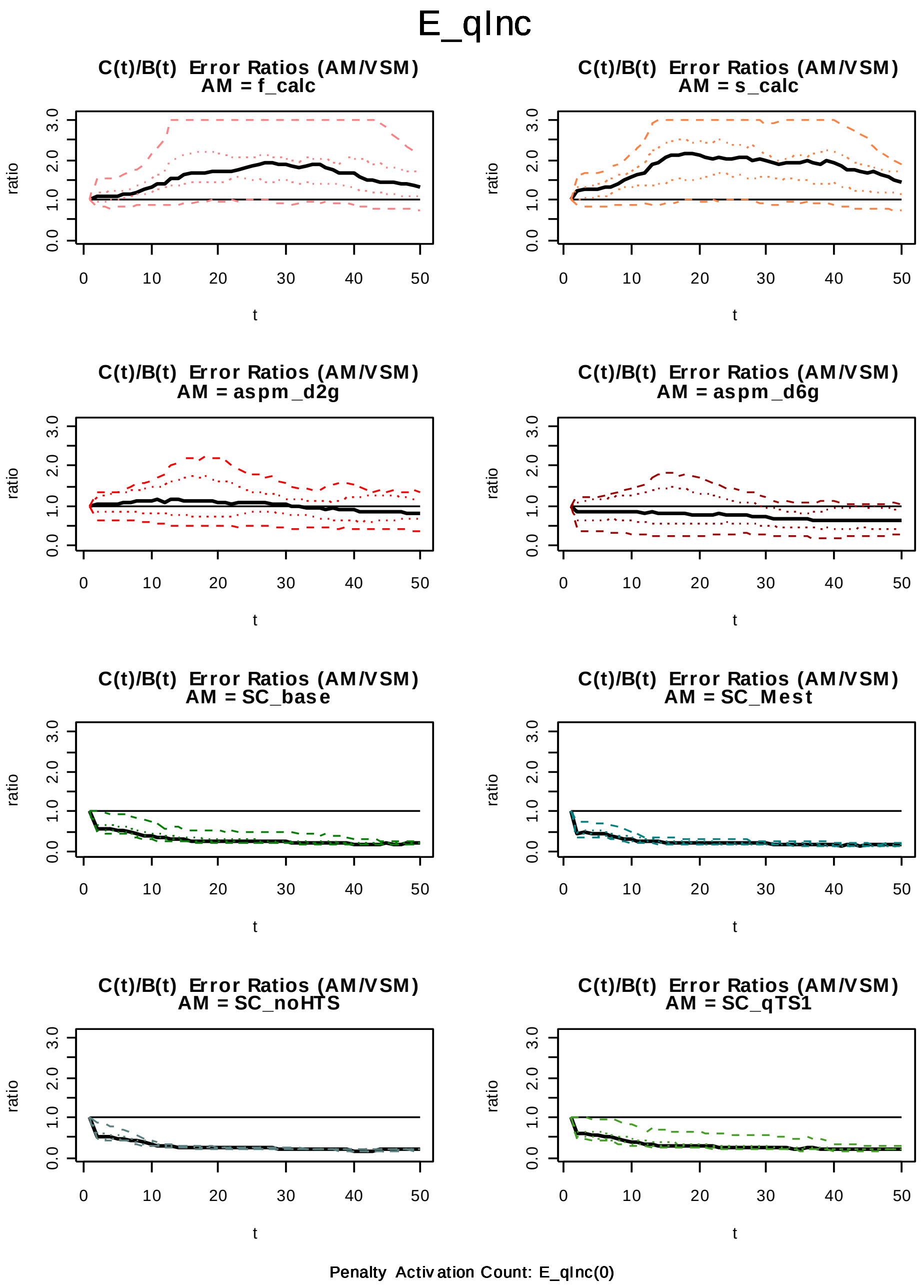

E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
AM = f_calc
C(t)/B(t) Error Ratios (AM/VSM)
AM = s_calc
C(t)/B(t) Error Ratios (AM/VSM)
AM = aspm_d2g
C(t)/B(t) Error Ratios (AM/VSM)
AM = aspm_d6g
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_base
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_Mest
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_noHTS
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)

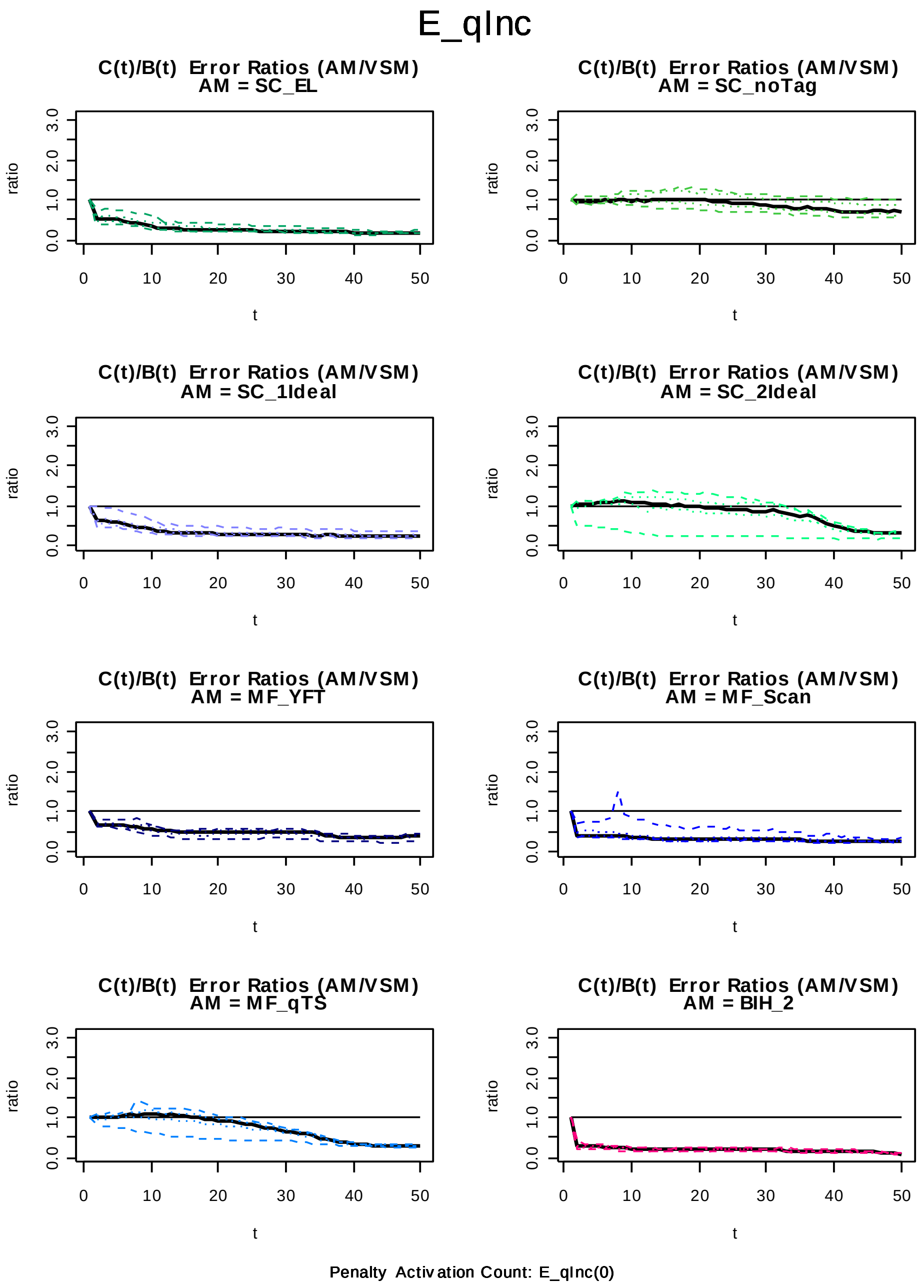

E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
AM = SC_EL
AM = SC_noTag
AM = SC_1Ideal
AM = SC_2Ideal
AM = MF_YFT
AM = MF_Scan
AM = MF_qTS
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)

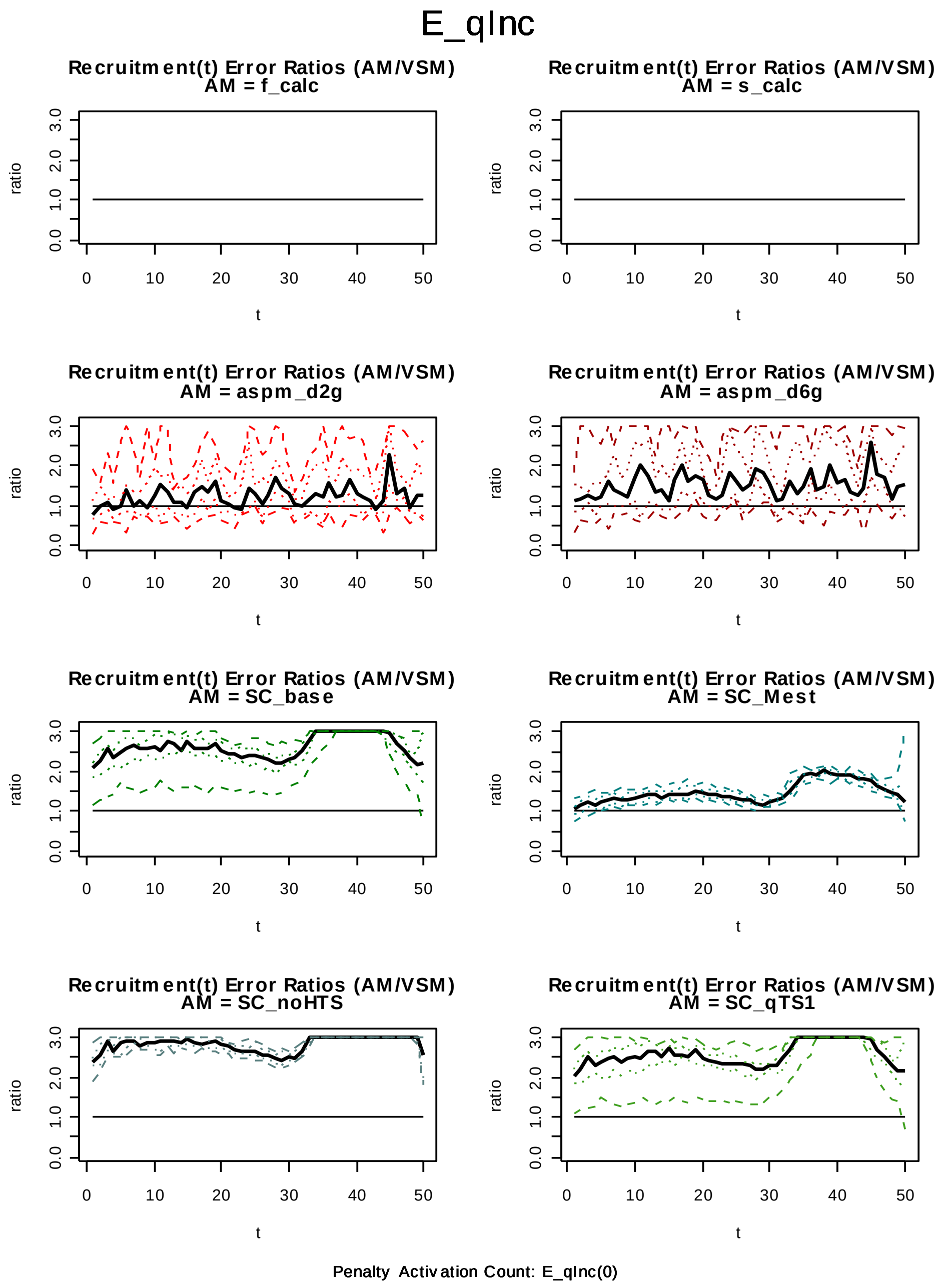

E_qInc
Recruitment(t) Error Ratios (AM/VSM)
AM = f_calc
Recruitment(t) Error Ratios (AM/VSM)
AM = s_calc
Recruitment(t) Error Ratios (AM/VSM)
AM = aspm_d2g
Recruitment(t) Error Ratios (AM/VSM)
AM = aspm_d6g
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_base
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_Mest
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_noHTS
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_qTS1
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)

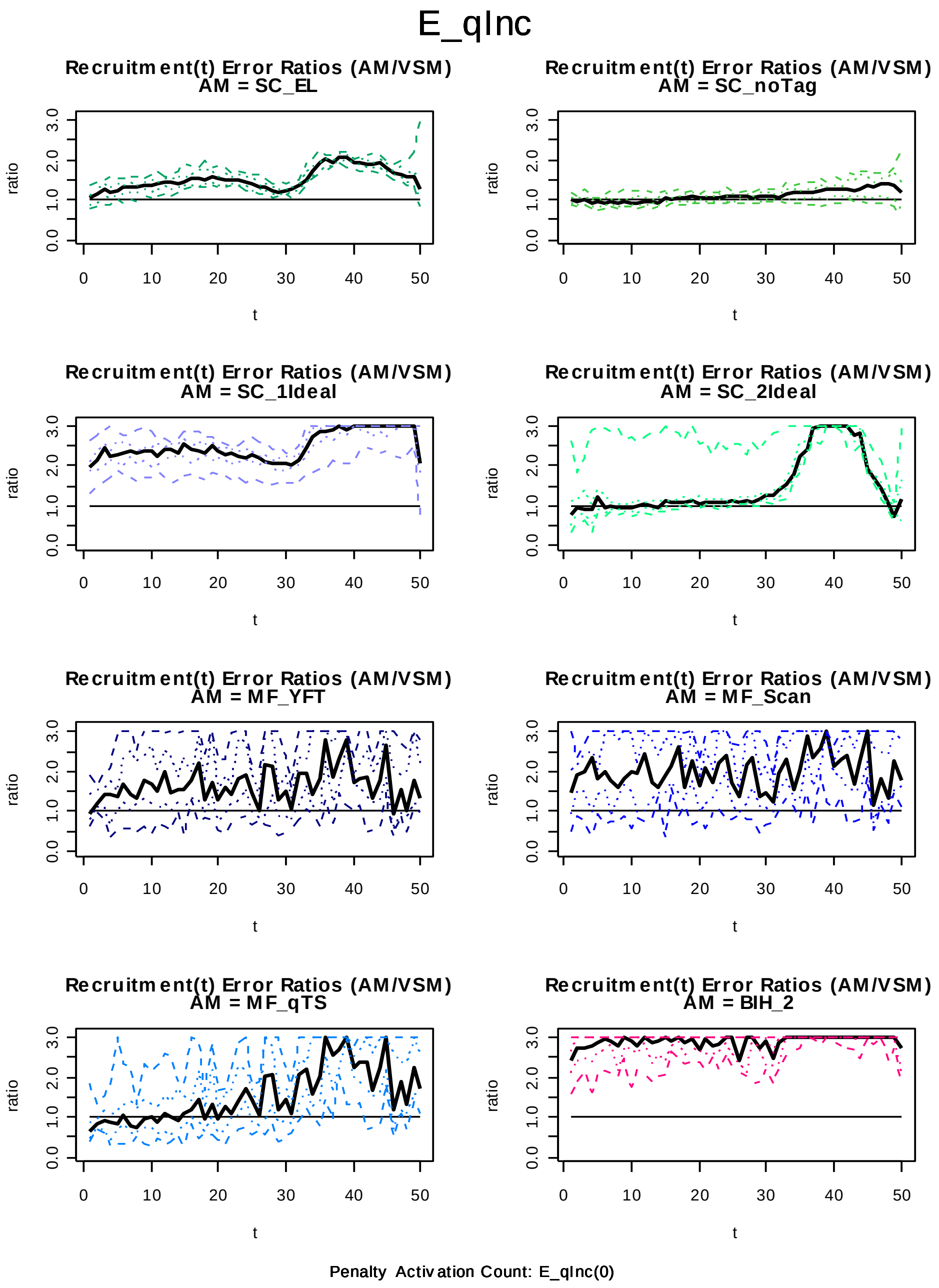

E_qInc
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_EL
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_noTag
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_1Ideal
Recruitment(t) Error Ratios (AM/VSM)
AM = SC_2Ideal
Recruitment(t) Error Ratios (AM/VSM)
AM = MF_YFT
Recruitment(t) Error Ratios (AM/VSM)
AM = MF_Scan
Recruitment(t) Error Ratios (AM/VSM)
AM = MF_qTS
Recruitment(t) Error Ratios (AM/VSM)
AM = BIH_2
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_qInc(0)

Fig. 21 (cont.)

The SCALIA models that attempted to estimate temporal variability in catchability (SC_qTS5, SC_qTS1 and SC_2ideal) were not very successful (Fig. 22):

- SC_qTS1 was so constrained (random walk CV = 0.01) that it was essentially the same as SC_base.

- SC_qTS5 (random walk CV = 0.05) showed slightly different behavior, but it actually seemed to be worse than SC_qTS1 and SC_base in terms of the relative biomass estimates.

- SC_2ideal differed from the other models in several ways, and appeared to perform slightly better than the others on the basis of the aggregate PI, but there were large estimation errors evident in the individual PIs.

Catchability trend estimation was not very successful in the other scenarios either (not shown). If this is true in general, it indicates that these stock assessment models are highly dependent on the quality of the relative abundance index, and the auxiliary data has limited capacity to improve the estimated abundance trends. It also follows that using the assessment model likelihood to choose among different effort standardization methods is probably not very useful for discriminating which effort series is the best.

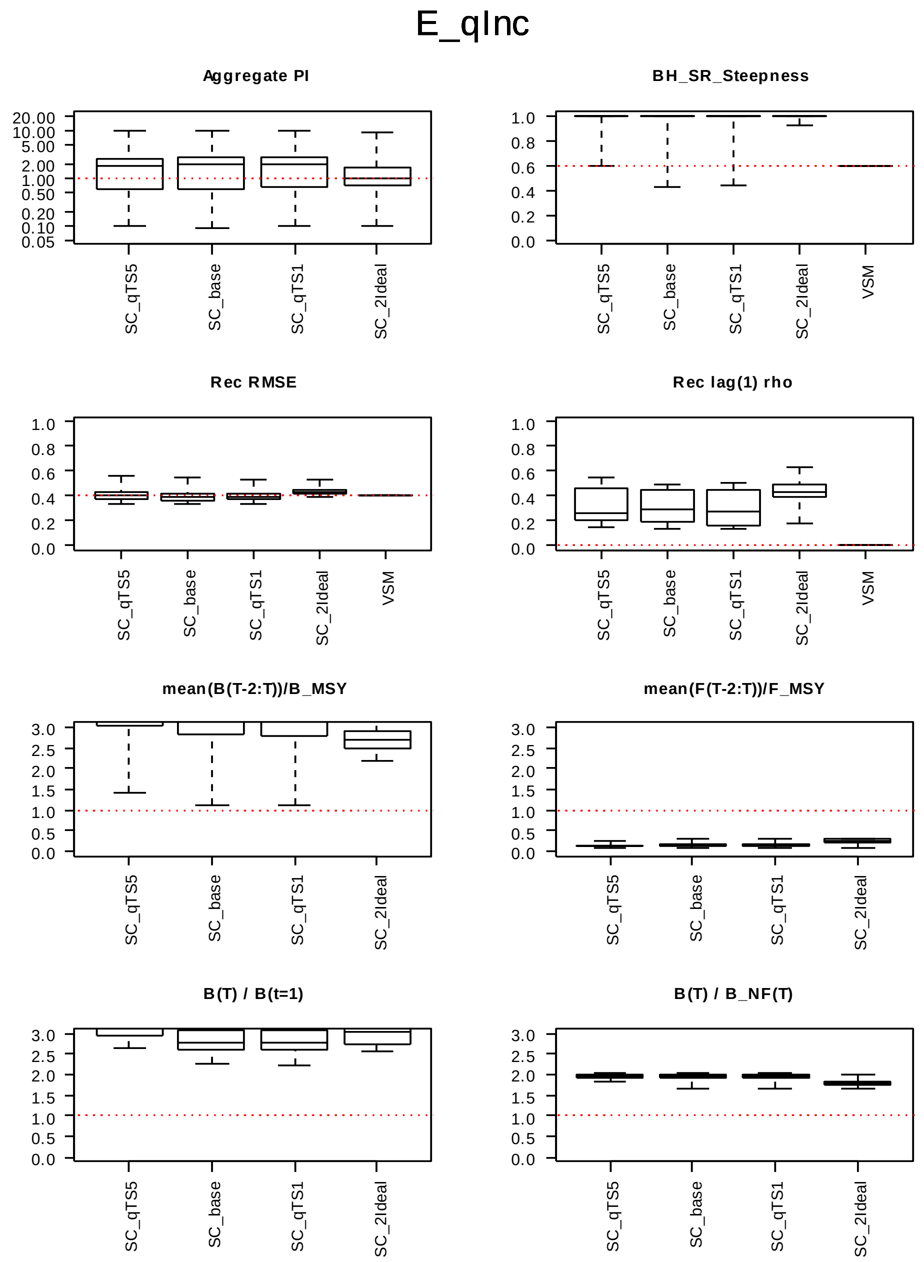


Fig. 22. **Comparison of AMs that attempt to estimate temporal variability in longline catchability with the baseline SCALIA model SC_base, when challenged by simulated data from an OM with a catchability trend. AMs were applied to 10 simulated data realizations. Each PI is a ratio of (AM estimated)/(OM actual), except for stock-recruitment related quantities that illustrate actual values. Boxplots describe the distribution of individual PIs; time series are represented by line plots of quantiles (median, 20th and 80th percentiles and range). OMs are defined in Table 1, AMs in Table 2 and PIs in Table 8.**

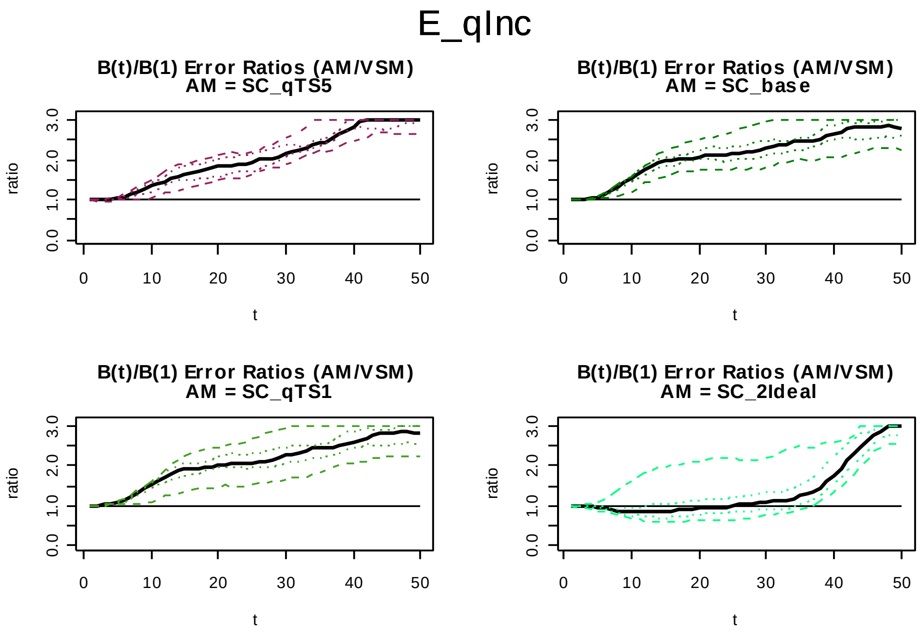

E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_qTS5
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_base
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_qTS1
B(t)/B(1) Error Ratios (AM/VSM)
AM = SC_2Ideal
ratio
0.0
1.0
2.0
3.0
t
0
10
20
30
40
50

Fig. 22. (cont.)

5.9 OBJECTIVE VII - THE SCTB METHODS WORKING GROUP AND ASSUMPTIONS ABOUT FISHERY SPATIAL STRUCTURE IN ASSESSMENT MODELS

The application of assessment models to the simulated YFT data has raised some interesting questions about the quality of inferences from different assessment models and how to evaluate them with simulation methods. At the time of writing this report, we can only make preliminary comments about model performance at a global population level. Hopefully the SCTB-MWG analysis will eventually provide more insight about the performance of MULTIFAN-CL at the level of the spatial sub-structure.

The 2002 MWG project resulted in the application of several assessment models from independent participants from several countries. Different models were applied with different levels of thoroughness, and the results were not all summarized in a form that was directly comparable with the operating model. Available results are summarized in MWG (2002). The most detailed results are available for MULTIFAN-CL (Labelle 2002; the YFT simulator is also described in this document). As part of the SESAME project, we applied SCALIA (Kolody 2002) and production models (Ricard and Kolody 2002). One clear result from this study was a recognition of the futility of attempting to estimate a catchability trend for a production model that uses only annual catch and CPUE as a relative abundance index. Results from A-SCALA (e.g. Maunder and Watters 2003), ADAPT (Bigelow 2002) and an independent application of MULTIFAN-CL were also presented. Discussions from the 2002 meeting were instrumental in defining the more comprehensive 2003 study, and are not discussed further here.

The 2003 YFT simulation study results were examined in the most detail for MULTIFAN-CL (Labelle 2003; updated YFT simulator description also provided therein). The study left the impression that MULTIFAN-CL generally made reasonable global inferences. The simulated data for the most complicated scenarios (7 fisheries X 7 regions and 16 F X 7 R) were not distributed in time for most other models to be applied. A-SCALA was applied to the first three scenarios (1 F X 1 R, 2 F X 1 R, 2 F X 2 R). As part of SESAME, we applied production models and SCALIA to all data sets (Kolody and Ricard 2003), but not all were completed in time for submission to the MWG in July 2003. There was limited comparison of assessment models at the 2003 meeting, and this was recognized as an objective for the MWG 2003-4 workplan.

Subsequent to the MWG 2003 meeting, we obtained a limited description of the YFT simulator biomass dynamics and prepared a number of preliminary summary graphics for the age-aggregated production model and SCALIA applications (plus a repetition of some of the MULTIFAN-CL (mfcl) results from Labelle 2003 for comparison). These are included in Fig. 23 - Fig 27. A number of points are suggested from this preliminary synthesis (note AM model definitions in Table 3):

- The performance of each assessment model differed considerably depending on the OM scenario. It is not clear which factors are responsible for the performance differences. From the summary details that we have seen, it seems as though variability among scenarios due to fishery selectivity/spatial patterns and spatial data aggregation units is probably greater than the variability due to biology and total global exploitation patterns.

- Absolute biomass estimates were often very poor, while relative estimates (depletion) were generally much better.

- MSY-related estimates were often poor for SCALIA models and generally very sensitive to the model specification. The AAPMs were generally better than SCALIA. We do not have the MULTIFAN-CL estimates for comparison, but they appear to be reasonable in Labelle (2003).

- ASPM results (including those submitted to the MWG) were found to be flawed for the majority of the MWG scenarios (detailed in the ASPM implementation section) and were withdrawn from further consideration.

- From scenario 1 (1F X 1R) we would probably conclude that the Fox model had the best performance (Fig. 23). The Schaefer model had high biomass estimate variance for both absolute and relative time series (perhaps unduly influenced by a few outliers). The SCALIA model SM_1921 had an absolute biomass bias, but relative estimates were good. MULTIFAN-CL had a substantial bias in both relative and absolute biomass over most of the final 20% of the time series.

- From scenario 2 (2F X 1R) we would probably again conclude that the Fox model was the best (Fig. 24). The Schaefer model again had high biomass and depletion variance, possibly due to outliers. SM_2930 had excellent biomass estimates (comparable with Fox); SM_2918 had a large fairly consistent bias in absolute biomass, and SM_2914 had bad temporal trends in bias in both the absolute and relative biomass estimates. The MULTIFAN-CL biomass estimates seemed to be more biased than the production models and SM_2930, but not as bad as SM_2914.

- Most of the assessment models seemed to have similar problems estimating relative biomass in Scenario 3 (4F X 2R), with an under-estimation bias towards the end of the time series (Fig. 25). MULTIFAN-CL biomass estimates were probably slightly better than the other models. The SCALIA and ASPM MSY estimates were all reasonable, but the F_MSY estimates were not as good.

- Fox_agg seemed to have the best biomass estimation performance in Scenario 4 (7F X 7R) (Fig. 26). The SCALIA models had very large biases in absolute biomass, and considerable biases in relative biomass. MULTIFAN-CL also had biases in relative biomass, though not as large, and with a different temporal pattern than the SCALIA models. The production models had much better MSY estimates than the SCALIA models, while they all had problems with F_MSY.

- All AMs had substantial absolute biomass estimation problems in Scenario 5 (16F X 7R) (Fig. 27). Fox_agg and Schaefer_agg probably had the best relative biomass estimation performance; SM_5950 and MULTIFAN-CL were reasonable and SM_5915 was particularly bad. The production model MSY estimates were better than the SCALIA estimates; it is not clear which F_MSY estimates were better.

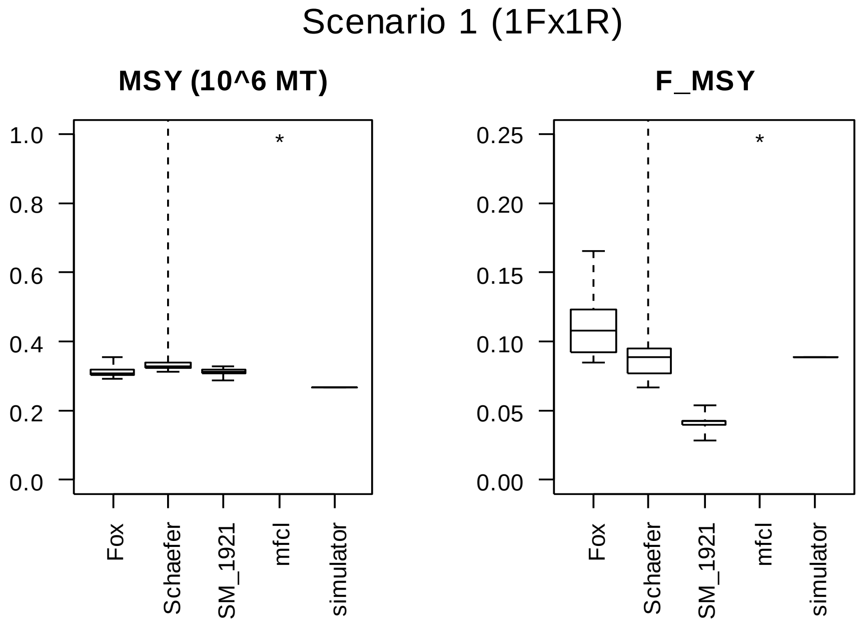


Fig. 23a. **Comparison of assessment model estimates with SPC-OFP YFT simulator values. Assessment results are a boxplot describing the distribution based on 40 simulated data realizations. Asterisk indicates that the results were not available for this model (mfcl). The simulator value is the median from the 40 realizations.**

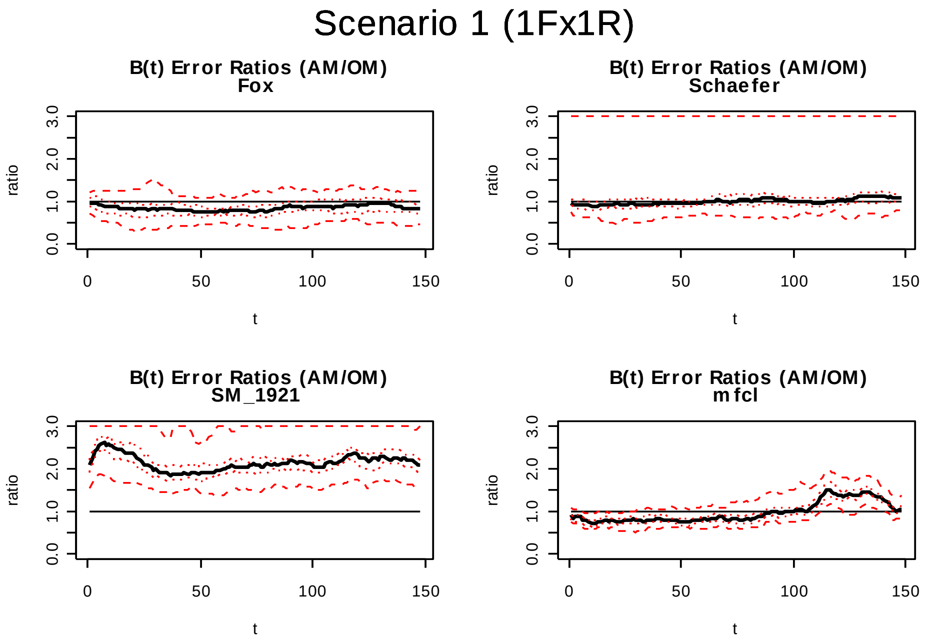
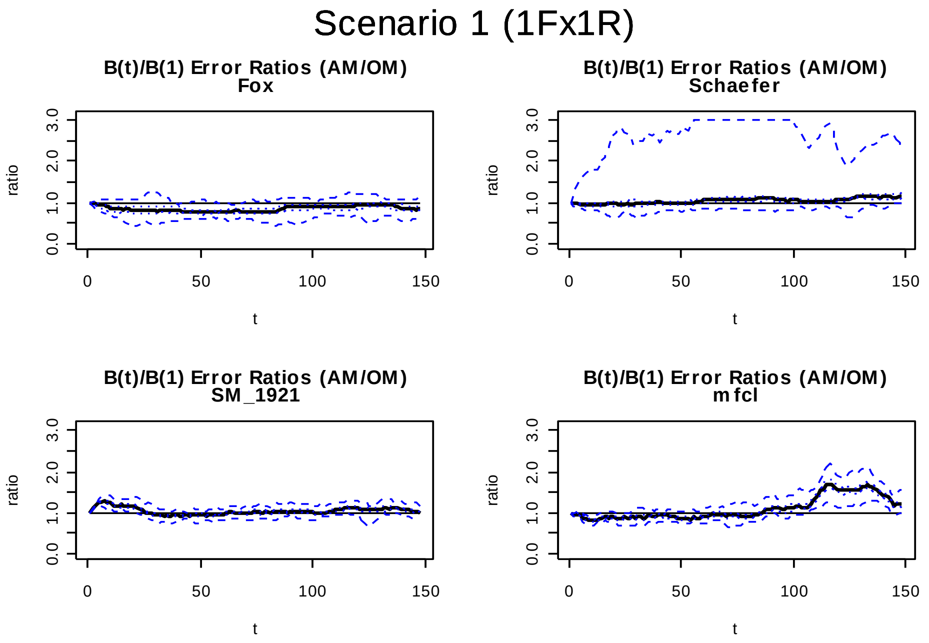

Scenario 1 (1Fx1R)
B(t) Error Ratios (AM/OM)
Fox
B(t) Error Ratios (AM/OM)
Schaefer
B(t) Error Ratios (AM/OM)
SM_1921
B(t) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
0
50
100
150
t

Fig. 23b.



Scenario 1 (1Fx1R)
B(t)/B(1) Error Ratios (AM/OM)
Fox
B(t)/B(1) Error Ratios (AM/OM)
Schaefer
B(t)/B(1) Error Ratios (AM/OM)
SM_1921
B(t)/B(1) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
0
50
100
150
t

Fig. 23c.

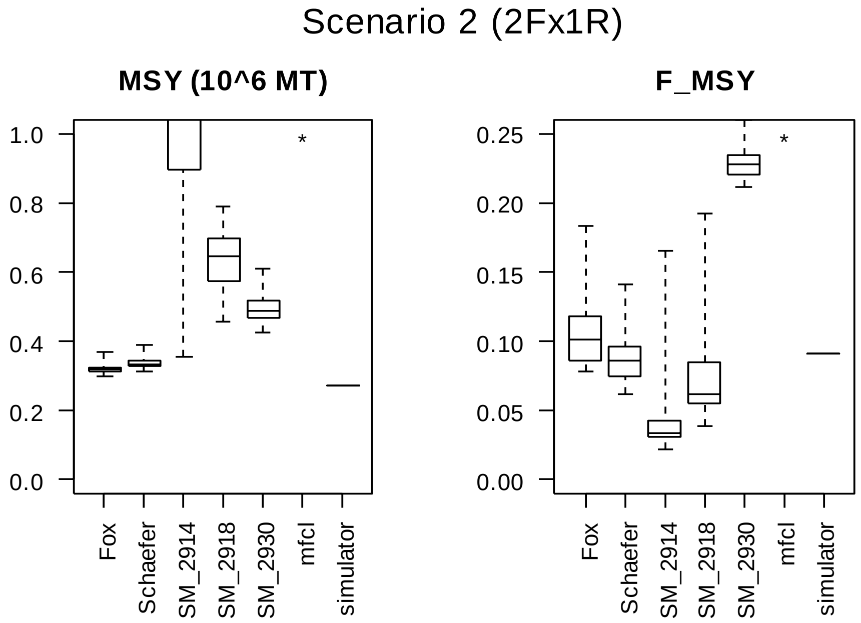


Fig. 24a. **Comparison of assessment model estimates with SPC-OFP YFT simulator values. Assessment results are a boxplot describing the distribution based on 40 simulated data realizations. Asterisk indicates that the results were not available for this model (mfcl). The simulator value is the median from the 40 realizations.**

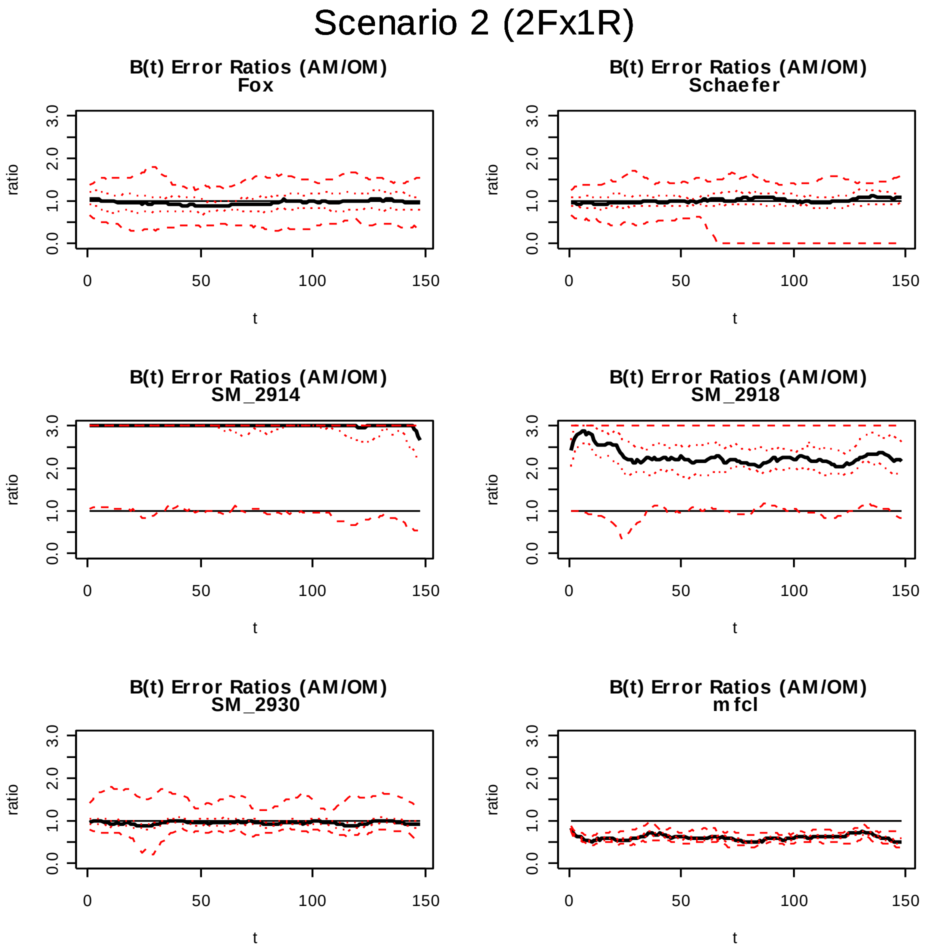
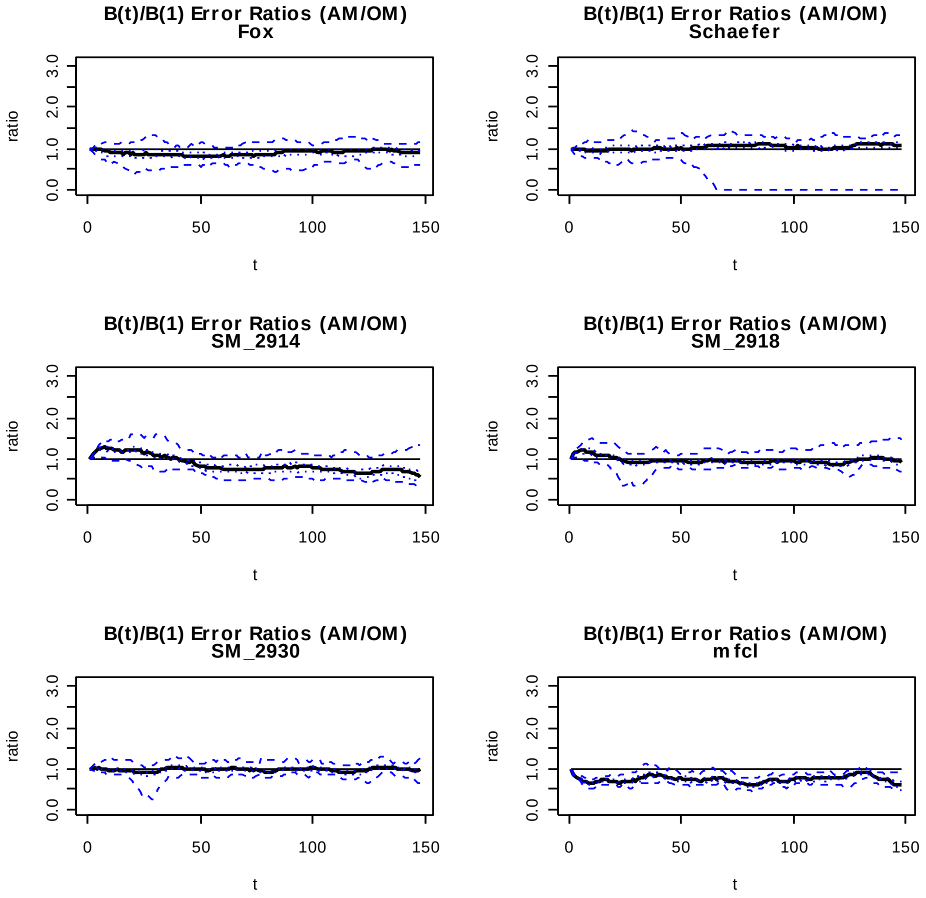

Scenario 2 (2Fx1R)
B(t) Error Ratios (AM/OM)
Fox
B(t) Error Ratios (AM/OM)
Schaefer
B(t) Error Ratios (AM/OM)
SM_2914
B(t) Error Ratios (AM/OM)
SM_2918
B(t) Error Ratios (AM/OM)
SM_2930
B(t) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
0
50
100
150
t

Fig. 24b.

Scenario 2 (2Fx1R)



B(t)/B(1) Error Ratios (AM/OM)
Fox

B(t)/B(1) Error Ratios (AM/OM)
Schaefer

B(t)/B(1) Error Ratios (AM/OM)
SM_2914

B(t)/B(1) Error Ratios (AM/OM)
SM_2918

B(t)/B(1) Error Ratios (AM/OM)
SM_2930

B(t)/B(1) Error Ratios (AM/OM)
mfcl

Fig. 24c.

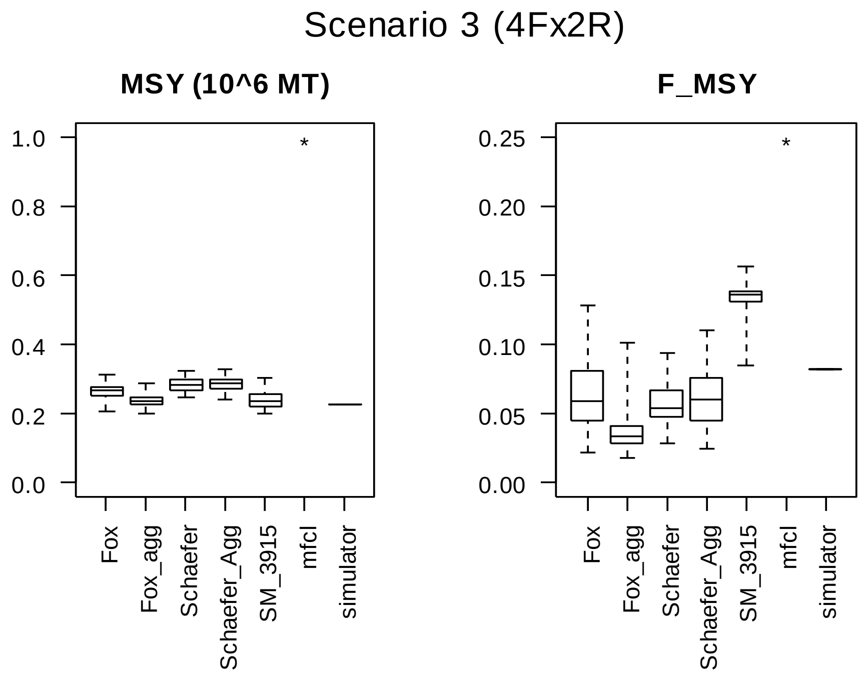
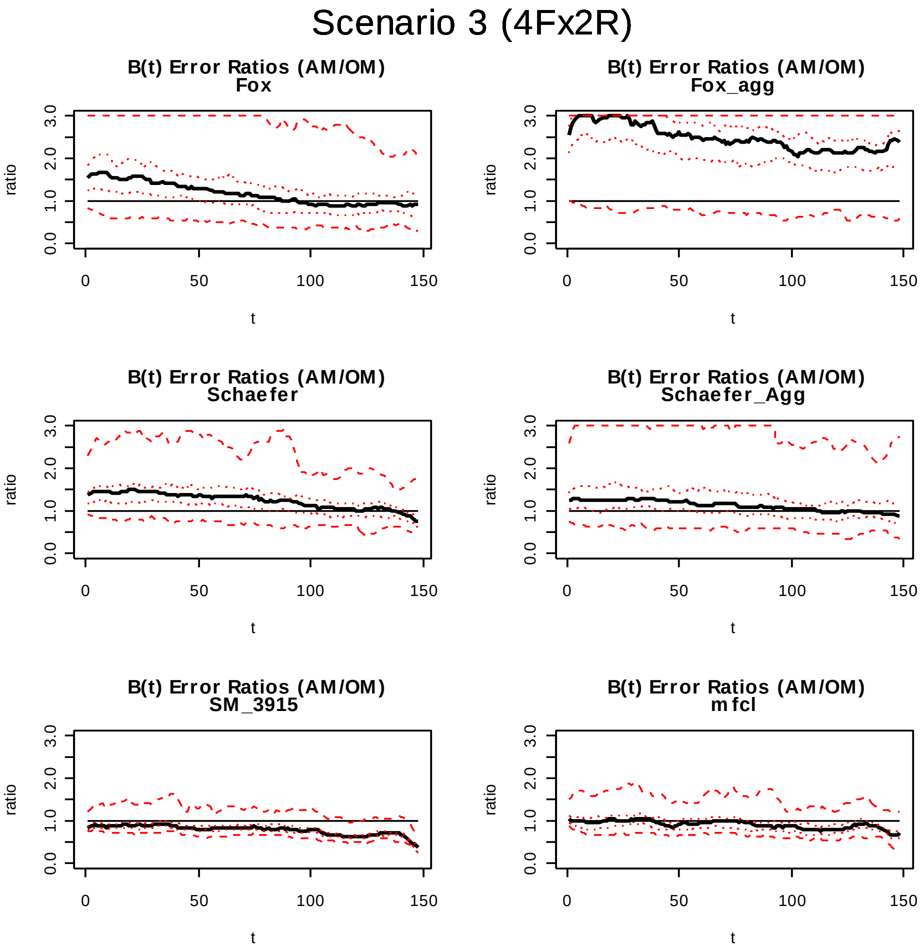


Fig. 25a. **Comparison of assessment model estimates with SPC-OFP YFT simulator values. Assessment results are a boxplot describing the distribution based on 40 simulated data realizations. Asterisk indicates that the results were not available for this model (mfcl). The simulator value is the median from the 40 realizations.**



Scenario 3 (4Fx2R)

B(t) Error Ratios (AM/OM) Fox

B(t) Error Ratios (AM/OM) Fox_agg

B(t) Error Ratios (AM/OM) Schaefer

B(t) Error Ratios (AM/OM) Schaefer_Agg

B(t) Error Ratios (AM/OM) SM_3915

B(t) Error Ratios (AM/OM) mfcl

Fig. 25b.

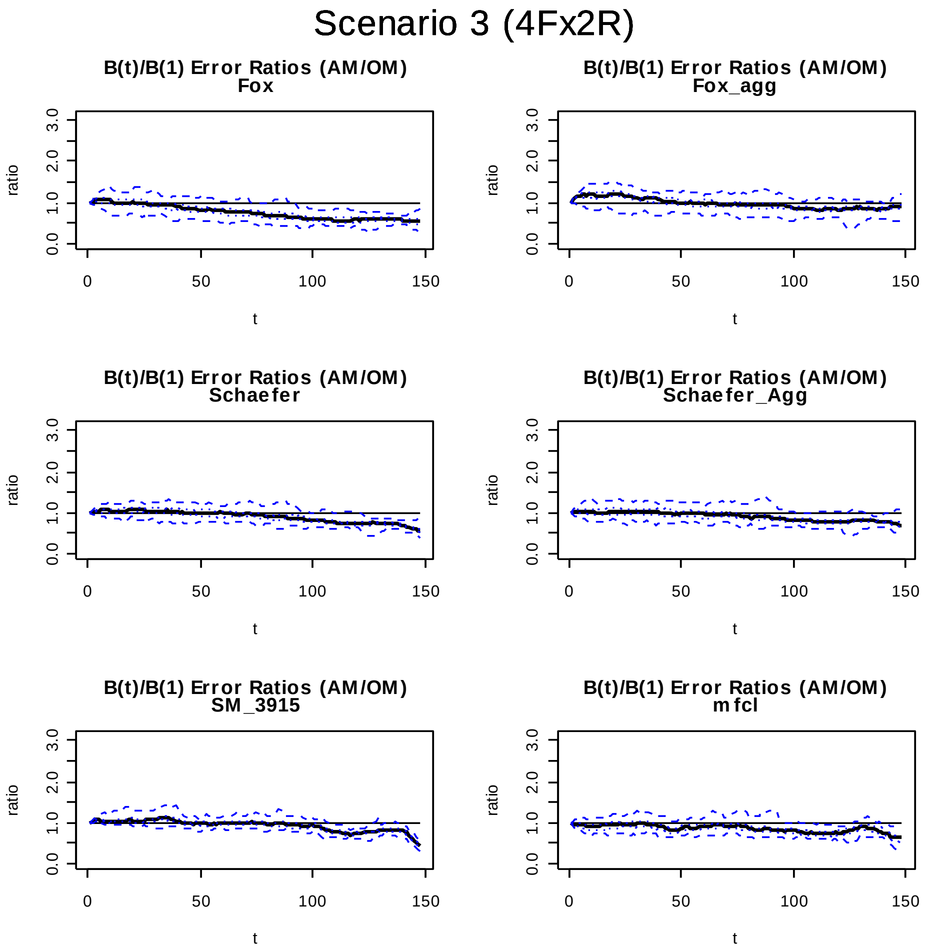

Scenario 3 (4Fx2R)
B(t)/B(1) Error Ratios (AM/OM)
Fox
B(t)/B(1) Error Ratios (AM/OM)
Fox_agg
B(t)/B(1) Error Ratios (AM/OM)
Schaefer
B(t)/B(1) Error Ratios (AM/OM)
Schaefer_Agg
B(t)/B(1) Error Ratios (AM/OM)
SM_3915
B(t)/B(1) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
t
0
50
100
150

Fig. 25c.

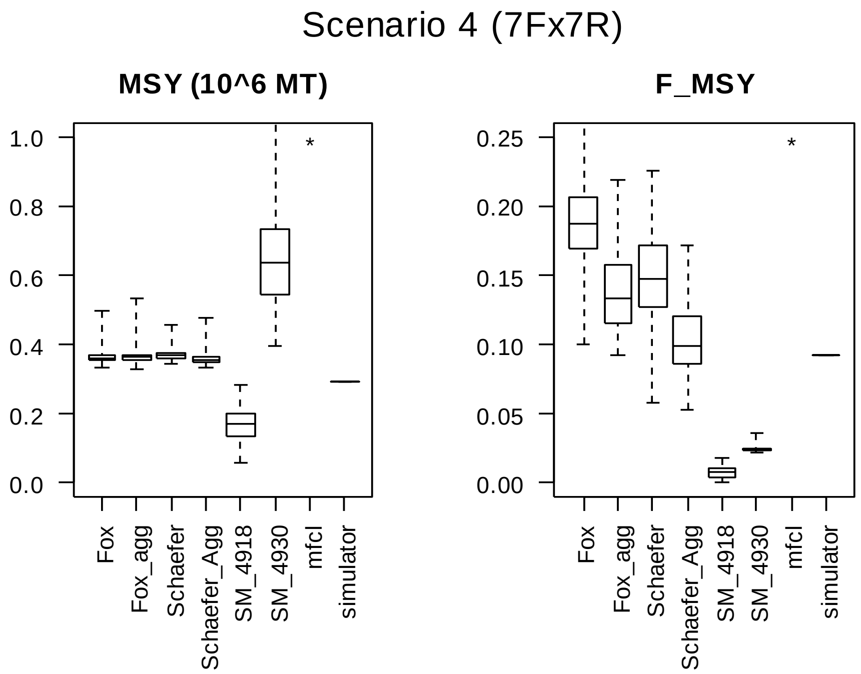


Fig. 26a. **Comparison of assessment model estimates with SPC-OFP YFT simulator values. Assessment results are a boxplot describing the distribution based on 40 simulated data realizations. Asterisk indicates that the results were not available for this model (mfcl). The simulator value is the median from the 40 realizations.**

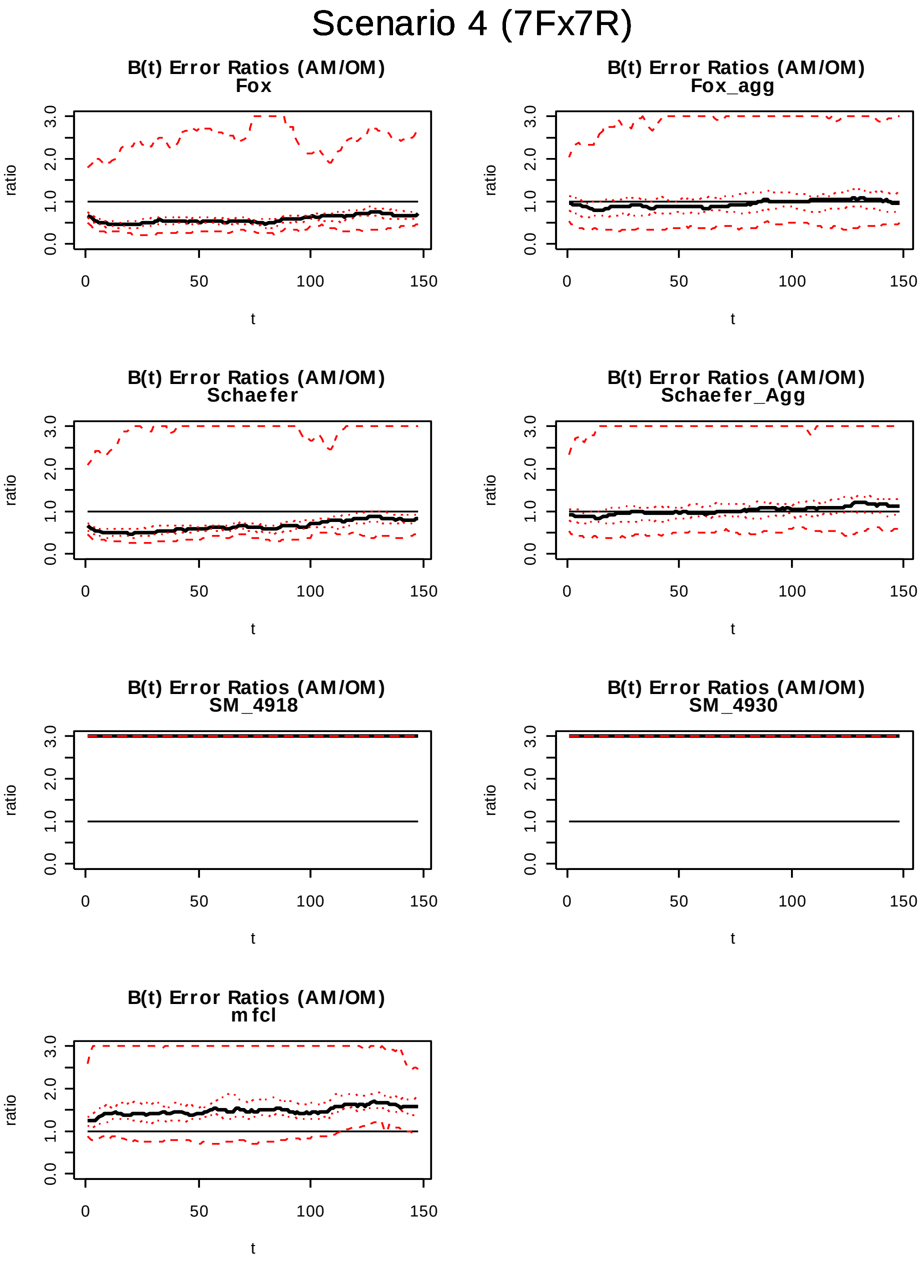

Scenario 4 (7Fx7R)
B(t) Error Ratios (AM/OM)
Fox
B(t) Error Ratios (AM/OM)
Fox_agg
B(t) Error Ratios (AM/OM)
Schaefer
B(t) Error Ratios (AM/OM)
Schaefer_Agg
B(t) Error Ratios (AM/OM)
SM_4918
B(t) Error Ratios (AM/OM)
SM_4930
B(t) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
t
0
50
100
150

Fig. 26b.

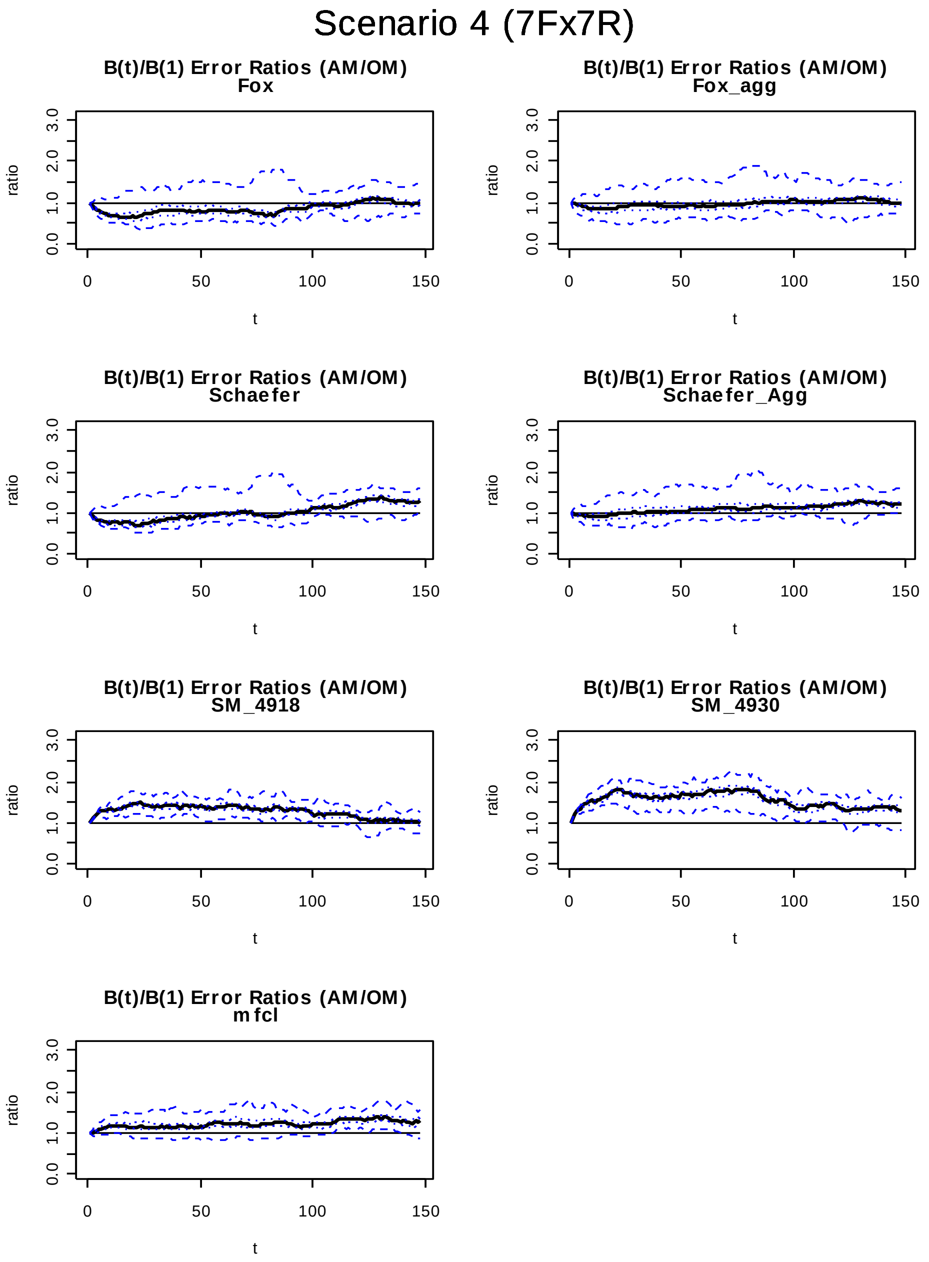

Scenario 4 (7Fx7R)
B(t)/B(1) Error Ratios (AM/OM)
Fox
B(t)/B(1) Error Ratios (AM/OM)
Fox_agg
B(t)/B(1) Error Ratios (AM/OM)
Schaefer
B(t)/B(1) Error Ratios (AM/OM)
Schaefer_Agg
B(t)/B(1) Error Ratios (AM/OM)
SM_4918
B(t)/B(1) Error Ratios (AM/OM)
SM_4930
B(t)/B(1) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
t
0
50
100
150

Fig. 26c.

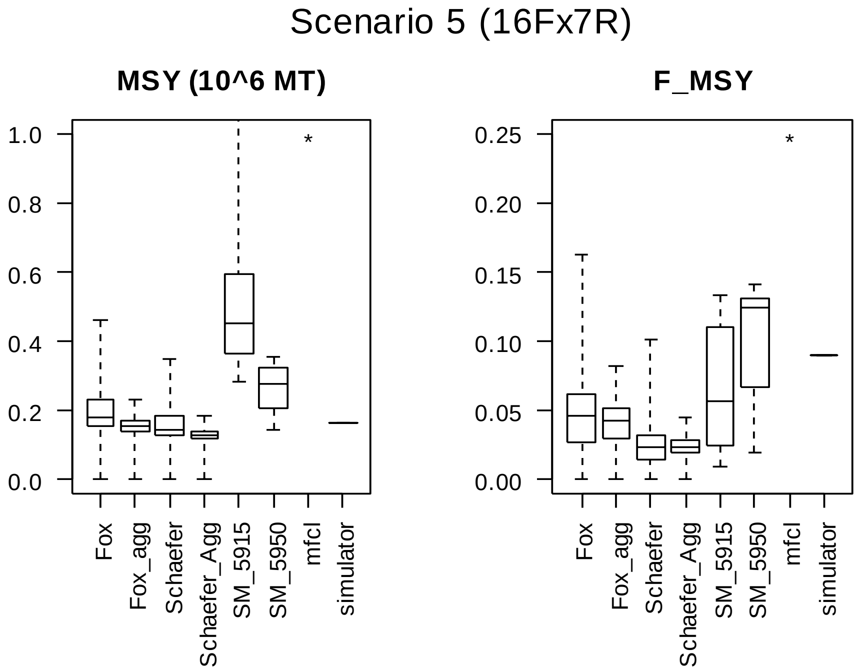


Fig. 27a. **Comparison of assessment model estimates with SPC-OFP YFT simulator values. Assessment results are a boxplot describing the distribution based on 40 simulated data realizations. Asterisk indicates that the results were not available for this model (mfcl). The simulator value is the median from the 40 realizations.**

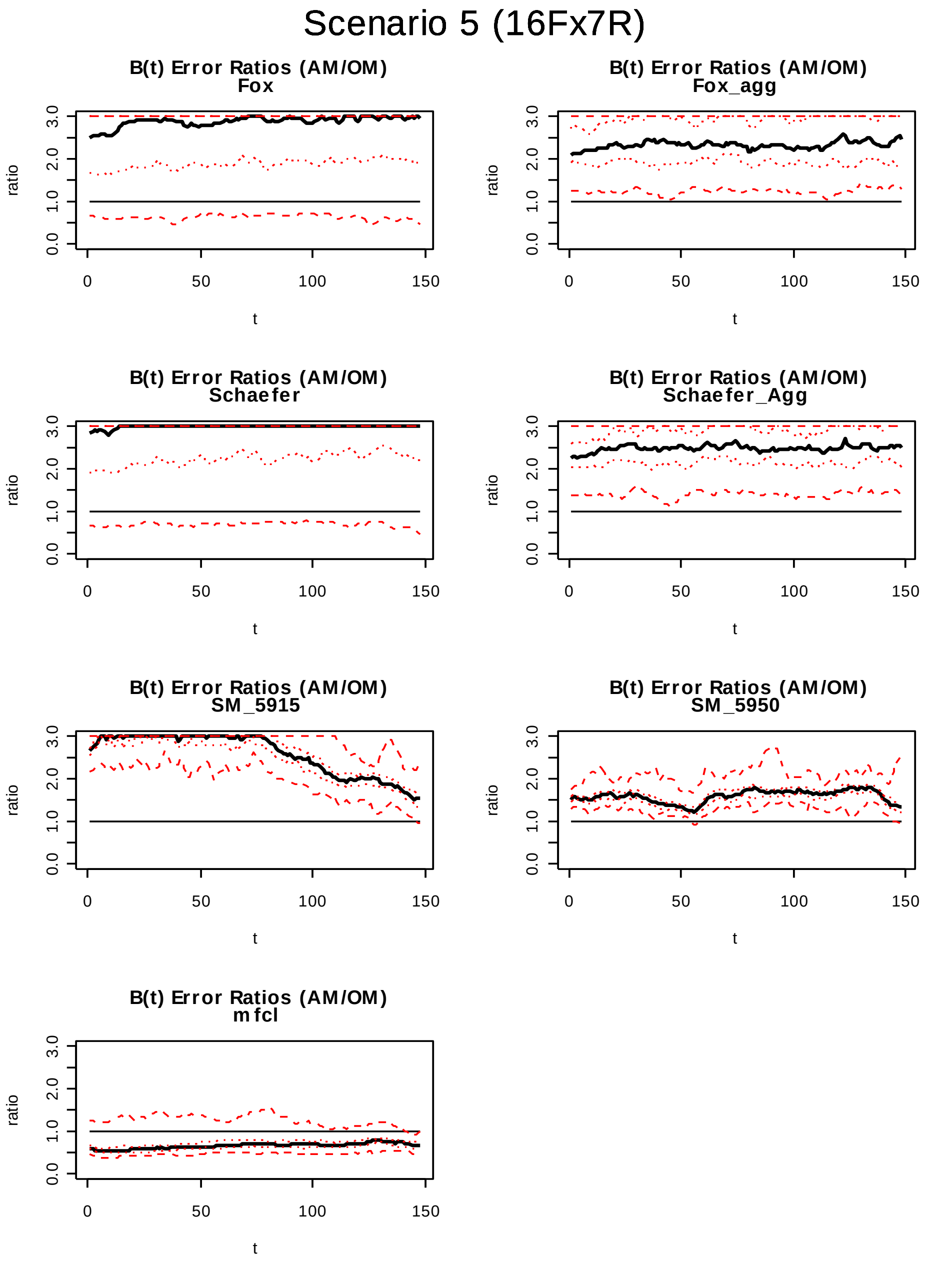

Scenario 5 (16Fx7R)
B(t) Error Ratios (AM/OM)
Fox
B(t) Error Ratios (AM/OM)
Fox_agg
B(t) Error Ratios (AM/OM)
Schaefer
B(t) Error Ratios (AM/OM)
Schaefer_Agg
B(t) Error Ratios (AM/OM)
SM_5915
B(t) Error Ratios (AM/OM)
SM_5950
B(t) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
t
0
50
100
150

Fig. 27b.

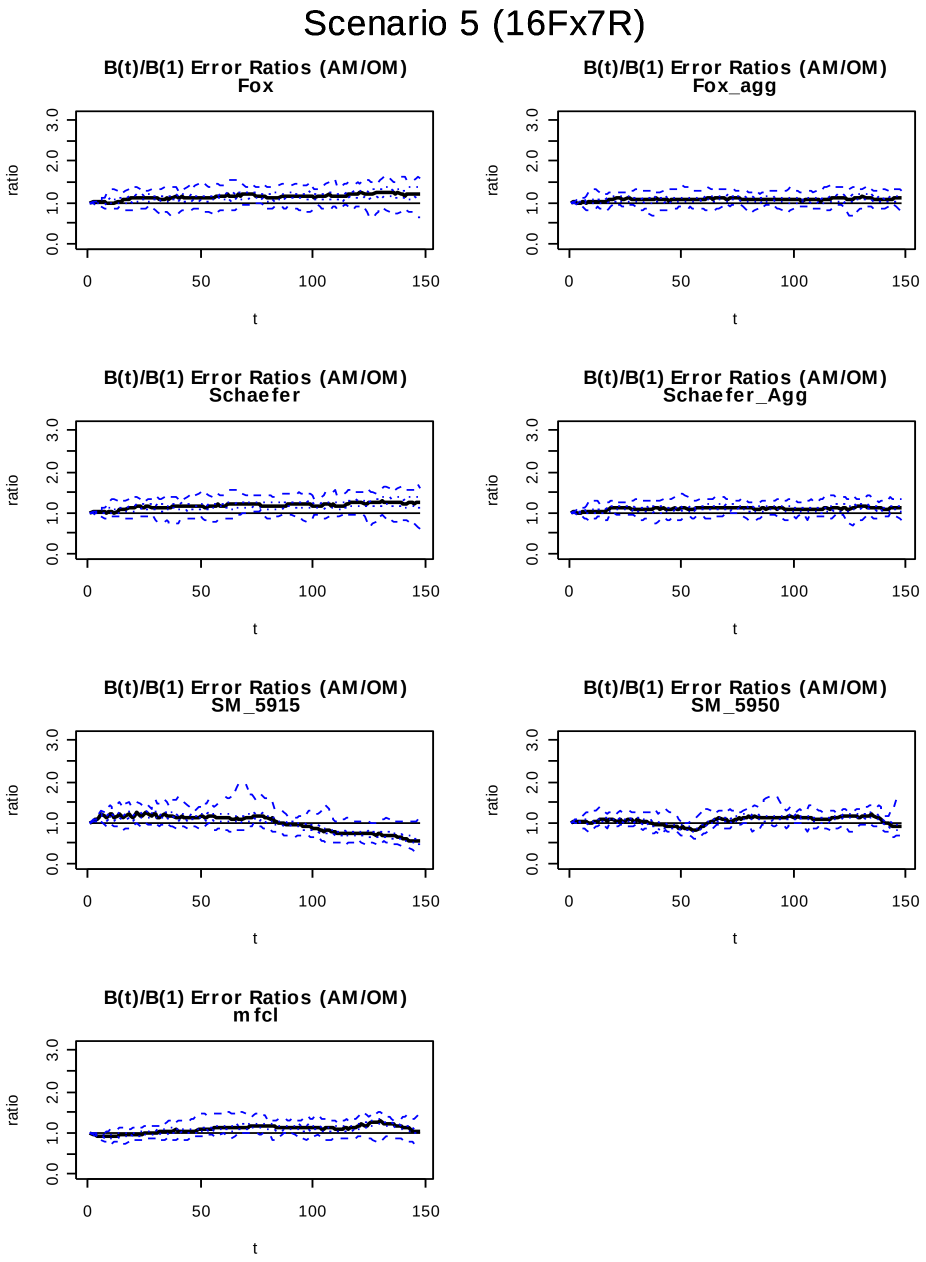

Scenario 5 (16Fx7R)
B(t)/B(1) Error Ratios (AM/OM)
Fox
B(t)/B(1) Error Ratios (AM/OM)
Fox_agg
B(t)/B(1) Error Ratios (AM/OM)
Schaefer
B(t)/B(1) Error Ratios (AM/OM)
Schaefer_Agg
B(t)/B(1) Error Ratios (AM/OM)
SM_5915
B(t)/B(1) Error Ratios (AM/OM)
SM_5950
B(t)/B(1) Error Ratios (AM/OM)
mfcl
ratio
0.0
1.0
2.0
3.0
t
0
50
100
150

Fig. 27c.

These AM applications explored a number of different approaches for dealing with the spatial dynamics in the YFT scenarios.

MULTIFAN-CL used the most complicated approach of dis-aggregating into regional units and explicitly modelling fish migration. From the limited results that we have available at this time, it is not clear that the MULTIFAN-CL global inferences are better than the spatially aggregated models. It would be interesting to know how well MULTIFAN-CL regional estimates of stock dynamics performed.

The SCALIA models use a global population, but dis-aggregated fishery fleets, and explored the spatial problem in 3 different ways. 1) the relationship between global abundance and regional CPUE was very relaxed (either via large CV on effort deviations and/or temporal variability in catchability), under the assumption that the proportion of global abundance in a particular region is likely to change as the fishery develops, 2) since each fishery in the SCALIA model is potentially fishing the global population, temporal variability in selectivity was allowed to admit the fact that temporal changes in the proportion of global abundance in a region will probably also differ by age over time. 3) Given the recognition that the relative abundance index tends to be the most important factor in an assessment, the longline effort data was given high weight and the potential spatial implications were ignored. The first approach received the most attention based on the analyst's prior expectation of the behavior of the YFT operating model. This produced rather disappointing results. The second approach resulted in even greater failure. Selectivity temporal variability resulted in excellent correspondence between predictions and data, but the biomass dynamics generally suggested rapid and sustained stock collapse in all cases tested (none of these trials were pursued across all realizations and are not presented here). The third approach (SC_4950 and SC_5950) resulted in the best performance, and was only implemented after the known values of the operating model were revealed, and it became apparent that the age-aggregated production models seemed to perform better than the SCALIA models.

The AAPMs used the simplest approach for the spatial problem, in that a single spatially- (and age-) aggregated population was defined and all fisheries were aggregated into one. For the most complicated OM scenarios (7F X 7R and 16F X 7R, two different approaches were compared for generating a relative abundance index: 1) Fox and Schaefer simply used CPUE from (one of) the largest longline fisheries (in terms of catch in numbers); and 2) Fox_agg and Schaefer_agg used the global nominal CPUE (total catch / total hooks). The two approaches were similar, but the global nominal CPUE performed somewhat better.

On the basis of this preliminary comparison of performance, it would be difficult to dismiss the production models (in particular Fox_agg) without further consideration. We note from Fig. 28 that the Fox biomass trajectories are rather smooth and "unrealistic"-looking because they cannot represent stochastic recruitment and transient age structure effects. However, despite missing certain short-term details, on average, the Fox model seemed to estimate relative exploitable biomass at least as well as the age-structured models in most of the YFT simulations.

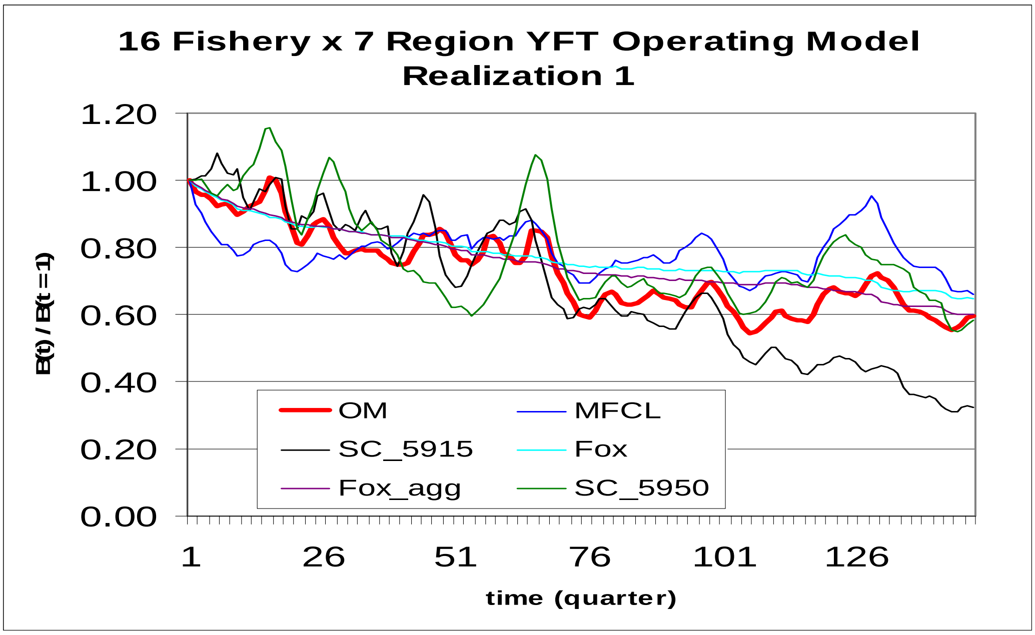


Fig. 28. **Comparison of the relative biomass estimates from 5 different assessment models with the actual YFT Operating Model values for the most complicated scenario (realization 1 from 16F X 7R).**

We are left with a number of questions regarding the MWG project:

1. Is there something particular about the YFT operating model scenarios that makes the Fox model (and to a lesser extent the Schaefer model) appear to be deceptively successful? If global CPUE and a Fox production model actually produce the best assessment inferences what have we gained from all the other complications? The age-aggregated production models were not generally superior to the SCALIA models in the SESAME SBT scenarios, and this demonstrates the importance of simulation testing under a wide range of plausible scenarios. It also suggests that we might want to give more consideration to the evaluation criteria. In this summary, we have tended to emphasize relative biomass estimates as the most important criteria (partly because it is all we had available at the time), but the criteria should probably be related to the type of advice required for effective management, and could well include recruitment, and spatial (or fishery) related objectives that production models cannot provide.

2. Were the spatial dynamics in the operating model parameterized in such a way that we actually gained useful insight about the spatial abilities of MULTIFAN-CL? Perhaps migration rates were so high (or fishery removals so uniform) that spatial dis-aggregation added nothing to global inferences. Alternatively, perhaps the YFT spatial dynamics were sufficiently different from the MULTIFAN-CL assumptions, or so difficult to estimate, that no net improvement was evident in the MULTIFAN-CL analyses over the spatially-aggregated models.

3. Did the simulations provide a sufficient challenge to test the implications of transient age-structure effects in the YFT population? It is curious that the age-aggregated production models seemed to perform as well as the complicated age-structured models in the majority of scenarios, despite the apparent diversity in the exploitation histories. Perhaps the exploitation rates were not sufficient to actually introduce much variation in the population age structure or the age-structure of the relatively short-lived YFT is simply not as much of an issue as with SBT. Alternatively, perhaps the age-structured models could not estimate the age-structure (mortality) reliably enough to represent an improvement in the understanding of global dynamics over that observed in the age-aggregated models (e.g. an updated example of Ludwig and Walters 1985).

4. Is the estimation of natural morality in MULTIFAN-CL and SCALIA the main limiting factor for model inferences (given a highly informative relative abundance index)? Qualitative inspection (not shown) indicated that both (but particularly SCALIA) had large biases in the estimates of mortality-at-age in the YFT scenarios. Similarly, in the SESAME SBT trials, the models with known M performed much better than those that attempted to estimate M. This probably also indicates that simulations should cover a range of plausible M assumptions to avoid the self-fulfilling prophecy (i.e. if an assessment model tends to estimate mortality-at-age with particular bias characteristics, it might not make sense to use the estimates of M from the assessment to feed into the operating model that is in turn used to test the assessment model).

5. If WCPO assessments are going to provide advice related to MSY and recruitment over-fishing, MULTIFAN-CL should also be evaluated against a range of stock recruitment scenarios. Our SBT simulations suggested that MULTIFAN-CL might have a bias that tends to over-estimate recruitment compensation. The YFT simulations indicated that SCALIA steepness estimates were highly sensitive to assessment model assumptions (although it is not clear that this remains true when CPUE is given a high weighting).

6. Advice to managers for the WCPO might require explicit consideration of spatial issues. In which case, there might be no alternative but to go down the route of a MULTIFAN-CL-type assessment. However, if CPUE truly is the main driving data for all models, perhaps there is merit in sacrificing age-structure complexity for improved spatial structure (or at least prioritizing catch rate analysis and interpretation as a major focus for assessment resources).

5.10 OBJECTIVE VIII - STOCK ASSESSMENT MODEL UNCERTAINTY QUANTIFICATION

This section of the report attempts to address the different facets of uncertainty quantification as we defined them in the introduction, and discusses how this should be considered in the context of stock assessment and the provision of advice to managers.

5.10.1.A Estimator Performance

It is clear from the results of the preceding sections, that there are likely to be substantial biases in many of the stock assessment model MPD estimates. Some of the stock assessment models had real problems estimating natural mortality, absolute biomass, stock recruitment curve steepness, and MSY, even when the data were unrealistically good. Models generally performed more reasonably with a higher degree of prior knowledge (e.g. knowing natural mortality, the absolute biomass estimates were generally good). Not surprisingly, the quality of the inferences degraded as the quality of the data decreased. Some key system features could not be reliably estimated under any of the simulation conditions (e.g. temporal variability in catchability for the main (CPUE) relative abundance index). Some quantities were estimated more reliably, including the relative biomass (the ratio of biomass at two points in time, e.g. B(t = 2005) / B(t = 1980)). In most cases, we would not place much faith in the best point estimates from any given assessment model. We would recommend that management advice should be focused on quantities that can be estimated relatively reliably (e.g. relative biomass, trends in historical recruitment), rather than quantities that may be easier to work with, but are more difficult to estimate (e.g. absolute biomass and absolute exploitation rates).

We would consider the relative abundance indices to be the most informative piece of data in most assessments. If the exploitation and data history resemble SBT, then the other data seem to have limited capacity to recognize problems in the relative abundance index, and are probably not going to be sufficient to produce reliable estimates of trends in catchability for the main CPUE series. This strongly suggests that quantifying the uncertainty in effort standardization and catch rate interpretation should be a primary focus for most pelagic fisheries assessments; and encourages further development of fishery independent abundance indices.

The implications of the limitations of estimator performance are described in further detail in the following sections on uncertainty estimation and model uncertainty.

5.10.1.B Statistical Uncertainty Estimation

A limited attempt to examine the reliability of statistical uncertainty estimation using the method commonly applied in SCALIA (and MULTIFAN-CL) analyses confirmed that it was not very reliable under the test conditions. We calculated the proportion of times that a known quantity from the operating model fell within the estimated 50% confidence intervals generated by SCALIA. The confidence intervals were calculated by AD Model Builder using the multi-variate normal approximation from the Inverse

Hessian matrix at the mode of the objective function (combined with the delta method if the quantity of interest is derived from the estimated parameters). Three example assessment models, SC_base, SC_noHTS and SC_1ideal were applied to 40 realizations of the most well-behaved operating model (E_base). Fig. 29 illustrates how the actual values in the operating models compare with the estimated 50% confidence intervals for 5 performance indicators (stock recruitment curve steepness, B(T), B(T)/B(1), C(T)/B(T), and Rec(T-9:T)/Rec(1:10)). Clearly the confidence intervals were too narrow in general. Exploitation rate estimates were the only estimates for which >10% of actual values ever fell within the 50% confidence intervals.

These confidence intervals were calculated conditional of the assessment model being “correct”. Of course, none of these models are perfectly correct in the sense that the dynamic equations of the operating model and assessment model are not identical. However, the three assessment models were specified with very good assumptions about the underlying dynamics, and are thus likely to be over-optimistic in terms of the reliability of the confidence intervals relative to a real assessment. These results are consistent with our observations of the MPD biases evident in the majority of the SESAME results, and our actual applications to SBT stock assessment, in which alternative plausible models often result in non-overlapping 95% confidence intervals (Kolody and Polacheck 2001).

There may be some confusion about the relative importance of bias and variance in the interpretation of these uncertainty estimates. Due to complicated non-linear interactions in these models, we are not surprised that there appear to be biases in many of the MPD estimators. One might argue that bias corrections could be applied to remove the worst effects. If there is a consistent 20% over-estimation bias in MSY, then it might be appropriate to apply a 20% adjustment to MSY estimates after fitting the model. After the relevant biases are removed, then variance estimators (e.g. from the inverse Hessian matrix) might provide a much more reasonable estimate of the confidence intervals. However, bias correction is not usually applied to these models, so our illustration of confidence intervals is representative of most applications. There may be merit in applying bias corrections, but the efficacy of these techniques would presumably be conditional on the assumptions under which they were developed. Given the nature of the MPD results observed throughout SESAME, we would expect that the magnitude of estimation biases would usually be sensitive to the simulation conditions and assessment model specification. If this is the case, it is not obvious that reliable bias correction methods could be developed.

Likelihood profiles and Bayesian posteriors (from a full Bayesian integration) might provide a better representation of the statistical uncertainty. e.g. Lewy and Nielsen (2003) used simulations to illustrate a Bayesian approach that does not seem to suffer from overly narrow confidence intervals. Seemingly all of their operating model parameters were within the estimated 95% intervals from the marginal posterior distributions. It is not clear that this apparent over-estimation of statistical uncertainty should be considered better performance, but it is worth further investigation. We also note that their simulation testing involved identical structure of operating model and assessment model, presumably eliminating model uncertainty as an issue.

All of the uncertainty estimation methods mentioned above (inverse Hessian, Bayesian posteriors and likelihood profiles) assume that the objective function in these models can be interpreted as a true likelihood. But there is a fundamental question of the appropriateness of applying likelihood theory to problems where the number of parameters increases in direct proportion to the amount of data. There is also an ambiguity between the definition of parameters and states (e.g. should an individual recruitment event be estimated as an individual parameter, or instead integrated out as an alternative state in a random effects model). Boot-strapping provides an alternative approach for estimating uncertainty that should be more robust to the objective function assumptions than the others. However, boot-strapping is also computationally intensive, and methods for dealing with time series data are not well developed.

While we have some optimism that alternative methods of uncertainty estimation might prove to be more effective than the Inverse Hessian approximation, we expect that more substantial improvements can be made in relation to model formulation. At one level, there is probably scope for improving the individual terms in the likelihood to more realistically represent the processes that govern the stochasticity in the system dynamics and observation/sampling methodologies (e.g. Stefansson 2003). However, we also expect that fisheries models will always contain arbitrary assumptions and that key assessment inferences will often be sensitive to these assumptions. This is discussed further under Model Uncertainty below.

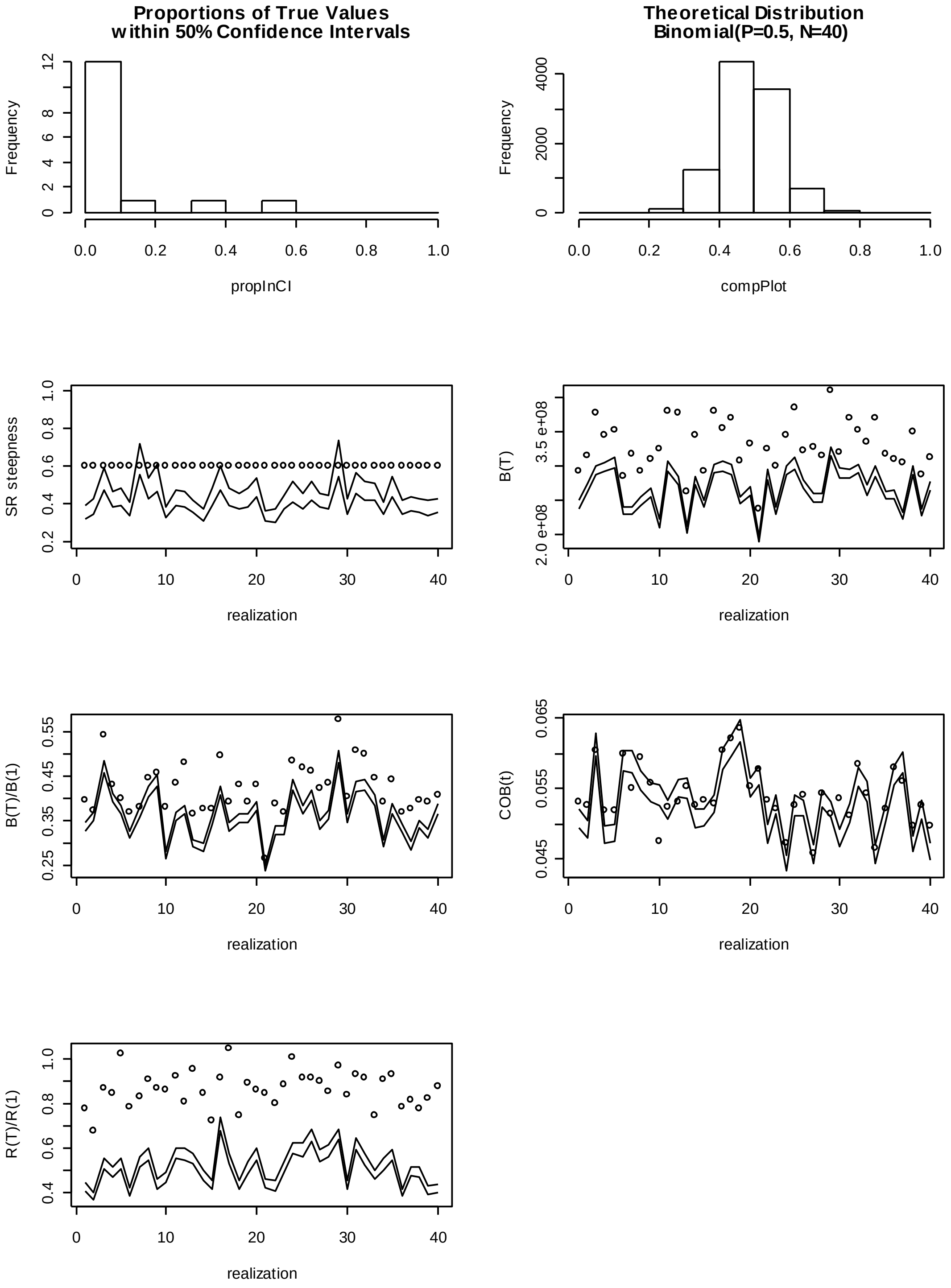


Fig. 29 Top left: frequency distribution of the proportion of times (out of 40) in which operating model state values (stock recruitment curve steepness, B(T), B(T)/B(1), C(T)/B(T), and Rec(T-9:T)/Rec(1:10)) fell within the estimated 50% confidence intervals for the SCALIA assessment models SC_base, SC_noHTS and SC_1ideal. Top Right: The corresponding theoretical distribution that would be expected if the analysis was repeated . Bottom 5 panels indicate the 50% confidence limits (lines) from SC_base and actual values from the operating model (E_base).

5.10.1.C Model Uncertainty

The SESAME SBT results strongly support the assertion that model uncertainty is a substantial issue in stock assessment. As indicated in most of the results addressing Objectives I - VII, assessment models had substantial bias and variance for many of the MPD estimates, and the nature of the biases differed among the different model specifications. The problem was evident to some degree when the simulated fishery dynamics and data were close to ideal, but became substantially worse as larger and more plausible process and observation errors were introduced. These results are generally consistent with our observations of model sensitivity in real assessment applications.

A shortfall in our considerations of model uncertainty is the absence of any analysis of diagnostics for examining the quality of fit between data and model predictions. Diagnostics are routinely used in actual assessments and can identify models that have obvious inconsistencies with the data. Implementing an automated expert system for analyzing diagnostics was beyond the scope of the SESAME project, and hence we might be somewhat over-emphasizing the importance of model uncertainty if the results include models that simply do not fit the data. However, we note that among the complicated models, the specifications that performed the worst were often the ones with the weakest constraining assumptions. Provided that the function minimization routine is working correctly, the models with the weaker constraints should actually fit the data better (presumably they perform more poorly because they are over-fitting to noise), in which case the diagnostics might not be very informative. There are successful examples of the use of the Aikake Information Criterion (AIC) and Bayesian Information Criterion (BIC) for model selection and the identification of optimal model complexity (e.g. Helu et al. 2000), these illustrations tend to look at a small subset of plausible models under simple simulation conditions. We also note that model diagnostics are often helpful for identifying conflicting trends in the data, but this does not necessarily help in selecting the best model or weighting the credibility of different models (unless there are objective reasons for believing one source of data over another).

We attempt to contrast some of the more consistent advantages and disadvantages of the different assessment models that we have examined as part of SESAME in the subsequent section on Relative Performance of Assessment Models. We are left with the impression that, in many cases, there is no easy way to reliably distinguish which model is the best for assessing a particular population. Using alternative specifications of the complicated integrative models to explore model uncertainty probably provides a reasonable means for illustrating the plausible states of nature that are consistent with the data. There is reduced scope for attempting this with the simpler models. We expect that the model uncertainty will generally be greater than the statistical uncertainty estimated assuming a particular model formulation is correct, and this is especially true when there are conflicts in the data. In some cases, models might be reformulated so that alternative structural assumptions can be encompassed as special cases of a more generalized model, and hence some model uncertainty might be subsumed into statistical uncertainty. However, in general, we think that ad hoc exploration of model uncertainty through sensitivity analyses (and informal comparison with independently implemented models if available) should be an important part of stock assessment. It seems inevitable that any synthesis of uncertainty quantification is going to have a large component of subjectivity, and it would be naive to interpret the probabilistic summary statements from most fisheries assessment endeavors too literally. An ad hoc admission of this uncertainty is probably more useful than a tidy mathematical synthesis that misses many plausible alternative interpretations of the data.

5.10.1.D Assessment Uncertainty and Fisheries Mangement

Our results are generally consistent with other recent literature that emphasizes the importance of admitting that there are substantial and seemingly unavoidable limitations to the quality of fisheries assessment model inferences (e.g. Patterson et al. (2001), Schnute and Richards (2001)). There is optimism that substantial improvements can still be made to assessment models by increasing the statistical rigor employed (e.g. Stefansson 2003, Maunder 2003). And there is a fear that increasing expectations (e.g. ecosystem management objectives) might require modellers to move in directions that are far less tractable, such that the successes that have been realized in the single species context to date might be undermined in the future (Quinn 2003). We tend to agree with the view that the greatest potential for improvements in assessment outcomes will probably be made at the interface of science and management. We should strive to express the plausible range of uncertainty about fisheries systems, and try to come up with management strategies that are robust to these uncertainties to the extent possible (e.g. Schnute and Richards (2001), Schnute 2003, Prager and Williams (2003)). The complicated statistical models provide an important tool to help achieve this, and operating models will probably play a larger role in the future, as will increased dialogue among scientists, managers and fishers.

The work that we have undertaken through SESAME is important for understanding the limitations of assessment models, however, it does not necessarily provide a good representation of the types of errors that are likely to occur in fisheries management. We attempted to evaluate the quality of several different model estimators that are routinely calculated and presented as advice for managers to consider. However, many of these quantities might be largely irrelevant for management decisions. And even when we conclude that a given estimator is poor when calculated under particular conditions, there is no guarantee that the estimator will remain poor as additional data becomes available. Different assessment model specifications might yield substantially different assessment estimates in any given year, but long term management performance based on the different models might be rather similar (e.g. Kolody and Patterson 1999), at least provided that the management decision rules are sensible and flexible enough to respond to apparent changes in stock status in a timely fashion.

These types of observations support the development of formal Management Procedures (MPs) as one possible means of achieving management objectives in the face of uncertainty (e.g. Butterworth and Bergh 1993, Punt 1996, CCSBT 2002). 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 constant catch or effort in future projections). MPs are also evaluated using performance measures that should be readily defined from management, and thus should avoid the need to simultaneously examine many values. In an MP context, complicated assessment models probably should play an important role in conditioning the operating model that is used to simulate the future fishery dynamics. But simple models, or even data-based decision rules are often as good or better than complicated models for making management decisions once they are "tuned" to be robust to the major uncertainties identified in the operating models.

MPs represent a promising method for dealing with assessment uncertainty, and reduce the need for applying complicated assessment models at frequent intervals, however, they will not resolve all of the issues currently facing stock assessment. The effectiveness of MPs will ultimately be affected to some extent by how well the underlying operating model represents reality, and this in turn is related to how well assessment models can be used to condition the operating models. We note at the time of writing, that the CCSBT has not been able to agree on the final conditioning of an operating model within 18 months of first proposing a structure. It remains to be seen whether an MP can be agreed by CCSBT commissioners, as the MP process in itself cannot resolve the difficult decisions required when management objectives conflict.

The successful development of an MP will be dependent on effective communication between scientists, managers and industry. Scientists need to gain an appreciation of the relative importance of conflicting management objectives. Fisheries managers need to understand the concepts of uncertainty and risk. At least in the initial phases of MP development, this is likely to require greater interactions among the participants than has traditionally occurred in Australia's pelagic fisheries in the past. MP development typically also requires an increased workload for scientific and technical staff than traditional assessments. However, to some extent this should be offset in subsequent years, as the burden of stock assessment is reduced to the implementation of a decision rule. Scientific staff will still be required to regularly evaluate whether the system remains within the realm that the MP was designed for, however, full model-based assessments with a comprehensive expressions of uncertainty should only need to be carried out at periodic intervals to check on MP performance, or redevelop MPs to address changing management objectives.

5.11 General Comments on the Relative Performance of Assessment Models

The lack of specific criteria for the overall evaluation of assessment models makes it difficult to provide definitive statements about model performance. However, we do attempt a simple synthesis here. We note the general impression that most of the assessment models performed better for the YFT simulations than the SBT simulations (although for the SCALIA models this is only true for the assessment models that assumed CPUE was highly informative). This probably represents a combination of factors, including:

- Although the SBT simulations generally experienced greater overall depletion (with potentially more informative contrast) than the YFT simulations, the highest exploitation rates and greatest depletion occurred before the relatively informative (in terms of CPUE) longline feeding grounds fishery was active.

- The long-lived nature of SBT, and absence of juvenile catch in the early part of the time series, limits the degree to which ages can be inferred from length frequency data.

- The SBT scenarios covered a greater range of simulation conditions, including many scenarios with dynamics and data characteristics that are more difficult than we generally assume when formulating stock assessment models for real applications. The major exception to this is the complicated spatial dynamics in the YFT simulations, but we question whether this was actually designed to represent an appropriate challenge for the assessment models.

The remainder of this section focuses only on the SBT scenarios, although we hope that more results will be forthcoming in relation to the YFT simulations from SCTB-17.

We focus on the aggregate performance indicator in attempting to make broad generalizations from the SBT simulations. We calculated the aggregate PI over several SBT operating model scenarios that we considered to be the most important (each scenario is implicitly given an equal weighting in the index), and for which a sufficient number of assessment models were run. In each case, this included the baseline specifications, plus other plausible models that were sufficiently different from the baseline. On the basis of the implementation problems described previously, the stochastic ASPM models were not included. Fig. 30 compares assessment models applied to the baseline operating models (E_base, D_base) and the particularly problematic operating models with trends in catchability (E_qInc, D_qInc). The largest range of assessment models were applied to these 4 scenarios. A number of points are suggested from these comparisons:

- On the basis of the aggregate performance indicator, one would probably conclude that SC_noTag had the best overall performance, followed by SC_2Ideal and aspm_d2g. Curiously, SC_BIH could arguably be considered the most robust model in that the range in performance seems to be the narrowest.

- The worst models would probably be the SCALIA models that estimated natural mortality (SC_Mest, SC_EL) and BIH_2; followed by s_calc and aspm_d6g.

These observations are not really in line with our general impressions throughout the study, as these aggregate indices are highly influenced by the troublesome E_qInc scenario. In Fig. 31, we calculate the aggregate PI based on a broader range of the operating models tested (E_base, D_base, E_h3, D_h3, E_h9, D_h9, E_h45, E_qC, E_qI, E_DDLinf), but exclude the E_qInc and D_qInc scenarios. The mf_x and BIH_2 assessment models were not included because they were not applied to most of these operating models. In this case, our impressions are somewhat different:

- Several of the SCALIA models (SC_base, SC_noHTS, SC_qTS1, SC_noTag, SC_1ideal, SC_2ideal and SC_CA60) and aspm_d2g had rather similar aggregate performance, and somewhat better than the remaining models. SC_2Ideal was possibly the most robust of all, in that the range was the narrowest.

- f_calc, s_calc, aspm_d2g, SC_Mest, SC_EL and SC_BIH, had substantially worse performance than the rest. The SCALIA models that attempted to estimate natural mortality were probably no better than the Fox model.

Not surprisingly, this summary does suggest that the model performance is influenced by the quality of prior information used in the model formulation. e.g. Perfect knowledge of natural mortality (and selectivity for the ASPMs) will usually result in improved performance over similar models that attempt to estimate these attributes. SC_2ideal seemed to have somewhat better performance over a range of operating model scenarios, which does suggest that a more relaxed model specification (e.g. larger variances, more structural flexibility) might give more robust results in general, but it was rarely (if ever) identified as the best performer for any specific operating model application. To some extent, we consider that this might be a deceptive result of the aggregate performance indicator. But there is probably merit in developing models that are robust to the most plausible assumption violations that we are likely to encounter.

From this study, we are not convinced that the estimation performance provided by complicated models is clearly better than the simple models, but we would argue that the complicated models provide a much more useful tool for exploring the range of dynamics supported by the data. It is likely that a complicated model will provide better estimates than a simple model if it is specified appropriately, but it also seems to be the case that complicated models might be less robust to certain types of assumption violations. This is obvious in some cases, e.g. all other things being equal, if tag dynamics assumptions are poor, a model without tags should perform better. It is interesting that the model specification that does yield the best estimates is often not the model that would be expected on the basis of prior knowledge of the individual components of the underlying dynamics. This could be related to subtle inconsistencies and un-anticipated interactions among the model components (e.g. in the SBT simulations it is not obvious how to specify catch-at-length effective sample sizes, given that there is a time-step mismatch, and an interaction between selectivity process errors and catch-at-length observation errors). The real advantage of the complicated models is probably realized from the expression of uncertainty through the exploration of alternative model structures. We consider this to be true even in the absence of a formal theoretical framework for integrating results. The simple models lack the structural richness to easily explore the implications of different model assumptions.

The following sections summarize our general impression of the different models.

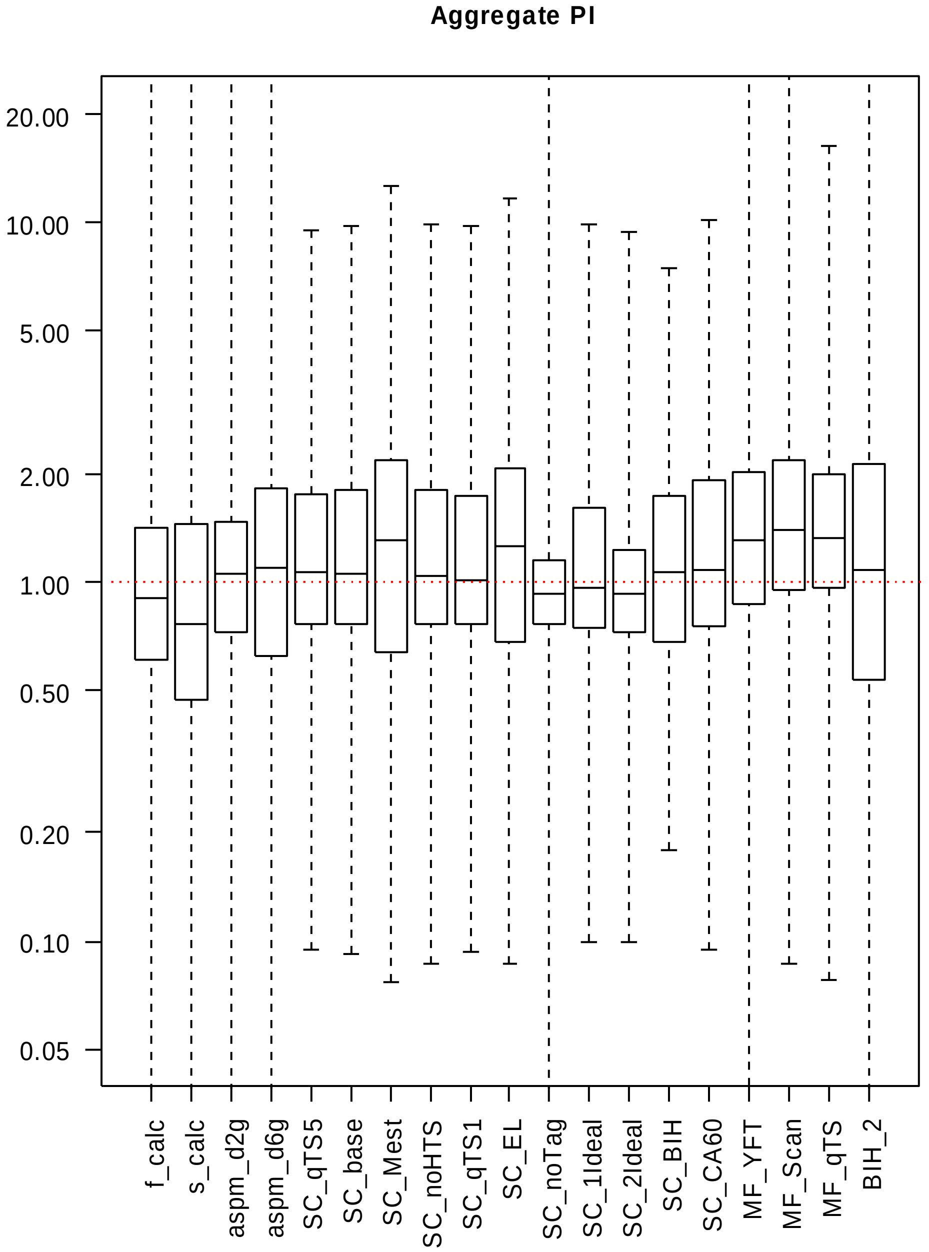


Fig. 30. Comparison of assessment models on the basis of the aggregate performance indicator (see Table 8) calculated across results from the simulated SBT operating model scenarios E_base, D_base, E_qInc and D_qInc. OMs are defined in Table 1, AMs inTable 2.

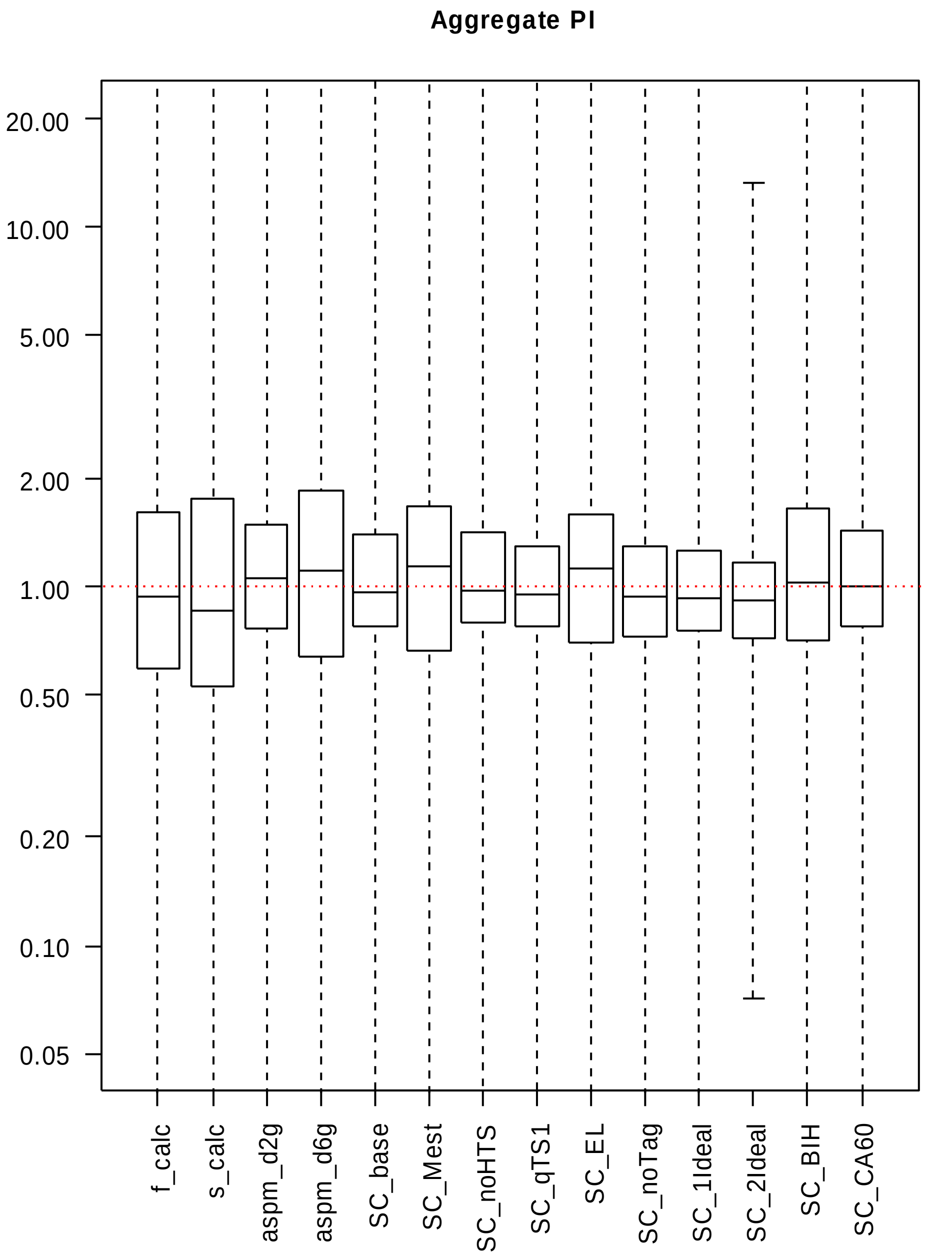


Fig. 31. Comparison of assessment models on the basis of the aggregate performance indicator (see Table 8) calculated across results from the simulated SBT operating model scenarios E_base, D_base, E_h3, D_h3, E_h9, D_h9, E_h45, E_qC, E_qI, and E_DDLinf. OMs are defined in Table 1, AMs inTable 2.

5.11.1 Age-Aggregated Production Models

AAPMs have a number of attractive features, and in both the SBT and YFT simulation results, we found it interesting that the Fox models often provided assessment results that seemed to be similar to, or better than, the complicated integrative models. The Fox model did not perform as well as the sophisticated models when the assumptions of the sophisticated models were in good agreement with the SBT simulators. However, the Fox model often seemed to be more robust when certain assumption violations were present (e.g. trends in catchability, unrecognized changes in length-at-age) and seemed to out-perform SCALIA when natural mortality was estimated. Many of these generalizations are true to a lesser extent for the Schaefer model, but the Fox model was virtually always the better of the two. We would expect the Fox model to be better than the Schaefer model because the underlying production dynamics more closely approximate the SBT and YFT simulated populations (e.g. B_MSY/B(unfished) < 0.5). However, in light of the general results that we have observed here, and the more general recognition of the limitations of assessment models in recent years, we do not find these models very attractive as the primary basis for providing stock assessment advice.

These models are often considered attractive because they have minimal data requirements, are quick and easy to implement and simple to interpret. However, these features are also limitations. If additional data are available, it makes sense to use them in some manner, but the scope is limited with a classical AAPM. We have also encountered some curious implementation problems. In the SBT and YFT simulations the automated fitting sometimes demonstrated a sensitivity to initial conditions, and in some cases we could not get reliable function minimization even with interactive fitting. This seemed to be more relevant for the YFT simulations, perhaps in part due to the problem of fitting a long time series with deterministic dynamics. We also observed that problematic likelihood surfaces can cause poor inferences in real applications (Ricard et al. 2002). It is usually presumed to be important to be able to examine the potential effects of changing fishery selectivity (particularly large changes related to different gear types) on the population dynamics, and the subsequent age-structured dynamics, but this is beyond the means of the AAPMs. The relatively fast dynamics of YFT relative to SBT might have contributed to the relative success of the AAPMs in the YFT context, and this might suggest that age structure is not always as important as we traditionally assume.

In light of the results presented here, we think that the quantification of uncertainty and development of robust management plans should be the main goals of stock assessment, but find that the AAPMs provide limited scope with which this can be achieved. There have been attempts to examine sensitivity in AAPMs (e.g. Butterworth and Plaganyi 2001), however, we felt that in the case of SBT, the ad hoc attempts to approximate more complicated structure actually made the models more difficult to interpret than age-structured models. Maunder (2002) suggests that it generally makes more sense to use a generalized form of AAPM such as the Pella-Tomlinson model. Although the shape parameter cannot usually be reliably estimated, it can be constrained in a manner that is consistent with auxillary information about the population biology. Some degree of uncertainty regarding productivity can be sensibly explored in this way. However, many sources of commonly available data and important structural features will still be left outside of the AAPM model framework.

But we do note that in the context of Management Procedures, we would consider these models to have considerable potential as the basis of a decision rule, particularly over short-medium time horizons.

5.11.2 Age-Structured Production Models

Overall, our experience with ASPMs was not very encouraging. The ASPM variants with deterministic recruitment, aspm_d2g, performed reasonably well in most of the SBT simulations (and often better than the more sophisticated models). However, this is not too surprising given that the fixed input natural mortality, selectivity and the functional form of the stock recruitment relationship were known perfectly (for the majority of OM scenarios). A simple attempt to analytically estimate selectivity from the catch-at-length data (aspm_d6g) was not satisfactory. Numerical problems in our implementation caused numerous failures in the automated applications, such that all ASPM results were withdrawn from the YFT studies. The ASPM variants with stochastic recruitment did not converge reliably. We did not really attempt to improve the implementation, because we see these models as actually a rather complicated transitional step to the fully integrated models. We could have attempted to improve the implementation of the ASPMs, but did not think the time was justified. If desired, any of the fully integrated models could be parameterized to work as a form of ASPM, by simply removing any superfluous data from the objective function and using fixed input for selectivity and mortality.

5.11.3 SCALIA

We were generally pleased with the SCALIA implementation and minimization reliability. But we were disappointed by the model sensitivity to assumptions, and the limited ability to reliably estimate some key stock characteristics even given unrealistically good data. Both the YFT and SBT simulations indicated that temporal variability in catchability is a real problem that probably cannot be resolved within the context of an assessment model (at least not without additional data). We would not be surprised if real applications resulted in rather poor estimates for absolute biomass, the stock recruitment curve or MSY. SCALIA failed to estimate natural mortality very well in the majority of cases. We recognize that a different approach for using tagging data should improve mortality estimation to some degree (i.e. even though tag recoveries are predicted by time and age, fitting to tag recoveries from individual release events is more informative than fitting to tag recoveries aggregated across release events as is currently done in SCALIA). Relative to MULTIFAN-CL, SCALIA did not seem to be as computationally efficient, but we could not conclude that the quality of inferences was different under the test conditions. In both cases, inference quality seemed to be driven primarily by user specifications.

From this study we are left with the seemingly inescapable conclusion that any serious application of SCALIA (or any similar model) for stock assessment, should involve a substantial exploration of model uncertainty (sensitivity to assumptions). This is also consistent with our observations from real SBT applications. The exploration of model structural uncertainty should be given a greater emphasis than the estimation of statistical uncertainty conditional on the model being correct.

These simulations have proved informative in identifying the limitations of SCALIA, and we expect that similar simulations will prove useful for guiding new developments to ensure that they actually represent improvements in the advice that we can provide to managers, as opposed to new complications that ultimately only impede our ability to effectively run and interpret the models. To avoid duplication of assessment model development effort, we expect that most new SCALIA innovations will be implemented in a manner that differs from the approach adopted in MULTIFAN-CL.

5.11.4 BIH_2

A major criticism of Butterworth et al. (2003) and Polacheck and Preece (2001) has been the use of cohort-slicing to estimate the catch-at-age composition. Under ideal assessment conditions, this did not seem to be a serious problem. BIH_2 recruitment estimates demonstrated high variance (estimation error as opposed to recruitment deviation CV) and auto-correlation, which are the expected problems for cohort-slicing, but the biomass-related and management-related estimates were not obviously worse than the equivalent models using catch-at-length data. There did seem to be some particular biomass bias trends that might have been related to the estimation of the initial age composition. BIH_2 performance seemed to be worse than many of the SCALIA models in the harder OM scenarios, but we did not explore why this was the case. It might have predominantly reflected “chance”, in that there were many SCALIA models and only one BIH_2. When the other plausible errors were present in the D_x scenarios, the recruitment estimation errors did not seem to be markedly worse than the SCALIA models. However, given modern computing power and the apparent success of catch-at-length methods, it is not clear why one would prefer cohort-slicing for age-estimation at this time.

5.11.5 MULTIFAN-CL

Our application of MULTIFAN-CL to the simulated SBT data demonstrated some similarities with the SCALIA applications, but overall the inferential performance seemed to not be as good. Notably, stock recruitment steepness estimates were usually biased high. This occurred despite a (very weak) prior with a mode slightly below the actual steepness value (in contrast, SCALIA had a uniform prior on steepness in all cases). We suggest that the MULTIFAN-CL problems in the SBT scenarios might have been related to the following:

- Some MULTIFAN-CL features are rumoured to exist but are not well documented at present, so we did not attempt to use them. In particular, the highly informative (but limited duration) catch-at-age data from the spawning ground fishery was not used in the objective function (catch-at-length from this fishery was used).

- The annual data aggregation and continuous fishing of the SBT simulator results in large variability in the length-at-age distribution for young ages due to within year growth. It is likely that MULTIFAN-CL would have handled this better by using a finer time-step (whereas SCALIA has an option to approximate some of the growth effects directly into the length-at-age distribution).

- Our lack of familiarity with the software

In contrast, MULTIFAN-CL seemed to perform reasonably well, and better than SCALIA in the YFT simulations. In part, it seems this might be largely related to assumptions about the relationship between effort and fishing mortality. But it is currently unclear whether the YFT simulations were actually appropriate to test the spatial dynamics capabilities of MULTIFAN-CL, given the apparent success of the production models using global nominal CPUE as a relative abundance index.

We recognize that MUTIFAN-CL is probably the most flexible assessment model of its type that is publicly available. However, there are some features thought to be important for SBT assessment that are not currently available (or perhaps not documented) that we would like to see before application to SBT:

- Catch-at-age data in the objective function

- the ability to incorporate variability in length-at-age over time (e.g. Polacheck et al. 2003a documents substantial changes in SBT length-at-age over time).

5.12 Methodological Limitations

The basic premise of the simulation-estimation methodology as outlined in the Methods (Fig. 1. Outline of simulation-estimation methodology for stock assessment model evaluation.) is very straightforward. However, as shown in Fig. 32, there are actually a large number of potential problems that can cause misleading results, or at least lead to inferences that potentially do not generalize as well as one would hope. In some cases, a serious flaw in one stage could invalidate all the results.

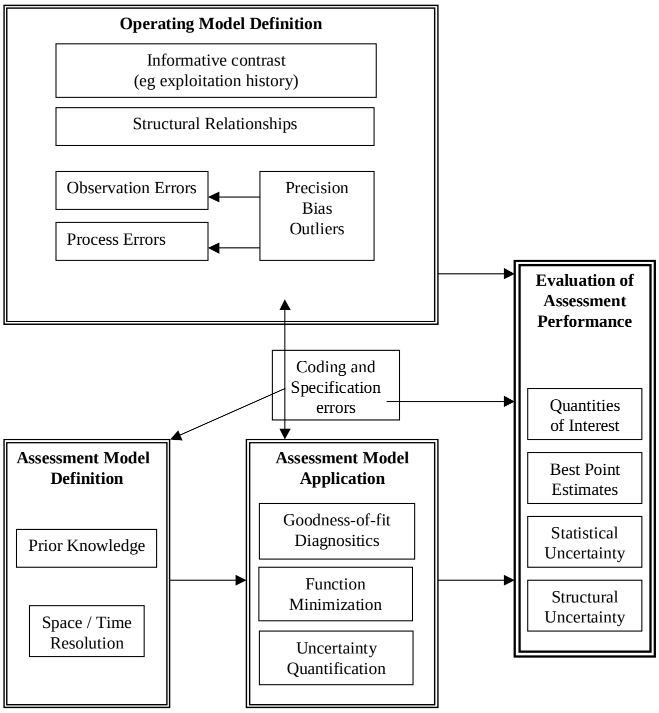

Operating Model Definition
Informative contrast (eg exploitation history)
Structural Relationships
Observation Errors
Process Errors
Precision
Bias
Outliers
Evaluation of Assessment Performance
Quantities of Interest
Best Point Estimates
Statistical Uncertainty
Structural Uncertainty
Coding and Specification errors
Assessment Model Definition
Prior Knowledge
Space / Time Resolution
Assessment Model Application
Goodness-of-fit Diagnositics
Function Minimization
Uncertainty Quantification

Fig. 32. Illustration of SESAME Simulation-Estimation methodology highlighting problematic issues.

The following list describes a number of potential problems that we encountered, our attempts to deal with them, and alternative solutions that might have been more appropriate.

1. Coding and specification errors – there is always the potential for an error in either the operating model simulator or assessment model that will cause misleading results. We attempted to minimize this potential by examining detailed graphical output and comparing dynamics between independently coded models under similar conditions. However, model structural incompatibilities and continuing evolution of the software limit the extent to which these latter comparisons can be done, and we will never be certain that all these errors are removed.

2. Operating Model specification – there is often a circularity in the parameterization of operating models. The SBT dynamics and biological characteristics were loosely based on assessments, as were many components of the YFT simulation (Labelle 2003). Within the SESAME participants, there were fairly divergent views about how to specify variances of the different SBT process and observation errors (e.g. recruitment variability, effort and fishing mortality relationship, tag dynamics, catch sample sizes, etc). Presumably, if we specify operating models that conform unrealistically well to the assumptions of the assessment models, we will get an overly optimistic impression of our analytical abilities. Conversely, if the operating model is impossibly difficult, all of the assessment models will fail badly and we will not gain any insight into the relative merit of different assessment approaches. In the SESAME simulations, we attempted to bracket the true SBT situation for many of the gross features of the system. The optimistic E_x scenarios probably approach the upper limits in terms of assessment data quality that we could hope for SBT. The difficult D_x scenarios define a lower plausible bound such that we can probably have some confidence that our inferences are reasonable if they succeed in this case (although even in this case, the potential perversity of other major structural considerations were never considered in combination). The diversity of scenarios allows us some insight into the relative performance of the different models under specific conditions, but our ability to make bold statements about absolute performance to be expected in the real world is rather limited.

3. Alternative exploitation histories. The SESAME SBT simulations were all limited to a rather narrow range of historical exploitation patterns that we believe resembles the main features of the real SBT situation; a “one way trip” usually with some recovery near the end of the time series. This time series might be less informative than many other fisheries, because the largest catches were taken on the spawning grounds in the first few years of the fishery, such that there are no observations of juveniles and no reliable relative abundance index from this critical time. In contrast, the SPC-OFP YFT simulations focused on situations with lower overall stock depletion and informative CPUE indices that extend throughout the time series. The YFT simulations had a range of fleet exploitation patterns among the 5 scenarios, (including a mixture of fisheries targeting adults and juveniles in different combinations) but the overall population dynamics were not highly variable among scenarios. We would be interested to know if the results observed here are actually highly dependent on these patterns. It is possible that the apparent success of the Fox model in the YFT simulations could be largely a chance occurrence driven by the particular exploitation dynamics of these scenarios (and the fact that global nominal CPUE seems to be a good index of the population). Given any range of assessment models, there will always be performance differences, and it is possible that the model that is best in a particular situation might appear to be so for reasons that might not be understood or repeatable in other situations. It is important that the model be tested under a range of simulated conditions to reach conclusions about how robust it is.

4. Assessment model specification – a balance in the independence between the operating model developer and the assessment analyst is required. We never did resolve the issue of prior knowledge satisfactorily. If one model is estimating M, and the other uses fixed input, how should the fixed input be specified? The performance of the assessment could be largely driven by the quality of this guess. This same principle applies to most aspects of model specification, but is probably particularly important for a few key issues (e.g. M, stock recruitment relationships, relationship between effort and fishing mortality). Our approach in the SBT simulations was to eliminate the guesswork and simply use the correct value of M when it was fixed input, and recognize that it is not fair to directly compare the performance of two models with different prior knowledge. This makes it rather difficult to meaningfully comment on the estimation of M – it seems to be generally poor, but is it worse than our ability to guess? The other approach that we might suggest for dealing with this issue would be for the operating model developer to provide exact priors on key parameters, and randomly draw the values from the priors for each simulated realization. But this does not really solve the problem, it just transfers it to the choice of priors.

5. Assessment automation – repeatedly fitting assessment models to different data sets requires a large degree of automation, and does not simulate what happens in a real stock assessment. Non-linear function minimization is somewhat of an art, and multiple minima are a distinct possibility. Curiously, we experienced greater problems with the production models (perhaps because they are constrained by less data, and/or minimization failures were more likely to be identified). We cannot be sure that the SCALIA models were consistently identifying the global minimum (a few convergence failures were obvious). Thus we would have to conclude that the assessment model evaluation is not really simply estimating the statistical properties of the model, but rather the combined statistical and implementation properties, including the quality of starting point guesses of the analyst. In the SESAME SBT study we fit several different models and were able to make inferences about the relative performance of each. If however, we were presenting results in the context of an assessment, we would also want to provide some commentary on the relative credibility of the different models. This would include discussion of the quality of agreement between model predictions and observations, and probably the qualitative agreement with our pre-conceived notions of the fishery, including the perspective of auxiliary data that were not integrated in the model. We are not currently in a position to provide this quality of fit evaluation in an automated way. Different criteria are routinely employed (e.g. Polacheck et al. 1999, Harley and Maunder 2003), and perhaps development of an effective expert system would be feasible, but we would always want interactive evaluation in a real assessment. In the context of simulations, attempting to estimate more parameters inevitably means a better fit to the data, and there is usually no auxiliary experience to draw upon. We are also skeptical of interpreting the objective function of these models too literally as a likelihood, so this has implications for the statistical significance attached to the addition of more parameters.

6. Performance Criteria – We defined several performance criteria that we have found to be of interest at one time or another in stock assessment, however, it really is not clear what their relative importance should be. We had hoped that different models would clearly be better or worse than others on the basis of all or a clear majority of indicators, however, we often found that this was not the case. In some situations, one model would be clearly superior in terms of one indicator, but clearly inferior in terms of another. It also is not clear how to trade-off good precision with moderate bias, against moderate precision with low bias, or average performance against robustness to outliers. It would be relatively straightforward to make performance judgments based on a single performance indicator that would be relevant to managers (e.g. how well can we estimate current fishing rates relative to over-fishing). However, if a model has done a poor job of estimating several stock attributes, it would not make sense to give the model high praise because it estimated a single quantity well under a relatively small set of test conditions. This also explains our reluctance to rely on the aggregate Performance Indicators too heavily, as the specific model failures are not evident (e.g. a consistent bias in biomass and exploitation rate estimates is probably less serious than excellent estimation characteristics except for a large bias in the last 5 years). The lack of agreement on specific evaluation objectives means that we were limited to making rather broad qualitative statements, and conclusive statements are generally restricted to large and obvious model failures. One of the possible solutions for these problems would be to evaluate model performance relative to specific management objectives (i.e. Management Procedures or Management Strategy Evaluation). While this is appealing in the context of simulations, it does not remove the fact that MP performance will ultimately depend on how well the operating model represents reality.

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.

7 REFERENCES

Adkison, M.D. 1992. Parameter estimation for models of chaotic time series. J. Math. Biol. 30: 839-852.

Bigelow, K.A. 2002. Application of an ADAPT VPA model to the simulated population data. Meeting of the Standing Committee on Tuna and Billfish 15, Working Paper MWG-8.

Butterworth, D.S., J.N. Ianelli and R. Hilborn. 2003. A statistical model for stock assessment of southern bluefin tuna with temporal changes in selectivity. Afr. J. Mar. Sci. 25: 331-361.

Butterworth, D.S. and E.E. Plaganyi. 2001. Exploratory analyses of southern bluefin tuna dynamics using production models (including separate addendum by D.S. Butterworth and S.J. Johnston). Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0108/24.

Butterworth, D.S. and M. Mori. 2003. Further investigations of a Fox model based management procedure for southern bluefin tuna. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-ESC/0309/37.

Butterworth, D.S. and M.O. Bergh. 1993. The development of a management procedure for the South African anchovy resource. p. 83-99. *In* S.J. Smith and D. Rivard [ed.] Risk evaluation and biological reference points for fisheries management. Can. Spec. Publ. Fish. Aquat. Sci. 120.

Caton, A.E. 1991. Review of aspects of southern bluefin tuna biology, population and fisheries. *In* Deriso, R.B. and Bayliff, W.H. (Eds.). World meeting on stock assessment of bluefin tunas: strengths and weaknesses. Inter-Am. Trop. Tuna Comm. Spec. Rep. 7: 181-357.

CCSBT 2000. Report of the special meeting. Canberra, Australia, 16-18 Nov 2000. Commission for the Conservation of Southern Bluefin Tuna doc.

CCSBT 2001. Report of the second meeting of the stock assessment group. Tokyo, Japan, 19-28 Aug 2001. Commission for the Conservation of Southern Bluefin Tuna doc.

CCSBT 2002. Report of the first Management Procedure Workshop, Tokyo, Japan, 3-4 & 6-8 March 2002. Commission for the Conservation of Southern Bluefin Tuna doc.

CCSBT 2003. Report of the second stock assessment group meeting 19-28 August 2001. Tokyo, Japan. Commission for the Conservation of Southern Bluefin Tuna doc.

FAO. 1996. FAO technical Guidelines for responsible fisheries 2 precautionary approach to capture fisheries and species introduction. Food and agriculture organization of the united nations. Rome. 54 p.

Fournier, D. and C.P. Archibald. 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39: 1195-207.

Fournier, D.A., J.R. Sibert, J. Majkowski, and J. Hampton. 1990. MULTIFAN: a likelihood based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data from southern bluefin tuna (*Thunnus maccoyii*). Can. J. Fish. Aquat. Sci. 47: 301-17.

Fournier, D.A., Hampton, J. and Sibert, J.R. 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat. Sci. 57: 1002-1010.

Haist, V., A. Parma and J. Ianelli. 2002. Initial specifications of operating models for southern bluefin tuna management procedure evaluation. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0209/7

Hampton, J. and D.A. Fournier. 2001. A spatially dis-aggregated, length-based, age-structured population model of yellowfin tuna (*Thunnus albacares*) in the western and central Pacific Ocean. Mar. Freshw. Res. 52: 937-963.

Hampton, J. and P. Kleiber. 2003. Stock assessment of yellowfin tuna in the western and central Pacific Ocean. Meeting of the Standing Committee on Tuna and Billfish 16 Working Paper YFT-1.

Hampton, J., P. Kleiber, Y. Takeuchi, H. Kurota, and M Maunder. 2003. Stock assessment of bigeye tuna in the western and central Pacific Ocean. Meeting of the Standing Committee on Tuna and Billfish 16 Working Paper BET-1.

Harley, J.H. and M.N. Maunder. 2003. Recommended diagnostics for large statistical stock assessment models. Meeting of the Standing Committee on Tuna and Billfish 16 Working Paper MWG-3.

Hilborn, R. and D.S. Butterworth. 1996. Fitting the 1994-1995 spawner age distribution data for SBT. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/96/34

Hilborn, R. and C.J. Walters. 1992. Quantitative fisheries stock assessment: choice, dynamics, and uncertainty. Chapman and Hall, New York.

Hilborn, R. and M. Mangel. 1997. The ecological detective confronting models with data. Monographs in population biology 28. Princeton University Press, Princeton.

Hiramatsu, K. and S. Tsuji. 2001. Stock assessment and future projection of the southern bluefin tuna based on the ADAPT VPA. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/01/08/31.

ICES. 1993. Report of the working group on methods of fish stock assessment. International Council for the Exploration of the Seas (ICES) Co-operative Research Report No. 191. Copenhagen.

Kolody, D. 2002. SCALIA: application of an integrated analysis stock assessment model to the 2002 SCTB methods working group simulated tuna fishery data. 15th meeting of the standing committee on tuna and billfish, Working Paper MWG-5.

Kolody, D., and D. Ricard. 2003. Application of SCALIA and production models (Fox, Schaefer, and age-structured) to the SCTB MWG 2003 simulated tuna fishery data. 16th meeting of the standing committee on tuna and billfish, Working Paper MWG-5.

Kolody, D. and K. Patterson. 1999. Evaluation of NE Atlantic mackerel stock assessment models on the basis of simulated long-term management performance. ICES Annual Science Conference CM 1999/S 1, 9 p.

Kolody, D. and P. Jumpannen. 2003. SCALIA simulation-estimation study results relevant to CCSBT management procedure development. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0304/10.

Kolody, D. and T. Polacheck. 2001. Application of a statistical catch-at-age and –length integrated analysis model for the assessment of southern bluefin tuna stock dynamics 1951-2000. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0108/13.

Kurota, H., S. Tsuji, N. Takahashi, K. Hiramatsu, and T. Itoh. 2001. Exploration of cohort analysis based on catch-at-length data for southern bluefin tuna. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0108/32.

Labelle, M. 2002. Testing the accuracy of MULTIFAN-CL assessments of the WCPO yellowfin tuna fishery conditions. Meeting of the Standing Committee on Tuna and Billfish 15, Working Paper MWG-1.

Labelle, M. 2003. Testing the accuracy of MULTIFAN-CL assessments using an operational model of yellowfin tuna (*Thunnus albacares*) fisheries in the western and central Pacific Ocean. Meeting of the Standing Committee on Tuna and Billfish 16, Working Paper MWG-1.

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

Langley, A., M. Ogura and J. Hampton. Stock assessment of skipjack tuna in the western and central Pacific Ocean. Meeting of the Standing Committee on Tuna and Billfish 16 Working Paper SKJ-1.

Lewy, P. and A. Nielsen. 2003. Modelling stochastic fish stock dynamics using Markov Chain Monte Carlo. ICES J. Mar. Sci., 60: 743-752.

Linhart, H. and W. Zuchini. 1986. Model selection. Wiley, NewYork.

Ludwig, D. and C.J. Walters. 1985. Are age-structured models appropriate for catch-effort data? Can. J. Fish. Aquat. Sci. 42:1066-1072.

Maunder, M.N. and G.M. Watters. 2003. A-SCALA: an age-structured statistical catch-at-length analysis for assessing tuna stocks in the Eastern Pacific Ocean. Inter-Am Trop. Tuna Com. Bull., Vol. 22, No. 5.

Maunder, M.N. 2002. Is it time to discard the Schaefer model from the stock assessment scientist's toolbox? Fish. Res., 61: 145-149.

Maunder, M.N. 2003. Paradigm shifts in fisheries stock assessment: from integrated analysis to Bayesian analysis and back again. Nat. Resour. Model., 16: 465-476.

McAllister, M.K. and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54: 284-300

MWG. 2002. Report of the methods working group. 15th meeting of the standing committee on tuna and billfish. Honolulu, U.S.A.

MWG. 2003. Report of the methods working group. 16th meeting of the standing committee on tuna and billfish. Mooloolaba, Australia.

NRC. 1998. Improving fish stock assessments / Committee on Fish Stock Assessment Methods, Ocean Studies Board, Commission on Geosciences, Environment, and Resources, National Research Council. National Academy Press. Washington, D.C.

Patterson, K., R. Cook, C. Darby, S. Gavaris, L. Kell, P.Lewy, B. Mesnil, A. Punt, V. Restrepo, D. Skagen and G. Stefansson. 2001. Estimating uncertainty in fish stock assessment and forecasting. Fish and Fisheries 2: 125-157.

Patterson, K.R. 1999. Evaluating uncertainty in harvest control law catches using Bayesian Markov chain Monte Carlo virtual population analysis with adaptive rejection sampling and including structural uncertainty. Can. J. Fish. Aquat. Sci. 56: 208-211.

Polacheck, T. 2002. Experimental catches and the precautionary approach: the southern bluefin tuna dispute. Mar. Policy 26: 283-294.

Polacheck, T., A. Preece, A. Betlehem, and N. Klaer. 1999. Treatment of Data and Model Uncertainties in the Assessment of Southern Bluefin Tuna Stocks. In: F. Funk, T.J Quinn II, J. Heifetz, J.N. Ianelli, J.E. Powers, J.F. Schweigert, P.J. Sullivan, and C.-I. Zhang (eds.), Fishery Stock

Assessment Models. Alaska Sea Grant College Program Report No. AK-SG-98-01, Univ. Alaska Fairbanks. pp 613-637.

Polacheck, T., A. Preece and D. Ricard. 2001. Assessment of the status of the southern bluefin tuna stock using virtual population analysis – 2001. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/01/08/20.

Polacheck, T., G.M. Laslett and J.P. Eveson. 2003a. An integrated analysis of the growth rates of southern bluefin tuna for use in estimating the catch at age matrix in the stock assessment. Final Report. FRDC Project 1999/104. ISBN 1 876996 38 2.

Polacheck, T., and A. Preece. 2001. An integrated statistical time series assessment of the southern bluefin tuna stock based on catch-at-age data. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/01/08/19.

Polacheck, T., D. Ricard, P. Eveson, M. Basson, D. Kolody and J. Hartog. 2003b. Results from further testing of candidate management procedures for southern bluefin tuna. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-ESC/0309/29.

Polacheck, T., D. Kolody and M. Basson. 2003c. Issues in the selection of final trials for testing SBT management procedures and for the process of synthesizing results from the simulation testing. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-ESC/0309/27.

Prager, M.H. and E.H. Williams. 2003. From the golden age to the new industrial age: fishery modeling in the early 21st century. Nat. Resour. Model., 16: 477-489.

Punt, A.E. 1996. The performance of VPA-based management. Fish. Res. 29: 217-243.

Quinn II, T.J. 2003. Ruminations on the development and future of population dynamics models in fisheries. Nat. Resour. Model., 16: 341-392.

Ricard, D. and D. Kolody. 2002. Application of production models to the assessment of the SCTB-MWG simulated tuna fishery data. 15th meeting of the standing committee on tuna and billfish, Working Paper MWG-8.

Ricard, D., D. Kolody, and M. Basson. 2002. Further exploration of biomass dynamics models for SBT stock assessment. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0209/28

Schnute, J.T. 2003. Designing fishery models: a personal adventure. Nat. Resour. Model., 16: 393-413.

Schnute, J.T. and L.J. Richards. 2001. Use and abuse of fishery models. Can. J. Fish. Aquat. Sci**. 58**: 10-17.

Schnute, J.T. and R. Hilborn. 1993. Analysis of contradictory data sources in fish stock assessment. Can. J. Fish. Aquat. Sci**. 50**: 1916-1923.

Schweder, T. 2001. Protecting whales by distorting uncertainty: non-precautionary mismanagement? Fish. Res., 52: 217-225.

Sibert, J. and J. Hampton. 2003. Mobility of tropical tunas and the implications for fisheries management. Mar. Policy 27: 87-95.

Stefansson, G. Issues in multispecies models. Nat. Resour. Model., 16: 415-437.

UN. 1994. The precautionary approach to fisheries with reference to straddling fish stocks and highly migratory fish stocks. In: United Nations conference on straddling fish stocks and highly migratory fish stocks (Proceedings of the UN conference New York, 14-31 Mar 1994. United Nations, New York. No. A/CONF 164 INF /8.

Walters, C.J. 1986. Adaptive management of renewable resources. Macmillian, New York.

8 APPENDICES

Note that some of the following appendices stand alone as self-contained documents, including their own references. Others cross-reference the main text of the SESAME report.

APPENDIX 1 VSM TECHNICAL DESCRIPTION

VSM (which is an abbreviation of *virtual stock model*) is a stand-alone Windows based command line application which simulates multi-species fisheries. It is a user specified operating model whose characteristics are user defined through a number of text input files. The VSM feature set is sufficiently rich to provide a wide range of possible model scenarios and behaviours. The central aim behind the development of VSM is the development of a flexible operating model well suited to critically evaluating the performance of stock assessment models. This aim is reflected in the model structure and feature set.

A1.1 MODEL OVERVIEW

VSM is a multi-area multi-species operating model. The model is split into two distinct components : the *system dynamics model* and the *observation model*. The system dynamics model produces data that reflects the true state of the fisheries model at any given point in time, or in other words the complete time history of the system behaviour. These data sets are referred to as the *state realisation*. The observation model acts as the observer, extracting the appropriate state realisation data and adding observation error consistent with the observation model specification. The observation model will typically report catch, effort and tagging statistics only[2]. These data sets are collectively referred to as a *data realisation*. The true state of the population is known within the state realisation but not within the data realisation. Assessment models applied to the data realisation attempt to determine the true population state and history.

A1.1.1 System Dynamics Model

The system dynamics model is designed to simulate the behaviour of a virtual or simulated (as opposed to real) fishery. The system dynamics model can take on a vast range of differing behaviours to reflect the range of *real* fisheries to which assessment models are currently being applied. VSM has been designed with flexibility in mind and is capable of supporting:

- Multiple areas
- Multiple species
- Multiple populations
- Area and sex specific biology
- Predation
- Fishing
- Tagging programs
- Migration

VSM allows tagging, fishing, migration and the like to be a function of time, fish age and fish length. Time variability may be specified as periodic which allows seasonal effects to be easily incorporated.

An *area* in VSM can be thought of as an arbitrary grouping (perhaps directly related to a geographical region) that contains populations of fish and can include fishing fleets and tagging programs. Although an area can be thought of as a physical space with a given size and position, within the model itself, it is simply a container without physical dimension. VSM is a bulk transfer model in which migration is controlled only by transfer coefficients (probabilities of moving from *A* to *B*). Any actual notion of relative size between differing areas is implied by the model specification and is not explicit. A clearer way to view areas is as independent boxes possibly connected via migratory flows as is illustrated in Figure A1.1, rather than a grid which implies a notion of physical dimension and place.

The number of areas defined within the system dynamics model is user specified. You may define as many areas as required, only limited by the amount of physical memory required to run the model. This in turn will depend upon the number of species, populations, fishing fleets and so forth defined within each area. Be aware that the more complex the model specification the longer it will take to run.

[2] Surveys were proposed as a feature of the model but never implemented. However, fishing gear surveys can be supported by defining a special type of fishery.

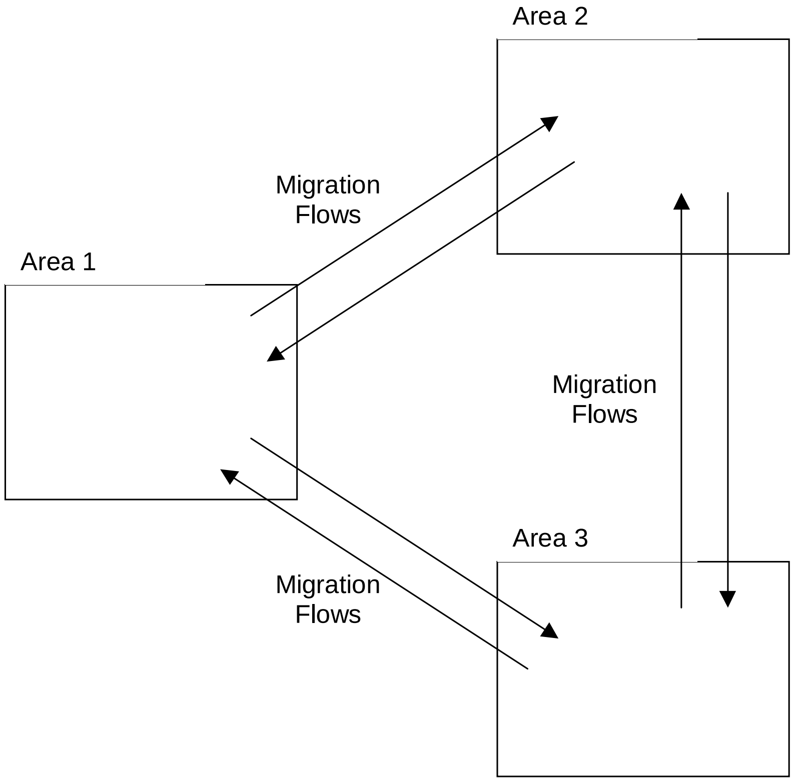


Figure A1.1: Graphical depiction of areas within VSM.

A *species* in VSM represents a specific kind of fish within the model. A *population* in VSM is a grouping of fish of the same species and populations generally exist within the *areas* of the model. The only case in which a population exists apart from an area is when it represents catch and mortality. Each population of fish is given a unique name (the *population name*) that identifies it. Individuals from a given population always remain within that population. In this way it is possible to represent multiple populations of the same species that are *traceable* (you can track the movement of a particular population). One useful application of this is to track the re-distribution of fish after migration. In the initial model state the populations in each area can be given unique names. Then when running the model the redistribution of fish from *Area 1* to the other areas is traceable. A simple illustration of this type of process is shown in Figure A1.2 which shows how the fish from an impulse tagging operation redistribute themselves over time.

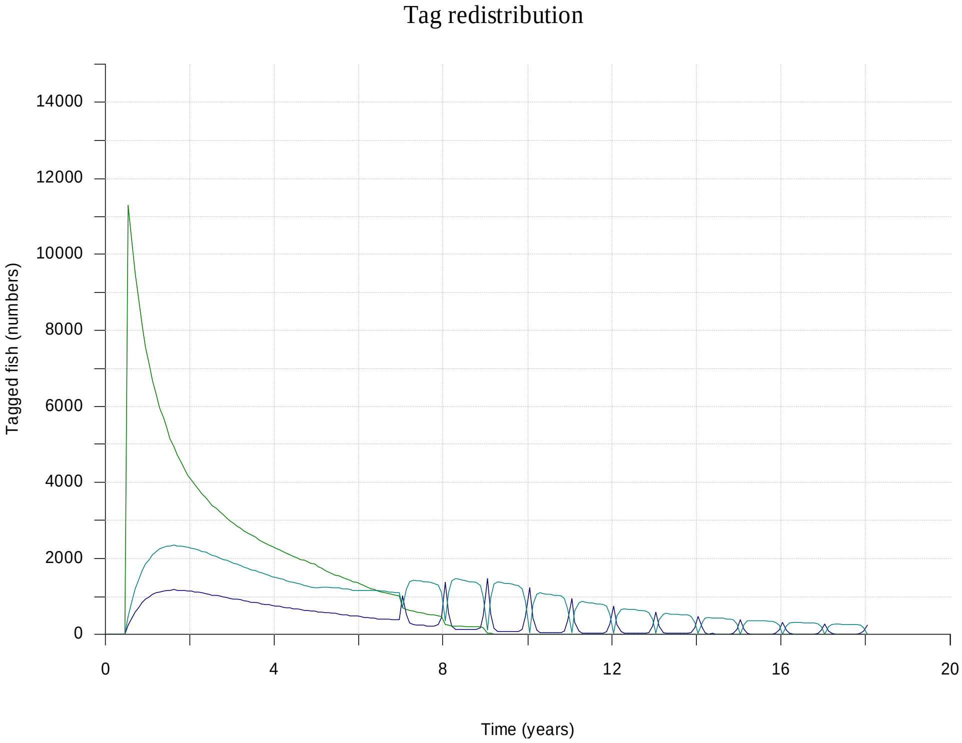


Figure A1.2: Population trajectory for a tagging event in a 3 area model (juvenile - green, spawning ground - blue, feeding ground - cyan) where tags released in the juvenile area only. The tagged fish then migrate out into the spawning ground and feeding ground areas. Note that the oscillation in migration after year 7 occurs because the tagged fish are reaching maturity and begin an annual migration between the feeding ground and spawning ground fisheries. No fishing mortality is present in this example. Being a naive migration model the transition to migratory behaviour is rather abrupt.

The act of fishing within the model is performed by *fishing fleets*. Fishing fleets exist within a given area and have a *fleet name* and a *fishery name* identifying the fleet and the fishery respectively. A fleet can target multiple species and may also have by-catch associated with targeted catch. Figure A1.3 shows the flows and relationships between populations and fleets within a given area. This basic relationship is common to all areas defined within a given system dynamics model.

Under the act of fishing it is worth noting that all identifiers are recorded in the state realisation (that is which *area*, which *species*, which *population*, which *fleet* and which *fishery* the catch came from) allowing a vast range of different information to be extracted using different filtering criteria. This theme is used extensively throughout VSM to provide flexibility in data extraction.

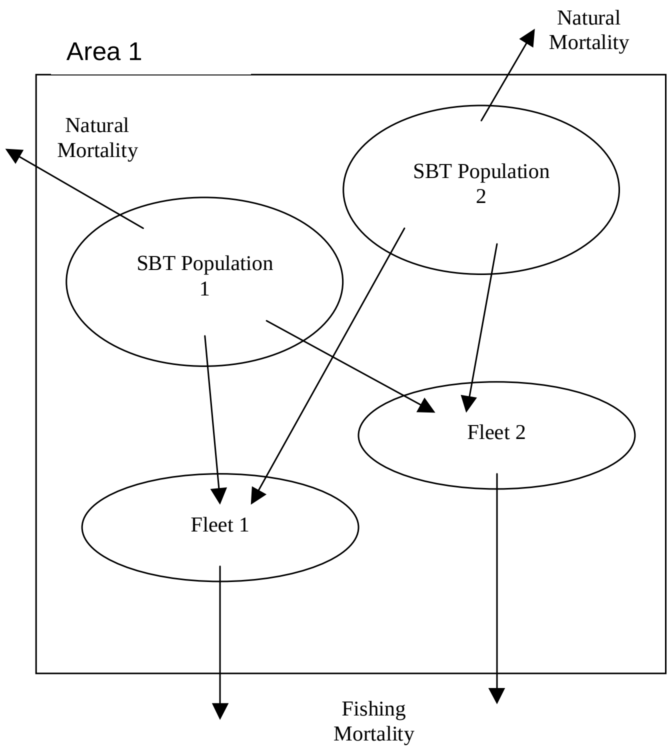


Figure A1.3: Population and fishing within a given area.

Tagging programs may also exist inside areas. Tagging programs are similar to fishing fleets in as much as they reside inside a given area and catch a target species, except in this case the catch is tagged and released back into the area. Once released the tagged population is then free to migrate to other areas.

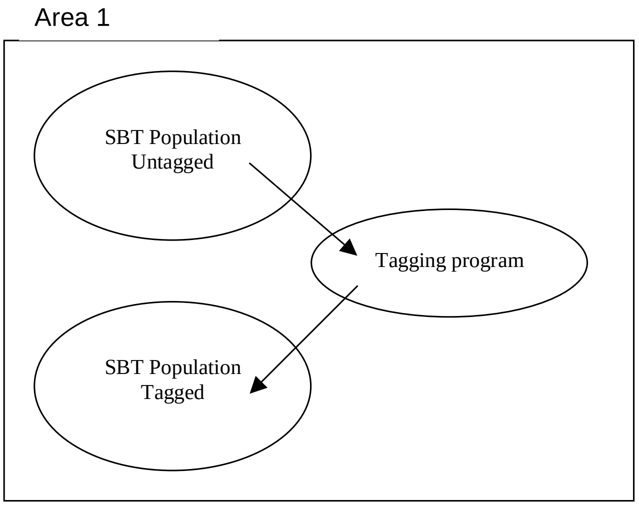


Figure A1.4: Tagging within a given area.

Tagged fish populations are created by extracting fish from an un-tagged population and placing them into a tagged population. The flows involved in this process are illustrated in Figure A1.4. Any number of tagging operations are allowed within a given area and each is identified by a unique *tagging operation name*. The un-tagged population name combined with the tagging operation name is used to synthesise a unique population name for the tagged fish. A tagging operation can only target a single species but given that multiple tagging operations are permissible this restriction imposes no problematic limitations.

Each population of a given species in a given area can have a unique biology. Furthermore, provided a population of the same name exists in other areas, any fish migrating to those areas will inherit the biology of those populations. Only when a given population does not exist in the migration destination does the migrating population keep its existing biology. For this reason if area specific biology is a requirement it is important that all *population classes* (species and population combination) are incorporated in every area, even if the initial population size is null. This action will ensure that the biology will change as expected under migration.

Area specific biology can be useful for implementing *productivity* effects and the like. For example, in the case of SBT, spawners on the spawning ground spend a brief but intense time breeding with minimal feeding, whereas on the feeding ground spend their time feeding. Therefore it is reasonable to expect growth to be significantly stronger on the feeding ground compared with the spawning ground. This effect can be modelled using area specific biology.

VSM is an aggregate population model rather than an individual based model. Aggregation can either be *age structured* (populations of a given species are grouped by age class) or *age and length structured* (populations are grouped by age and length class). In the *age and length structured* case the length classes are held in bins of fixed width and the length bins are user defined. Furthermore, it is a requirement in the *age and length structured* case that the length bins are identical for each compatible population class. More shall be said on this subject in the Model Specification section.

A1.1.2 Observation Model

The *observation model* is responsible for re-sampling relevant portions of the state realisation to produce a data realisation. In the re-sampling process, errors are introduced to simulate the sampling error found in real data sets. The observation model includes a number of mechanisms to introduce sampling error, some of which overlap. These mechanisms include:

- total catch numbers deviations
- effort deviations
- age and length distribution errors
- Age estimation via cohort slicing
- tag reporting rate errors

The outputs from the observation process can be directed to *comma separated text files* or *SCALIA / Multifan CL compatible text files*. Furthermore, summary statistics can be reported to text files or uploaded to an ODBC compliant database. All sampling errors are performed on a given species on a fleet by fleet basis. Thus individual fleets can have individual error regimes when targeting specific species.

Total catch numbers deviations are implemented through a log normal deviate. The CV of the log normal deviate can be an arbitrary function of time. Effort deviations are applied to each individual effort record recorded in the state realisation. The CV of the effort deviations can also be an arbitrary function of time.

Age and length sampling are simulated using a multinomial selection process. At any given time step the true age and length distribution is measured. This distribution combined with a user specified sample size is then used to synthesise a new distribution using the following process. The age and length distributions (in absolute catch numbers) are converted into probability coefficients. Using multinomial selection with these coefficients an artificial population distribution (in absolute numbers) is created through sampling the user defined sample size. The length distribution sample is then reported as-is, whereas the age distribution is re-scaled to sum to the total catch. This discrepancy in behaviour for age data is a consequence of the file format used to report the catch at age data (total catch is reported through a catch at age matrix that reports the catch age distribution).

Tag reporting rate errors are introduced through a binomial selection process. As with the total catch numbers and the effort deviations, the reporting rate errors can be a function of time.

A1.2 System Dynamics Model Implementation Details

A major aim in designing VSM was to produce an operating model with sufficient flexibility to allow for a diverse range of possible models with varying degrees of complexity. To considerably simplify the process of coding this model we chose to split the model into a series of independent processes (recruitment, natural mortality, tagging, fishing, predation, migration and aging/growth) that could be executed serially.

The model uses a finite difference approximation for these continuous processes and may be substantially biased[3]. In particular, the processes of fishing mortality, natural mortality and migration, which are executed serially, present a potential source of bias. However, if the model time step is sufficiently small then these biases are insignificant from an assessment point of view. For annually reported data it was found that a monthly model time step was sufficient to avoid the order bias problem.

The processing order within the model is illustrated in Figure A1.5 below.

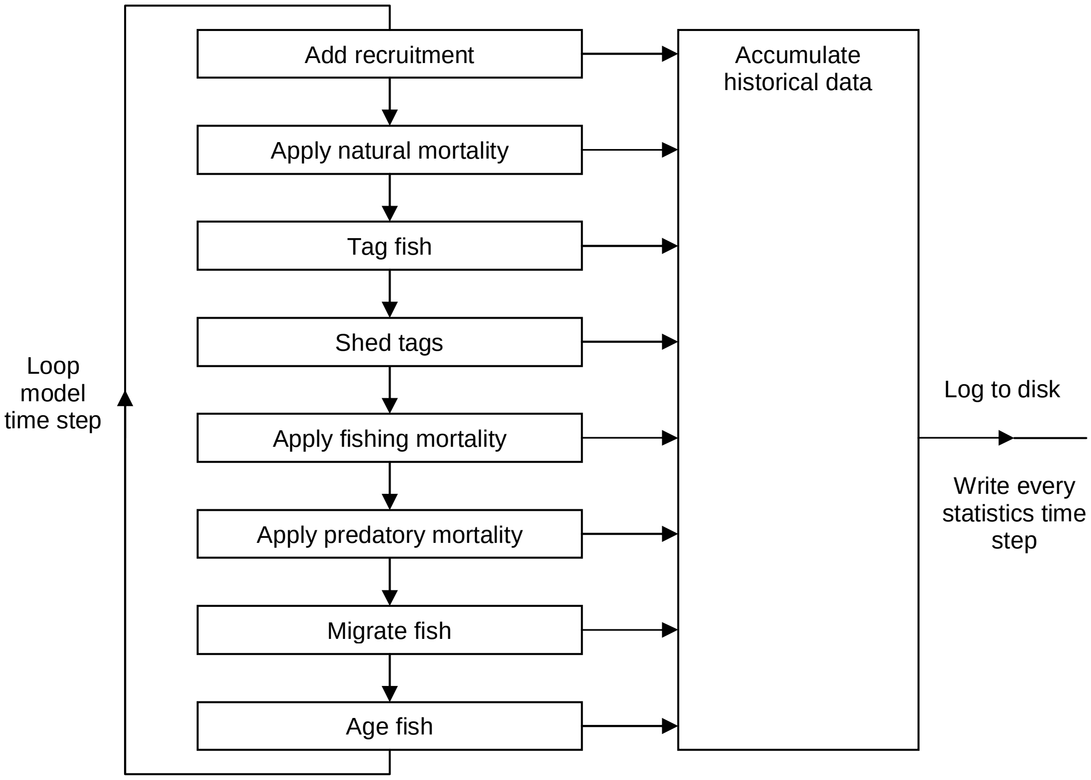


Figure A1.5: An illustration of the processing order within VSM

[3] The bias manifests itself as the first to run removal process taking a disproportionately large portion of the population compared with the processes to follow. This occurs because the first removal process depletes the population, reducing the number returned in subsequent binomial selection. If the impact on the total population is small enough such that the relative population size does not change significantly in a given time step then the bias is negligible.

During a given model iteration (one model time step) historical data from each of the processes is collated in memory. When the duration of time that the historical data has been accumulated equals the statistics time step, the collated statistics are then written to file. Generally speaking, the model time step should be less than the statistics time step. In the models developed for SBT assessment model testing the model time step was chosen to be monthly and the statistics time step yearly.

A1.2.1 Recruitment

VSM supports a number of different recruitment relationships. The recruitment relationship predominantly used in the documented scenarios is the Beverton-Holt stock recruitment relationship, and is defined as,

$$R = \frac{d\,SSB}{b + SSB} \qquad \text{(A1.1)}$$

where *d* and *b* are the Beverton-Holt parameters, *R* is the number of recruits and *SSB* is the spawning stock biomass. Within the model specification the parameters *d* and *b* are re-parameterised in terms of the virgin recruitment R_0 and the steepness *h*. During the start-up phase, the model performs a stable stock projection using the virgin recruitment and the natural mortality specification to provide an estimate for the virgin spawning stock biomass SSB_0. The Beverton-Holt parameters are then found using,

$$d = \frac{4hR_0}{5h-1} \qquad \text{(A1.2)}$$

and,

$$b = \frac{SSB_0(1-h)}{5h-1} \qquad \text{(A1.3)}$$

In the projection the spawning stock biomass is calculated from the growth equation and length-weight relationship combined with an age based maturity vector. That is,

$$SSB = \sum_{a=0}^{Max\,a} w(l_a)Q_a N_a \qquad \text{(A1.4)}$$

where l_a is the length at age *a*, N_a is the numbers of fish at age *a*, *w(l)* is the mass-length relationship and Q_a is the maturity at age *a*. A power mass-length relationship is used and is defined by,

$$w(l) = q\,l^v \qquad \text{(A1.5)}$$

where *v* is the allometric growth parameter and *q* is a scaling factor. The length at age relationship used in VSM is the VB log-k growth equation (Laslett, G.M., Eveson, J.P., and Polacheck, T. 2002), which is defined as,

$$l(t) = L_\infty\left(1 - e^{-k_2(t-t_0)}\right)\left[\frac{1 + e^{-\beta(t-t_0-\alpha)}}{1 + e^{\alpha\beta}}\right]^{-\frac{k_1-k_2}{\beta}} \;;\; \sigma(t) = \sigma_\infty \frac{l(t)}{L_\infty} \qquad \text{(A1.6)}$$

where σ_∞ is the standard deviation of the asymptotic average length, L_∞ is the asymptotic average length, k_1 and k_2 are growth rate coefficients, α and β are transitional parameters and t_0 is the hypothetical age at which mean length is zero. Note that if $k_1 = k_2$ then the V-B log *k* equation degenerates into the Von-Bertalanffy growth equation. The length at age is assumed to be normally distributed. Within the model we often deal with the age on a discrete basis, in which case we define,

$$l_a = l(at_m) \ ; \ \sigma_a = \sigma(at_m) \qquad \text{(A1.7)}$$

where t_m is the model time step and a is the age index. Within the model run itself, the mean recruitment of equation A1.1 is perturbed by an optional normal or log-normal deviate. This deviate also has the possibility of including correlation effects through a first order difference equation as detailed below.

A1.2.2 Generalized Correlated Deviate

Within VSM, normal and log-normal deviates, either correlated or uncorrelated, are used in a number of places to add process noise to the operating model. Here we define the exact nature of these deviates.

Let ρ be the parameter governing degree of auto-correlation, let n_k be a gaussian white sequence with zero mean and standard deviation σ, let β be the bias of the deviate (i.e. non-zero mean), let μ be the mean of the deviate and let m_k be an intermediate correlated gaussian sequence defined by,

$$m_k = \rho\, m_{k-1} + \sqrt{1-\rho^2}\; n_k \qquad \text{(A1.8)}$$

Then we can define the normal deviate case Δ_k as,

$$\Delta_k = m_k + \mu + \beta \qquad \text{(A1.9)}$$

and the log normal deviate case as,

$$\Delta_k = \mu\, e^{m_k+\beta} \qquad \text{(A1.10)}$$

Furthermore, if the log normal deviate is bias compensated then the log normal deviate case becomes,

$$\Delta_k = \mu\, e^{m_k+\beta-\frac{\sigma^2}{2}} \qquad \text{(A1.11)}$$

Due to the scaling applied to n_k in m_k, the standard deviation of m_k is σ.

A1.2.3 Natural Mortality

The processes of natural mortality, predatory mortality, fishing mortality, and migration are all modelled using binomial selection (binomial random variates) driven by the probability of selection. The nature of that selection probability is tied to the particular process: in natural mortality it is the probability of death, whereas in fishing it is the probability of capture.

In the model, natural mortality is specified through a mortality vector, which is the equivalent probability of death in the time period of one year for a particular age, length and sex category at a particular point in time (the mortality vector can change with time). Internally VSM will re-scale the user specified probabilities to correspond to the internal model time step by using the transforming function,

$$\Omega_{t_m,t_y}(P) = 1-(1-P)^{\frac{t_m}{t_y}} \qquad \text{(A1.12)}$$

where P is the probability based on a time period of one year, t_m is the model time step and t_y is the time period corresponding to one year. Note that in the case of user specified catch probabilities the re-scaling process is delayed until after the baseline probability is scaled by effort and perturbed by effort deviations. Note that t_y is an independent parameter because the time unit specification is user defined, and by doing so flexibility is added to the process of defining the model. For example, the model could be parameterised using months as the fundamental unit of time rather than years. In our case the models developed using VSM were all parameterised using years as the unit of time running under a monthly model time step, giving,

$$\Omega_{\frac{1}{12},1}(P) = 1 - (1-P)^{\frac{1}{12}} \qquad \text{(A1.13)}$$

If $P^{m}_{a,l,s,t}$ is the probability of death of fish age *a*, length *l*, time *t* and sex *s*, and $N_{a,l,s,t}$ is the number of fish of age *a*, length *l* and sex *s* in the population, then the expected number of fish deaths at time *t* and of sex *s* is given by,

$$\overline{M}_{s,t} = \sum_{a=0}^{Max\,a} \sum_{l=0}^{Max\,l} N_{a,l,s,t} \Omega_{t_m,t_y}\left(P^{m}_{a,l,s,t}\right) \qquad \text{(A1.14)}$$

For the binomial variate that implements this mortality process the number of independent trials is $N_{a,l,s,t}$. In the fishing process the number of independent trials can be reduced by an effective sample size to add noise to the process (see below).

A1.2.4 Fishing Mortality

For fishing mortality the process is similar to natural mortality except in this case the probability of capture is scaled through the exertion of fishing effort[4]. Process noise is also included in the form of minimum effective sample sizes and catchability deviates. Also note that fishing mortality in the model is implemented as a predation process, so the analysis described here for fishing mortality applies equally to predator-prey derived mortality. The main difference between predation and fishing mortality is that the deaths due to predation are not reported / available to assessment models. Although predation is supported it is largely untested as this feature was not used within this project.

In the model, fishing mortality is specified through an effort series E_t, a fishing catchability series q_t, fishing catchability deviate standard deviation σ_t, bias β_t and auto-correlation parameter ρ, and a baseline catch probability vector $S_{a,l,s,t}$ which is a function of age *a*, length *l*, sex *s* and time *t*. The baseline catch probability is notionally the probability of catch for a fishing period of 1 year with unity fishing effort and efficiency. The baseline catch probability vector can be altered by an internal selectivity altering function $\zeta_{a,t}$ such that,

$$S'_{a,l,s,t} = \zeta_{a,t}\left(S_{a,l,t,s}\right) \qquad \text{(A1.15)}$$

The transformed value of *S* is used to determine the unscaled probability of catch through,

$$P^{f}_{a,l,s,t} = \left(\Delta_{E_t q_t,\sigma_t,\beta_t,\rho}\right)^{\eta} S'_{a,l,s,t} \qquad \text{(A1.16)}$$

where Δ is the generalised correlated deviate (defined above) with a mean of $E_t q_t$, a standard deviation of σ_t, a bias of β_t and an auto-correlation parameter of ρ, and η is a non-linearity parameter. This unscaled probability is based upon a fishing period of one year. In the implementation of fishing mortality the probability of capture is re-scaled according to equation A1.12 to yield the probability of capture for a single model time step. In other words,

$$P'^{f}_{a,l,s,t} = \Omega_{t_m,t_y}\left(\left(\Delta_{E_t q_t,\sigma_t,\beta_t,\rho}\right)^{\eta} S'_{a,l,s,t}\right) \qquad \text{(A1.17)}$$

where t_m is the model time step, t_y is the time period corresponding to one year and P'^{f} is the scaled probability of capture. The fishing mortality is applied using binomial selection. The catch (in numbers) at time *t* and of sex *s* has an expected value given by,

[4] Note that under very high fishing mortality this discrete time approximation will show significant departure from the continuous time process it is modeling. Under normal circumstances, provided the model time step is sufficiently small, this problem should not eventuate.

$$\overline{C}_{t,s} = \sum_{a=0}^{Max\,a} \sum_{l=0}^{Max\,l} N_{a,l,s,t} P'^{f}_{a,l,s,t} \qquad \text{(A1.18)}$$

The binomial selection process for fishing selects from the full population $N_{a,l,s,t}$. If an effective sampling size is specified then the effect is simulated using the following process. Let us assume that the model specifies an effective sample size of ψ and a minimum effective sample size of ψ_{min}. Then if $\Phi(p,N)$ is a binomial random deviate with probability p and number of independent trials N, a new probability of capture is calculated using,

$$\psi'_{a,l,s,t} = Max\left\{ \psi \frac{N_{a,l,s,t}}{N} + 1; \psi_{\min} \right\}$$

$$P''^{f}_{a,l,s,t} = \frac{\Phi\left(P'^{f}_{a,l,s,t}, \psi'_{a,l,s,t}\right)}{\psi'_{a,l,s,t}} \qquad \text{(A1.19)}$$

where N is the total fish in the population and ψ' is the scaled effective sample size. By using this approach the effective sample size nominally scales according to the population structure. The total fish actually caught by this stochastic binomial process at time t of sex s is therefore,

$$C_{t,s} = \sum_{a=0}^{Max\,a} \sum_{l=0}^{Max\,l} \Phi\left(P''^{f}_{a,l,s,t}, N_{a,l,s,t}\right) \qquad \text{(A1.20)}$$

Note that effective sample size is used in part to represent the variability in selectivity due to stochastic processes (eg. spatial effects at a resolution finer than the spatial structure).

A1.2.4.1 Selectivity Changes

As mentioned before, a mechanism is in place to incorporate selectivity changes into the fishing through a special function $\zeta_{a,t}$, which transforms the baseline catch probability into a new one that incorporates a selectivity change based on age structure. Two change mechanisms are supported: a *constant catch proportion* mechanism and a *cohort targeting* mechanism.

The constant catch proportion mechanism operates around the notion that a particular catch age distribution is desirable and the fleet can modify the age selectivity in order to obtain it (eg. The Australian tuna farming industry claims preferential targeting of 3 year olds). Rather than directly specify a target catch distribution VSM obtains the distribution through the baseline catch probability by nominating a fixed period (say two years for example) of unregulated fishing to establish a catch profile statistic. Once established, this baseline distribution is then used to regulate the catch profile.

In the constant catch proportion case the selectivity transform $\zeta_{a,t}$ is defined as,

$$\zeta_{a,t}(p) = p\left(\lambda \frac{H_{a,t}}{\overline{H}_{a,t}}\right)^{\alpha} \qquad \text{(A1.21)}$$

where λ is a scaling factor, $H_{a,t}$ is the compensating selectivity at age a and time t, $\overline{H}_{a,t}$ is its mean, p is the catch probability being transformed, and α is a controlling parameter that dictates the degree of constant catch proportion control. With α set to 1, selectivity is fully controlled and with α set to 0 selectivity is uncontrolled (i.e. constant). To define how $H_{a,t}$ is actually arrived at let us define the total catch at age a over a period of one year starting at sample time t as $C^{y}_{a,t}$. Then the reference catch, which is used to arrive at a new selectivity, is defined as,

$$C_a^{yref} = \sum_{x=\xi(a)}^{\xi(a)+r-1} \sum_{n\in\tau} C_{x,nr}^{y} \tag{A1.22}$$

where τ is the set of years with non-zero catch used in compiling the reference catch,

$$r = \frac{t_y}{t_m} \tag{A1.23}$$

where t_m divides exactly into t_y (ie. r is an integer) and,

$$\xi(n) = a \bmod r \tag{A1.24}$$

where the **mod** operator returns the whole part of the division of two numbers. The added complexity of Equation A1.22 comes about because two successive age indices, say a and a+1, represent a physical age difference of t_m whereas $C_{a,t}^{yref}$ represents the catch for one whole year aggregated on yearly age intervals, hence the outer summation. Following on, the compensating selectivity is then defined as,

$$H_{a,t} = \frac{C_a^{yref}}{C_{a,t}^{yest}} \tag{A1.25}$$

where $C_{a,t}^{yest}$ is the estimated total catch at age based on the average baseline catch probability (that is with α= 0 in equation A1.21) over a year, the average natural mortality over a year and the initial population structure at time t, or in other words,

$$C_{a,t}^{yest} = N_{a,t}^{tot}\left(1 - M_{a,t}^{tot}\right)S_{a,t}^{tot} \tag{A1.26}$$

where,

$$N_{a,t}^{tot} = \sum_{n=\xi(a)}^{\xi(a)+r-1} N_{n,t} \tag{A1.27}$$

$N_{n,t}$ is the number of fish of age n at time t,

$$M_{a,t}^{tot} = \frac{1}{2} \sum_{s\in\left\{\begin{matrix}Male,\\ Female\end{matrix}\right\}} \sum_{n=\xi(a)}^{\xi(a)+r-1} P_{n,l_n,s,t+n-\xi(a)}^{m} \tag{A1.28}$$

and,

$$S_{a,t}^{tot} = \frac{1}{2} \sum_{s\in\left\{\begin{matrix}Male,\\ Female\end{matrix}\right\}} \sum_{n=\xi(a)}^{\xi(a)+r-1} S_{n,l_n,s,t+n-\xi(a)} \tag{A1.29}$$

where r is the number of model time steps in a year, P^m is the un-scaled natural mortality probability, S is the un-scaled baseline catch probability and l_a is the mean length at age as defined by equation A1.7.

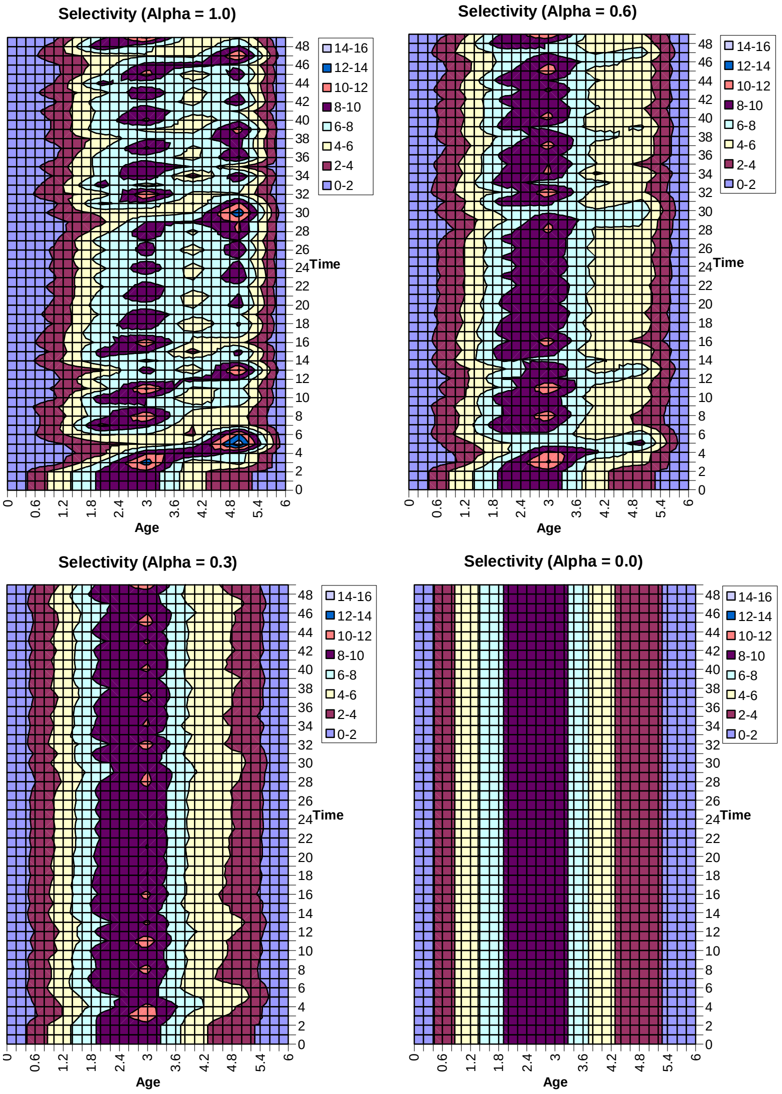


Figure A1.6: Selectivity contours arising from a typical model run with an α values of 0.0, 0.3, 0.6 and 1.0 using the constant catch proportion case.

Figures A1.6 and A1.7 show illustrative examples of this mechanism at work. Figure A1.6 shows the changing selectivity used to maintain a constant catch proportion for a range of values of α. Figure A1.7 shows the correspond catch proportions (standardised catch distribution) actually caught using the selectivity of Figure A1.6. From Figure A1.7 it can seen that for small α the catch proportion varies significantly whereas for α approaching 1 the catch proportion is well regulated. Note that the regulation is not perfect. The imperfection stems from the fact that the *estimated required selectivity* for next years catch does not predict the stochasticity in recruitment, fishing and natural mortality. As such the estimated catch of Equation A1.26 is imperfect and results in an imperfectly regulated catch distribution. If the model is run without stochasticity the regulation is precise.

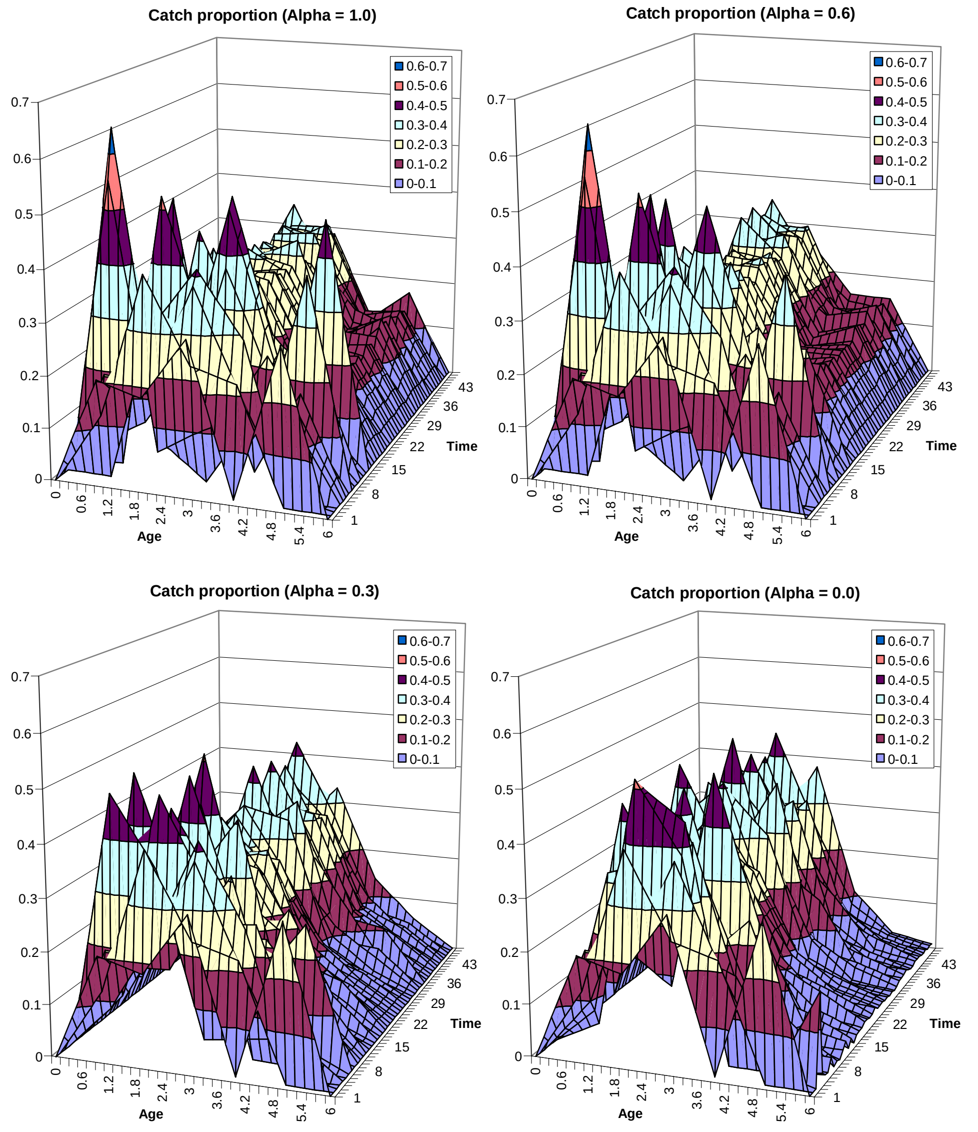


Figure A1.7: Normalised catch proportion by age arising from a typical model run with an α values of 0.0, 0.3, 0.6 and 1.0 using the constant catch proportion case.

In contrast to constant catch proportion, the cohort targeting mechanism works by weighting the baseline selectivity with a weighting function derived from the population biomass distribution in the fishery. This idea is intended to represent the fact that desirability of age/size classes change in relation to the age structure and more fish can likely be caught with less effort if abundant ages/sizes are disproportionately targeted. To give added flexibility the weighting function is also shaped by a user specified double sigmoid (see equation A1.34), giving the model the ability to restrict targeting changes to certain age groups.

In the cohort targeting case the selectivity transform $\zeta_{a,t}$ is defined as,

$$\zeta_{a,t}(p) = p^{1-\alpha}\left(\lambda \frac{H_{a,t}}{\overline{H}_{a,t}}\right)^{\alpha} \tag{A1.30}$$

where λ is a scaling factor, $H_{a,t}$ is the targeting selectivity at age *a* and time *t*, $\overline{H}_{a,t}$ is its mean and α is a controlling parameter that dictates the degree of cohort targeting. With α set to 1, selectivity is entirely determined by the cohort targeting mechanism and with α set to 0 selectivity is entirely determined by the baseline catch probability vector. $H_{a,t}$ is defined as,

$$H_{a,t} = \left(\omega(a) \frac{\sum_{n=0}^{A} B_{a,t}\, e^{-\frac{\left(\frac{l(\xi(a)t_y)-l(nt_y)}{l_{avg}}\right)^2}{\eta}}}{\sum_{n=0}^{A} e^{-\frac{\left(\frac{l(\xi(a)t_y)-l(nt_y)}{l_{avg}}\right)^2}{\eta}}} \cdot \frac{1}{Max\{B_{a,t}\}} \right)^{\delta} \qquad \text{(A1.31)}$$

where $l(a)$ is the mean fish length at age described by equation A1.6, A is the maximum age in years within the targeting selectivity vector $H_{a,t}$,

$$B_{a,t} = \sum_{s \in \left\{ \begin{matrix} Male, \\ Female \end{matrix} \right\}} \sum_{n=\xi(a)}^{\xi(a)+r-1} w(l_n) N_{n,l_n,s,t} \qquad \text{(A1.32)}$$

is the total population biomass for stock of $\xi(a)$ years old, $w(l)$ is the length-mass relationship of equation A1.5,

$$l_{avg} = \frac{1}{A+1} \sum_{n=0}^{A} l(nt_y) \qquad \text{(A1.33)}$$

is the average length of all lengths in an age cohort and,

$$\omega(a) = \frac{1}{\left(1+e^{\frac{(\phi_1-\xi(a)t_y)}{\eta_1}}\right)\left(1+e^{\frac{-(\phi_2-\xi(a)t_y)}{\eta_2}}\right)} \qquad \text{(A1.34)}$$

is a double sigmoid shaping function where ϕ_1 and ϕ_2 are the low and high cutoff age parameters respectively and η_1 and η_2 are the low and high transition parameters respectively.

Figure A1.8 shows illustrative examples of this mechanism at work. This figure shows the selectivity as a function of age and time for a range of α values. With α set to 0, the selectivity shows no variation in time except for an explicit baseline selectivity change between year 43 and year 44 (for contrast). With increasing values of α the selectivity shows progressively more selectivity variation, with high selectivity coinciding with strong cohorts. This effect is strongest for α equal to 1, where the selectivity is entirely driven by the population structure. For α values in between the effect is a combination of both the baseline selectivity and cohort targeting. Even with α equal to 0.6 the baseline selectivity change is clearly noticeable.

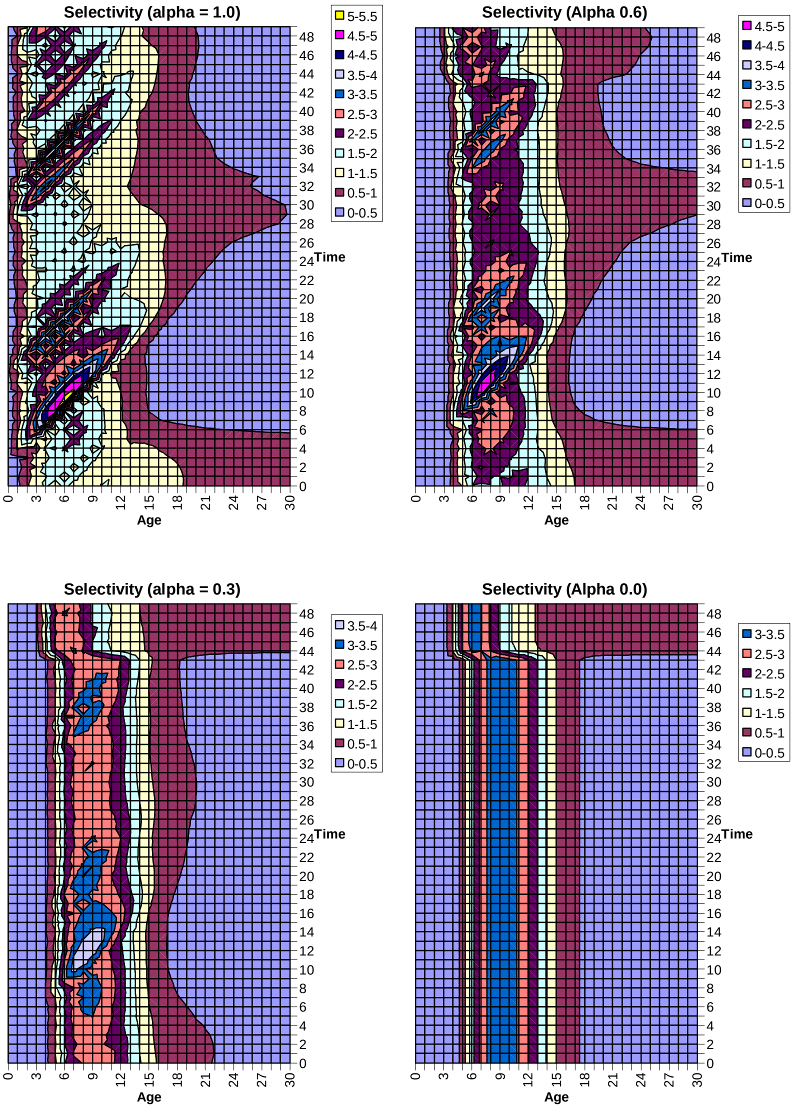


Figure A1.8: Selectivity contours arising from a typical model run with an α values of 0.0, 0.3, 0.6 and 1.0 using the cohort targeting case.

A1.2.5 Tagging

Tagging operations in VSM are implemented deterministically for the sake of implementation simplicity (it is easier to implement the code that will tag the exact population size requested if implemented as a deterministic process). For this reason the raw tagging data in VSM gives perfect knowledge of the stock structure of the tagged ages/sizes irrespective of the observational error specification in the observational model. Note that tagged fish are actually removed from the main population allowing very large tagging programs (or small populations) to be modelled. The tagged fish are then managed in their own population identifying them as being tagged.

Tagging releases are specified by naming the total number of fish in a particular tagging category that should be tagged (user specified). A tagging category can include age, length, sex and time as classifying variables and a given tagging operation is bound to a given region. Region specific tagging requires a tagging specification for each tagging region. At a minimum, a tagging operation will usually classify by age and time. Tagging is then implemented by collating the population that fits into the category and using this population data to create a tagged population structure with the same distribution. Mathematically then,

$$T_{a,l,s,t} = \frac{T_t^0 N_{a,l,s,t}}{N_t^T} \tag{A1.35}$$

where T is the number of tags at age a, length l, sex s and time t, T_t^0 is the total number of tags requested in the tagging operation at time t, $N_{a,l,s,t}$ is the number of fish in the population of age a, length l, sex s at time t, and N_t^T is the total number of fish in the population that fit into the tagging operation category. That is,

$$N_t^T = \sum_{a \in A_T ; l \in L_T ; s \in S_T} N_{a,l,s,t} \tag{A1.36}$$

where A_T, L_T and S_T are the sets of all ages, lengths and sexes that simultaneously satisfying the tagging operation at time t.

Once fish are tagged they are tracked internally as being tagged by assigning these fish to a distinct population which is subject to the same fishing pressures as the parent population. In this way the tags can and will be recovered through the act of fishing. Furthermore, tag shedding may also be specified, in which case a certain proportion of the tagged population will lose their tags over time and be returned to the parent population. In other words, if P'^{shed} is the re-scaled probability of tag shedding (re-scaled from the baseline probability according to equation A1.12) and $T_{a,l,s,t}$ is the number of tagged fish of age a, length l and sex s at time t, then the number of tags shed of sex s at time t is given by,

$$N_{s,t}^{shed} = \sum_{a=0}^{Max\,a} \sum_{l=0}^{Max\,l} T_{a,l,s,t} P'^{shed}_{a,l,s,t} \tag{A1.37}$$

Tagged populations can migrate to other areas and will remained within their tagged populations in each destination area. Should the biology be area specific and you wish to have this area specific biology reflected in the tagging data then it is necessary to define empty tagged populations with appropriate biology within each destination area. Only then will the tagged population biology change accordingly with area.

A1.2.6 Aging

To describe the process of aging and growth we first need to define how VSM maintains age and length structure internally. For any given age a, the model maintains a fixed size vector containing the numbers of fish in a particular length bin. The position and size of the length bins are pre-determined in the model specification files and are the same for each age class within the model.

More formally, let us assume that the length bin boundaries are specified by the vector **D** containing m elements in numerically ascending order, the n^{th} element of which we represent as d_n. Then the length boundary vector (transposed) is defined as,

$$D^T = [d_0, d_1, d_2, ..., d_{m-1}] \tag{A1.38}$$

Now let $\mathbf{N}_{a,s,t}$ be the vector that contains the total number of fish within each bin of a given cohort or age a, sex s at time t, where the range of each bin lies between two successive lengths. That is,

$$\mathbf{N}_{a,s,t}^T = [N_{a,0,s,t}, N_{a,1,s,t}, N_{a,2,s,t}, ..., N_{a,l,s,t}, ..., N_{a,m,s,t}] \tag{A1.39}$$

where $N_{a,l,s,t}$ is the number of fish whose length lies in the range $[d_l, d_{l+1})$ and the last element $N_{a,m,s,t}$, is the number of fish whose length lies in the range $[d_m, \infty)$. For the deterministic case, growth within the model is represented by,

$$\mathbf{N}_{a+1,s,t+1} = \mathbf{G}\mathbf{N}_{a,s,t} \tag{A1.40}$$

where **G** is a square growth transition matrix of dimension $m+1$. From equation A1.40 the numbers at length n is the inner product,

$$N_{a+1,n,s,t+1} = \sum_{l=0}^{m} g_{n,l} N_{a,l,s,t} \tag{A1.41}$$

where $g_{n,l}$ is the growth transition matrix coefficient at row n and column l.

The model can also operate in a stochastic mode whereby the growth is still represented by the transition matrix **G** but that the new length distribution is determined using multinomial selection instead of vector multiplication. The establishment of growth then reduces to determining suitable entries for the growth transition matrix, **G**.

The specification of **G** is handled in one of two ways : (1) strict adherence to the length distribution at any given age, as specified by the growth equation of equation A1.6 or (2) an incremental approach whereby we attempt to approximate the change in length and variance between two successive ages, as determined through the growth equation. For method 1 the model has no memory of length selective depletion since after each aging cycle the length distribution returns to the distribution specified in the growth equation. Method 2, on the other hand, has some memory of selective depletion, as growth is handled incrementally using the previous length distribution as a starting point. Method 2 is useful for exploring size selective mortality effects.

A1.2.6.1 Method 1: The absolute growth method

In the absolute growth method prior history of the population is ignored. This is achieved through having the column vectors of **G** being identical and only determined by the length distribution prescribed by the growth equation. In other words,

$$g_{n,l} = g_{n,l+i} = g_n \text{ for all } i \tag{A1.42}$$

where,

$$g_n = \int_{d_n}^{d_{n+1}} \Phi(l)dl \tag{A1.43}$$

and $\Phi(l)$ is the probability density function that describes the length distribution of the cohort. In the model this distribution is defined by the growth equation (equation A1.6) or by a user defined length distribution variance specified as a function of age. This is the most common method of representing growth-at-age structure within an assessment model.

A1.2.6.2 Method 2 : The differential growth method

In the differential growth method, a heuristic approach is taken to implement an algorithm that includes prior growth history. It is only an approximation of what would happen in an individual-based representation, and the approximation can be poor if length bins are coarse relative to the expected growth in one time step. The approach revolves around the notion that a change in distribution mean and variance can be simulated through geometrically transforming the parent distribution (assuming that the length at age is normally distributed). For example, an increase in mean and variance can be simulated by simultaneously translating and stretching the current distribution and then re-assigning the bin totals, as illustrated in Figure A1.9.

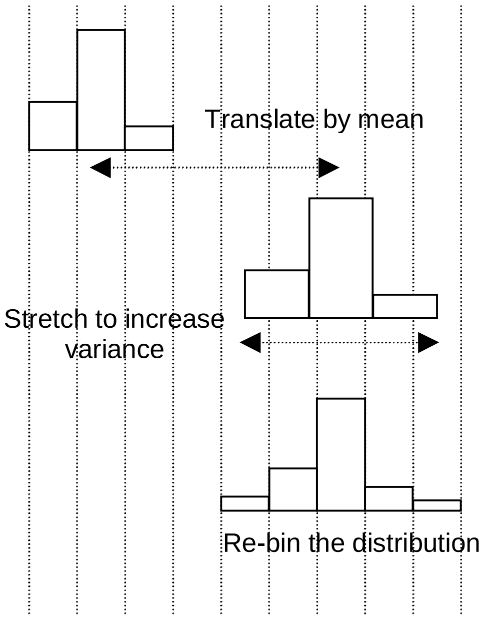


Figure A1.9: A graphical illustration of the differential growth method used in VSM

More rigorously, let δ_a^l and δ_a^σ be the change in mean length and variance experience by a fish aging from *a* to *a*+1. That is,

$$\delta_a^l = l_{a+1} - l_a \qquad \text{(A1.44)}$$

and

$$\delta_a^\sigma = \sigma_{a+1} - \sigma_a \qquad \text{(A1.45)}$$

where l_a and σ_a are defined in equation A1.7. Then the translated and stretched length bin boundaries are described by,

$$d'_n = l_a + \delta_a^l + \left(d_n - l_a\right)\left(\frac{\sigma_a + \delta_a^\sigma}{\sigma_a}\right)\theta_{n,a} \qquad \text{(A1.46)}$$

where,

$$\theta_{n,a} = \frac{1}{\dfrac{\nu\kappa_{n,a}}{\left(1-\kappa_{n,a}\right)} + 1} \qquad \text{(A1.47)}$$

and,

$$\kappa_{n,a} = \left|\frac{\delta_a^l}{d_{n+1} - d_n}\right| \bmod 1 \qquad \text{(A1.48)}$$

Equation A1.47 is a variance compensation term added to reduce the effect of the variance expansion that occurs when using this approach. The term ν is a scaling factor used to *fine tune* the variance compensation. In the simulated SBT case a ν value of 0.1 was found (experimentally) to give good performance (low variance creep).

Given transformed length bin boundaries we can determine the growth transition matrix coefficients by overlaying the transformed bin onto the length bin boundaries in **D** and proportioning each bin total into the bins underneath (assuming a uniform distribution within each bin). That is,

$$g_{n,l} = \begin{cases} \dfrac{\min(d_{l+1}, d'_{n+1}) - \max(d_l, d'_n)}{d'_{n+1} - d'_n}; \min(d_{l+1}, d'_{n+1}) > \max(d_l, d'_n) \\ 0; otherwise \end{cases} \tag{A1.49}$$

A1.2.7 Migration

Up to this point all the analysis has dealt with fish populating a single region. In models supporting migration an added dimension is needed to represent the region and on that basis we add an extra subscript to our definitions. That is $N_{i,a,l,s,t}$ is the number of fish of age *a*, length *l*, sex *s* at time *t* in region *i*. For the deterministic case let us define $N^-_{i,a,l,s,t}$ as the population state immediately prior to migration and $N^+_{i,a,l,s,t}$ as the population state immediately after migration, and define the population (transposed) vectors pre and post migration as,

$$\mathbf{N}^{-\ \mathbf{T}}_{a,l,s,t} = \left[N^-_{0,a,l,s,t}, N^-_{1,a,l,s,t}, N^-_{2,a,l,s,t}, \ldots, N^-_{Z-1,a,l,s,t}\right] \tag{A1.50}$$

$$\mathbf{N}^{+\ \mathbf{T}}_{a,l,s,t} = \left[N^+_{0,a,l,s,t}, N^+_{1,a,l,s,t}, N^+_{2,a,l,s,t}, \ldots, N^+_{Z-1,a,l,s,t}\right] \tag{A1.51}$$

where Z is the number of regions in the model. Then the deterministic migration process is defined by,

$$\mathbf{N}^+_{a,l,s,t} = \mathbf{M}_{a,l,s,t}\mathbf{N}^-_{a,l,s,t} \tag{A1.52}$$

where $\mathbf{M}_{a,l,s,t}$ is a square migration transition matrix of dimension Z. From equation A1.52 the numbers migrating to region *j* is the inner product,

$$N^+_{j,a,l,s,t} = \sum_{i=0}^{Z-1} m_{j,i,a,l,s,t} N^-_{i,a,l,s,t} \tag{A1.53}$$

where $m_{j,i,a,l,s,t}$ is the migration transition matrix coefficient at row *j* and column *i* that describes the proportion of fish that will migrate from region *i* to region *j* of age *a*, length *l*, sex *s* at time *t*.

Migration can also be implemented as a stochastic process. To illustrate, let us define $\overline{\mathbf{\Phi}}_Z$ as the Z dimensional multinomial deviate which gives output deviates u_0 through $u_{Z\text{-}1}$ corresponding to probabilities P_0 through $P_{Z\text{-}1}$ when selecting from a sample size of *U*. Using a vector notation for brevity we have,

$$\mathbf{U} = \overline{\mathbf{\Phi}}_Z(\mathbf{P}, U) \tag{A1.54}$$

where,

$$\mathbf{U}^{\mathbf{T}} = \left[u_0, u_{1,} u_2, \ldots, u_{Z-1}\right] \tag{A1.55}$$

and,

$$\mathbf{P}^{\mathbf{T}} = \left[P_0, P_{1,} P_2, ..., P_{Z-1}\right] \tag{A1.56}$$

Then the stochastic migration process can be represented by,

$$\mathbf{N}^{+}_{a,l,s,t} = \sum_{i=0}^{Z-1} \overline{\Phi}_Z \left(\mathbf{P}^{\nu}_{i,a,l,s,t}, N^{-}_{i,a,l,s,t}\right) \tag{A1.57}$$

where,

$$\mathbf{P}^{\nu}_{i,a,l,s,t}{}^{\mathbf{T}} = \left[P^{\nu}_{0,i,a,l,s,t}, P^{\nu}_{1,i,a,l,s,t}, P^{\nu}_{2,i,a,l,s,t}, ... P^{\nu}_{Z-1,i,a,l,s,t}\right] \tag{A1.58}$$

and $P^{\nu}_{j,i,a,l,s,t}$ is the probability of fish age *a*, length *l*, sex *s* at time *t* migrating from region *i* to region *j*. The probability coefficient $P^{\nu}_{j,i,a,l,s,t}$ is equivalent to the migration transition matrix coefficient $m_{j,i,a,l,s,t}$. Note that, unlike the fishing and natural mortality probability vectors, the migration probabilities specified in the model definition files are not re-scaled according to equation A1.12. Furthermore, since the migration probabilities can be a function of time it is possible to develop models in VSM that include seasonal migration patterns.

A1.2.8 Summary Statistics

Some form of metrics are required for the purposes of cross checking assessment model results against the operating model data. These metrics are provided by VSM in the form of both summary time-series and single point management indicators. These indicators include:

Summary time-series Indicators

Indicator Name	**Description**
R[5]	Recruitment
COB[6]	(1^+ y.o. Catch biomass)/(1^+ y.o. Population biomass prior to catch removal)
B[5,6]	1^+ y.o. Population biomass
SSB[5]	Spawning stock biomass

Single Point Management Indicators

Indicator Name	**Description**
MSY	Maximum sustainable yield
$\mathbf{B_{MSY}}$[6]	1^+ y.o. Population biomass at maximum sustainable yield
$\mathbf{SSB_{MSY}}$	Spawning stock biomass at maximum sustainable yield
$\mathbf{COB_{MSY}}$[6]	(1^+ y.o. Catch biomass)/(1^+ y.o. Population biomass prior to catch removal) at maximum sustainable yield
$\mathbf{COB_{0.1}}$[6]	(1^+ y.o. Catch biomass)/(1^+ y.o. Population biomass prior to catch removal) at the COB value corresponding to a slope of 0.1 of the origin on the Yield versus COB curve.

Exploitation ratios (*catch over biomass*) are used as indicators of fishing mortality since they provide a unified measure of fishing mortality not sensitive to population structure or fishing selectivity. Since we have a number of competing fisheries targeting different age groups it is difficult to use instantaneous fishing mortality as an indicator because the population age structure changes with fishing pressure. This is further complicated by the requirement of constant catch ratios between fisheries when performing the MSY projection. *Catch over biomass* gives a simple indicator of fishing pressure that is independent of the age structure.

[5] Summary time-series for this indicator include both the fished and unfished cases. The process noise (recruitment deviations) are identical in both cases.

[6] Biomass estimates used in performance indicators have a low age cut-off that is user selectable. By default it ignores the biomass of fish less than 1 year old.

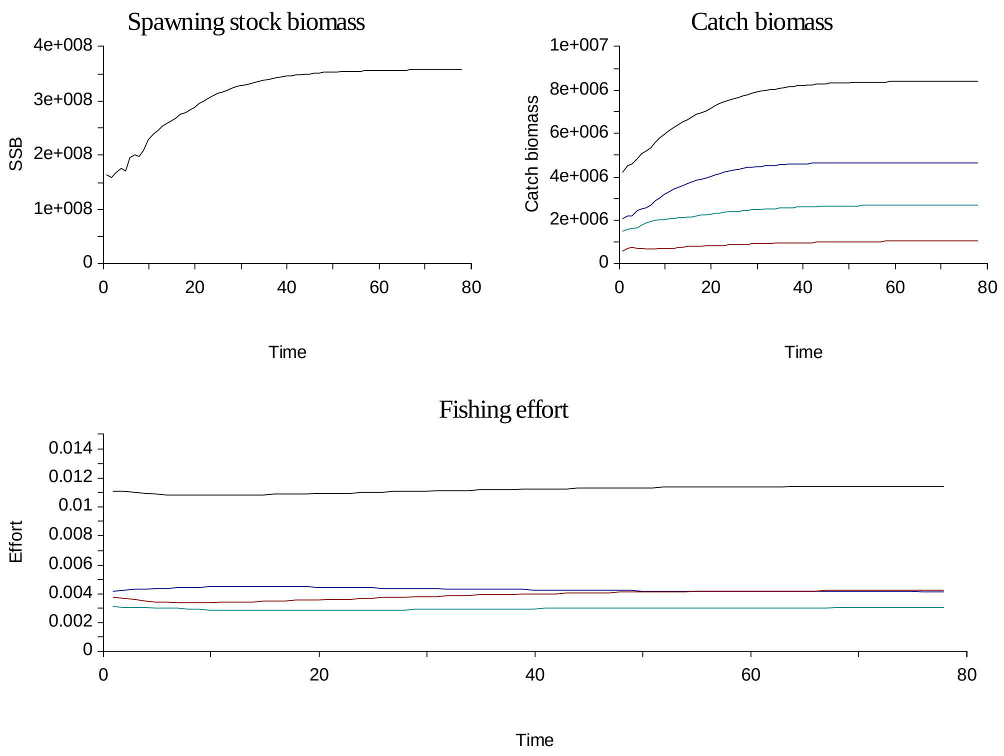
In some cases (eg. with non-stationary recruitment) it can be interesting to compare the state of the population that would have been observed with and without fishing. To synthesize *with* and *without fishing* histories VSM employs multiple random number generators, ensuring that the recruitment deviations and other stochastic processes are identical between fished and unfished cases. Each class of stochastic process has its own random number generator and each random number generator is seeded from a random number sourced from the *master random number generator*, which in turn, is seeded by the user specified seed. This ensures that each random number generator will be, for all practical purposes, statistically independent.

The MSY and $COB_{0.1}$ indicators require a more involved process. To accurately estimate these indicators VSM builds a *exploitation-yield* plot which it then searches for MSY and $COB_{0.1}$ using a *bisection* method. In the first stage of the projection VSM increments the overall effort (in a power series) until the steady state catch biomass decreases. This effort and an effort of zero then form the range of effort used to build the exploitation-yield plot. VSM nominally chooses 10 different levels of effort to build the exploitation-yield plot and does so in such a manner as to obtain points approximately uniformly distributed along the exploitation axis. These points are then fitted with a cubic spline and a *golden section search* performed to find the point of zero slope (MSY) and a *bisection search* used to find the point of 10% of the slope at the origin ($COB_{0.1}$). The slope is estimated using a first order difference approximation to the derivative with a Δ of COB_{MSY} / 10^6. That is, the single derivative of a function *f*(*x*) with respect to *x* is approximated by,

$$\frac{df(x)}{dx} \approx \frac{df(x+\Delta) - df(x-\Delta)}{2\Delta} \qquad \text{(A1.59)}$$

A cubic spline is also fitted to the points of exploitation versus effort which is used to map the exploitation at MSY and COB0.1 to corresponding efforts. These efforts are then used in steady state projections to determine accurate values of B_{MSY}, SSB_{MSY}, $B_{0.1}$, $SSB_{0.1}$ etc.

It is worth noting that in performing steady state projections the catch ratios between fisheries is regulated to remain constant via an internal adaptive feedback mechanism. The catch ratios are determined by the historical catch in the final year. This behaviour is currently hard coded into VSM. Furthermore, during all projections the stochastic processes are switched off, being replaced by their deterministic equivalents. Not doing so would make it difficult to both, determine when steady state is reached, and the true mean values of the various state variables (ie. Catch biomass, Biomass, Recruitment etc). Note that this differs from the usual assumptions of constant effort (fishing mortality) ratios among fleets. The simulation was intended to explore fishery systems like SBT, in which TACs are likely to be regulated with constant allocation proportions. Figure A1.10 shows a typical *steady state projection* in progress. Note that the catch ratio is well regulated by constant adjustment of the relative effort between fisheries. Figure A1.11 illustrates a typical *exploitation-yield* plot.



Spawning stock biomass

Catch biomass

Fishing effort

Figure A1.10: An example of a VSM steady state projection in progress. Note how the fishing effort is constantly adjusted to ensure constant catch ratios.

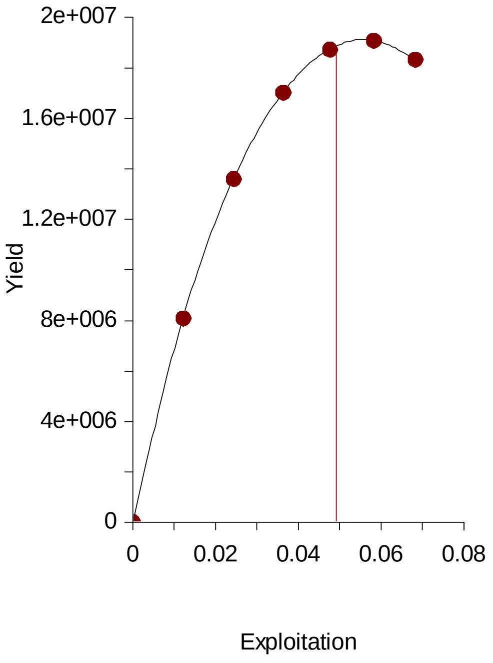


Figure A1.11: An example of a VSM Exploitation versus Yield plot with $COB_{0.1}$ shown

A1.3 Observation Model Implementation Details

VSM treats the process of reporting catch and effort data separately from the operating model. The observation model reports the actual catch and effort data with various observational errors imposed on that data. By separating the operating model and the observation model it then is possible to create multiple observations of the one true data set that have statistically independent error deviations while not having to re-run the operating model. Furthermore, by separating the two processes software maintenance issues are improved by virtue of the fact that if there is a logical error in the observational model code it can be fixed without the need to re-run the operating model to create a given state realisation.

VSM has the ability to introduce the following sources of error into observation data:

- Effort reporting errors
- Total catch errors (by number)
- Age and length distribution errors (by number and excluding tags)
- Tag reporting rate errors
- Age estimation by cohort slicing

All these error mechanisms operate on a fishery basis, meaning that individual fisheries have their own error regimes when targeting specific species. Furthermore the recorded data is aggregated over a statistical time period which, in the models used in this study, corresponds to one year.

A1.3.1 Effort errors

Effort deviations are applied to each individual effort record in a given state realization using the generalized correlated deviate with the deviate CV optionally being a function of time. In the observation models used to date, effort has been reported without error. But note that effective effort process deviations are highly confounded with effort reporting errors.

A1.3.2 Total catch errors

Similarly, total catch deviations are also applied using the generalized correlated deviate with the deviate CV optionally being a function of time. The deviate is applied to the total catch at age with the numbers at length re-scaled accordingly (to preserve the length at age distribution). In the observation models used to date, total catch has been reported with a log normal deviate applied whose variance is not a function of time.

A1.3.3 Age and length distribution errors

Age and length distribution errors are synthesised using a multinomial selection process combined with *catch at age* and *catch at length* effective sample size specifications. The true length distribution is combined with an effective sample size to synthesise a new distribution using a multinomial sampling process. The age / length distributions (in absolute catch numbers) are converted into probability coefficients. A sample population distribution is created through multinomial selection driven by these probability coefficients and the effective sample size. For the length distribution data the sample population distribution is reported as is and for the age distribution data the sample distribution is scaled up to equal total catch in numbers. The need to scale up the age data arises from the output file format of the catch at age data as it has no entry for total catch but assumes the sum of all entries in the catch at age matrix is the total catch. This file format was used for historical reasons. Future revisions of VSM may alter the reporting of age based data to be consistent with the length based data through a modification of the file format.

More rigorously, if there are A ages classes and L length classes at time t, the effective sample size applied to the age distribution is A_{ess} and $C_{a,l,s,t}$ is the total catch in numbers of age a, length l, sex s at time t, then the modified catch-age distribution that includes age distribution errors is given by,

$$\mathbf{C}'_{s,t} = \overline{\Phi}_A\left(\frac{1}{A_{ess}}\overline{\Phi}_A\left(\mathbf{R}_{s,t}, A_{ess}\right), C_{s,t}\right) \tag{A1.60}$$

where,

$$\mathbf{C}'^{\,\mathbf{T}}_{s,t} = \left[C'_{0,s,t}, C'_{1,s,t}, C'_{2,s,t}, \ldots C'_{a,s,t}, \ldots, C'_{A-1,s,t}\right] \tag{A1.61}$$

$$\mathbf{R}_{s,t}^{\mathbf{T}} = \left[\frac{C_{0,s,t}}{C_{s,t}}, \frac{C_{1,s,t}}{C_{s,t}}, \frac{C_{2,s,t}}{C_{s,t}}, \ldots \frac{C_{a,s,t}}{C_{s,t}}, \ldots, \frac{C_{A-1,s,t}}{C_{s,t}}\right] \tag{A1.62}$$

$$C_{a,s,t} = \sum_{l=0}^{L-1} C_{a,l,s,t} \tag{A1.63}$$

$$C_{s,t} = \sum_{a=0}^{A-1} C_{a,s,t} \tag{A1.64}$$

and $\overline{\Phi}_A$ is a multinomial deviate as defined by Equation A1.54. As mentioned earlier, the reported catch-age distribution is scaled to correspond to the total catch, hence the second level of multinomial selection in Equation A1.60. In the case of the reported catch-length distribution the total numbers in the distribution sums to the length effective sample size L_{ess} and the modified catch-length distribution is given by,

$$\mathbf{C}''_{s,t} = \overline{\Phi}_L\left(\mathbf{R}'_{s,t}, L_{ess}\right) \tag{A1.65}$$

where,

$$\mathbf{C}''^{\,\mathbf{T}}_{s,t} = \left[C''_{0,s,t}, C''_{1,s,t}, C''_{2,s,t}, \ldots, C''_{l,s,t} \ldots, C''_{L-1,s,t}\right] \tag{A1.66}$$

$$\mathbf{R}'^{\,\mathbf{T}}_{s,t} = \left[\frac{C'_{0,s,t}}{C_{s,t}}, \frac{C'_{1,s,t}}{C_{s,t}}, \frac{C'_{2,s,t}}{C_{s,t}}, \ldots, \frac{C'_{l,s,t}}{C_{s,t}}, \ldots, \frac{C'_{L-1,s,t}}{C_{s,t}}\right] \tag{A1.67}$$

$$C'_{l,s,t} = \sum_{a=0}^{A-1} C'_{a,l,s,t} \tag{A1.68}$$

A1.3.4 Tag reporting rate errors

Tag reporting rate errors are introduced through a binomial selection process. As with the total catch numbers and the effort deviations the reporting rate errors can be a function of time. In other words, if $C^{tag}_{a,l,s,t}$ is the true number of tagged fish age *a*, length *l*, sex *s* at time *t* and P^{rep}_t is the probability of reporting the catch as tagged then the reported tagged catch is given by,

$$C^{rep}_{a,l,s,t} = \Phi\left(P^{rep}_t, C^{tag}_{a,l,s,t}\right) \tag{A1.69}$$

where $\Phi(p,N)$ is a binomial random deviate with probability *p* and number of independent trials *N*. The true number of tagged fish is known and results from the fishing mortality process within VSM. Remember that as tagged fish are managed internally in a separate sub population, VSM automatically keeps track of recaptures using the same predation process that generates non-tagged catch.

A1.3.5 Age errors through cohort slicing

Age estimation through the process of cohort slicing can also be simulated. VSM supports cohort slicing over an arbitrary number of cut-points through a fixed age step between cut-points. In other words, if δ_a is the age step between cut-points and $l(a)$ is the length at age as specified by Equation A1.6 then the cut-point age and cut-point length are defined by,

$$a_n^{cutpoint} = \left(n + \frac{1}{2}\right)\delta_a \qquad \text{(A1.70)}$$

$$l_n^{cutpoint} = l\left(a_{n+1}^{cutpoint}\right) \qquad \text{(A1.71)}$$

Catch is then assigned an age by finding the smallest cut-point length that is greater than the catch length and assigning the corresponding cut-point age. To obtain realistic cohort slicing the age step, δ_a, should be the same as the model time step, t_m.

A1.4 LIST OF SYMBOLS

R	Recruitment
R_0	Virgin reruitment
SSB	Spawning stock biomass
d,b	Beverton-Holt recruitment parameters
h	Steepness (Beverton-Holt recruitment relationship)
B	Biomass
C	Catch in numbers
T	Tag releases in numbers
N	Population size in numbers
M	Natural mortality in numbers
P	Probability
Ω	Probability scaling function
s	Fish sex
a	Fish age
l	Fish length
t	Time
g	Growth transition index
$l(a)$	Growth equation
$w(a)$	Mass-length relationship
q,ν	Mass-length relationship parameters
α,β,k_1,k_2	vb log k growth equation parameters
σ	Distribution standard deviation
μ	Distribution mean
ρ	Lag 1 auto-correlation of an AR1 process
Δ_k	Random deviate sequence
β	Random deviate sequence bias
S	Baseline catch probability / selectivity
ς	Selectivity altering function
λ	scaling factor for selectivity change mechanism
α	controlling parameter for selectivity change mechanism
E	Effort
q	Catchability
Φ(p,N)	Binomial random deviate with probability p in N trials

A1.5 REFERENCES

1. Laslett, G.M., Eveson, J.P., and Polacheck, T. 2002. A flexible maximum likelihood approach for fitting growth curves to tag-recapture data. Can. J. Fish. Aquat. Sci. 59: 976-986.

APPENDIX 2 VSM PARAMETERIZATION FOR A FISHERY RESEMBLING SBT

This appendix describes the different VSM specifications used to simulate alternative representations of the Southern Bluefin Tuna population for the SESAME project. Biological parameters were mostly adopted from input to actual SBT assessments (see Preece et al. 2001 and references therein). Production dynamics and exploitation history were intended to approximate the general perceptions about the stock from assessment results (e.g. Butterworth et al. 2003, Polacheck et al. 2001). However, there was no explicit conditioning of the operating models to the real SBT data. A number of details were also simplified (e.g. reduced number of fisheries, consistency of data collection methods over time).

The operating models are defined relative to one of two baseline scenarios. The stage 1 scenarios are designated E_x in the text of the SESAME report and tend to be *easy* in terms of the nature of the model process and the observation errors. The stage 2 scenarios are designated D_x and are *difficult*, with substantial process errors and small sample sizes. A number of intermediate scenarios are also defined, and a range of models with additional characteristics that are likely to be assumption violations for most of the assessment models. A summary of the models used in this study is provided in Table A2.1. Each model provides 50 years of simulated data generated through VSM models running on a monthly time step and aggregating and reporting statistics on an annual basis. Fishing, when active, occurs continuously throughout the year. There are four fisheries: *juvenile, long-line spawning grounds* and *long-line feeding grounds* (split into an early and late operating periods). The juvenile fishery predominantly targets 3-5 year old fish (Figure A2.4), the long-line feeding grounds fishery targets 5-15 year old fish (Figure A2.5) and the long-line spawning grounds 10+ year old fish (Figure A2.6). In all scenarios with the exception of E_HL and D_HL the baseline selectivity is age based. For E_HL & D_HL the baseline selectivity is length based with the selectivity vectors designed to give roughly the same catch-age distribution as the age based case when all else is the same. The spawning ground long-line fishery actually represents two distinct fisheries, the early Japanese fishery that primarily targeted SBT, and the recent Indonesian fishery that takes SBT mostly as by-catch. In the operating models, they are the same, except for the data collection.

Targeting in stage 2 scenarios (D_x) uses the same baseline selectivities as with the stage 1 (E_x) scenarios along with some population structure dependent selectivity changes. The long-line spawning ground fishery selectivity in stage 2 (Figure A2.20) is the same as that of stage 2. The long-line feeding ground fishery (Figure A2.21) has cohort targeting selectivity changes applied to the baseline case obtained from stage 1 scenarios. This is intended to simulate a tendency for fishers to disproportionately target relatively abundant age classes of variable size. Similarly, the juvenile fishery (Figure A2.22) has constant catch proportion selectivity changes applied to the baseline case. This is intended to simulate the apparent behaviour of the Australian domestic purse seine fishery, which seems to have some capacity to select schools of a certain age class best suited for aquaculture purposes, regardless of the overall age composition in the bight.

Similar fishery effort series were used for all Stage 1 scenarios and were chosen to loosely mimic the recorded exploitation history of the SBT fishery since the 1950's. Generally there is high exploitation in the early spawning grounds fishery and increasing exploitation in the juvenile and long-line feeding grounds fisheries over a period of around 35 years followed by a dramatic effort reduction about 10-15 years before the end of the time series (due to management restrictions). All of the fisheries have informative effort time series for some Stage 1 scenarios (Figure A2.1). However, only the long-line feeding grounds fishery was intended to be interpreted as informative in an assessment. In Stage 2 scenarios only the long-line feeding grounds fishery has an informative effort time series (Figure A2.17) and it is much less informative than stage 1. The effort series of the other fisheries is intentionally misleading in the majority of cases so that extra information cannot be extracted (intentionally or otherwise).

Depletion levels for the stock range from between 15-50% of virgin biomass depending primarily upon the steepness in the Beverton-Holt recruitment relationship. For high steepness the level of depletion is somewhat lower than the low steepness cases owing to the insensitivity of recruitment to depletion. We desired substantial depletion in all cases, but did not attempt to ensure that they were all comparable.

Complete histories of the biomass and recruitment are maintained for every realisation, including the values that would have occurred without fishing (Figures A2.3 & A2.19).

Catch age data is provided in all scenarios with the juvenile and long-line feeding grounds fisheries providing age estimates from cohort-slicing. The early long-line spawning grounds fishery provides cohort-sliced age data, while the late spawning grounds fishery provides simulated direct ageing data.

All scenarios use the same length-mass relationship (Figure A2.15) and age-maturity relationship (Figure A2.16) with 100% of SBT becoming mature at age 10, and spawning every year (fecundity is directly proportional to mass). Stage 1 scenarios use the age-mortality relationship of Figure A2.2 whilst stage 2 scenarios use that of Figure A2.18.

Most stage 1 scenarios release between 6000 and 12000 tags per annum attached to 1-3 year old fish between the 40th and the 44th years (Figure A2.1). The exception is E_CA60 in which 600 to 1200 tags per annum are released. In all stage 2 scenarios, the release numbers drop to 300-600 tags for the same age range (Figure A2.17). Zero tag shedding and 100% reporting of recaptures is implemented across all fisheries.

All scenarios use a Beverton-Holt recruitment relationship with a range of steepness and recruitment variability, with the exception of E_HSSR & D_HSSR which both use a double linear *"hockey stick"* function. The recruitment variability is provided by a log-normal deviate with or without lag-1 auto-correlation (Figures A2.3 & A2.19). Figure A2.36 shows the recruitment time series for five different realisations of scenario E_h3. Similarly Figure A2.37 shows the recruitment time series for five different realisations of scenario E_h4_r8, which has a lag 1 year auto-correlation of 0.8. Scenario E_h3 and E_h4_r8 show similar trends although the trend in E_h4_r8 is smoother on account of the high recruitment auto-correlation. Figure A2.33 shows the recruitment relationships used in stage 1 scenarios whilst Figure A2.34 shows the same for stage 2.

The same growth equation was used throughout except for E_DDLinf and D_DDLinf which included changes in the asymptotic length (of the growth equation) over time to simulate density dependent changes in growth rates. This change in asymptotic length is highlighted in Figure A2.40. In all other cases this resulted in identical length-at-age relationships, except for the scenarios that had purely length-based fishing mortality (E_HL, D_HL). The growth equation is the same for stage 1 and Stage 2 scenarios although the stage 2 scenarios have a higher variance on the length-at-age distributions (compare Figures A2.14 & A2.30). In the length-based selectivity scenarios (E_HL, D_HL), the growth rates are the same as the other scenarios, but length-at-age potentially changes over time and among cohorts, depending on the size selective exploitation history that each cohort experiences. This is one possible implementation of size selective mortality, though it might be argued that maintaining a variety of growth curves within the population would be preferable.

All scenarios include log normal deviations in catchability. Stage 1 scenarios have an annual CV of 14% on the spawning ground fishery, 10% on the feeding ground fishery and 6% on the juvenile fishery, all with no auto-correlation. Stage 2 scenarios have an annual CV of 16% on the long-line spawning ground and juvenile fisheries with no auto-correlation whilst the long-line feeding ground fishery has an annual CV of 40% and a lag 1 auto-correlation of 0.5. Example time series highlighting the deviations in catchability are shown in Figures A2.11, A2.12, A2.13, A2.27, A2.28 & A2.29. Note that for the purposes of assessment, only the longline feeding grounds fishery effort – fishing mortality relationship was to be considered informative as a relative abundance index (Figures A2.12, A2.28).

Catchability trends for the informative long line fishery are included in scenarios E_qInc, E_qC, E_qI, D_qInc, D_qC and D_qI. In the case of E_qInc and D_qInc these models include the explicitly defined catchability trends of Figure A2.38. Cases E_qC, E_qI, D_qC and D_qI include implicit catchability trends that stem from the non-linear effort/fishing mortality relationship defined in those cases. These effort/fishing mortality relationships are shown in Figure A2.35 and give rise to the catchability trends shown in Figure A2.39. Note that the implied nature of this catchability relationship stems from the assumption, in the assessment models, that effort and fishing mortality are linearly related.

Stochastic variation in fishery selectivity is intended to mimic noise from a range of processes (e.g. spatial heterogeneity in the fish population and fishing fleets), and is implemented with an "effective sample size" which introduces additional noise to the binomial selection process. Note that this is a process error and is completely distinct from the sample sizes in the catch observation process (and effective sample sizes that are commonly used to downweight catch data within assessment models). The typical effect of this approach is demonstrated in Figure A2.41, which shows the catch-length distribution for total catch in the final year of the model run with and without the effective sample size implemented. When implemented in the D_x scenarios, the process noise increased with lower effort (but was constant in the E_x scenarios).

In the observation model, sample sizes are used to introduce distribution errors in the catch-at-age and catch-at-length observed data. For stage 1 scenarios an effective sample size of 1000 was typically used whereas for the stage 2 scenarios this was reduced to 50. In stage 1 the exception is model E_CA60 which has an effective sample size of 60. The typical effect this has upon observed data is demonstrated in Figure A2.42, which shows the standardised reported catch-length distribution in the final year of the model run with sample sizes of 1000 and 60.

Finally, a number of scenarios are included to test the effect of catch under-reporting on assessment model performance. Models E_C10ju, E_C10llf and E_C10lls have 10% under-reporting of catch in the *juvenile*, *long-line feeding* and *long-line spawning ground* fisheries respectively. Likewise, models E_C20ju, E_C20llf and E_C20lls have 20% under-reporting of catch in the *juvenile*, *long-line feeding* and *long-line spawning ground* fisheries respectively. In all cases, the under-reporting was constant over time and had no effect on the age and length frequency sampling.

REFERENCES

Butterworth, D.S., J.N. Ianelli and R. Hilborn. 2003. A statistical model for stock assessment of southern bluefin tuna with temporal changes in selectivity. Afr. J. Mar. Sci. 25: 331-361.

Polacheck, T., A. Preece and D. Ricard. 2001. Assessment of the status of the southern bluefin tuna stock using virtual population analysis – 2001. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/01/08/20.

Preece, A. T. Polacheck, D. Kolody, P. Eveson, D. Ricard, P. Jumppanen, J. Farley, and T. Davis. 2001. Summary of the primary inputs to CSIRO’s 2001 stock assessment models. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/01/08/21.

Table A2.1: Summary of Operating Model Definitions used in this study. Stage 1 scenarios, designated E_x, are considered relatively easy in terms of the nature of the model process and the observation errors. Stage 2 scenarios, designated D_x, are considered difficult and have substantial process errors and small sample sizes.

Scenario	Recruitment	Selectivity	Catchability	Effort Fishing Mortality relationship	Observation Error	Growth	Tagging
E_h3 1_1n	Beverton-Holt, h = 0.3, ρ = 0.0, σ = 0.4	Age based	Catchability includes log normal deviations with no auto-correlation or bias. Annual CV 14% LL *Spawning* fishery, 10% *LL Feeding* fishery and 6% *Juvenile* fishery.	Linear	Catch-at-age effective sample size 1000 Catch-at-length effective sample size 1000	VB log k growth equation, $L\infty=182$, $t0=0$, $k1=k2=0.18$, $\alpha=2.9$, $\beta=30.0$ Length standard deviation 5 irrespective of age. Absolute growth method.	Tagging in years 40 to 44 inclusive mid year for one month. Tagging 1-3 year olds in equal proportion. Total releases per year {12000,6000, 12000,6000, 12000}
E_base 1_2n	Same as 1_1n but with, h = 0.6	Age based	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_h9 1_3n	Same as 1_1n but with, h = 0.9	Age based	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_h4_r8 1_4n	Same as 1_1n but with, h = 0.4, ρ = 0.8	Age based	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_h8_r8 1_5n	Same as 1_1n but with, h = 0.8, ρ = 0.8	Age based	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_qInc 1_6n	Same as 1_2n	Age based	Same as 1_1n plus *LL Feeding* increasing catchability trend 1% per year from 10th year onwards, *LL spawning* catchability increasing exponentially in fishing years, *Juvenile* catchability exponential decay with noise.	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_CA60 1_9n	Same as 1_2n	Age based	Same as 1_1n	Linear	Catch-at-age effective sample size 60 Catch-at-length effective sample size 60	Same as 1_2n	Tagging in years 40 to 44 inclusive mid year for one month. Tagging 1-3 year olds in equal proportion. Total releases per year {1200,600, 1200,600, 1200}

Scenario	Recruitment	Selectivity	Catchability	Effort Fishing Mortality relationship	Observation Error	Growth	Tagging
E_HTS **1_10n**	Same as 1_2n	Age based with selectivity changes: *const. catch proportion* on *Juvenile* and *cohort targeting* on *LL feeding.*	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_HL **1_11n**	Same as 1_2n	Length based	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n except uses the differential growth method.	Same as 1_1n
E_H45 **1_12n**	Same as 1_2n	Age based with explicit selectivity change to younger fish at the 45th year on *LL Feeding*	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_qC **1_14n**	Same as 1_2n	Age based	Same as 1_1n	Cobb-Douglas type model where effective effort accelerates with increasing effort, ie. co-operation Effort = Hooks1.5	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_qI **1_15n**	Same as 1_2n	Age based	Same as 1_1n	Cobb-Douglas type model where effective effort decelerates with increasing effort, ie. interference Effort = Hooks0.67	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_DDLinf **1_16n**	Same as 1_2n	Age based	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n except $L\infty$ changes from 182 to 162cm from 10th to the 20th year mimicking density dependence effects.	Same as 1_1n
E_DRq **1_17n**	Same as 1_2n	Age based	Same as 1_1n except *LL Feeding* catchability has lag 1 year autocorrelation of 0.5 and an annual CV of 40%	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_HSSR **1_19n**	Hockey stick, h = 0.6, ρ = 0.0, σ = 0.4	Age based	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
E_H40 **1_20n**	Same as *1_2n*	Age based with explicit selectivity change to younger fish at the 40th year on *LL Feeding*	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n

Scenario	**Recruitment**	**Selectivity**	**Catchability**	**Effort Fishing Mortality relationship**	**Observation Error**	**Growth**	**Tagging**
E_stoH **1_21n**	Same as 1_2n	Age based with effective sample size, Ψ = 500 Ψmin = 50	Same as 1_1n	Linear	Same as 1_1n	Same as 1_2n	Same as 1_1n
D_h3 **2_1n**	Beverton-Holt, h = 0.3, ρ = 0.0, σ = 0.6 Spawning spread over 4 months with weights: {0.1,0.4,0.4,0.1}	Age based with effective sample size, Ψ = 500 Ψmin = 50 plus selectivity changes: *const. catch proportion* on *Juvenile* and *cohort targeting* on *LL feeding.*	Catchability includes log normal deviations with no auto-correlation or bias. Annual CV 16% *Spawning ground* and *Juvenile* fisheries, *LL Feeding* catchability has lag 1 year autocorrelation of 0.5 and an annual CV of 40%	Linear	Catch-at-age effective sample size 50 Catch-at-length effective sample size 50	VB log k growth equation, $L\infty=182$, $t0=0$, $k1=k2=0.18$, $\alpha=2.9$, $\beta=30.0$ Length standard deviation 8 irrespective of age. Absolute growth method.	Tagging in years 40 to 44 inclusive mid year for one month. Tagging 1-3 year olds in equal proportion. Total releases per year {600,300, 600,300, 600}
D_base **2_2n**	Same *as* 2_1n but *with*, h = *0.6*	Same as 2_1n	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_h9 **2_3n**	Same as 2_1n but with, *h* = 0.9	Same as 2_1n	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_h4_r8 **2_4n**	Same as 2_1n but with, *h* = 0.4, ρ = 0.8	Same as 2_1n	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_h8_r8 **2_5n**	Same as 2_1n but with, *h* = 0.8, ρ = 0.8	Same as 2_1n	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_qInc **2_6n**	Same as 2_2n	Same as 2_1n	Same as 2_1n plus *LL Feeding* increasing catchability trend 1% per year from 10th year onwards, *LL spawning* catchability increasing exponentially in fishing years, *Juvenile* catchability exponential decay with noise.	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_HL **2_11n**	Same as 2_2n	Length based with effective sample size, Ψ = 500 Ψmin = 50 plus selectivity changes: *const. catch proportion* on *Juvenile* and *cohort targeting* on *LL feeding.*	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n except using differential growth method.	Same as 2_1n

Scenario	Recruitment	Selectivity	Catchability	Effort Fishing Mortality relationship	Observation Error	Growth	Tagging
D_H45 2_12n	Same as 2_2n	Same as 2_1n plus explicit selectivity *change* to younger *fish* at the 45th year on *LL Feeding*	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_qC 2_14n	Same as 2_2n	Same as 2_1n	Same as 2_1n	Cobb-Douglas type model where effective effort accelerates with increasing effort, ie. co-operation Effort = Hooks1.5	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_qI 2_15n	Same as 2_2n	Same as 2_1n	Same as 2_1n	Cobb-Douglas type model where effective effort decelerates with increasing effort, ie. interference Effort = Hooks0.67	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_DDLinf 2_16n	Same as 2_2n	Same as 2_1n	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n except $L\infty$ changes from 182 to 162cm from 10th to the 20th year mimicking density dependence effects. Uses the differential growth method.	Same as 2_1n
D_HSSR 2_19n	Hockey stick, h = 0.6, $\rho = 0.0$, $\sigma = 0.6$ Spawning spread over 4 months with weights: {0.1,0.4,0.4,0.1}	Same as 2_1n	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
D_H40 2_20n	Same as *2_2n*	Same as 2_1n plus explicit selectivity change to younger fish at the 40th year on *LL Feeding*	Same as 2_1n	Linear	Same as 2_1n	Same as 2_1n	Same as 2_1n
E_C10ju 1_2nuj	Same as 1_2n	Same as 1_2n	Same as 1_2n except with a nominal 10% under-reporting of catch in the *juvenile* fishery	Linear	Same as 1_2n	Same as 1_2n	Same as 1_2n

Scenario	Recruitment	Selectivity	Catchability	Effort Fishing Mortality relationship	Observation Error	Growth	Tagging
E_C10llf 1_2nullf	Same as 1_2n	Same as 1_2n	Same as 1_2n except with a nominal 10% under-reporting of catch in the *long-line feeding* fishery	Linear	Same as 1_2n	Same as 1_2n	Same as 1_2n
E_C10lls 1_2nulls	Same as 1_2n	Same as 1_2n	Same as 1_2n except with a nominal 10% under-reporting of catch in the *long-line spawning ground* fishery	Linear	Same as 1_2n	Same as 1_2n	Same as 1_2n
E_C20ju 1_2nuj2	Same as 1_2n	Same as 1_2n	Same as 1_2n except with a nominal 20% under-reporting of catch in the juvenile fishery	Linear	Same as 1_2n	Same as 1_2n	Same as 1_2n
E_C20llf 1_2nullf2	Same as 1_2n	Same as 1_2n	Same as 1_2n except with a nominal 20% under-reporting of catch in the *long-line feeding* fishery	Linear	Same as 1_2n	Same as 1_2n	Same as 1_2n
E_C20lls 1_2nulls2	Same as 1_2n	Same as 1_2n	Same as 1_2n except with a nominal 20% under-reporting of catch in the *long-line spawning ground* fishery	Linear	Same as 1_2n	Same as 1_2n	Same as 1_2n

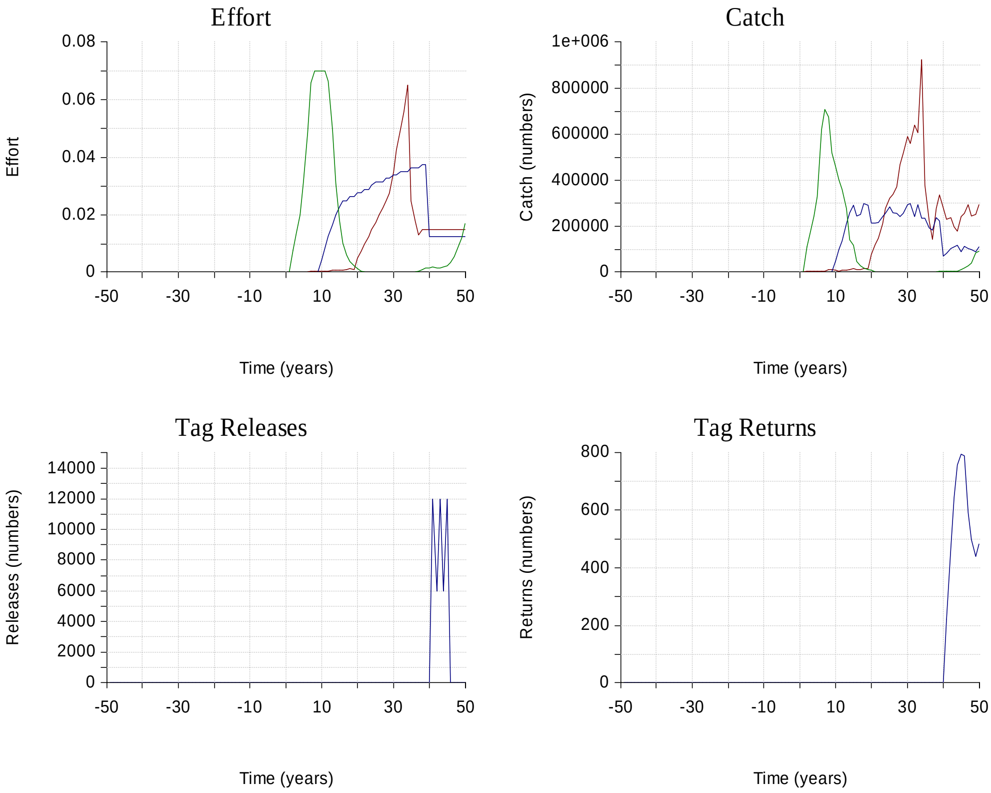


Figure A2.1: Model E_base fishery and tag dynamics from one stochastic state realisation: red - *juvenile* fishery, blue - *long-line feeding ground* fishery, green - *long-line spawning ground* fishery

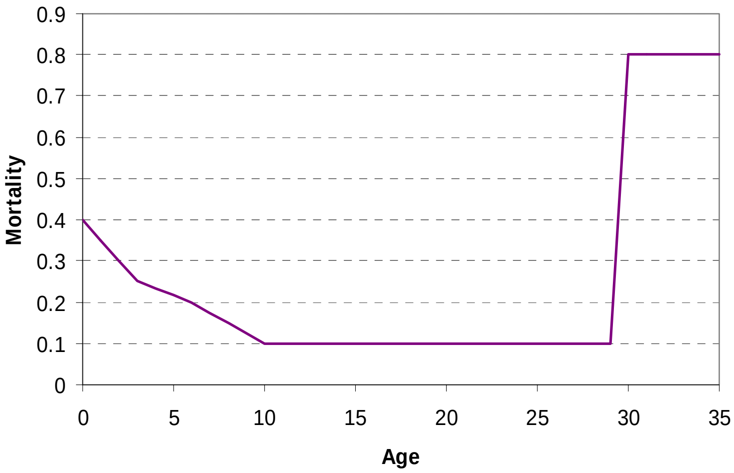


Figure A2.2: Model E_base natural mortality-at-age relationship

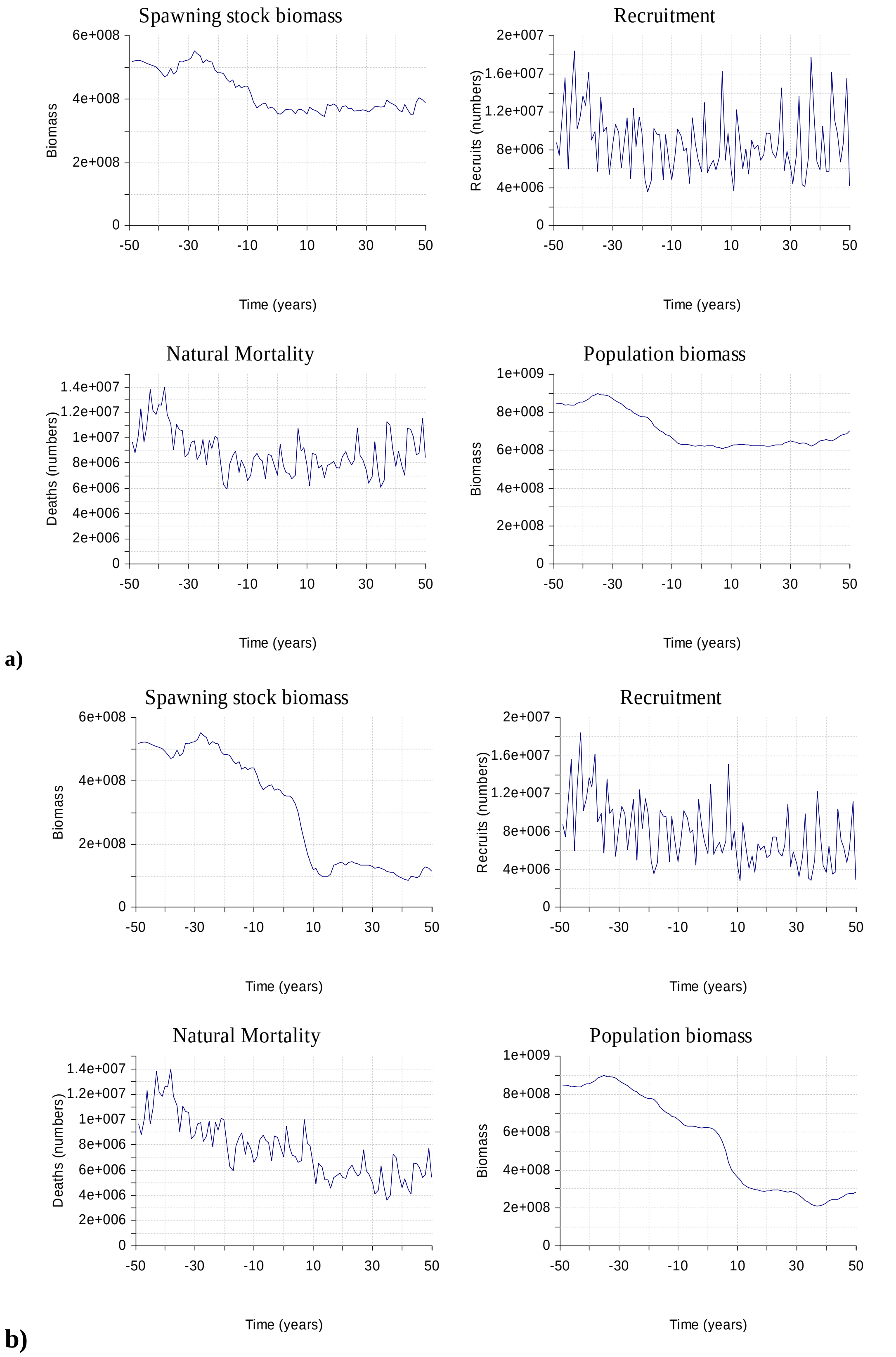


Figure A2.3: Model E_base population dynamics under (a) unfished and (b) fished conditions

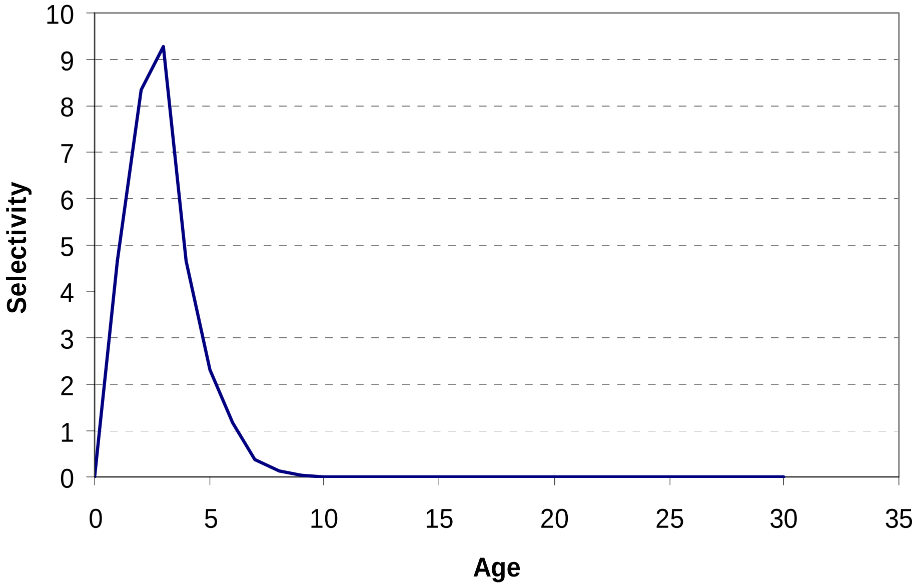


Figure A2.4: Model E_base selectivity: *juvenile* fishery, constant over time.

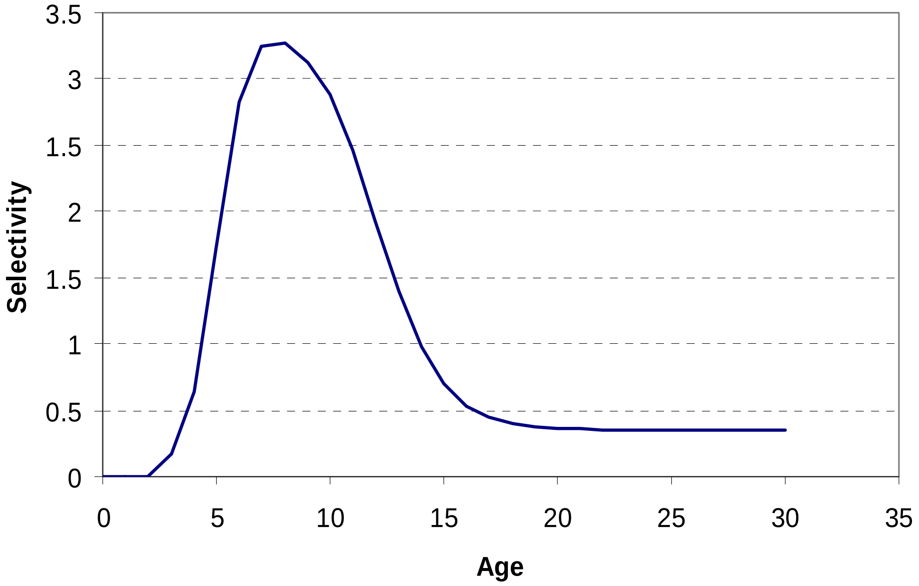


Figure A2.5: Model E_base selectivity: *long-line feeding ground* fishery, constant over time.

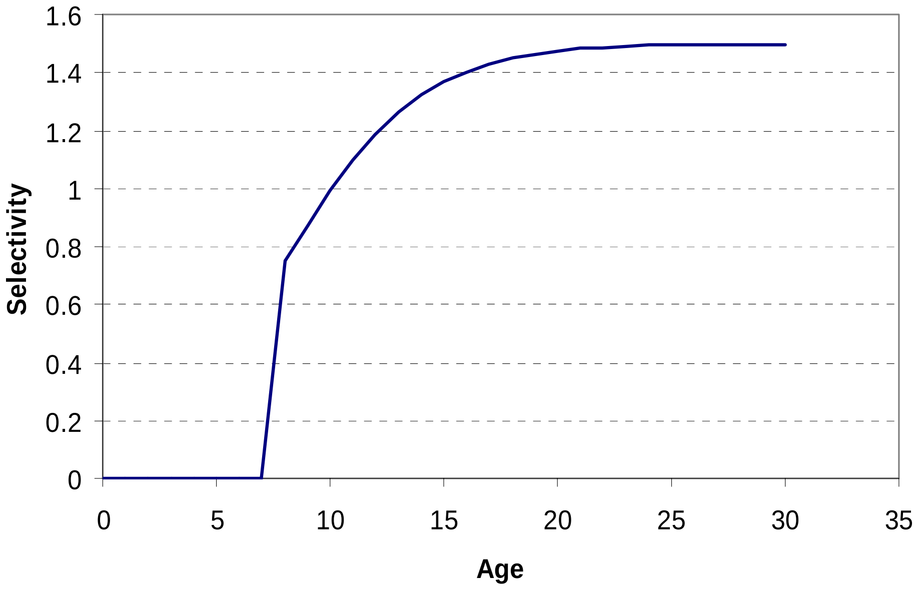


Figure A2.6: Model E_base selectivity: *long-line spawning ground* fishery, constant over time.

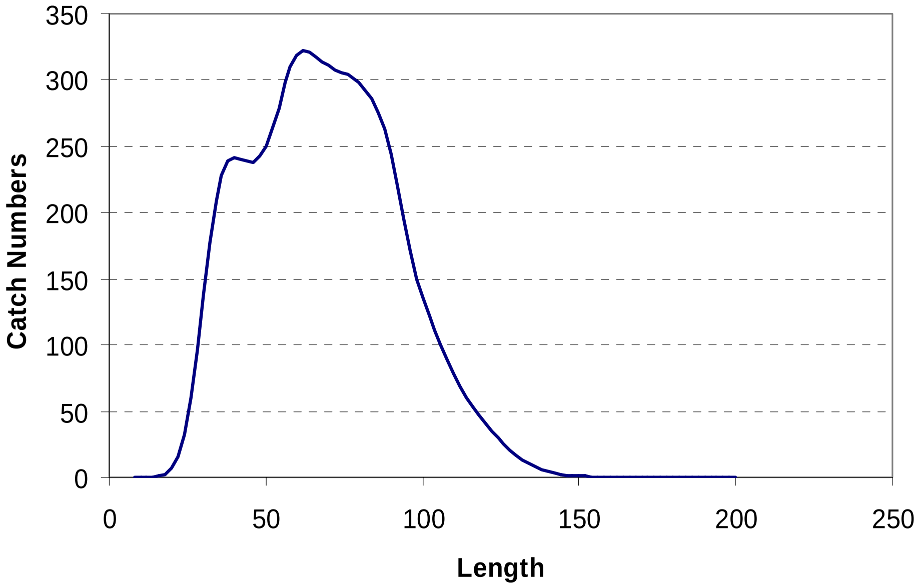


Figure A2.7: Model E_base aggregate (over all time) catch-at-length frequency distribution: *juvenile* fishery

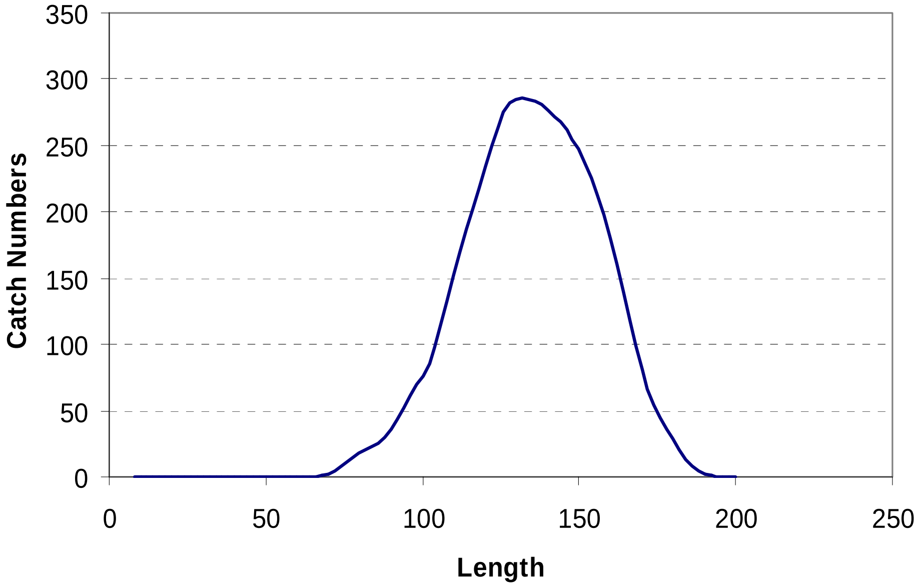


Figure A2.8: Model E_base aggregate (over all time) catch-at-length frequency distribution: *long-line feeding ground* fishery

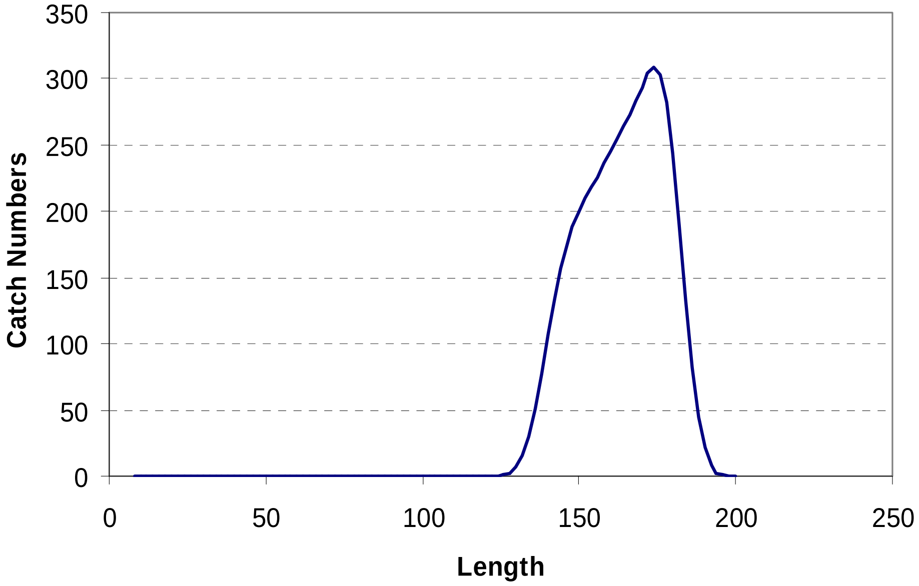


Figure A2.9: Model E_base aggregate (over all time) catch-at-length frequency distribution: *long-line spawning ground* fishery

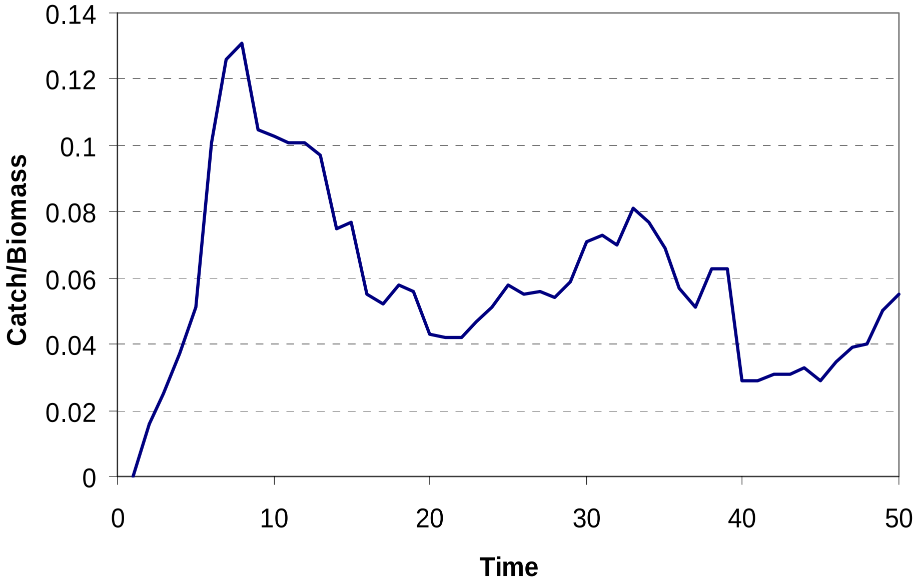


Figure A2.10: Model E_base total exploitation rate time series

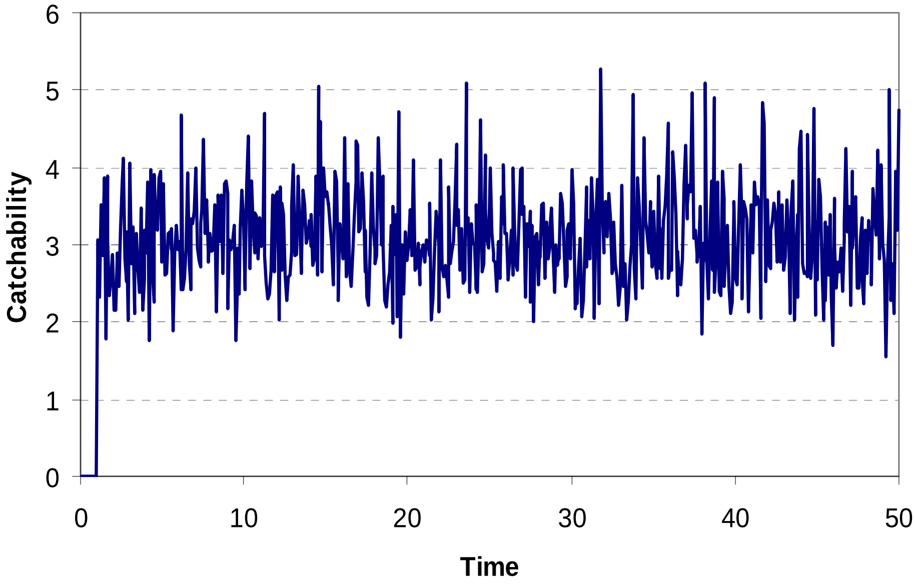


Figure A2.11: Model E_base catchability series (including effective effort deviations): *juvenile* fishery. The fishery operates from the first year onwards.

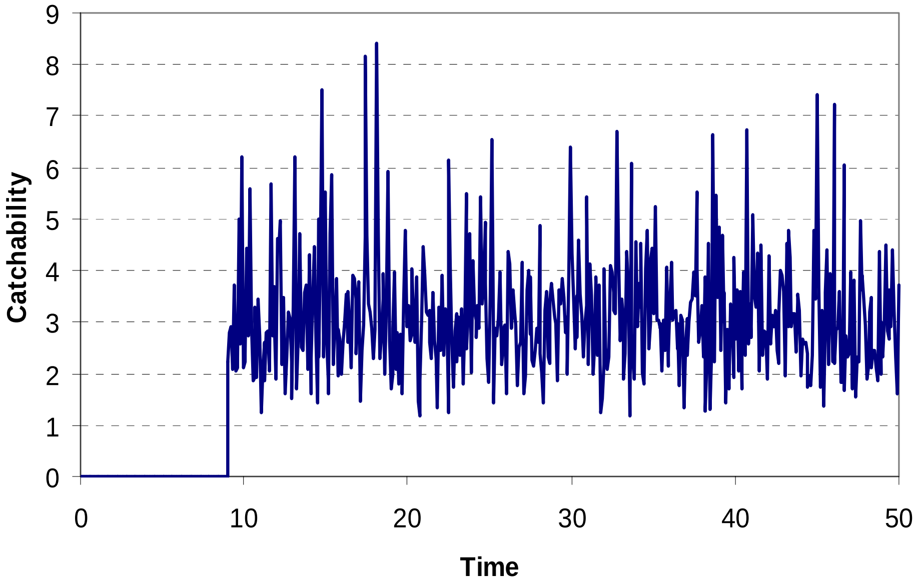


Figure A2.12: Model E_base catchability series (including effective effort deviations): *long-line feeding ground* fishery, The fishery operates from the ninth year onwards.

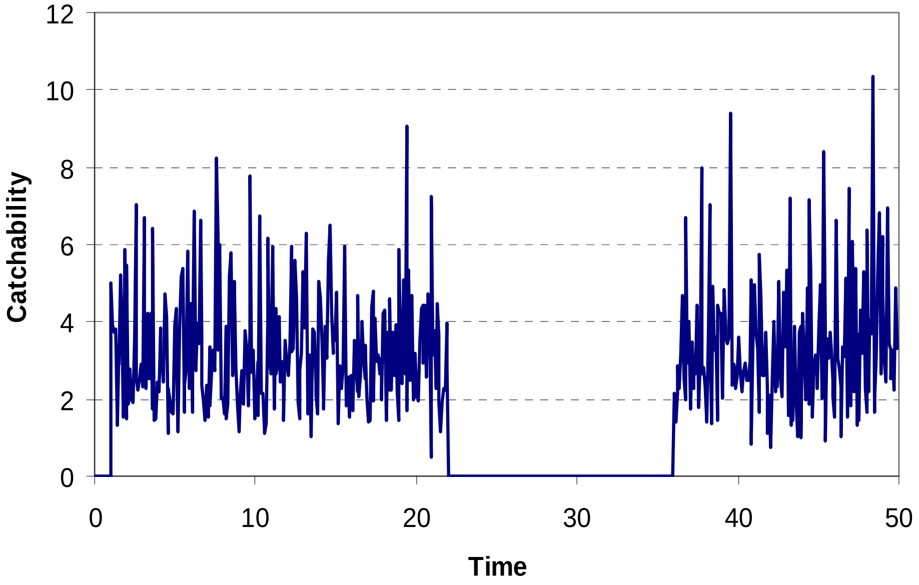


Figure A2.13: Model E_base catchability series (including effective effort deviations): *long-line spawning ground* fishery. The fishery is closed from the 23rd to the 36th year inclusive.

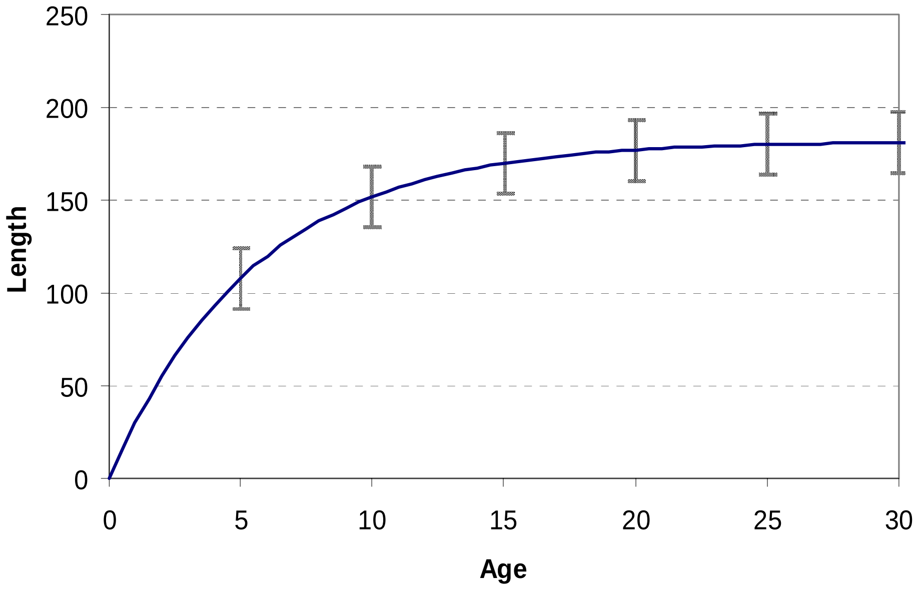


Figure A2.14: Model E_base instantaneous age-length relationship (error bars indicate 5% and 95% distribution limits)

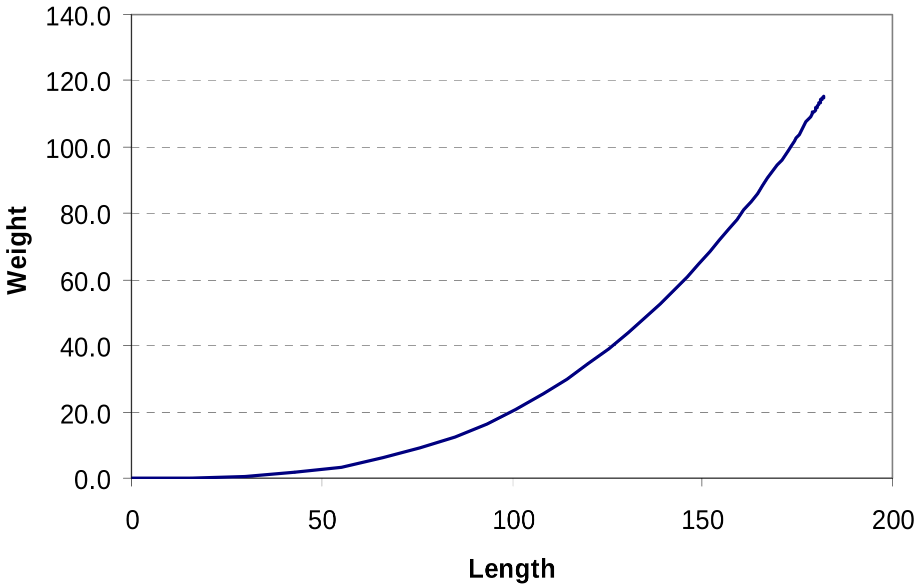


Figure A2.15: Model E_base length-mass relationship

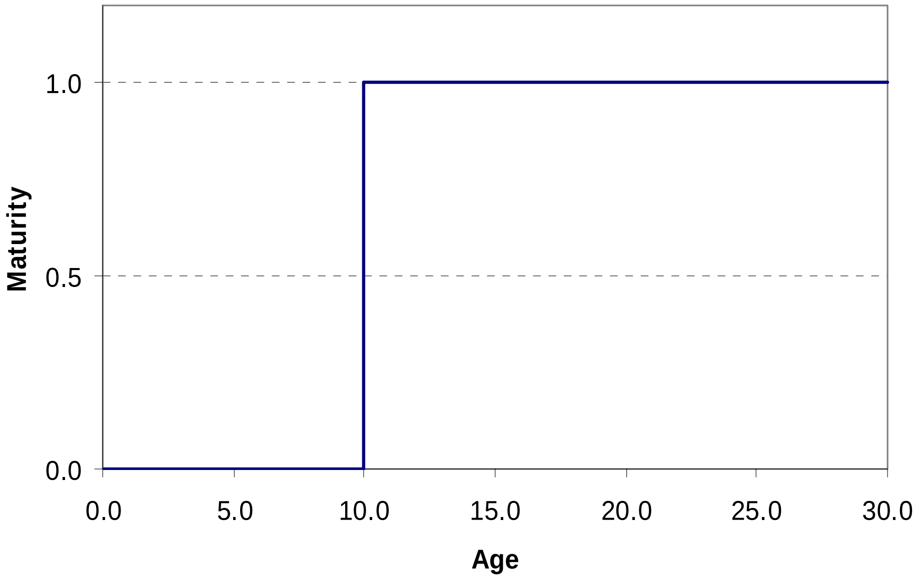


Figure A2.16: Model E_base maturity

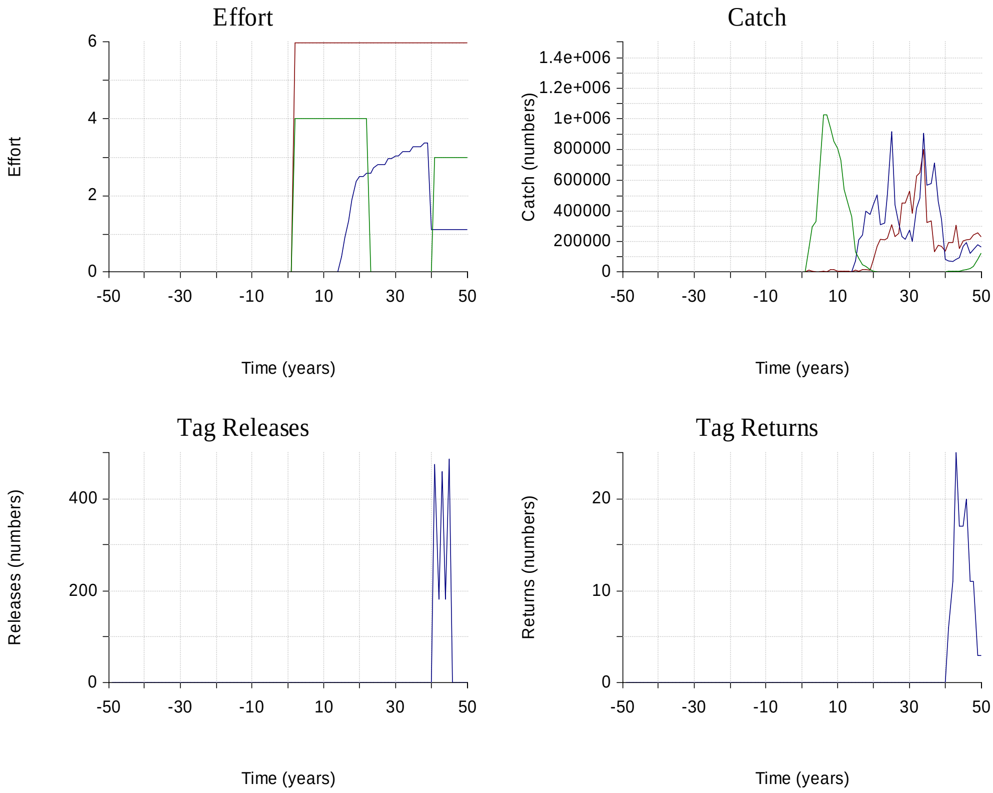


Figure A2.17: Model D_base fishery and tag dynamics from one stochastic state realisation: red - *juvenile* fishery, blue - *long-line feeding ground* fishery, green - *long-line spawning ground* fishery

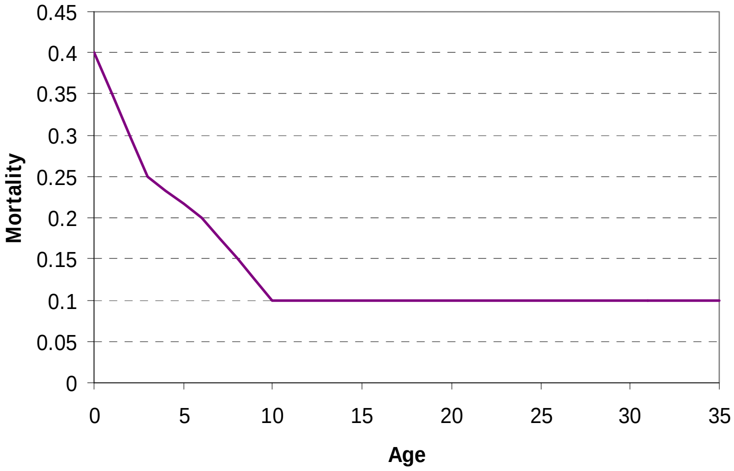


Figure A2.18: Model D_base natural mortality-at-age relationship

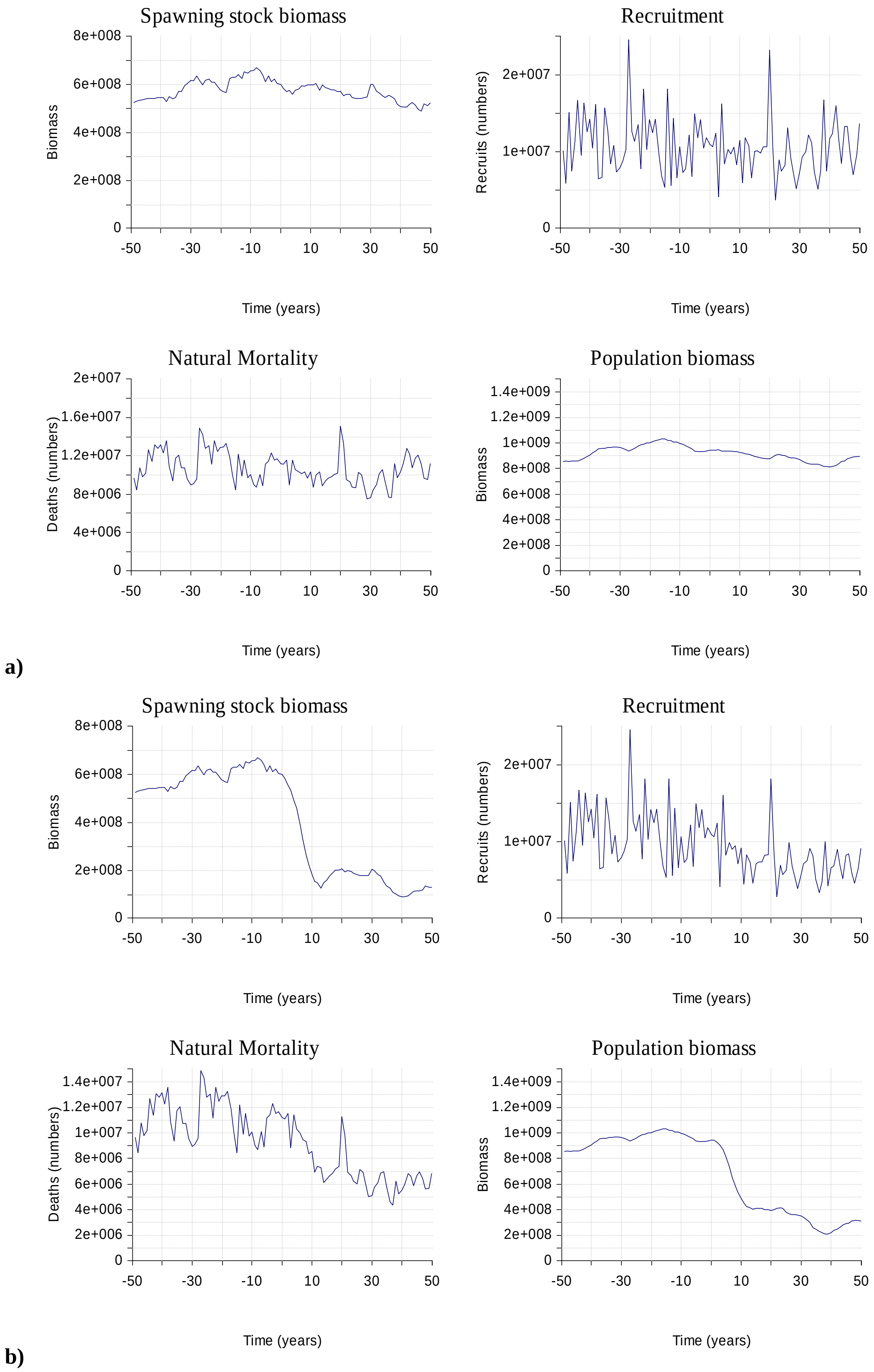


Figure A2.19: Model D_base population dynamics under (a) unfished and (b) fished conditions

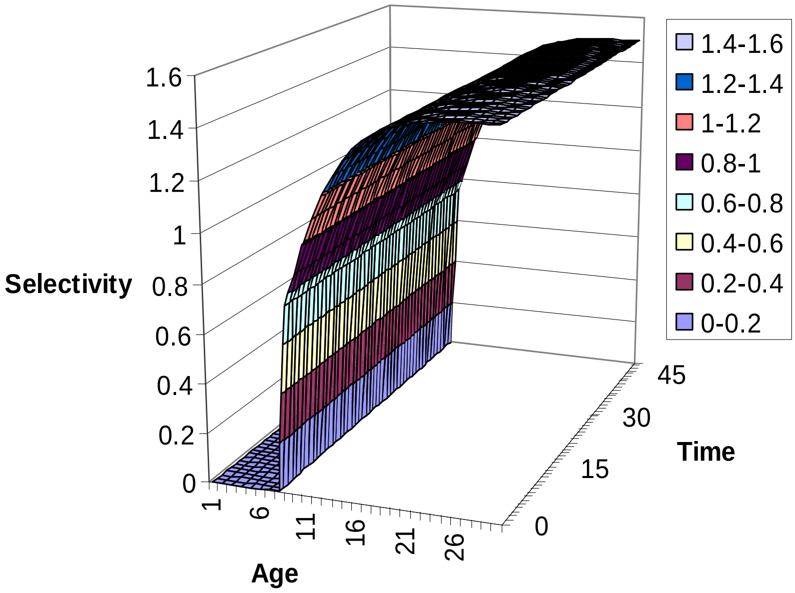


Figure A2.20: Model D_base selectivity: *long-line spawning ground* fishery

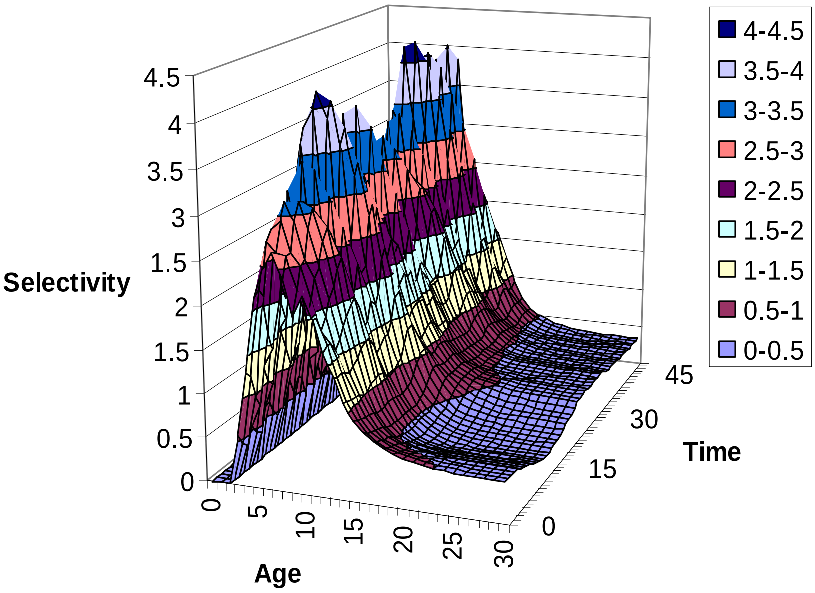


Figure A2.21: Model D_base selectivity: Long-line feeding ground fishery

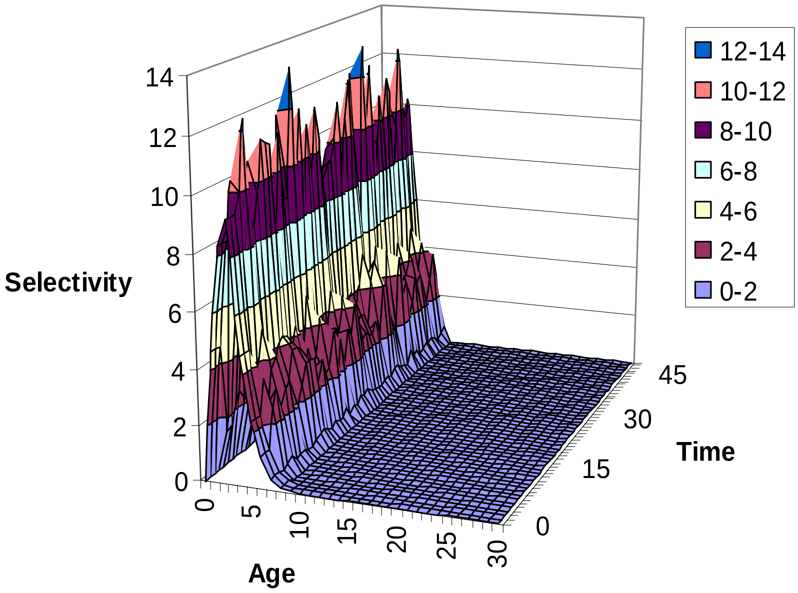


Figure A2.22: Model D_base selectivity: *juvenile* fishery

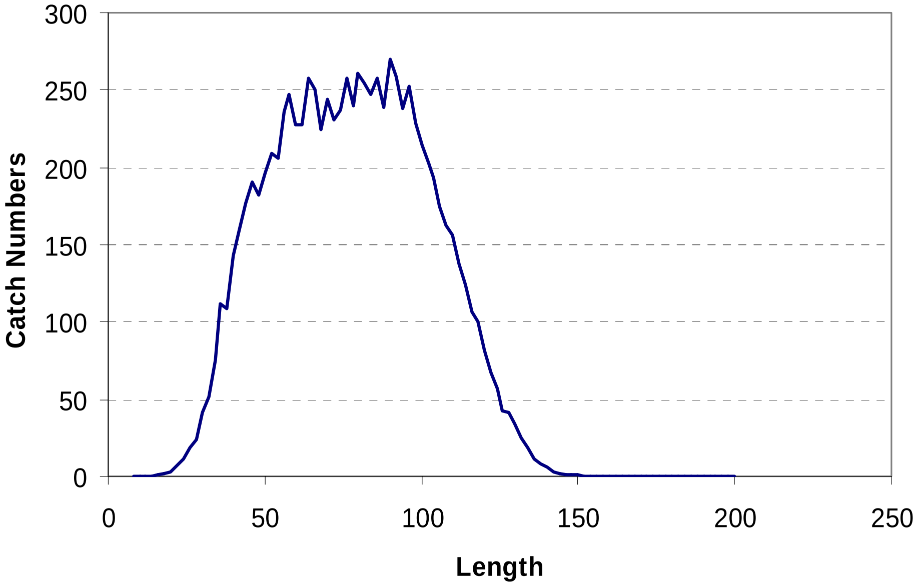


Figure A2.23: Model D_base aggregate (over all time) catch-at-length frequency distribution: *juvenile* fishery

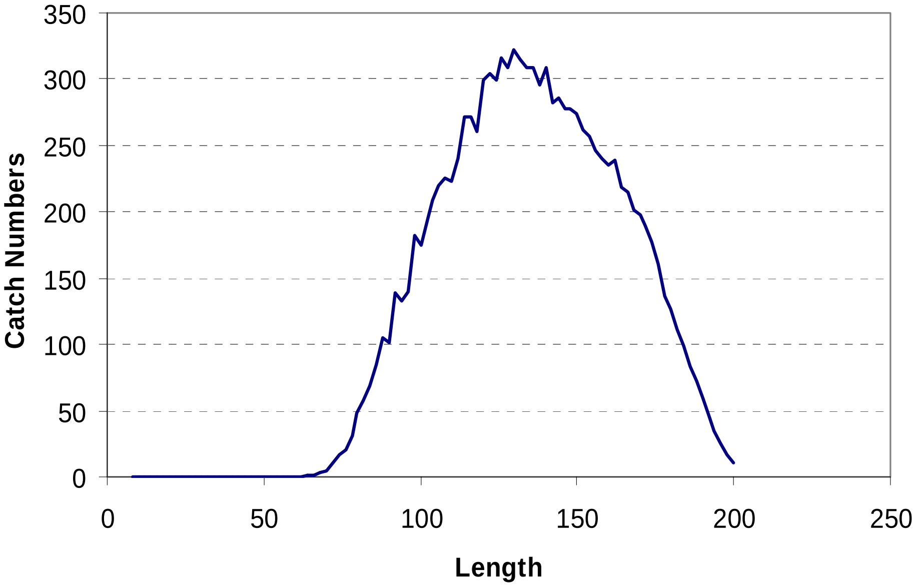


Figure A2.24: Model D_base aggregate (over all time) catch-at-length frequency distribution: *long-line feeding ground* fishery

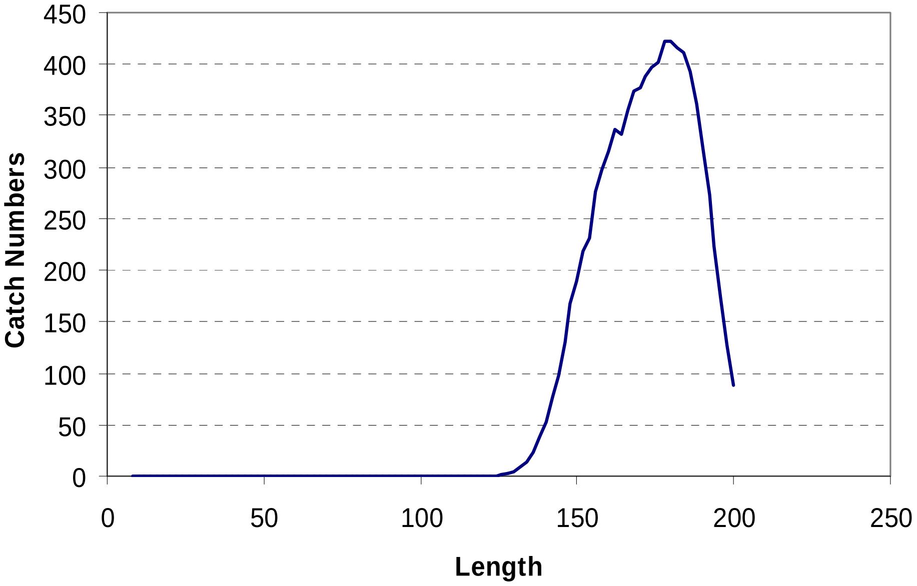


Figure A2.25: Model D_base aggregate (over all time) catch-at-length frequency distribution: *long-line spawning ground* fishery

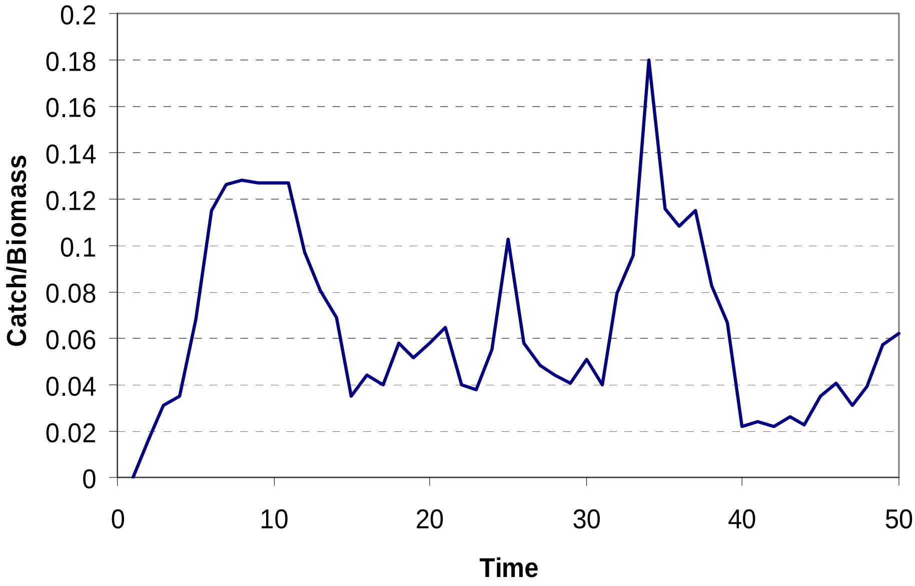


Figure A2.26: Model D_base total exploitation rate time series

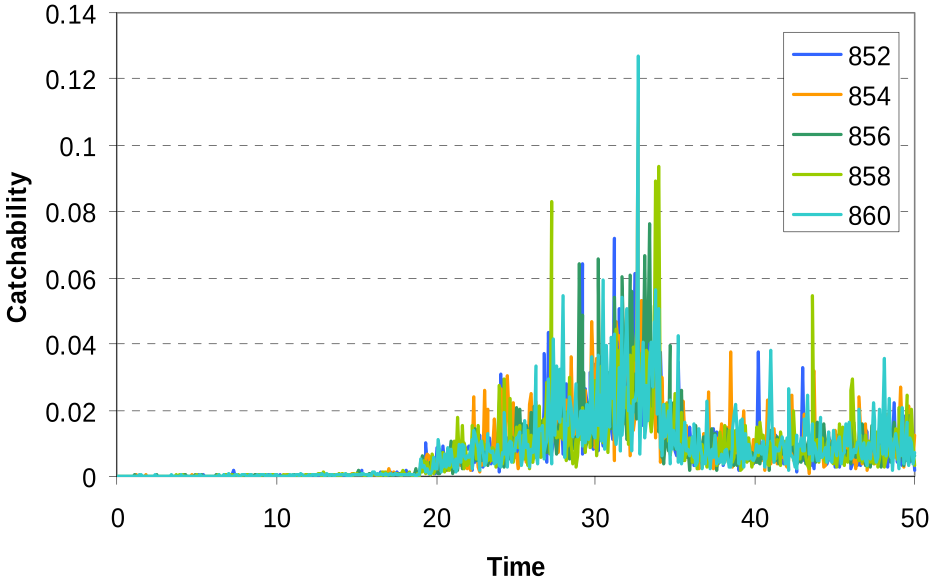


Figure A2.27: Overlayed catchability series (including effective effort deviations) for five different realisations of model D_base: *juvenile* fishery

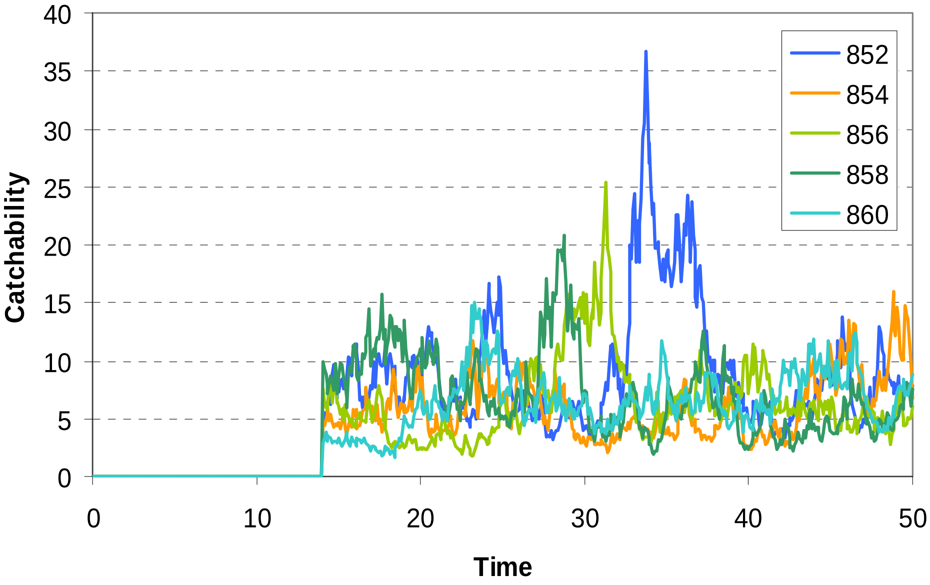


Figure A2.28: Overlayed catchability series (including effective effort deviations) for five different realisations of model D_base: *long-line feeding ground* fishery

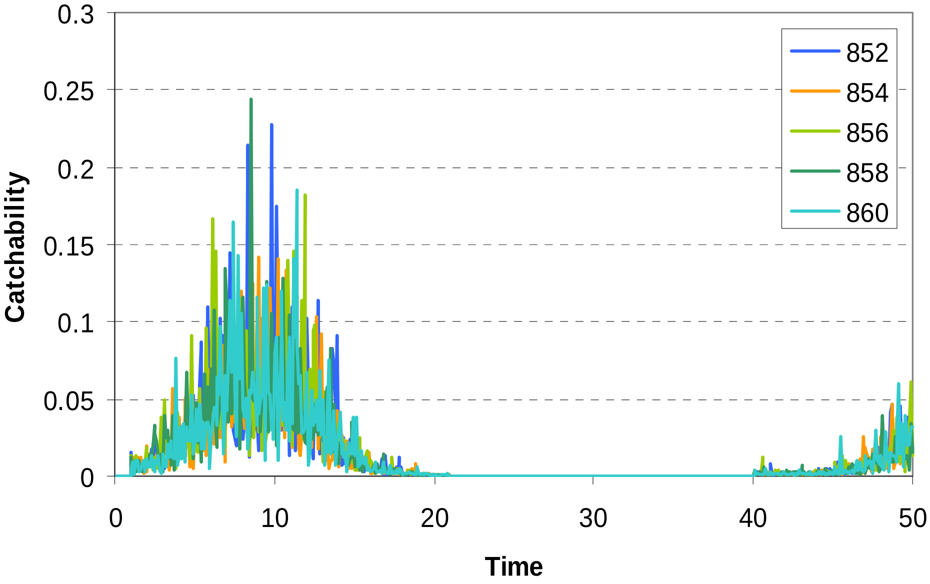


Figure A2.29: Overlayed catchability series (including effective effort deviations) for five different realisations of model D_base: *long-line spawning ground* fishery

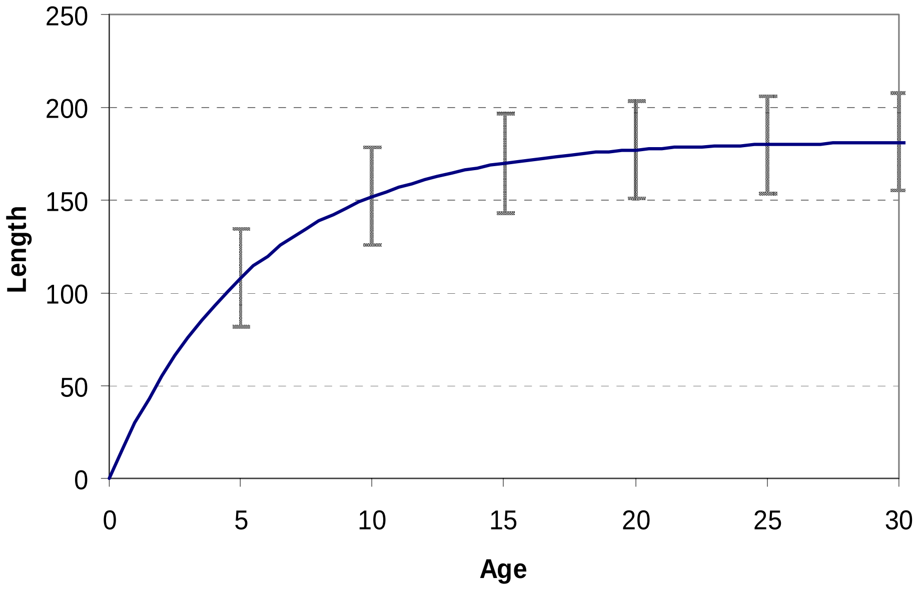


Figure A2.30: Model D_base instantaneous age-length relationship (error bars indicate 5% and 95% distribution limits)

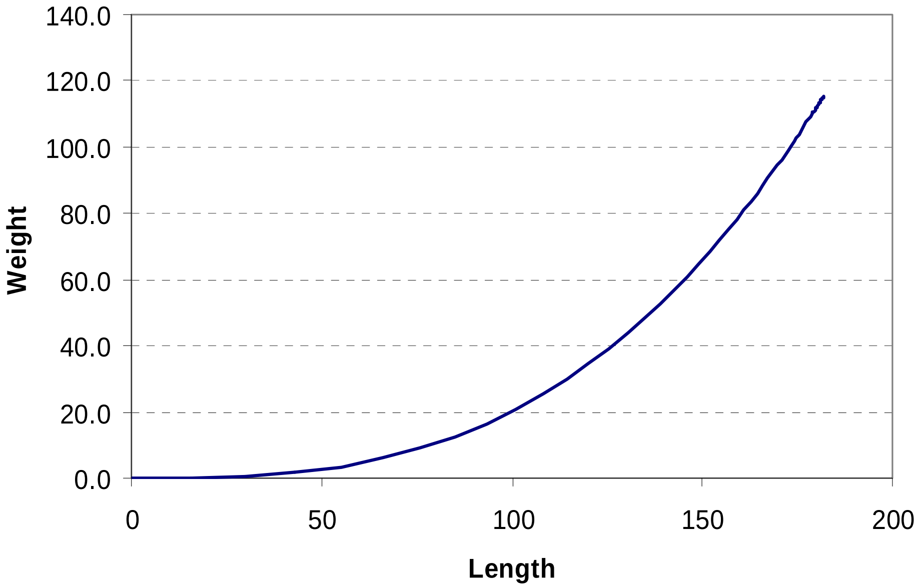


Figure A2.31: Model D_base length-mass relationship

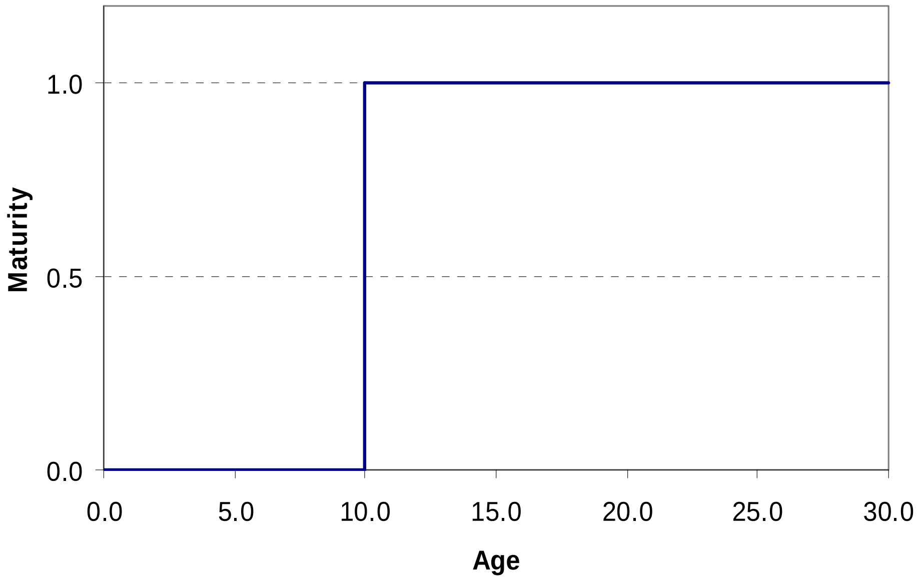


Figure A2.32: Model D_base maturity

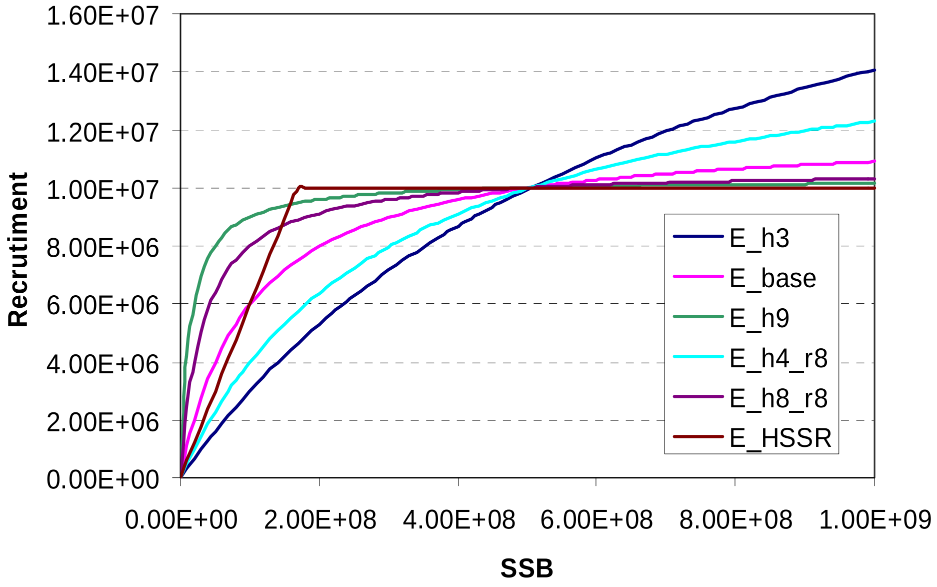


Figure A2.33: Recruitment relationships for models E_h3, E_base, E_h9, E_h4_r8, E_h8_r8 and E_HSSR

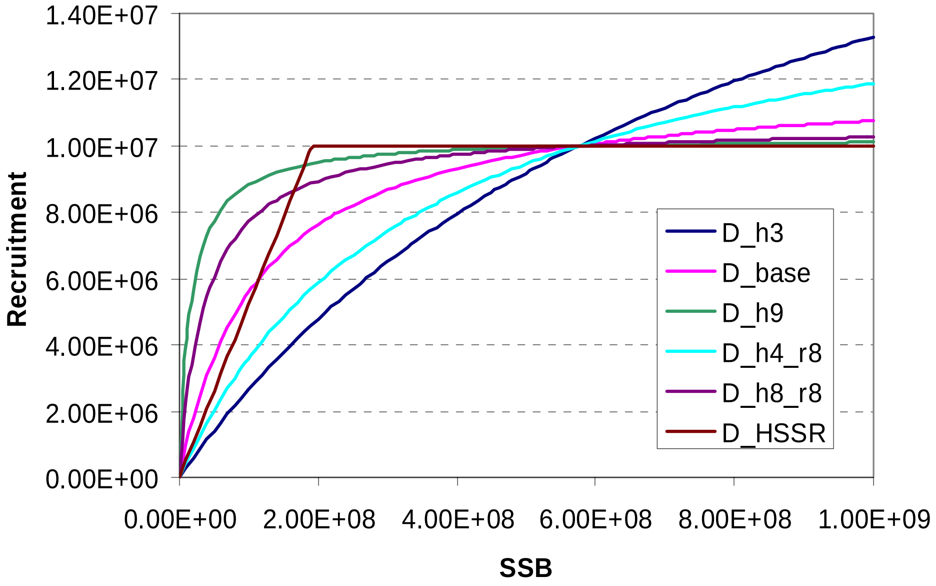


Figure A2.34: Recruitment relationships for models D_h3, D_base, D_h9, D_h4_r8, D_h8_r8 and D_HSSR

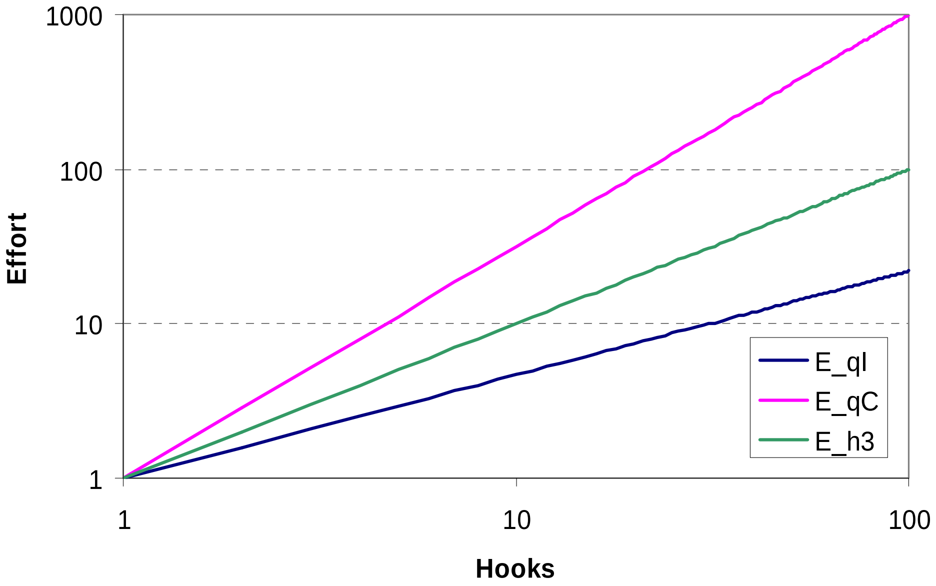


Figure A2.35: Effort-Fishing Mortality Relationship: Model E_h3 Effort = Hooks; Model E_qC Effort = Hooks1.5; Model E_qI Effort = Hooks0.67

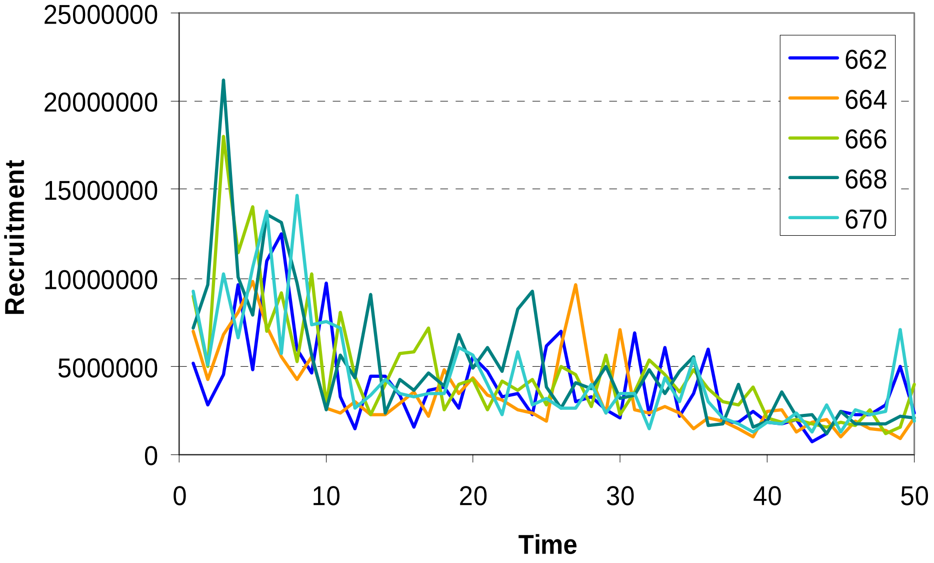


Figure A2.36: Overlayed recruitment time series for five different realisations of E_h3. Recruitment in E_h3 has no auto-correlation.

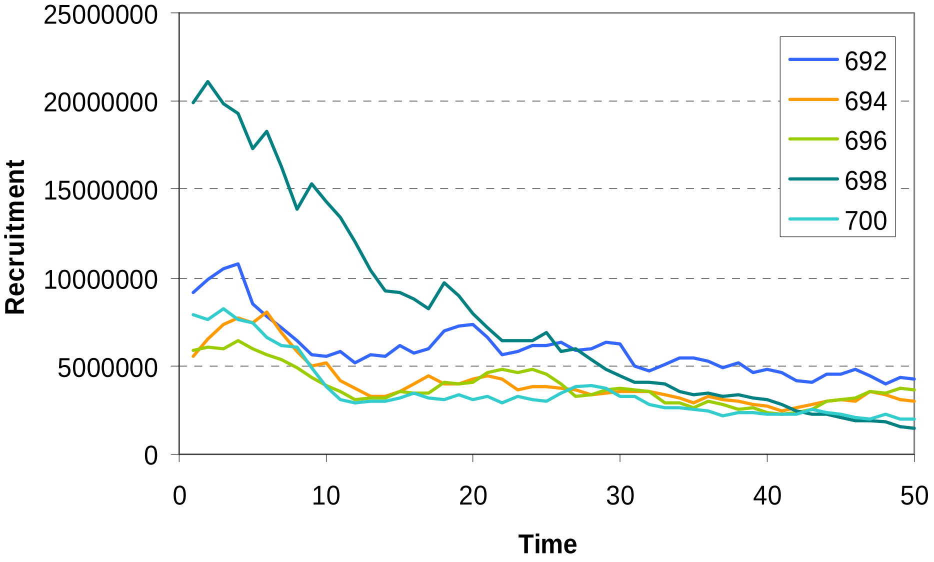


Figure A2.37: Overlayed recruitment time series for five different realisations of E_h4_r8. Recruitment in E_h4_r8 has a lag one year auto-correlation of 0.8.

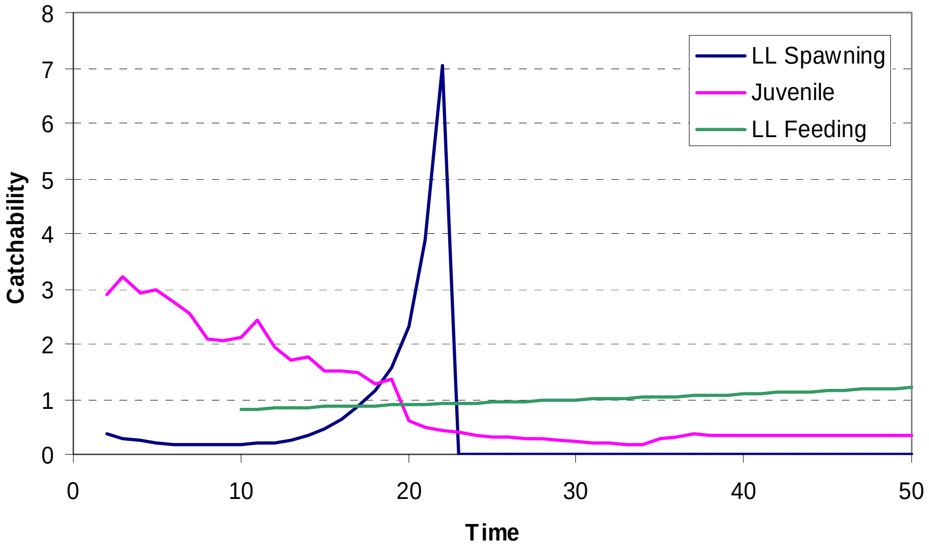


Figure A2.38: Standardised catchability trends used in models E_qInc and D_qInc: *Long-line feeding ground* fishery (LL Feeding) has increasing catchability trend 1% per year from the 10th year onwards; *Long-line spawning ground* fishery (LL Spawning) has catchability increasing exponentially in fishing years; *Juvenile* fishery has catchability exponentially decreasing with superimposed noise

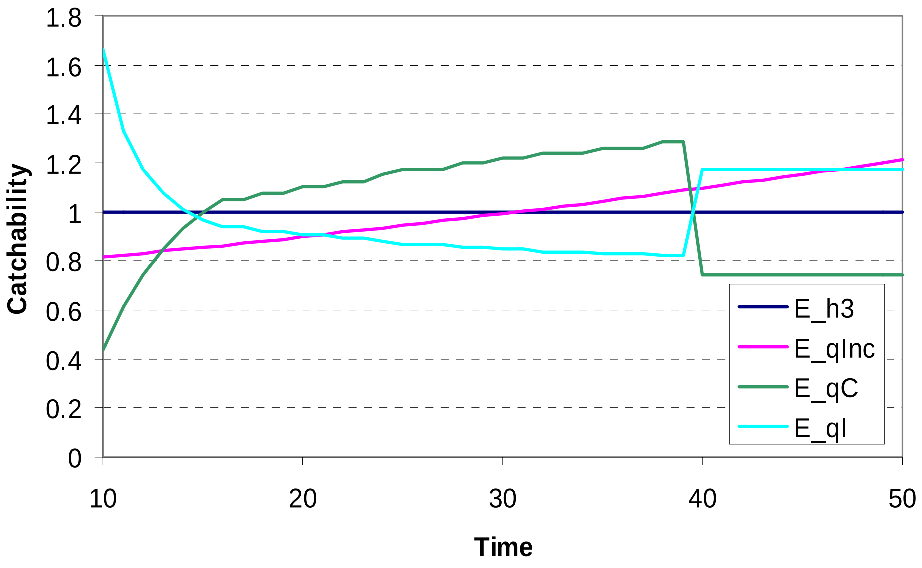


Figure A2.39: Comparison of standardised catchability trends for the *Long-line feeding ground* fishery in models E_h3, E_qInc, E_qC and E_qI. Note that the catchability trends in E_qC and E_qI stem from the non-linear hooks-effort relationship used in these cases

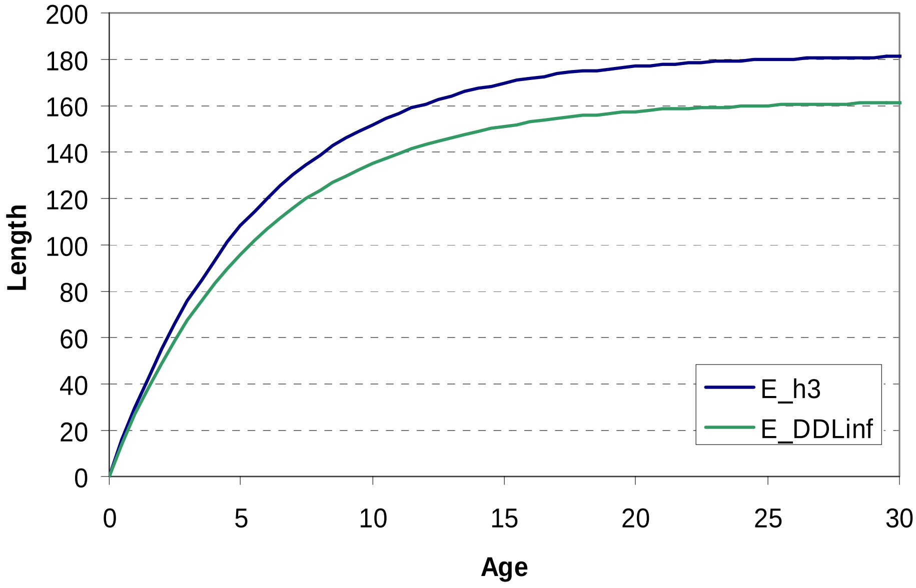


Figure A2.40: Growth equation changes in models E_DDLinf and D_DDLinf. *L_∞* changes from 182 cm (E_h3 plot) to 162 cm (E_DDLinf plot) in 10 years from the 10th year onwards

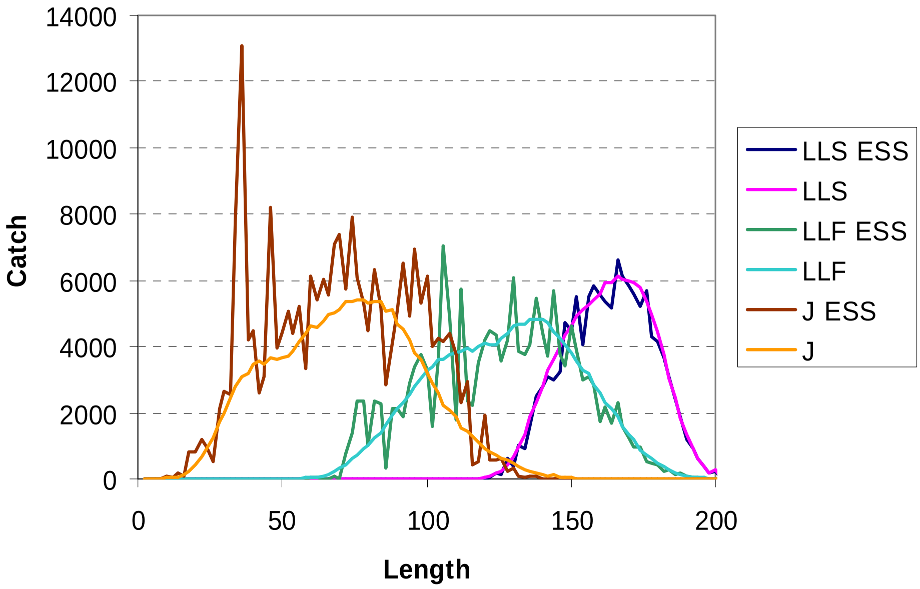


Figure A2.41: Catch in final year for model D_h3 with and without the use of stochastic variation in the system dynamics model. *LLS - Long-line spawning ground* fishery; *LLF - Long-line feeding ground* fishery; *J - Juvenile* fishery; *ESS* - with an effective sample size specification *Ψ* = 500 and *Ψmin* = 50

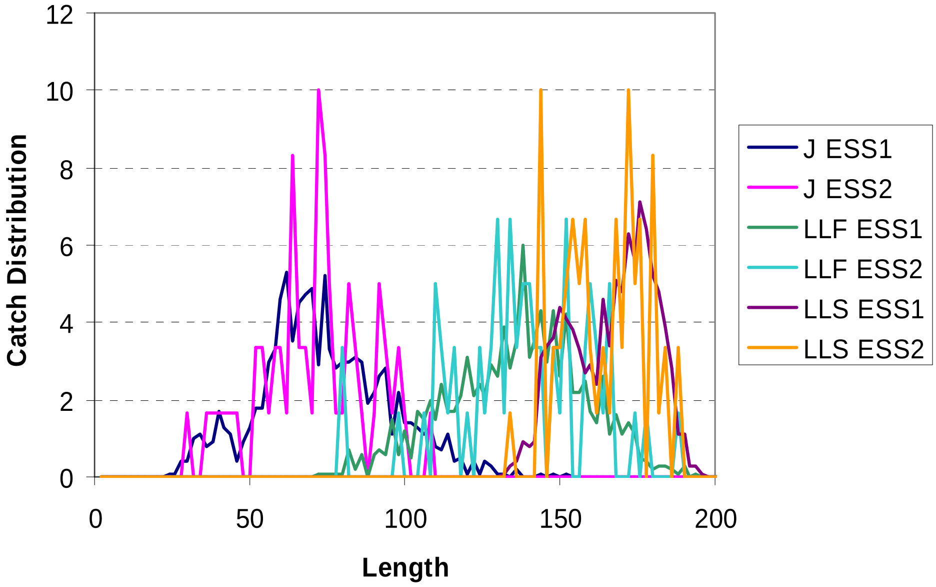


Figure A2.42: Standardised reported catch distribution in the final year for model E_CA60 with the use of two different effective sample sizes in the observation model. *LLS - Long-line spawning ground* fishery; *LLF - Long-line feeding ground* fishery; *J - Juvenile* fishery; *ESS1* - with length based effective sample size specification L_{ess} = 1000; *ESS2* - with length based effective sample size specification L_{ess} = 60

APPENDIX 3 AGE-AGGREGATED AND AGE-STRUCTURED PRODUCTION MODELS TECHNICAL DESCRIPTION

This document describes the application of age-aggregated (Schaefer and Fox) and Age-Structured Production Models (ASPMs) to the simulated data generated by the VSM SBT and SPC-OFP YFT operating models. The Schaefer, Fox and deterministic ASPM implementations are adapted from actual applications to SBT (Butterworth and Plaganyi 2001; Ricard et al. 2002). This document provides brief technical details including data processing requirements, and describes difficulties encountered in the automated application of these models to the simulated data. All models were implemented using AD Model Builder software (Otter Research, Victoria, Canada).

A 3.1 AGE-AGGREGATED PRODUCTION MODELS (AAPMS)

Table A 3-1. AAPM Variations

Model name	"Free" Parameters estimated with function minimizer*	Details
Applied to simulated VSM SBT data		
f_calc	r, K	Fox model; uses longline feeding grounds CPUE as relative abundance index
s_calc	r, K	Schaefer model; uses longline feeding grounds CPUE as relative abundance index
Applied to simulated SPC-OFP YFT data		
Fox	r, K	Fox model; uses CPUE from the longline fishery with the largest catch as a relative abundance index
Schaefer	r, K	Schaefer model; uses CPUE from the longline fishery with the largest catch as a relative abundance index
Fox_Agg	r, K	Fox model; uses global nominal CPUE as a relative abundance index
Schaefer_Agg	r, K	Schaefer model; uses global nominal CPUE as a relative abundance index

* note that the actual parameters estimated may have undergone various transformations to improve stability during function minimization.

A 3.1.1 AAPM Biomass Dynamics

The Fox and Schaefer model dynamics are identical except for the logs in the density dependent terms:

Schaefer

$$B_{t+1} = B_t + rB_t\left(1 - \frac{B_t}{K}\right) - C_t \qquad \text{(Eq A 3-1)}$$

Fox

$$B_{t+1} = B_t + rB_t\left(1 - \frac{\log_e(B_t)}{\log_e(K)}\right) - C_t \qquad \text{(Eq A 3-2)}$$

where:

t = time-step; annual for SBT and quarterly for YFT.
B_t = biomass at time t
C_t = total catch in mass at time t summed over all fisheries
r = intrinsic population growth rate parameter
K = carrying capacity parameter

A 3.1.2 AAPM Objective Function

The parameters from both AAPMs are estimated using the same objective function that assumes that the relationship between the population abundance and the CPUE is:

$$CPUE_t = q\left(\frac{B_t + B_{t+1}}{2}\right)e^{\varepsilon_t} \qquad \text{(Eq A 3-3)}$$

$$-\log_e(L) = \sum_t\left[\log_e(\sigma) + \left(\frac{(\varepsilon_t)^2}{2\sigma^2}\right)\right] \qquad \text{(Eq A 3-4)}$$

where:

ε_t are the observation errors, assumed to be normally distributed $N(0,(\sigma_t)^2)$), and
σ is the estimated standard deviation of ε_t.

A 3.2 AGE-STRUCTURED PRODUCTION MODELS (ASPMS)

Table A 3-2. **ASPM variations**

Model name	# parameters estimated	"Free" Parameters estimated with function minimizer *	Details
Applied to simulated VSM SBT data			
aspm_d2g	2	h, K^{sp}	ASPM with deterministic recruitment; correct selectivity and mortality used as fixed input
aspm_d6g	2	h, K^{sp}	ASPM with deterministic recruitment; known mortality used as fixed input; selectivity analytically calculated from length frequency approximation
aspm_sto (results withdrawn)	52 (50 annual recruitment deviations)	h, K^{sp}, ϕ	ASPM with stochastic recruitment
Applied to simulated SPC-OFP YFT data			
aspm_det (results withdrawn)	2	h, K^{sp}	ASPM with deterministic recruitment; known mortality used as fixed input; selectivity analytically calculated from length frequency approximation
aspm_sto (results withdrawn)	150 (148 quarterly recruitment deviations)	h, K^{sp}, ϕ	ASPM with stochastic recruitment

* note that the actual parameters estimated may have undergone various transformations to improve stability during function minimization.

A 3.2.1 ASPM Population dynamics

The dynamics of the fish population are described by three equations:

$$N_{t+1,0} = R\left(B_{t+1}^{sp}\right) \qquad \text{(Eq A 3-5)}$$

$$N_{t+1,t+1} = \left(N_{t,a} e^{-\frac{M_a}{2}} - C_{t,a} \right) e^{-\frac{M_a}{2}} \qquad \text{(Eq A 3-6)}$$

$$N_{t+1,m} = \left(N_{t,m} e^{-\frac{M_m}{2}} - C_{t,m} \right) e^{-\frac{M_m}{2}} + \left(N_{t,m-1} e^{-\frac{M_{m-1}}{2}} - C_{t,m-1} \right) e^{-\frac{M_{m-1}}{2}} \qquad \text{(Eq A 3-7)}$$

where,

t = time-step; annual for SBT and quarterly for YFT.

$N_{t,a}$ is the number of tuna age *a* at the start of time-step *t*

$R\left(B_{t+1}^{sp}\right)$ is the stock recruitment relationship assumed

$C_{t,a}$ is the total number of tuna age *a* taken by the fishery in year *t*

M_a is the natural mortality rate for fish age *a*

m is the largest age considered (the "plus" group)

The fishery is assumed to occur as a pulse catch in the middle of the year. The total number of tuna of age *a* caught each year ($C_{t,a}$) is given by:

$$C_{t,a} = \sum_f C_{t,a}^f \qquad \text{(Eq A 3-8)}$$

where,

f is fishery/fleet concerned

The mass of the fleet-specific annual catch (C_t^f) is given by:

$$\begin{aligned} C_t^f &= \sum_{a=0}^{m} w_{a+\frac{1}{2}} C_{t,a}^f \\ &= \sum_{a=0}^{m} w_{a+\frac{1}{2}} S_a^f F_t^f N_{t,a} e^{-\frac{M_a}{2}} \end{aligned} \qquad \text{(Eq A 3-9)}$$

where,

S_a^f is the fleet-specific selectivity for tuna of age *a*

F_t^f is the fleet-specific fishing mortality for year t

$w_{a+\frac{1}{2}}$ is the weight at mid-time-step

The fleet-specific exploitable biomass is calculated as:

$$B_t^f = \sum_{a=0}^{m} w_{a+\frac{1}{2}} S_a^f N_{t,a} e^{-\frac{M_a}{2}} \qquad \text{(Eq A 3-10)}$$

The proportion of the resource harvested each year (F_t^f) by fleet *f* is given by:

$$F_t^f = C_t^f / B_t^f \qquad \text{(Eq A 3-11)}$$

and

$$C_{t,a}^f = S_a^f F_t^f N_{t,a} e^{-\frac{M_a}{2}} \qquad \text{(Eq A 3-12)}$$

A 3.2.2 Stock recruitment relationship

The spawning biomass in year *y* is:

$$B_t^{sp} = \sum_{a=0}^{m} f_a w_a N_{t,a} \qquad \text{(Eq A 3-13)}$$

where,

f_a is the proportion of sexually mature tuna at age *a*

In the simplest case, recruitment is calculated using a Beverton-Holt relationship:

$$R(B_t^{sp}) = \frac{\alpha B_t^{sp}}{\beta + B_t^{sp}} \qquad \text{(Eq A 3-14)}$$

To facilitate the biological interpretation of the stock-recruitment parameters we reparameterise the Beverton-Holt relationship in terms of the pre-exploitation equilibrium spawning biomass (*Ksp*) and the "steepness" (*h*) of the stock-recruitment relationship. Steepness is defined as the fraction of the pristine recruitment (*R*0) that results when the spawning biomass drops to 20% of its pristine level:

$$hR_0 = R(0.2K^{sp}) \qquad \text{(Eq A 3-15)}$$

from which it follows that:

$$h = 0.2[\beta + K^{sp}]/[\beta + 0.2K^{sp}] \qquad \text{(Eq A 3-16)}$$

and hence:

$$\alpha = \frac{4hR_0}{5h-1} \qquad \text{(Eq A 3-17)}$$

and

$$\beta = \frac{K^{sp}(1-h)}{5h-1} \qquad \text{(Eq A 3-18)}$$

The stochastic version of the ASPM requires the estimation of additional parameters. We estimate recruitment as deviations from the deterministic recruitment.

$$R_t^* = R_t * e^{\tau_t + (\sigma_r^2/2)} = \frac{\alpha B_t^{sp} e^{\tau_t + (\sigma_r^2/2)}}{\beta + B_t^{sp}} \qquad \text{(Eq A 3-19)}$$

$$\tau_t = \log_e\left(\frac{R_t^*}{R_t}\right) - (\sigma_r^2/2) = \log_e(R_t^*) - \log_e(R_t) - (\sigma_r^2/2) \qquad \text{(Eq A 3-20)}$$

where,

R_t^* is the estimated recruitment at time *t*

R_t is the deterministic recruitment at time *t*

τ is a vector of recruitment deviations

σ_r is the standard deviation around recruitment deviations

We note that the stochastic recruitment implementation generally required user interaction to obtain seemingly satisfactory convergence, and was not adequate for the automated SESAME simulation testing.

A 3.2.3 Biomass trajectories

Given a value for the pre-exploitation equilibrium spawning biomass K_{sp} and assuming that the initial age structure is at equilibrium, the initial recruitment R_0 can be estimated as:

$$R_0 = K^{sp} /[\sum_{a=1}^{m-1} f_a w_a e^{-\sum_{a=0}^{m-1} M_a} + (f_m w_m e^{-\sum_{a=0}^{m-1} M_a} /(1 - e^{-M_m}))] \qquad \text{(Eq A 3-21)}$$

An additional parameter (γ) can be estimated to allow the stock to be at a state other than the unfished equilibrium at the onset of fishing. Note that the population structure, as represented by the proportion of fish in each age class, will be similar to that of the unfished equilibrium. The extra parameter simply scales the initial population:

$$N_{0,0} = R_0 e^{\gamma} \qquad \text{(Eq A 3-22)}$$

The notation $N_{0,0}$ means the number of recruits at the onset of fishing. Initial abundance of older age classes are calculated as:

$$N_{a,0} = N_{a-1,0} e^{-M_{a-1}} \qquad \text{(Eq A 3-23)}$$

Once the numbers-at-age of the population at the onset of fishing have been calculated, the population dynamics can be obtained through equations 3 through 14.

A 3.2.4 Objective function

To estimate the stock recruitment parameters *h* and K_{sp}, the model is fitted to an index of abundance by maximizing an associated likelihood function. The likelihood is calculated assuming that the observed index of abundance is log-normally distributed about its expected value:

$$I_t^l = \hat{I}_t^l e^{\varepsilon_t^l} \text{ or } \varepsilon_t^l = \ln(I_t^l) - \ln(\hat{I}_t^l) \qquad \text{(Eq A 3-24)}$$

where,

I_t^l is the longline fleet abundance index for time *t*

$\hat{I}_t^l = q^l N_t^l$ is the corresponding model estimated value, where N_t^l is the model value for the longline exploitable resource abundance (Eq A 3-10)

q^l is the constant longline catchability coefficient

ε_t^l is assumed to be normally distributed $N(0,(\sigma_t^l)^2))$

The simplified log-likelihood function for the indices of abundance is given by:

$$\log_e(L_1) = -\sum_y \left[\log_e \sigma_t^l + \left(\frac{(\varepsilon_t^l)^2}{2(\sigma_t^l)^2} \right) \right] \qquad \text{(Eq A 3-25)}$$

Independent estimates of $N(0,(\sigma_t^l)^2)$ are not available so they are assumed not to be dependent on year (σ_t^l is simplified to σ^l). σ^l is estimated in the likelihood maximization process as:

$$\sigma^l = \sqrt{\frac{\sum_t (\varepsilon_t^l)^2}{n}} \qquad \text{(Eq A 3-26)}$$

where n is the number of data points in the abundance time series. The indices of abundance component of the log-likelihood can be further simplified to:

$$\log_e(L_2) = -n \cdot \log_e(\sigma^l) - \frac{n}{2} \qquad \text{(Eq A 3-27)}$$

Under this assumption, the maximum likelihood estimate of q^l is given by:

$$\hat{q}^l = \exp\left[\sum_t \left(\log_e(I_t^l) - \log_e(N_t^l)\right)\right] \qquad \text{(Eq A 3-28)}$$

The deterministic version of the ASPM estimates parameters h and K_{sp} by maximizing $\log_e(L_1)$ (Eq A 3-25). The stochastic version of the ASPM estimates parameters h, K_{sp}, *and* ϕ by maximizing

$$\log_e(L) = \log_e(L_1) + \log_e(L_2) \qquad \text{(Eq A 3-29)}$$

In either case, parameter γ is estimated if the stock is assumed to be at a state other than the unfished equilibrium at the onset of fishing. Note that minimization on the negative log likelihood is in fact used in the software (i.e. minimise: -log$_e$(L)).

A 3.3 AAPM and ASPM Data Processing and Biological Assumptions

A 3.3.1 Computation of total catch biomass time-series

The SBT and YFT simulations provided total catches in mass or numbers depending on the fleet. Catch in numbers was converted to catch in mass by calculating the mass frequency distribution corresponding to the length frequency distribution using the mass-length relationship that was made available with the simulated data. The total catch biomass is the sum of catch biomass from the long-line and purse seine fisheries.

A 3.3.2 Computation of nominal LL CPUE

For the SBT simulations, the nominal CPUE from the longline feeding grounds was used as the relative abundance index.

For the YFT simulations, two different relative abundance indices were tested. In the base case, we used the nominal longline CPUE from one of the regions with the highest average catch. In the scenarios where there were more than one LL fishery (scenarios 3-5), we used the global nominal CPUE from all longline fisheries combined (total catch / total hooks).

A 3.3.3 ASPM Biological parameters

The amount of prior information available to the analysts varied in the two studies. Both the SBT and YFT simulations provided some information about length-at-age, maturity-at-age and length-mass relationships. For at least one of the YFT scenarios, sexual dimorphism was present but there was no prior information that this was the case.

The SBT ASPM applications used perfect knowledge of natural mortality-at-age, while the prior knowledge provided with the YFT data only suggested that the maximum longevity of a fish is 28 quarters. For the YFT, we assumed a constant natural mortality rate of 0.15 per quarterly timestep, such that the abundance of fish of age 28 quarters represents 1.7% of the abundance of fish aged 1 quarter. Figure A 3-1 shows the abundance-at-age with and without the fishing mortality.

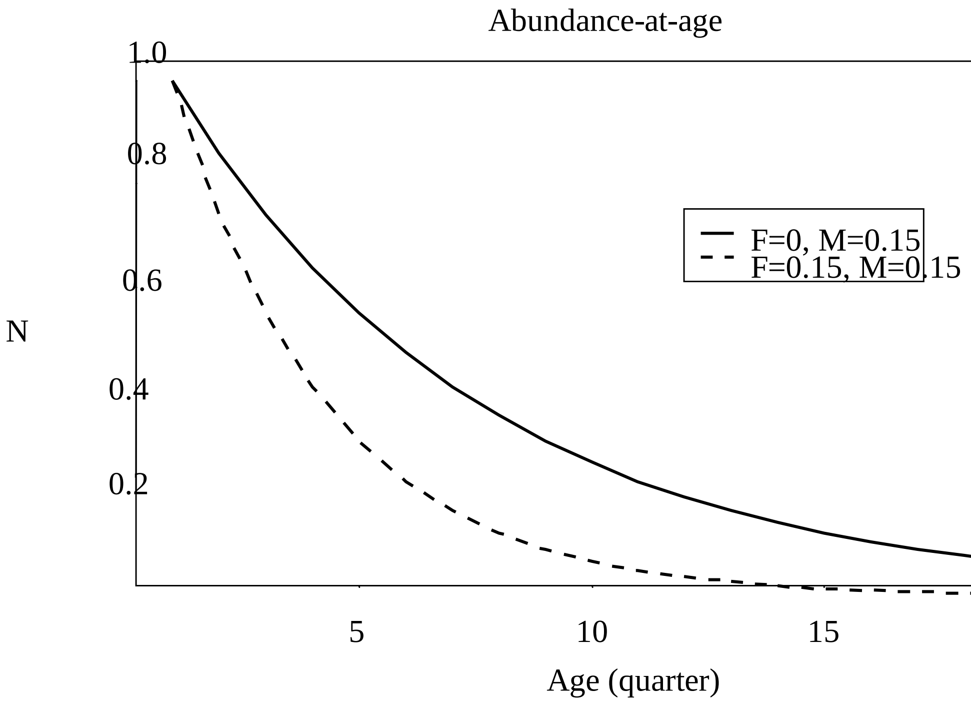


Figure A 3-1 YFT assumed equilibrium abundance-at-age with constant natural mortality = 0.15 and fishing mortality = 0 or 0.15 for all ages.

A 3.3.4 Selectivity Calculation

The ASPMs required fixed input for fleet-specific selectivity. The true values were available for the SBT simulations and used in model aspm_d2g. However, we also made a simple attempt to calculate selectivity based on the observed catch length frequency distributions and equilibrium age structure assumptions (aspm_d6g). For the SBT simulations, selectivities were set constant at ages 13+ years. This is a gross approximation at best, but we did not want to invoke any additional complicated analyses that would detract from the simplicity that formed the basis for testing ASPMs in the first place. We used this approach for the YFT simulations because there was no selectivity information provided in this case.

The equilibrium age-composition of the population assuming that F = M is shown in Figure A 3-1 with the corresponding length frequency distribution shown in Figure A 3-2. We assumed that the ratio of the catch length frequency distribution (total for each fleet over the whole time series) to the equilibrium population length frequency distributions provides some measure of the relative selectivity of the different fisheries (Figure A 3-3). This approximate selectivity-at-length was converted to relative selectivity-at-age by taking the ratio from Figure A 3-3 that corresponds to the mean length-at-age for each age-class. The small number of catch samples at larger sizes and the errors introduced by this "cohort-slicing" approach would be expected to yield particularly bad estimates for older age classes. For the YFT scenarios, we set the longline selectivity on age classes 12 quarters and older equal to the selectivity at age 12 quarters and the purse seine selectivity on age classes 15 quarters and older equal to the selectivity at age 15 quarters. The selectivity calculation was applied independently for each data realization, and a typical result is illustrated in Figure A 3-4.

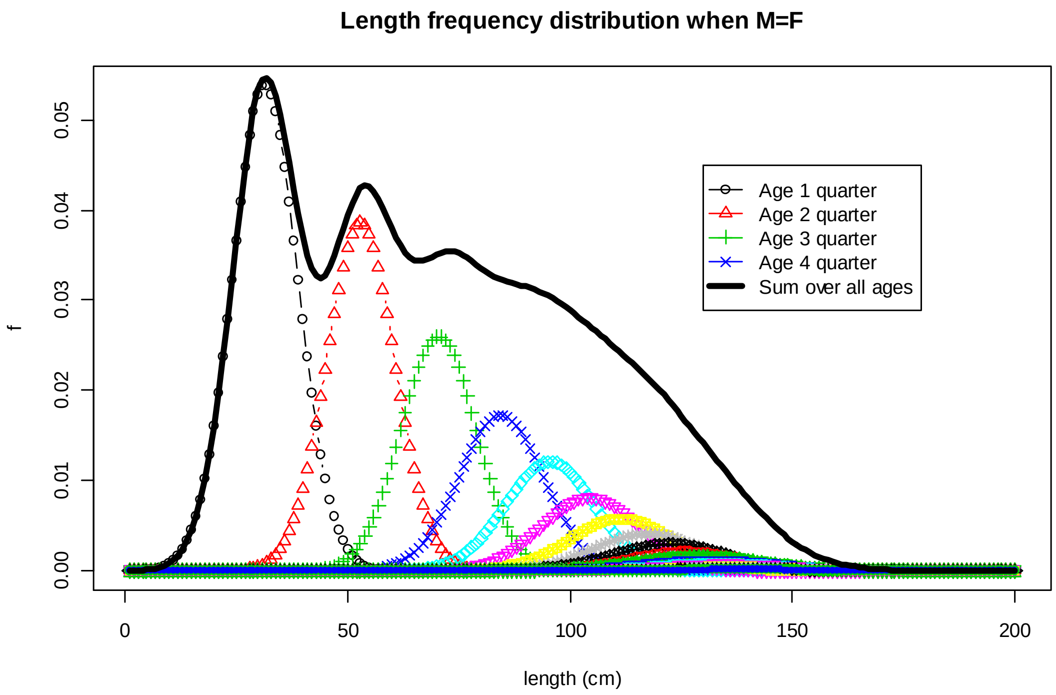


Figure A 3-2 Equilibrium length frequency distribution of the fish population when M=F=0.15 for YFT.

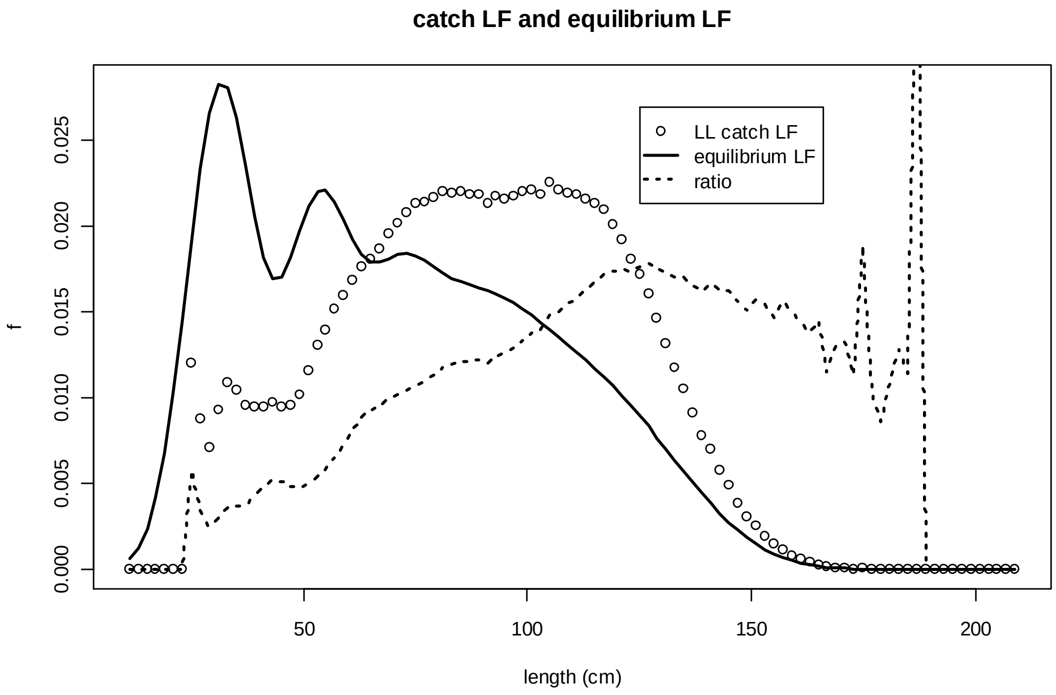


Figure A 3-3 Equilibrium length frequency and longline catch length frequency for YFT. The ratio of these two distributions is used to estimate the selectivity of the longline fleet.

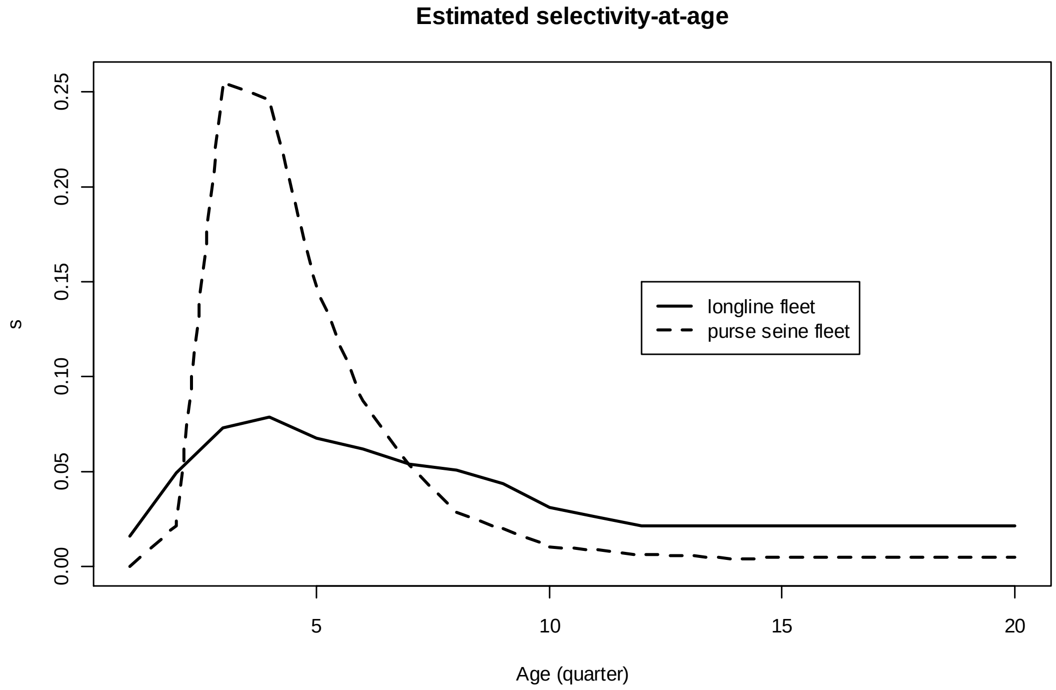


Figure A 3-4 Estimated YFT selectivity-at-age for the longline and purse seine fleets.

A 3.3.5 Comments on Parameter estimation

Both the AAPMs and ASPMs were problematic in the automated fitting context required in the simulation testing. Without good starting parameter estimates, these models frequently failed to converge to the global minimum, sometimes yielding bizarre results that were obviously identifiable as flawed if examined. We were able to avoid the initial sensitivity problem for the deterministic ASPM by introducing an automated two dimensional grid search of starting values to get near the global minimum prior to running the function minimizer. The plot in Figure A 3-5 shows an example likelihood surface.

The ASPMs were subject to additional implementation problems. The ASPM with stochastic recruitment failed to converge reliably in the majority of cases in both the SBT and YFT scenarios. This is not too surprising given the limited data and large number of parameters that were being estimated. We did not expend much effort trying to improve the automated stochastic ASPM convergence performance because these models failed to meet the original objective of representing a simple alternative to the fully integrated models. The deterministic ASPMs were also prone to a numerical problem that affected the majority of the YFT simulations and a few of the SBT simulations. The problem arises from Eq A 3-6, when the model attempts to extract more than the existing number of fish (i.e. C(a,t) > N(a,t)). We used penalties to try to prevent the function minimizer from wandering into the parameter space where this occurs, however, in some cases the likelihood surface is minimized arbitrarily close to the point where C(a,t) approaches N(a,t) and hence is ultimately determined by the arbitrary nature of this penalty. This is illustrated in Figure A 3-6, and we do not consider these results to be valid. Because of these problems, we withdrew all of the ASPM results from the YFT study, and have flagged the problem in the SBT results.

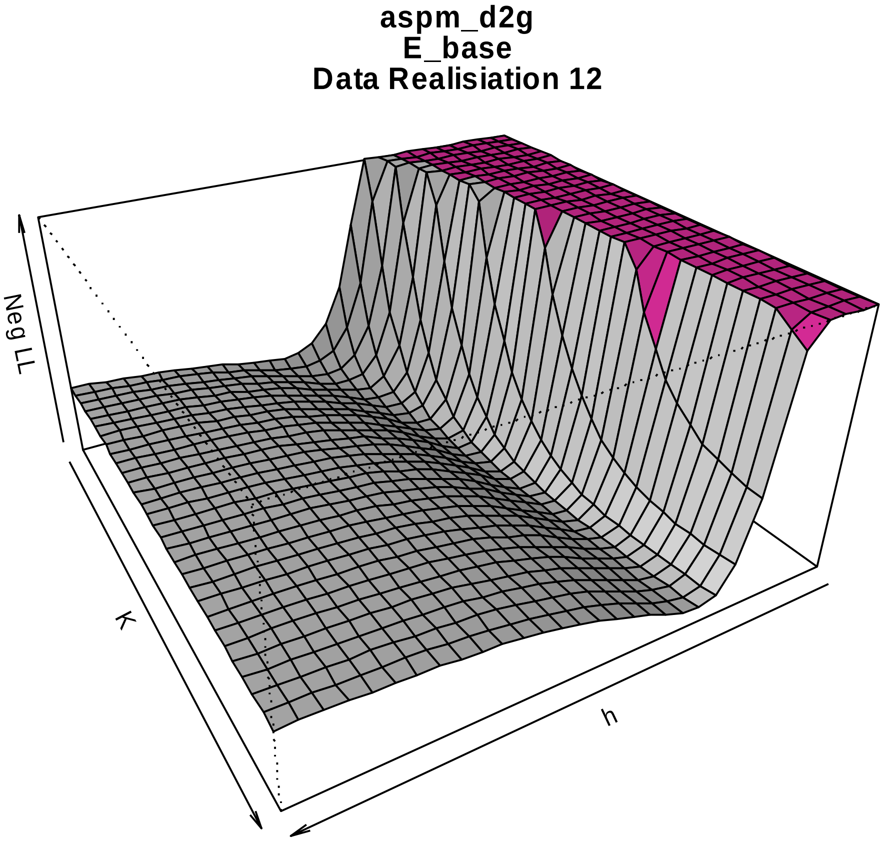


Figure A 3-5 Negative Log-Likelihood surface obtained from grid search procedure on an E_Base realisation. The purple facets of the surface plot indicate where the penalties were applied as C(a,t) approached N(a,t). The flat areas on the surface plot correspond to penalties, and have been truncated for display purposes.

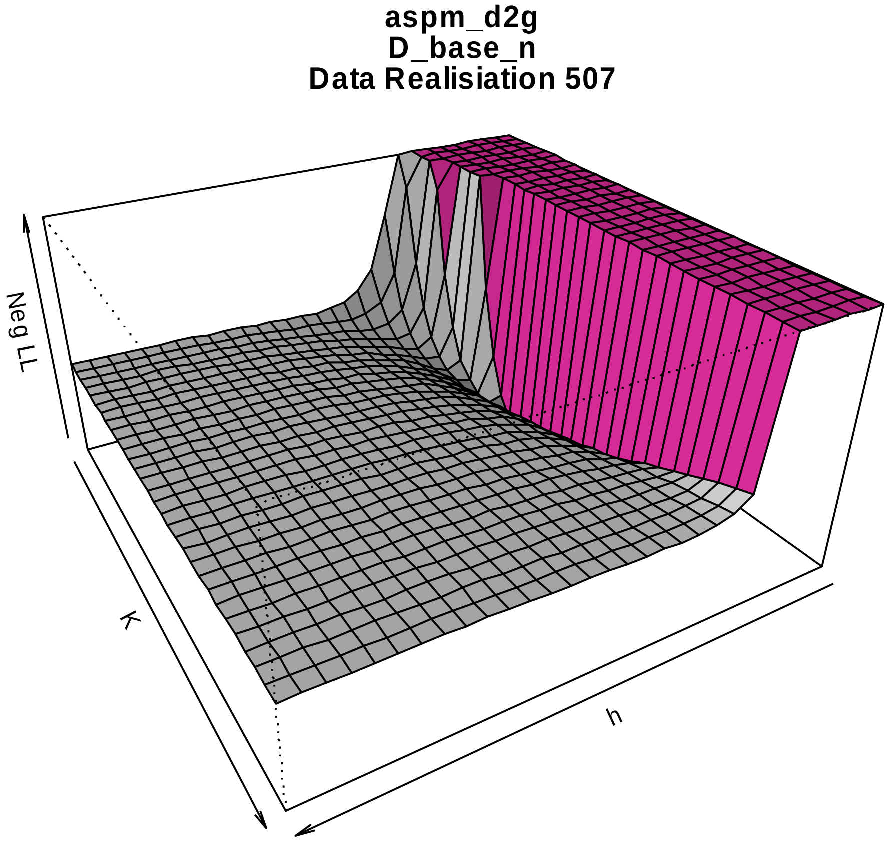


Figure A 3-6 Example ASPM negative Log-Likelihood surface illustrating a numerical failure. The purple facets of the surface plot indicate where the penalties were applied (as C(a,t) approached N(a,t)). The flat areas on the surface plot correspond to penalties, and have been truncated for display purposes.

A 3.4 REFERENCES

Butterworth, D.S. and E.E. Plaganyi. 2001. Exploratory analyses of southern bluefin tuna dynamics using production models (including separate addendum by D.S. Butterworth and S.J. Johnston). Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0108/24.

Butterworth, D.S., J.N. Ianelli and R. Hilborn. 2003. A statistical model for stock assessment of southern bluefin tuna with temporal changes in selectivity. Afr. J. Mar. Sci. 25: 331-361.

Ricard, D., D. Kolody, and M. Basson. 2002. Further exploration of biomass dynamics models for SBT stock assessment. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0209/28.

APPENDIX 4 SCALIA TECHNICAL DESCRIPTION

SCALIA (Statistical Catch-at-Age/Length Integrated Analysis) is a continuously evolving modelling framework initially implemented for the assessment of SBT (Kolody and Polacheck 2001), and adapted for application to the simulated YFT data generated by the Secretariat of the Pacific Community - Oceanic Fisheries Programme (Labelle 2002, 2003), under the co-ordination of the Standing Committee on Tuna and Billfish Methods Working Group. The model is largely an amalgamation of features from other stock assessment models, notably Butterworth et al. (2003) and MULTIFAN-CL (e.g. Hampton and Fournier 2001). Model parameters are estimated using a likelihood-based objective function, fitting to data that include some or all of: total catch by fishery (in numbers or mass), fishery catch-at-age distributions, fishery catch-at-length distributions, fishing effort, and tag releases/recoveries.

The following document provides a description of the most commonly used model structural assumptions and equations, and general comments about application. Model notation is summarized in Table A4 - 1. Subscripts and superscripts may be omitted in some of the following for clarity, but should be implicit from the context.

A 4.1 POPULATION DYNAMICS

Population dynamics are represented by the difference form of the standard (Baranov) catch equations, including the usual accumulator for the plus-group:

$$N_{t,a=0} = R_t$$

$$N_{t,a} = s_{t-1,a-1} N_{t-1,a-1}; \quad \text{for } 0 < a < A,$$

$$N_{t,A} = s_{t-1,A-1} N_{t-1,A-1} + s_{t-1,A} N_{t-1,A},$$

$$s_{t,a} = \exp(-F_{t,a} - M_a),$$

where $N_{t,a}$ is the population size at time t (usually annual or quarterly units) of age-class a (maximum age-class A), and new recruits, R, enter the population at age 0. Survival, s, of an age-class through time t is a function of age-specific natural mortality, M, and fishing mortality, F. Natural mortality is user-defined input, or estimated for each age subject to constraints. If M_a is estimated, variability among age-classes is assumed to be a random normal deviate, $\delta^M \sim \mathrm{N}(\mu = 0, \sigma^M)$, from the mean across all ages and/or a third difference curvature penalty is applied (objective function terms O_8 and O_9 in A 4.4).

A Beverton-Holt Stock Recruitment (SR) relationship, and individual deviations around this relationship are estimated as part of the overall parameter estimation. The SR curve is parameterized in terms of mean unfished recruitment R^*, and steepness, h (the ratio of mean recruitment at SSB(unfished)/5 over R^*). Lognormally-distributed deviations from the SR are estimated, potentially including a lag(1) autoregressive process as in Butterworth et al. (2003):

$$R_t = \frac{\alpha SSB_t}{\beta + SSB_t}\exp(\tau_t - \tfrac{1}{2}(\sigma_t^{SR})^2),$$

$$\beta = \frac{SSB_{unfished}(h-1)}{1-5h},$$

$$\alpha = \frac{R^*(\beta + SSB_{unfished})}{SSB_{unfished}},$$

$$\tau_t = \rho\tau_{t-1} + \varpi_t\sqrt{1-\rho^2} \text{ , and}$$

$$\omega_t \sim Normal(\mu = 0, \sigma_t^{SR}) .$$

The objective function term relating to the stock recruitment relationship is O_5 in section A 4.4. The initial population age structure is estimated, subject to constraint by the stock-recruitment relationship. The recruitment deviation CV is user-defined, and can be specified to change over time. We have often found it useful to reduce the CV on the early recruitment, particularly the initial age structure. Spawning stock biomass is calculated:

$$SSB_t = \sum_a Maturity_a N_{t,a} m_{t,a}^{Fec}$$

where $Maturity_a$ is the vector of maturity-at-age, *m* is the mean mass of an individual of age *a* in year *t*, *Fec* is an exponent reflecting the fact that larger individuals are usually dis-proportionately more fecund than smaller individuals.

A 4.2 FISHERY DYNAMICS

Fishing mortality follows a separable assumption in that *F* is composed of a time-step component and age component for each fishery:

$$F_{f,t,a} = G_{f,t}H_{f,t,a} ,$$

and total fishing mortality is given by

$$F_{total,t,a} = \sum_f F_{f,t,a}$$

where, for a single fishery, G_t is the time component of the fishing mortality term, $H_{t.a}$ is the age-specific fishery selectivity term (if selectivity does not change over time then the *t* sub-script is redundant). G_t is further partitioned into a number of components that have attractive mechanistic interpretations, but are in practice rather confounded in the estimation process:

$$G_{f,t} = q_{f,t}Q_{f,season}E_{f,t}\exp(\gamma_{f,t})$$
$$\gamma_{f,t} \sim Normal(\mu = 0, \sigma_f^{E_t}) ,$$

where, *q* is the fishery catchability (average scaling factor relating effort to fishing mortality), *Q* is a seasonal catchability effect (For a quarterly timestep, *Q* consists of 4 parameters per fishery corresponding to the 4 quarters within a year; one of which is defined as unity to avoid confounding with *q; Q* is not used with an annual time-step), *E* is the observed effort in relative units, and $\gamma$is an effective effort deviation (an error term that describes a potentially large temporary deviation from the mean relationship between effort and fishing mortality, e.g. due to inter-annual variability in fish distributions), giving rise to objective function term O_4 in section A 4.4. It is optional to assume that effort deviations tend to be larger when effort is low (e.g. catch rates at a fine spatio-temporal scale are highly variable, but the CV is likely to be lower if aggregated over more effort units):

$$\sigma_f^{E_t} = \left(\frac{\max(E_f)}{E_{f,t}} \right)^{\nu} \sigma_f^{E\max} .$$

Thus the CV of the prior distribution for the effort deviation in a given time-step is inversely proportional to the ratio of the maximum observed effort over effort in the given quarter (all raised to the power of υ). In practice, if effort is always large, or effort standardization is believed to be reliable, υ is set to 0, and the CV is constant over time for any given fishery. Catchability can be given some freedom to change over time, via a random walk process:

$$q_{f,t+b^q} = q_{f,t} \exp(\delta_{f,t}^q),$$
$$\delta_f^q \sim Normal(\mu = 0, \sigma_f^q)$$

where *b* indicates the number of time-steps in which *q* is assumed to remain constant between changes (*b* can be as small as 1, but in applications to SBT, we have found larger time blocks yield similar results with fewer estimated parameters). This forms the basis for objective function term O_7 in section A 4.4. In contrast to the effort deviations, this process is intended to describe gradual, systematic changes in the system which affect the relationship between effort and fishing mortality (e.g. due to cumulative improvements in fishing technology). If there is a strongly auto-correlated pattern in the effort deviations, this is usually interpreted as evidence for a change in catchability.

Fishery selectivity is represented as a purely age-based process. There are two methods of constraining the shape of the selectivity-at-age vector. We usually assume that the degree of similarity in fish vulnerability to fishing gear is influenced by the degree of similarity in size. In this case, the vector *H* is actually derived from a length-based concept:

$$H_a^* = \sum_l P_{a,l*}^{AL} \Lambda_{l*} ,$$

where Λ is a length-based selectivity parameter, $P_{a,l*}^{AL}$ is the proportion of age *a* fish in length-class *l**, (in this case, *l** is used to indicate that there are usually far fewer length-based parameters estimated than are used for the catch-at-length frequency distributions used elsewhere in the model; the actual number is user-defined, and we have generally used 6-12). Each age-based selectivity parameter is a weighted sum of the length-based selectivity parameters where the weighting is equal to the proportion of fish age *a* in length-class *l**. Thus consecutive ages must have similar selectivity, depending on the degree of length overlap. We also apply a third-difference curvature penalty to smooth out *H* across adjacent age-classes (objective function term O_9 in section A 4.4). When we use the length-based selectivity parameterization in SCALIA, none of the curvature penalties seem to be required to produce a visually satisfactory curve, but we have not actually compared the performance implications of the different constraints. Age-based selectivity vectors are always re-scaled to a mean of unity:

$$H_a = A \cdot \left(\frac{H_a^*}{\sum_a H_a^*} \right).$$

Selectivity can also be implemented with temporal variability, using a random walk process similar to catchability:

$$H_{f,t+b^H,a} = H_{f,t,a} \exp(\delta_{f,t}^H),$$
$$\delta_f^H \sim Normal(\mu = 0, \sigma_f^H).$$

This gives rise to objective function term O_6 in section A 4.4.

Catch, *C*, is the proportion of total mortality attributed to fishing. For a given age class and fishery:

$$C_{f,t,a} = \frac{F_{f,t,a}}{F_{total,t,a} + M_{t,a}} N_{t,a}(1 - s_{t,a}),$$

and the predicted catch-at-age (CA) composition for each fishery is expressed by the proportions of catch (P^{CA}) in each age-class:

$$P_{t,a}^{CA} = \frac{C_{f,t,a}}{\sum_a C_{f,t,a}}.$$

For fisheries that only have length samples, the corresponding predicted catch-at-length composition for each fishery (P^{CL}) is calculated from the age composition weighted by the length-at-age distribution:

$$P_{t,l}^{CL} = \sum_a P_{t,a}^{CA} \cdot P_{t,a,l}^{AL},$$

where P^{AL} is the proportion of age *a* fish in length-class *l*. It is assumed that CA and CL observations from the commercial catches are random samples, giving rise to the multinomial CA/CL likelihoods (objective function term O_2 in section A 4.4). To partially compensate for potential sampling problems, SCALIA reduces the observed sample sizes in the catch-at-length likelihoods by a constant proportion (usually 0.1-1) and a maximum effective sample size is specified (usually 30-200). Thus the effective sample size (η) potentially differs for each fishery at each time-step (but in practice is usually set to the maximum).

P^{AL} is derived in one of two ways. Assuming that length distributions are normally distributed for each age-class, the mean and variance can be a user-defined input. This is the preferred approach for SBT, in which independent analyses indicate that there is variability in the length-at-age relationship over time. Alternatively, the mean (instantaneous) length-at-age is assumed to be constant over time, defined by the following growth equation (Laslett et al. 2003):

$$\mu_a = L_\infty \left[1 - \exp(-k_2(a-a_0)) \left\{ \frac{1+\exp(-\beta(a-a_0-\Phi))}{1+\exp(\Phi\Theta)} \right\}^{-(k_2-k_1)/\Theta} \right].$$

This is a weighted mean of two von Bertalanffy curves, where the weighting is a logistic function. This function is well suited for describing an overall growth curve in which younger ages and older ages seem to follow substantially different von Bertalanffy curves (this relationship produces a smooth transition between the two). The (instantaneous) distribution of length-at-age is assumed to be normally distributed, and the standard deviation is a linear function of the mean length-at-age:

$$\sigma_a = slope\mu_a + intercept\,.$$

The instantaneous length-at-age distributions described by the above two equations might be a poor approximation to the observed catch length frequency distributions if spawning occurs in a very narrow time window, and time-steps are large relative to growth rates. In this case, the length frequency distribution of a given age is effectively a sum of distributions with different means (a flat-topped platykurtic distribution). SCALIA has an option to partially account for this effect by calculating each length-at-age distribution as the sum (of a user-defined number of) instantaneous length-at-age distributions evenly spaced within a time-step. We have not tested whether this feature actually has a significant effect on assessment inferences, but we expect that it could be useful in some circumstances. We also note that mortality within a time-step will cause a related error in the length-at-age distribution, but this has been ignored. In principle all of the length-at-age parameters (*Linf, k1, k2, a(0), slope, intercept,Φ,Θ*) could be estimated, but in most applications to date, some or all have been taken as fixed input. The method used to determine the length-at-age distribution also has potential implications for the analysis of tagging data (see A 4.3 below).

The total catch in numbers, by fishery, is estimated in the model:

$$C_{f,t}^{numbers} = \sum_a C_{f,t,a}\,,$$

or in mass as

$$C_{f,t}^{mass} = \sum_a C_{t,a} m_{t,a} \,,$$

And we assume that the total catch for each fishery (numbers or mass) is measured with log-normal errors:

$$C_{f,t}^{obs} = C_{f,t} \exp(\delta_{f,t}^{C})$$
$$\delta_f^C \sim Normal(\mu = 0, \sigma_f^C) \,,$$

giving rise to objective function term O_1 (see A 4.4 below). In most applications to date, we have assumed that total catch is essentially known perfectly for all fisheries (σ^C ~ 0.01), and have not tested the reliability of estimating total catch errors.

A 4.3 Tag Dynamics

Population dynamics of fully-mixed tagged fish are assumed to be identical to the general population. Predicted recaptures, *Tags(rec,pred)*, of age *a* at time *t* depend on the number of tagged fish that are fully mixed in the general population, *Tags(mixed)*, in the same manner as the catch is related to the total population:

$$Tags_{t,a}^{rec,\,pred} = r_t \cdot \left(\frac{F_{t,a}}{Z_{t,a}}\right) Tags_{t,a}^{mixed} (1 - s_{t,a}) \,,$$

where *r* is the tag reporting rate (preferably a fixed input, but in principle can be admitted as an estimated parameter; independent for each fishery but assumed constant over time). There is a user-defined mixing period in which tags are assumed to not be representative of the general population. The dynamics of unmixed tags *Tags(unmixed)* from release group *g* are described by Pope's approximation to the catch equation, and enter the fully-mixed tag population (*mixed*) after a period of *mixTime* time-steps (we have usually applied values of 0-4):

$$Tags_{t,a}^{mixed} = Tags_{t-1,a-1}^{mixed} \cdot s_{t-1,a-1} + Tags_{t,a}^{g,newlyMixed}$$

$$Tags_{t,a}^{g,newlyMixed} = \begin{cases} Tags_{t,a}^{g,released} ; mixTime = 0 \\ Tags_{t-1,a-1}^{g,unmixed} \exp(-\tfrac{1}{2} M_{a-1}) - \dfrac{Tags_{t-1,a-1}^{g,recaptured,obs}}{r}) \exp(-\tfrac{1}{2} M_{a-1}); mixTime > 0 \text{ and } t = t_g + mixTime \end{cases}$$

$$Tags_{t,a}^{g,unmixed} = \begin{cases} Tags_{t,a}^{g,released} ; mixTime > 0 \\ Tags_{t-1,a-1}^{g,unmixed} \exp(-\tfrac{1}{2} M_{a-1}) - \dfrac{Tags_{t-1,a-1}^{g,recaptured,obs}}{r}) \exp(-\tfrac{1}{2} M_{a-1}); t_g < t < t_g + mixTime \end{cases}$$

where each release group has a unique release time, t_g, and must be tracked independently. Thus, the number of fully mixed tags is dependent on the surviving fully mixed tags from the previous time step, plus the number of tags that have just achieved fully-mixed status (*newlyMixed*). For the unmixed individuals, fishing mortality is applied as a pulse fishery in the middle of the timestep. It is assumed that recapture probabilities are described by the Poisson distribution (giving rise to O_3 below). As a simple means of admitting that over-dispersion is probable when using the Poisson likelihood for tag analyses, we have added an effective tag release co-efficient (analogous to the effective sample size in the multinomial CL likelihood, the effective tag release co-efficient, η^{tags}, downweights the tagging term). The negative binomial distribution is gaining popularity for this purpose, but we have not compared the two approaches. We note that the tag dynamics should be implemented with independent analysis of each release event, rather than predicting the aggregated tag recoveries in the equations above. Using the dis-aggregated data in a Brownie-type model potentially improves the estimation of natural mortality for cohorts that are repeatedly tagged at successive ages.

All of the tag dynamics work in an age-structured context, but in most applications, only the length of released tags is measured. In the cases that we have worked with to date, young fish have predominantly been tagged, and this allows us to be reasonably confident of the age distribution. For actual SBT assessments, tag release ages have been input from external analyses. In the YFT and SBT simulations, two different options have been implemented for estimating ages of tagged fish from lengths. Cohort slicing is the simplest approach, but this is potentially problematic if the length-at-age distributions are being estimated. Tag ages are assigned integer values, and can change in a discontinuous fashion as the growth curve changes, causing instability in the function minimization. An alternative approach that we have used for ageing tags is to assign all ages a partial weighting in proportion to the likelihood of the fish having come from each length-at-age distribution. If the probability of an age 2 fish being 60cm is 3 times as high as the probability of an age 3 fish being 60 cm (and there is no probability of fish of any other age being 60cm), then the tag is assigned to age 2 with a weight of 0.75, and age 3 with a weight of 0.25. SCALIA has an additional option to place the age assignment into a more Bayesian context, and admit that the prior probabilities of being age 2 or age 3 should actually be proportional to their abundance in the population at a given time, e.g.:

$$\Pr(age = a \mid length = l) = \frac{\Pr(length = l \mid age = a) \cdot N_a}{\sum_a \Pr(length = l \mid age = a) \cdot N_a}$$

We have not observed much difference in results due to the different tag ageing methods, except that the minimization can fail with cohort slicing if the growth curve is estimated.

A 4.4 Objective Function

The objective function is likelihood-based and Bayesian in the sense that prior probabilities are assigned to some terms. There are a number of somewhat arbitrary penalties usually applied to the model fitting as well. We do not view these models as statistically rigorous, and are skeptical of interpreting the objective function as a true likelihood for the purposes of statistical uncertainty quantification or hypothesis testing.

The objective function, O_{total} consists of several components. In the following list of objective function terms, O_1-O_4 directly quantify the degree of agreement between observations (*obs*) and model predictions (the superscript *pred* is used here for clarity, while it is implicit in most of the preceding text). O_5-O_{11} are not directly dependent on the data:

$$O_{total} = \sum_{i=1}^{8} O_i \text{ , where:}$$

$$O_6 = \sum_f \frac{1}{2(\sigma^{C_f})^2} \sum_t (\delta^C_{f,t})^2$$

(total catch, where *X* indicates numbers or mass),

$$O_2 = \sum_f -\eta_f^{CACL} \sum_t \sum_X P^{CACL,obs}_{f,t,X} \log_e (P^{CACL,pred}_{f,t,X})$$

(catch composition, where *X* indicates age or length frequency bins as appropriate),

$$O_3 = \eta^{tags} \sum_t \sum_a \left(T^{pred}_{a,t} - T^{obs}_{a,t} \log_e (T^{pred}_{a,t}) \right)$$

(tag recoveries)

$$O_4 = \sum_f \frac{1}{2(\sigma^{E_f})^2} \sum_t (\gamma_{f,t})^2$$

(effort deviations, analogous to residuals in the relationship between CPUE and abundance)

$$O_5 = \frac{1}{2(\sigma_t^{SR})^2} \sum_t \omega_t{}^2$$

(recruitment deviations and the stock recruitment relationship)

$$O_6 = \sum_f \frac{1}{2\sigma^2_{H_f}} \sum_t \sum_a (\delta^H_{f,t,a})^2$$

(constraint on temporal variability in fishery selectivity)

$$O_7 = \sum_f \frac{1}{2(\sigma^{q_f})^2} \sum_t (\delta^q_{f,t})^2$$

(constraint on temporal variability in fishery catchability)

$$O_8 = \frac{1}{2(\sigma^M)^2} \sum_a \delta_a^2$$

(constraint on mortality-at-age estimates)

$$O_9 = \sum_{X=(H_f;M)} \eta^X \sum_f \sum_t \sum_{a=1}^{A-3} \frac{\left[\log_e(X_{t,a+3}) - 3\log_e(X_{t,a+2}) + 3\log_e(X_{t,a+1}) - \log_e(X_{t,a})\right]^2}{2(\sigma^X)^2}$$

(third difference curvature penalty with respect to age, where *X* indicates mortality or selectivity; summation over *f* not relevant for mortality)

$$O_{10} = \eta_f^{qmono} \sum_f \sum_t \begin{cases} \left(q_{f,t+1} - q_{f,t}\right)^2; q_{f,t+1} < q_{f,t} \\ 0; q_{f,t+1} \geq q_{f,t} \end{cases}$$

(penalty to encourage monotonically increasing catchability over time)

$$O_{11} = \eta_f^{Hmono} \sum_f \sum_t \sum_a \begin{cases} \left(H_{f,t,a+1} - H_{f,t,a}\right)^2; H_{f,t,a+1} < H_{f,t,a} \\ 0; H_{f,t,a+1} \geq H_{f,t,a} \end{cases}$$

(penalty to encourage selectivity monotonically increases with age).

Not all terms are relevant for all applications; some terms (e.g. O_{10},O_{11}) might be useful in intermediate phases of parameter estimation but removed in the final phase.

A 4.5 Parameter Estimation and statistical uncertainty quantification

SCALIA is implemented with AD Model Builder software (Otter Research, Victoria, Canada), which uses automatic differentiation and efficient function minimization routines to identify the maximum posterior density (MPD) of the parameter estimates. The reliability of the function minimization can be sensitive to initial parameter specifications, and the manner in which the parameters are constrained. We use a phased approach to minimization, in which assumptions are strong and relatively few parameters of greatest influence are initially estimated (e.g. mean recruitment, mean fishery catchability). In subsequent phases, the assumptions are relaxed and the parameters perceived to be of lesser global importance are estimated (e.g. recruitment deviations, effort deviations). In the final phase, all parameters are estimated simultaneously. In real applications to SBT, we have usually been content with the estimated confidence limits provided by the multi-variate normal approximation from the inverse Hessian matrix at the MPD. AD Model Builder can also calculate likelihood profiles for parameters of interest, or approximate Bayesian Posteriors using an MCMC routine. We have not routinely applied these methods of statistical uncertainty quantification, because we are usually more concerned with the greater uncertainty that generally arises from sensitivity to model structural assumptions.

A 4.6 OUTPUT VISUALIZATION AND GOODNESS-OF-FIT DIAGNOSTICS

Applications of SCALIA for real assessments involved fitting multiple model specifications, and examining the quality of fit for irregular behaviour, which could indicate minimization problems, mis-specification issues and model sensitivity. SCALIA outputs the MPD estimates of the entire stock dynamics history, including population numbers, fishing mortality, natural mortality, etc., in a format that can easily be visualized using an R software (e.g. http://www.r-project.org/) script. The quality of fit between predictions and observations (catch-at-length, catch-at-age, tag recaptures, effort deviations, stock recruitment relationship) can also be visually examined in different ways. Some of the typical graphical output that we usually examine is indicated in Fig. A4 - 1to Fig. A4 - 11.

The value of all components of the objective function are recorded, as are a number of other goodness-of-fit summary statistics. The empirical effective sample size is used to calculate the approximate catch-at-length sample size that would on average produce the indicated quality of fit between observations and model predictions (McAllister and Ianelli 1997):

$$ESS_l = \frac{\sum_l p_{t,l}(1-p_{t,l})}{\sum_l (o_{t,l}-p_{t,l})^2},$$

where p and o are the predicted and observed proportions of catch in each length (or age) class frequency distribution in each timestep. The Root Mean Square Error (RMSE), is used to compare the input variance specifications with the empirical output of the MPD estimates for recruitment deviations:

$$RMSE = \sqrt{\tfrac{1}{n}(\ln(obs/pred))^2}\,.$$

Auto-correlation (lag(1)) in time series of recruitment and effort deviations are also calculated for evidence of systematic lack of fit.

A 4.7 PROJECTIONS AND REFERENCE POINT CALCULATIONS

Separate executable code is invoked to produce SCALIA projections and reference point estimates. To date, these calculations have been based only on the MPD parameter estimates and corresponding stock dynamics, so the model statistical uncertainty is not maintained from the original SCALIA analysis. The projections are deterministic, and invoked with the main intention of generating MSY-related estimates for model comparison. MSY and related quantities are calculated in two different ways. MSY_F is the traditional approach, in which the aggregate selectivity across all fisheries in the last (user-defined number of) timesteps is held constant, and the sustainable yield corresponding to a range of constant effort multipliers is calculated by projecting forward in time until the population equilibrates. In contrast, MSY_C assumes that the catch ratio among fleets remains constant over time, in a manner consistent with the CCSBT management objectives of maintaining constant catch allocations among member nations. In this case, the global selectivity changes, depending on the age structure and the total catch. In both cases, we constrain the minimum value of surplus production (i.e. steepness ≥ 0.3 with a Beverton Holt curve), to avoid numerical problems. MSY_C, MSY_F, B_MSY_C, B_MSY_F and SSB_MSY_C and SSB_MSY_F are output.

A 4.8 SCALIA EVOLUTION

SCALIA has gone through several phases of development in an unsystematic fashion and it is not clear what the future of the model will be. Various SCALIA features have been explored but may not be fully implemented or documented. These include:

- An approximation to purely length-based selectivity has been tested. The implementation allows for length-at-age distributions to change over time as influenced by size selective fishing mortality. The growth curve remains constant over time, but the mean length-at-age is only relevant up to the youngest age selected by the fishery. For subsequent ages, the growth rate is determined by the growth curve, but the length-at-age changes depending on the exploitation history. This is only one of many possible implementations of size selective mortality, and is flawed in the sense that growth rate variability among individuals is poorly admitted. There was not much evidence that it made much difference for SBT assessment, and was not pursued further.

- A form of spatial structure was examined in which different fisheries were able to access only portions of the global population, and inter-annual variability in the fish spatial distribution was estimated. In the SBT context, the net effect of these spatial dynamics could be considered an additional constraint on the effort deviations - such that a large positive deviation in one area should be balanced by a negative deviation in another area. However, it was not clear that this added anything new to our interpretation of SBT dynamics, and also was not pursued further.

In the SBT context, an operating model for the evaluation of candidate Management Procedures (MPs) was jointly developed by the participants of the CCSBT SC, including the external scientific advisory panel (e.g. Haist et al. 2002). This operating model is conditioned to historical data and thus has an assessment model at its core. As the relevant features of this model are generally the same as SCALIA, we are not sure what role SCALIA should play in future SBT assessments. Furthermore, MULTIFAN-CL has recently become publicly available, and in most respects SCALIA is a sub-set of this model. MULTIFAN-CL is more versatile and has been widely applied to the assessment of a number of tuna populations. It does lack some of the features of interest for SBT (e.g. catch-at-age data, changes in length-at-age over time), but some of these are currently being added or documented. This leaves SCALIA with an uncertain future, unless additional features of interest are required, and it can be argued that SCALIA is the best platform with which these extensions should be implemented.

A 4.9 References

Butterworth D. S., J. N. Ianelli, and R. Hilborn. 2002. A statistical model for stock assessment of southern bluefin tuna with temporal changes in selectivity. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-MP/0203/4.

Haist, V., A. Parma and J. Ianelli. 2002. Initial specifications of operating models for southern bluefin tuna management procedure evaluation. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0209/7.

Hampton, J. and D.A. Fournier. 2001. A spatially dis-aggregated, length-based, age-structured population model of yellowfin tuna (*Thunnus albacares*) in the western and central Pacific Ocean. Mar. Freshw. Res. 52: 937-963.

Kolody, D. and T.Polacheck. 2001. Application of a statistical catch-at-age and –length integrated analysis model for the assessment of southern bluefin tuna stock dynamics 1951-2000. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0108/13.

Laslett,G.M., J.P.Eveson and T. Polacheck. 2003. A flexible maximum likelihood approach for fitting growth curves to tag-recpature data. Can. J. Fish. Aquat. Sci. 59: 976-986.

McAllister, M.K. and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54: 284-300

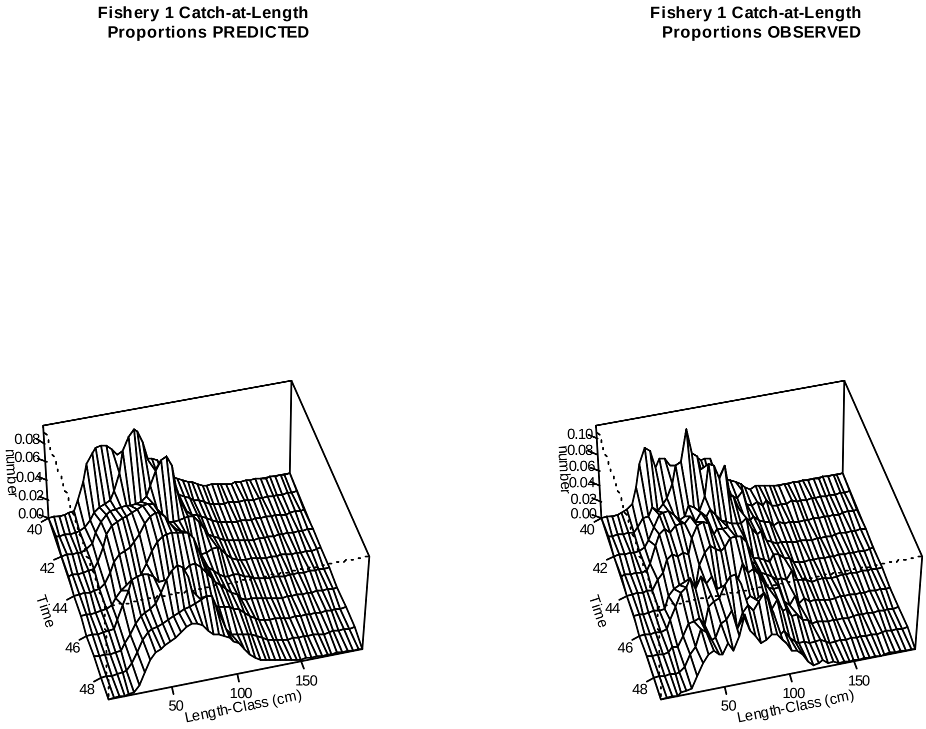
Table A4 - 1. SCALIA notation.

Subscripts and Superscripts:

a	= age (in time-step increments)
A	= maximum age in the population (plus-group accumulator)
b	= number of time-steps H or q remains constant between changes
f	= fishery
g	= identifier for tag release groups
l	= length frequency distribution bins
t	= time-step (usually annual or quarterly in practice)
tt	= time (increments of b time-steps)
mixed	= tagged fish assumed to be representative of general population
obs	= observed (data)
pred	= predicted (deterministic function of the model parameters)
recaptured	= tagged fish caught in a fishery
released	= newly released tagged fish
unmixed	= tagged fish not fully mixed in the general population
newlyMixed	= tagged fish making the transition from *unmixed* to *mixed*

States, variables, parameters and weighting factors:

C	= catch (numbers)
N	= numbers
F	= instantaneous fishing mortality (time-step units)
M	= instantaneous natural mortality (time-step units)
Z	= $F + M$ = total mortality
G	= time component of fishing mortality
H	= fishery selectivity
Tags	= tagged fish
R	= recruitment
s	= survival
m	= mass
E	= effort (hooks)
η	= weighting factor for likelihood components
P	= Proportions of catch-at-age, catch-at-length or length-at-age
q	= fishery catchability
Q	= seasonal component of fishery catchability
r	= tag recovery reporting rate
SSB	= spawning stock biomass
SR	= related to stock recruitment relationship
μ	= distribution mean
σ	= distribution standard deviation
$\omega, \gamma, \varepsilon$	= random deviate from specified distribution
τ	= an auto-correlated deviate from the stock recruitment relationship
ρ	= lag(1) correlation co-efficient for SR deviations
$Linf,k,\alpha,\beta$	= parameters describing length-at-age frequency distributions
Λ	= a pseudo-length-based selectivity parameter


Fishery 1 Catch-at-Length Proportions PREDICTED

Fishery 1 Catch-at-Length Proportions OBSERVED

Fig. A4 - 1. Example comparison of predicted and observed Catch Length Frequency distributions for SCALIA model fit to simulated SBT fishery with good data characteristics.

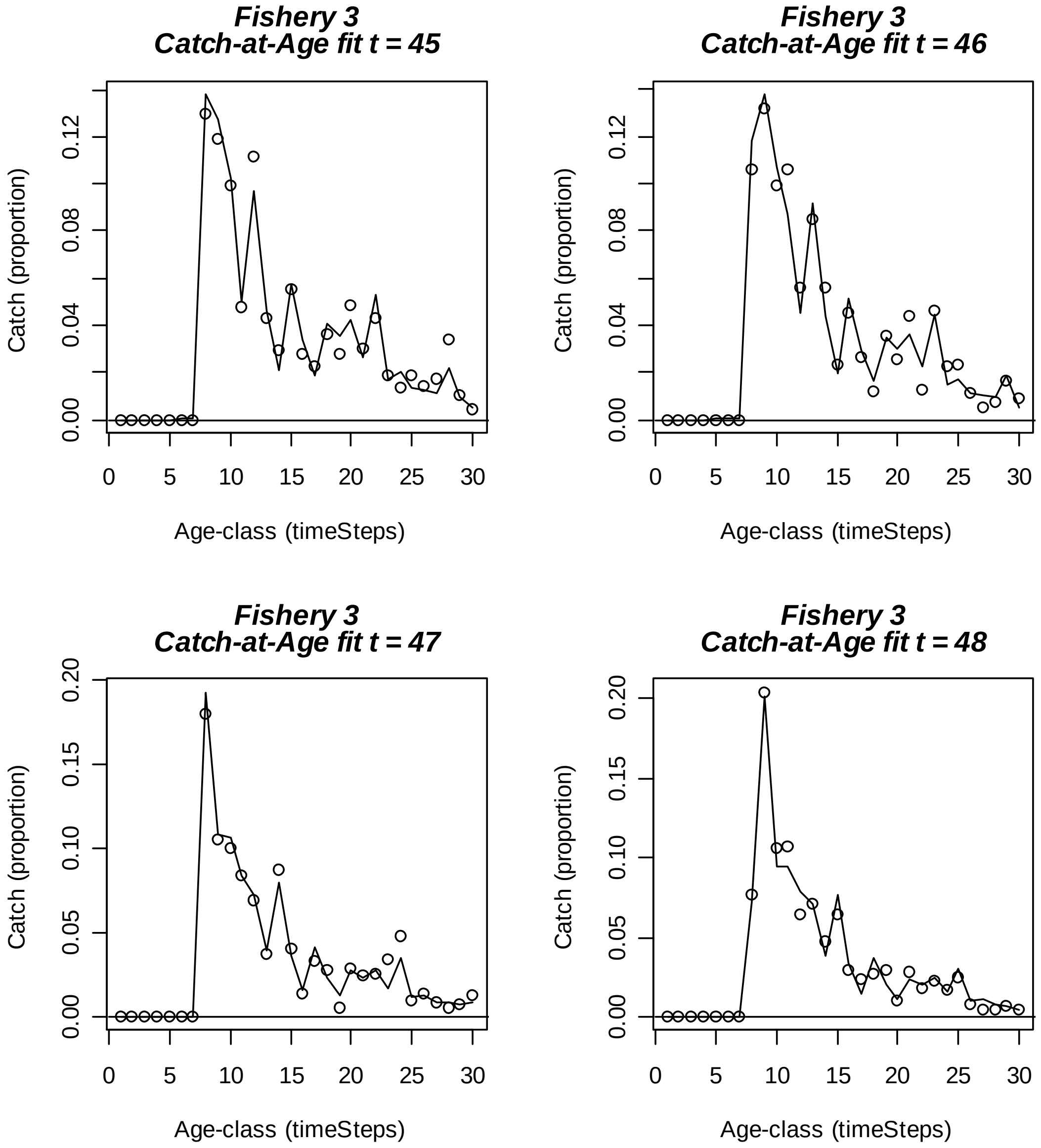


Fig. A4 - 2. Example comparison of predicted (lines) and observed (circles) Catch Age Frequency distributions for SCALIA model fit to simulated SBT fishery with good data characteristics.

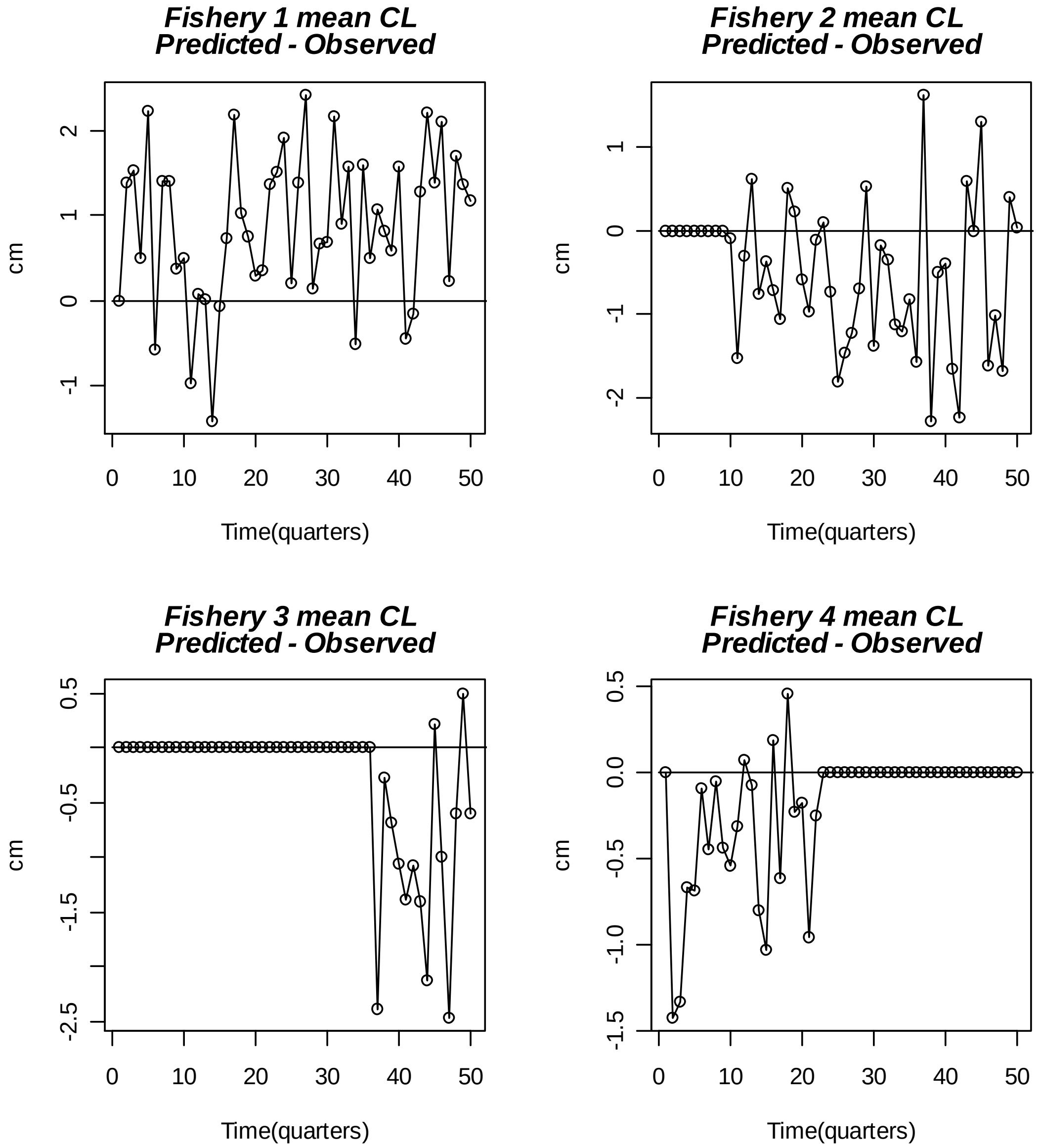


Fig. A4 - 3. Example comparison of mean predicted and observed Catch Length Frequency distributions for SCALIA model fit to simulated SBT fishery with good data characteristics (consecutive 0s indicate no catch).

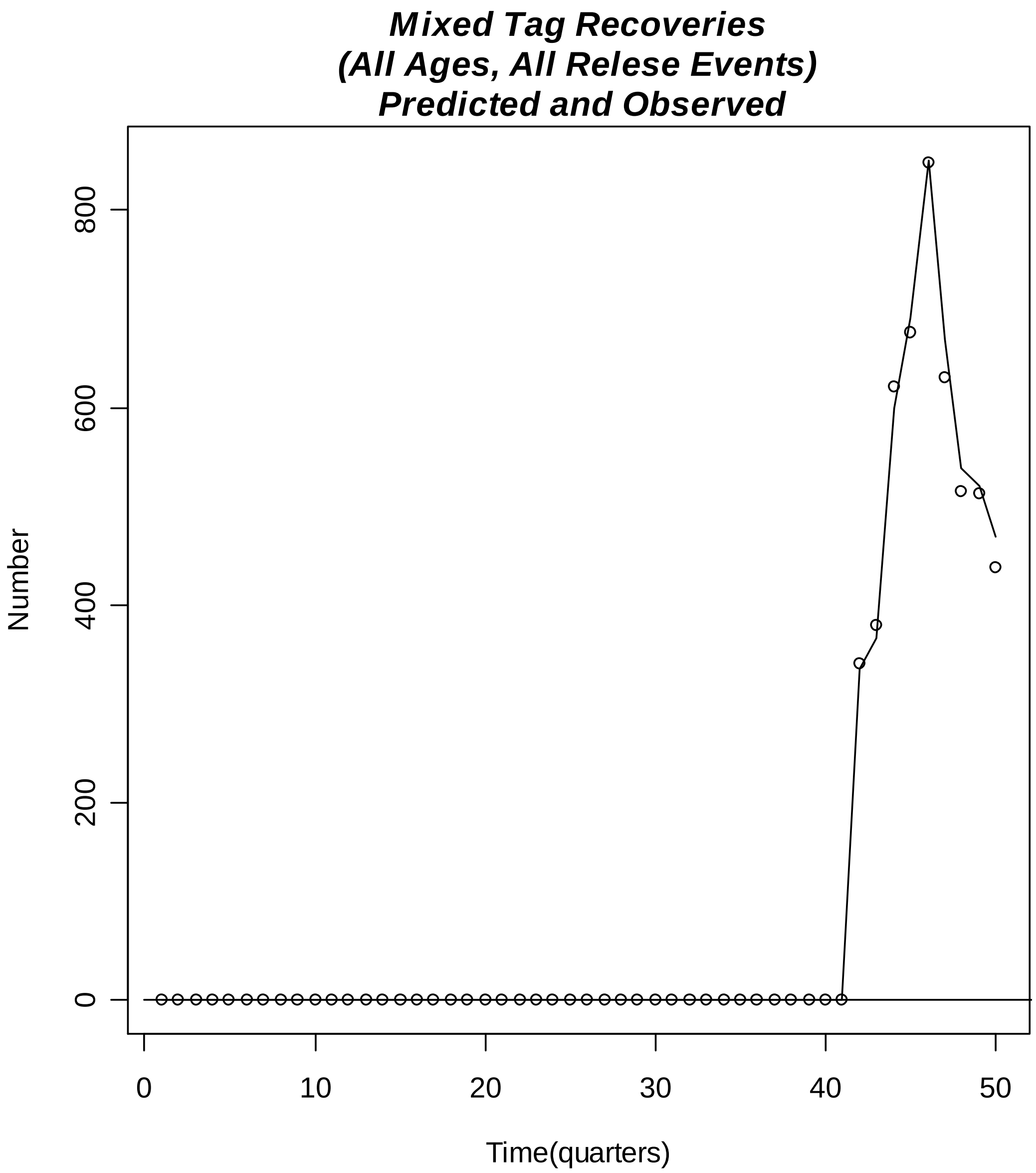


Fig. A4 - 4. Example comparison of mean predicted (lines) and observed (circles) tag recoveries (aggregated over all ages) for SCALIA model fit to simulated SBT fishery with good data characteristics.

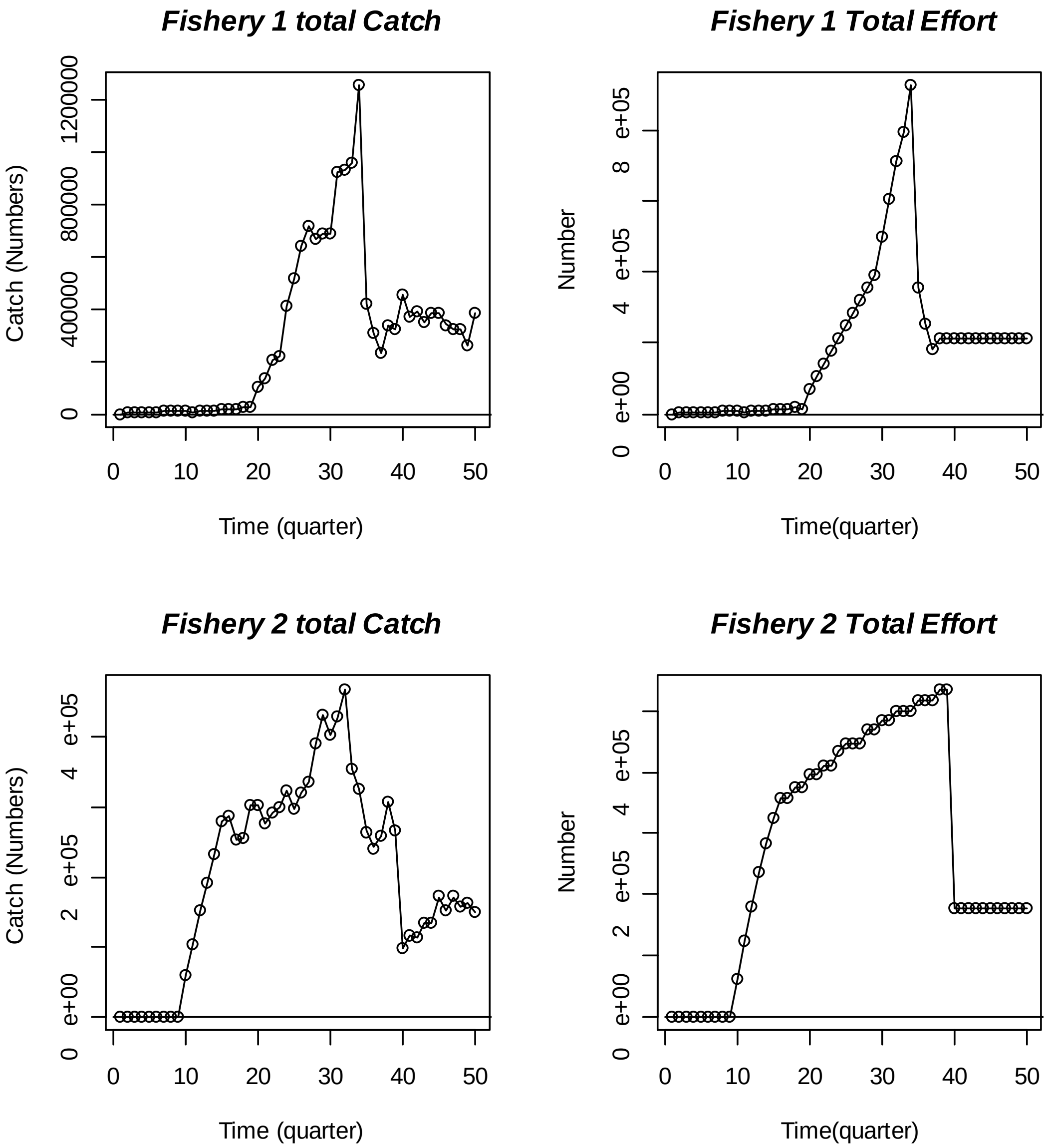


Fig. A4 - 5. Example comparison of total catch and effort by fishery for SCALIA model fit to simulated SBT fishery with good data characteristics. Catch consists of predicted (lines) and observed (circles) which are almost the same in this case; effort is observations only.

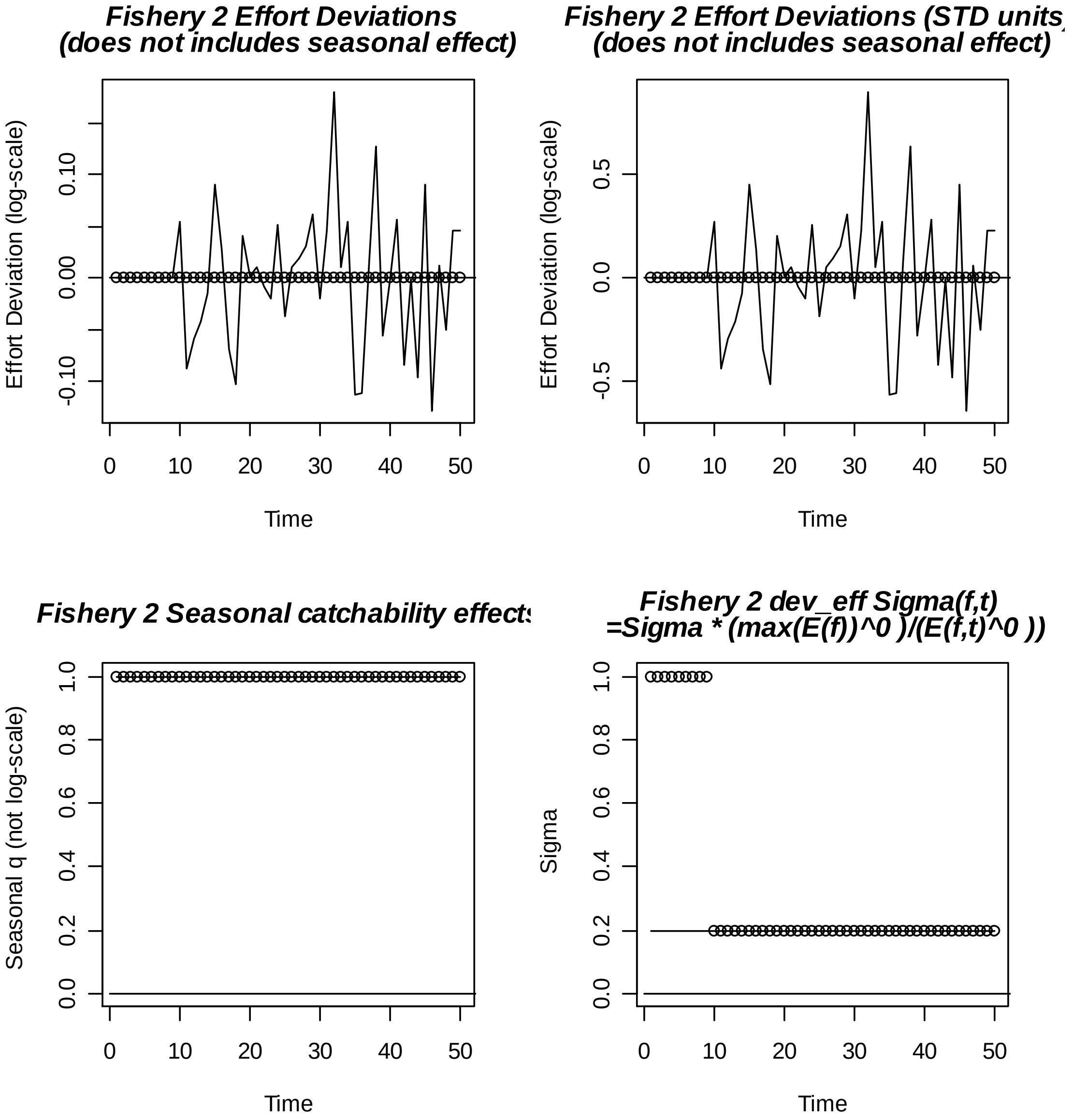


Fig. A4 - 6. Example of output related to the effort-fishing mortality relationship for SCALIA model fit to simulated SBT fishery with good data characteristics. Effort deviations are analogous to residuals around the CPUE-abundance relationship. In this case there is no seasonal catchability effect estimated, and the variance on the (log) effort deviations is assumed constant regardless of the magnitude of the observed effort. The fishery is only active starting in year 10.

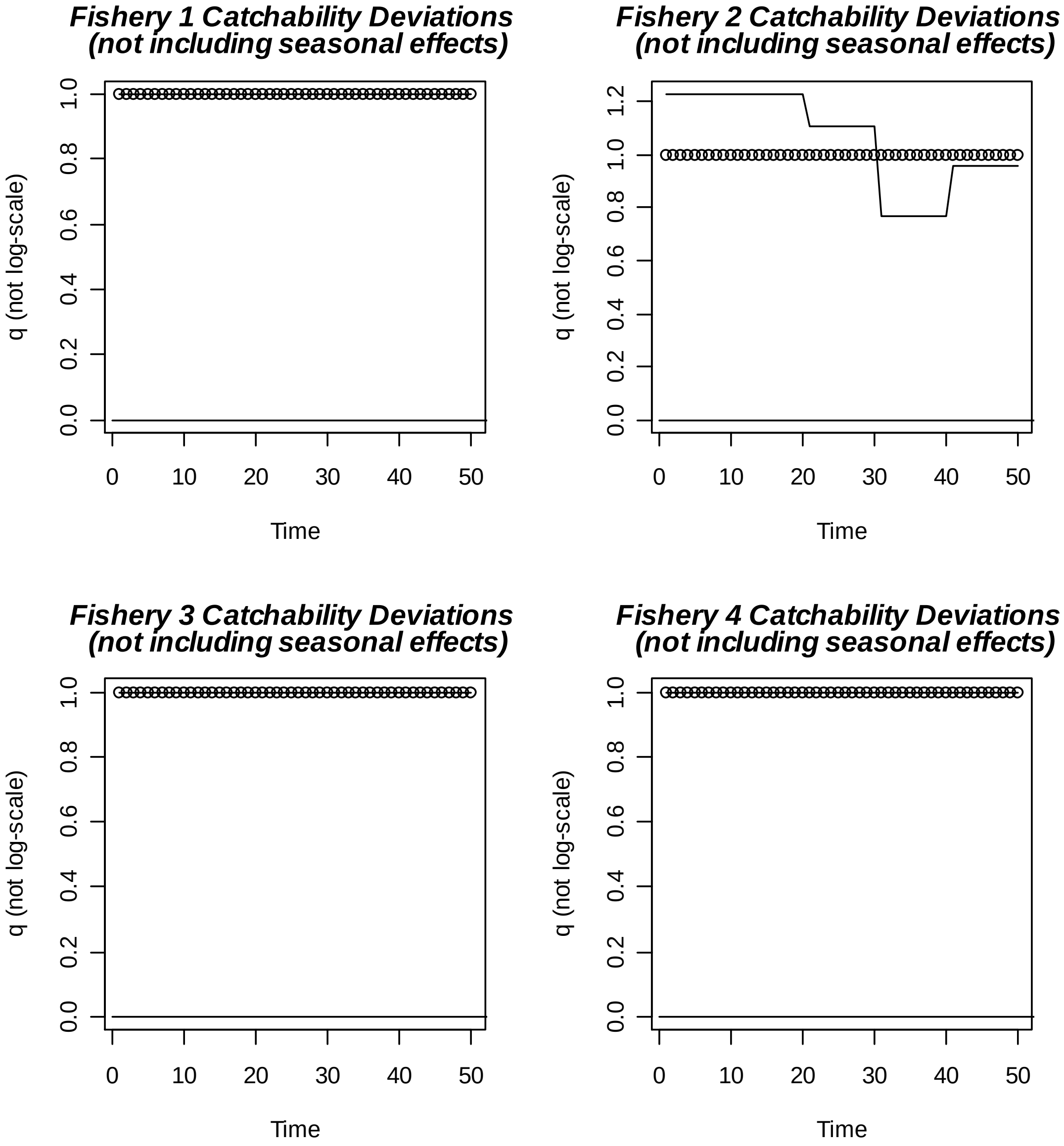


Fig. A4 - 7. Example of SCALIA output from an application to a simulated SBT data set. In this case, the effort series is only considered informative for fishery 2, and temporal variability in catchability is estimated in blocks of 10 timesteps. For presentation, catchability is re-scaled to a mean of unity.

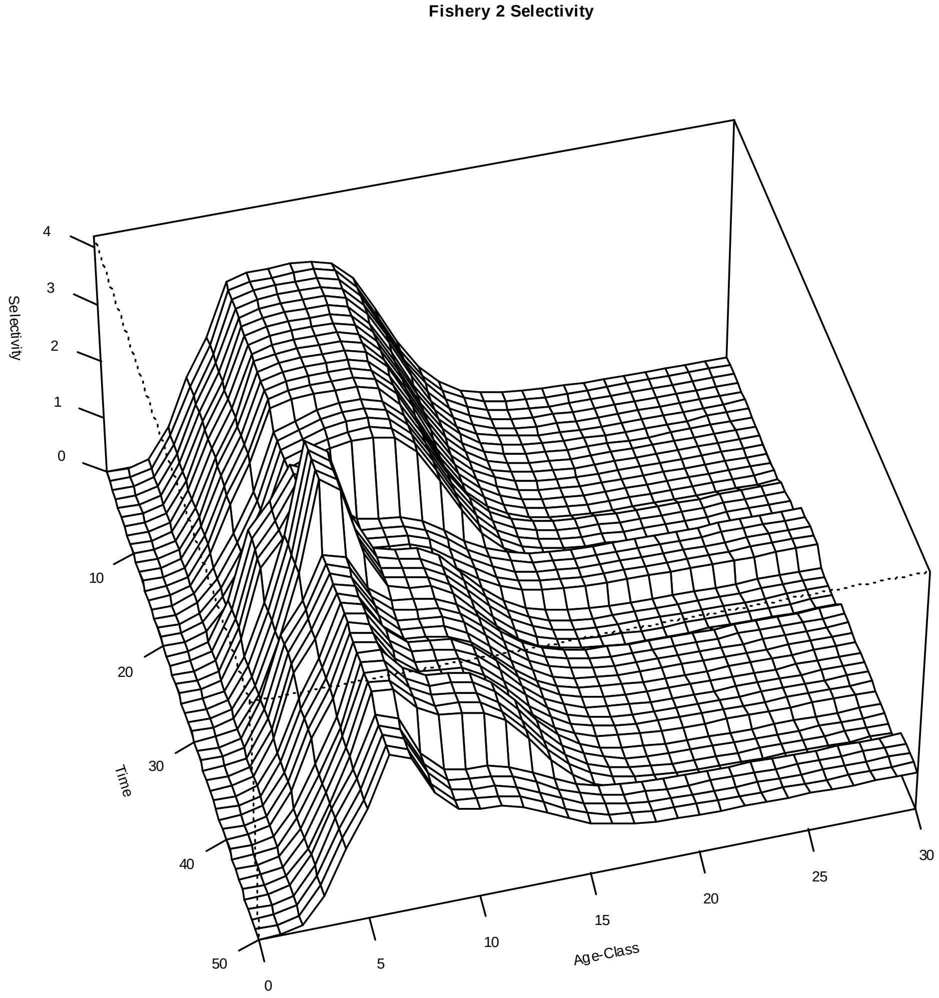


Fig. A4 - 8. Example of SCALIA selectivity estimates from an application to a simulated SBT data set. In this case, selectivity is constant for blocks of 5 consecutive timesteps.

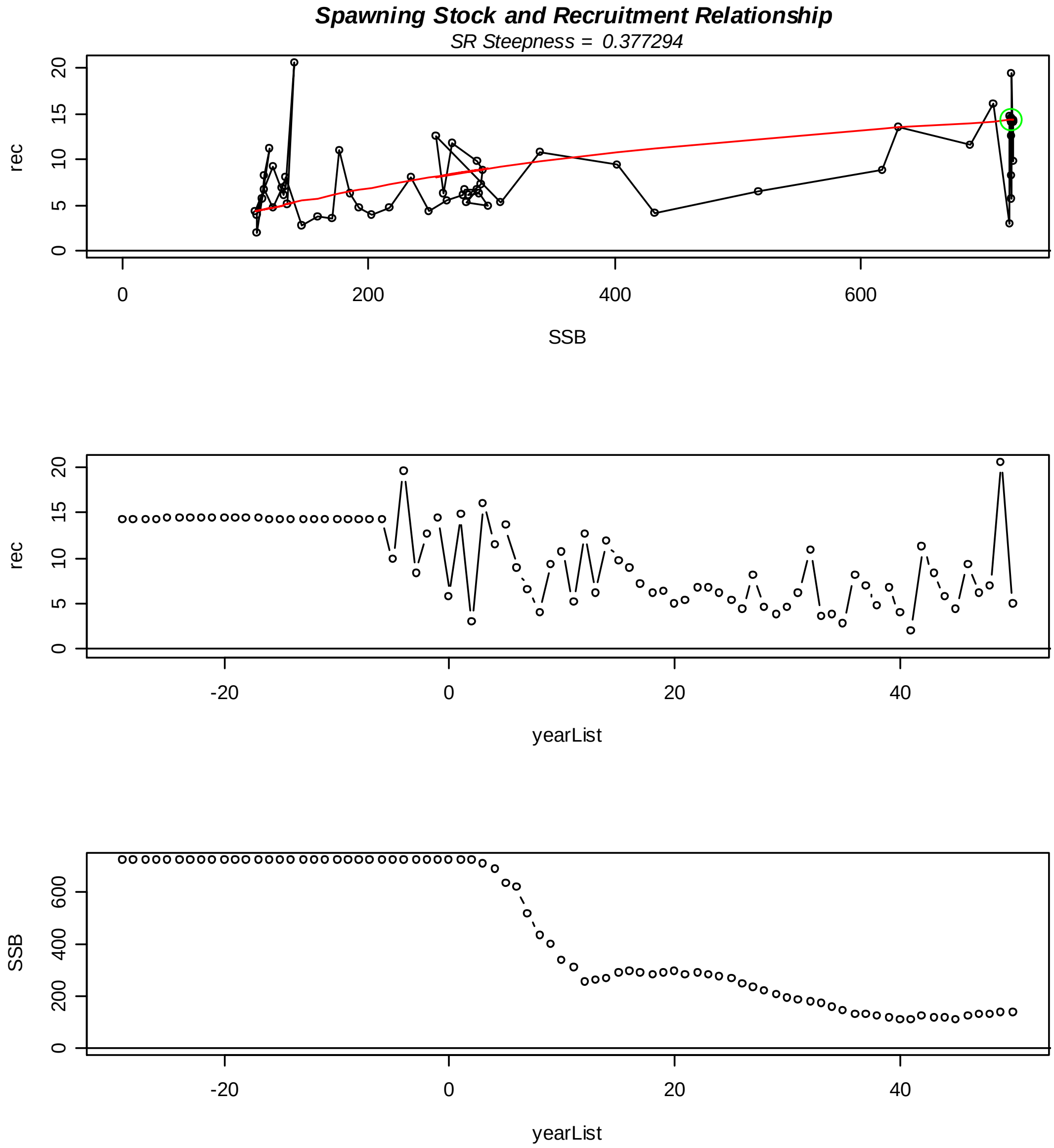


Fig. A4 - 9. Example of SCALIA stock and recruitment estimates from an application to simulated SBT data. The smooth line in the top panel indicates the estimated mean stock recruitment relationship; the large circle indicates the mean unfished biomass and recruitment.

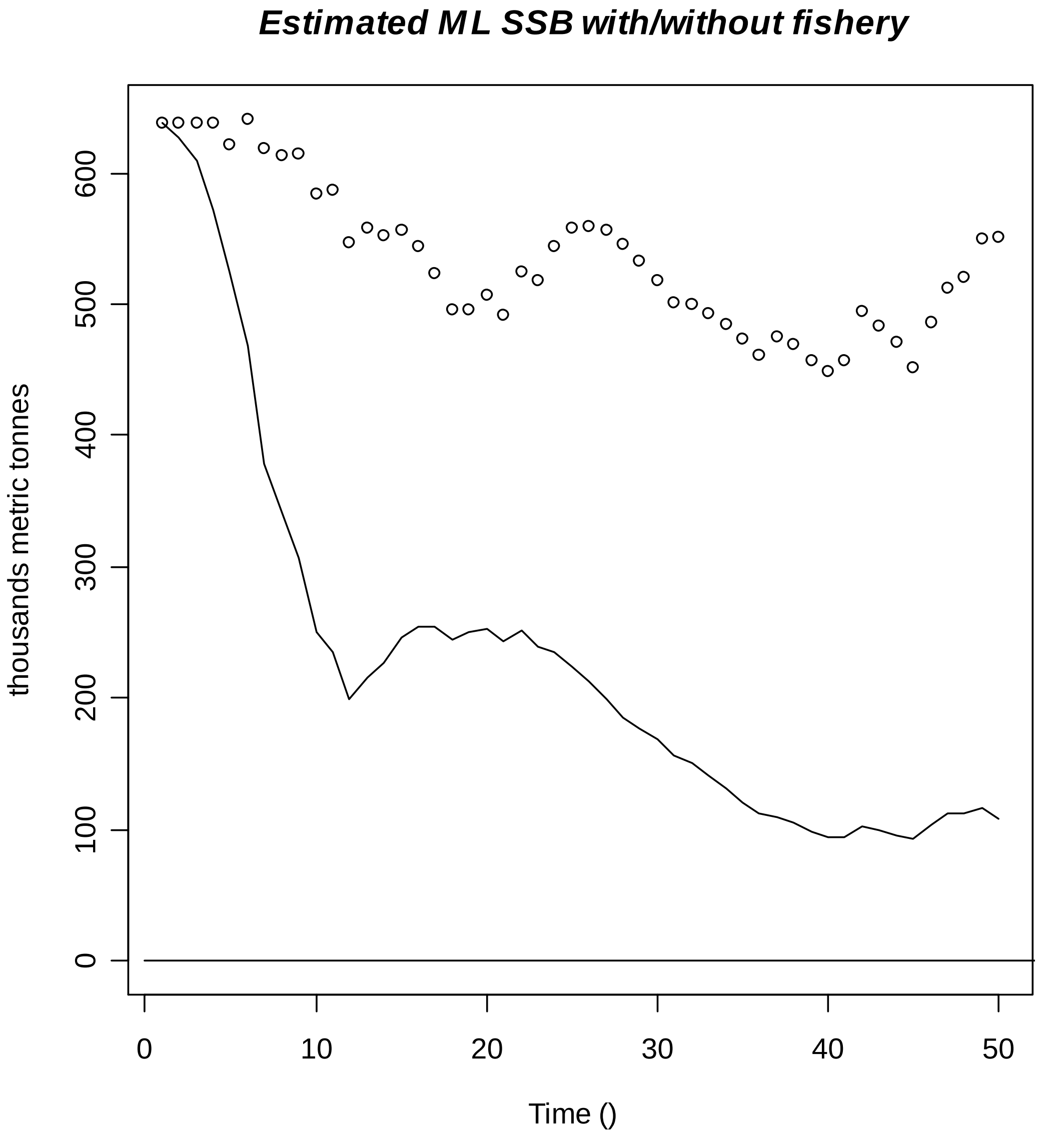


Fig. A4 - 10. Example of SCALIA Spawning stock biomass estimates from an application to simulated SBT data. The line indicates the MPD estimates; circles indicate the SSB that is predicted would have occurred in the absence of fishing.

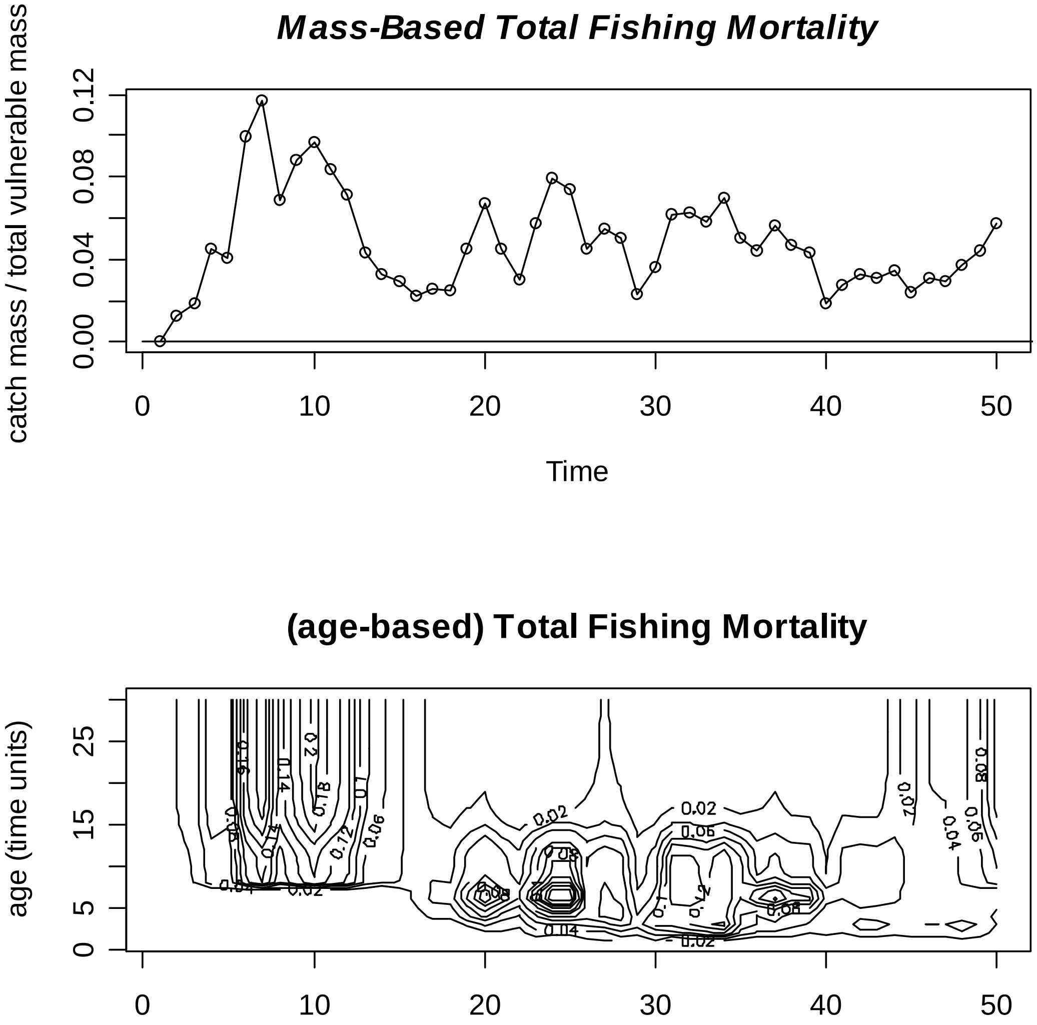


Fig. A4 - 11.Example of SCALIA fishing mortality estimates from an application to simulated SBT data. The top panel is the exploitation rate over time (catch mass over exploitable biomass); bottom panel is the instantaneous *F* by time and age.

APPENDIX 5 MULTIFAN-CL SPECIFICATIONS USED IN THE SBT SIMULATION TESTING

The attached file MF_scan.script is the control file used for the MULTIFAN-CL (e.g. Hampton and Fournier 2001) analysis, applied to the simulated SBT data as part of the SESAME project. The MF_scan specification was derived from the YFT example specification file from http://www.multifan-cl.org, and was modified to resemble the baseline SCALIA specification SC_base. Differences between MF_scan and SC_base included:

- the catch-at-length objective function component was multinomial for SC_base, and a robustified chi-square for MF_scan

- MF_scan used catch-at-length for all fisheries; SC_base used catch-at-age for the late spawning ground fishery and catch-at-length for all others

- SC_base estimated temporal variability in selectivity, while MF_scan assumed constant selectivity over time (but differing by fishery)

- MF_scan had a weak prior on stock recruitment curve steepness with a mode near the actual operating model value; SC_base had no explicit prior constraint.

- MF_scan had an effort deviation CV ~ 0.07; SC_base ~ 0.2

In addition to MF_scan, two other MULTIFAN-CL models were tested. MF_qTS differed from MF_scan in that temporal variability in the main longline fishery catchability was estimated (in 10 year blocks with a CV~0.1; effort deviation CV ~ 0.2). The third specification, MF_yft, was adapted from the MULTIFAN-CL example application to simulated yellowfin tuna fishery data. Differences from MF_scan included:

- MF_yft natural mortality is estimated

- MF_yft length-at-age is estimated (with the correct SBT simulator values as starting points)

- MF_yft tag recovery likelihood is negative binomial; (MF_scan is Poisson)

```
# MF_scan.script
#!/bin/sh
#  -----------------------
#  PHASE 0 - create initial par file
#  -----------------------
#
if [ ! -f 00.par ]; then
  mfcl base.frq base.ini 00.par -makepar
fi
```

```
#  -----------------------
#  PHASE 1 - initial par
#  -----------------------
#
if [ ! -f 01.par ]; then
  mfcl base.frq 00.par 01.par -file - <<PHASE1
  2 113 1         # estimate initpop/totpop scaling parameter
  1 32 2          # sets standard initial estimation scheme
  ###1 111 4          # sets likelihood function for tags to negative
binomial
  1 111 2         ### sets likelihood function for tags to Poisson
  1 141 3         # sets likelihood function for LF data to normal
  2 57 1          # sets no. of recruitments per year to 1
  2 69 1          # sets generic movement option (now default)
  2  94  1  2  95  10   #  initial  age  structure  based  on  estimated  M
(assume virgin)
  -999  41  1        #  sets  penalty  weight  for  2nd  diff  smoothing  -
selectivity
  -9999  1  1         #  sets  no.  mixing  periods  for  all  tag  release
groups to 1
# sets non-decreasing selectivity for spaawning longline fisheries
#  -3 16 1 -4 16 1
# set penalty weight on effort devs (note could have used 0 for 1,3,4
#                            and negative means sqrt(effort) invoked)
  -1 13 1
  -2 13 100
  -3 13 1
  -4 13 1
# grouping of fisheries with common selectivity
   -1 24 1
   -2 24 2
   -3 24 3
   -4 24 4
# grouping of fisheries with common tag-reporting rates
    -1 34 1
    -2 34 2
    -3 34 3
    -4 34 4
# sets penalties on tag-reporting rate priors
    -1 35 1
    -2 35 1
    -3 35 1
    -4 35 1
# sets prior means for tag-reporting rates
    -1 36 100
    -2 36 100
    -3 36 100
    -4 36 100
#  -999 33 1       # estimate tag-reporting rates
  1 33 100         # maximum tag reporting rate for all fisheries is
1
PHASE1
fi
#  ---------
#   PHASE 2
#  ---------
if [ ! -f 02.par ]; then
  mfcl base.frq 01.par 02.par -file - <<PHASE2
  1 149 -1       # set penalty on recruitment devs to 200/10
  -999 3 15      # all selectivities equal for age class 15 and older
  -1 3 10    # set selectivity to 0
  -1 16 2    # for ages 10 in fishery 1
  1 189 1          # write length.fit and weight.fit (obs. and pred.
LF data)
  1 190 1         # write plot.rep
  1 1 100         # set max. number of function evaluations per phase
to 100
  1 50 0         # set convergence criterion to 1E+00
PHASE2
fi
#  ---------
#   PHASE 3
#  ---------
if [ ! -f 03.par ]; then
  mfcl base.frq 02.par 03.par -file - <<PHASE3
  2 70 1          # activate parameters and turn on
  2 71 1             # estimation of temporal changes in recruitment
distribution
  2 110 5         # penalty weight for deviations
PHASE3
fi
#  ---------
#   PHASE 4
#  ---------
if [ ! -f 04.par ]; then
  mfcl base.frq 03.par 04.par -file - <<PHASE4
  2 68 1          # estimate movement coefficients
PHASE4
fi
#  ---------
#   PHASE 5
#  ---------
if [ ! -f 05.par ]; then
  mfcl base.frq 04.par 05.par -file - <<PHASE5
  1 16 1          # estimate length dependent SD
PHASE5
fi
#  ---------
#   PHASE 6
#  ---------
if [ ! -f 06.par ]; then
  mfcl base.frq 05.par 06.par -file - <<PHASE6
#  1 14 1          # estimate K
PHASE6
fi
#  ---------
#   PHASE 7
#  ---------
if [ ! -f 07.par ]; then
  mfcl base.frq 06.par 07.par -file - <<PHASE7
  ###1 173 8          # estimate independent mean lengths for 1st 8
age classes
  ###1 182 10        # penalty weight for deviations
PHASE7
fi
#  ---------
#   PHASE 8
#  ---------
#if [ ! -f 08.par ]; then
#  mfcl base.frq 07.par 08.par -file - <<PHASE8
#  -999 27 1       # estimate seasonal catchability for all fisheries
#    1 14 0        # de-activate K for the time being
```

```
#PHASE8
#fi
#  ---------
#   PHASE 9
#  ---------
if [ ! -f 09.par ]; then
  mfcl base.frq 07.par 09.par -file - <<PHASE9
  -1 10 1       # estimate catchability time series for fishery 1
  -3 10 1       # estimate catchability time series for fishery 3
  -4 10 1       # estimate catchability time series for fishery 4
  -999 23 23      # and do a random-walk step every 23+1 months
PHASE9
fi
#  ---------
#   PHASE 10
#  ---------
if [ ! -f 10.par ]; then
  mfcl base.frq 09.par 10.par -file - <<PHASE10
  ###2 33 1          # estimate constant natural mortality rate
PHASE10
fi
#  ---------
#   PHASE 11
#  ---------
#if [ ! -f 11.par ]; then
#  mfcl base.frq 10.par 11.par -file - <<PHASE11
#  2 88 1          # activate parameters
#  2 89 1          # and estimate age-dependent movement
#PHASE11
#fi
#  ---------
#   PHASE 12
#  ---------
if [ ! -f 12.par ]; then
  mfcl base.frq 10.par 12.par -file - <<PHASE12
  ###2 73 1          # estimate age-dependent M
PHASE12
fi
#  ---------
#   PHASE 13
#  ---------
if [ ! -f 13.par ]; then
  mfcl base.frq 12.par 13.par -file - <<PHASE13
  1 14 1          # estimate von Bertalanffy K
PHASE13
fi
#  ---------
#   PHASE 14
#  ---------
if [ ! -f 14.par ]; then
  mfcl base.frq 13.par 14.par -file - <<PHASE14
# estimation of negative binomial parameter a
 -999 43 1        # estimate a for all fisheries
PHASE14
fi
#  ---------
#   PHASE 15
#  ---------
if [ ! -f 15.par ]; then
  mfcl base.frq 14.par 15.par -file - <<PHASE15
  -100000  1  1         #  estimate  time-invariant  distribution  of
recruitment
PHASE15
fi
#  ---------
#   PHASE 16
#  ---------
if [ ! -f 16.par ]; then
  mfcl base.frq 15.par 16.par -file - <<PHASE16
  1  12 1         # estimate age 1 length-at-age
  -1 15 1         # q time series penalty weakest possible constraint
  -3 15 1         # q time series penalty weakest possible constraint
  -4 15 1         # q time series penalty weakest possible constraint
  2 145 1         # estimate Beverton Holt SRR with small penalty
  2 146 1         # SRR parameter active
  2 147 1         # recruitment lag is 1 year
  2 148 10        # base F is average over last 10 years
  2 155 2         # base F average does not include last 2 year
  2 153 1           # parameters of beta distribution defining prior
for
  2  154  1             #  steepness  -  mode  =  0.5,  sd  =  "broad  or
unconstrained"
  1 149 0           # reduce pens on devs from av. recr (to avoid 2
penalties)
  1 1 3000         # set no. function evaluations for final phase to
3000
  1 50 4          # set convergence criterion to 1E-04
  -999  55  1           #  compute  biomass  with  catchability  for  all
fisheries set to 0
PHASE16
fi
```

APPENDIX 6 GRAPHICAL SUMMARY OF SESAME SIMULATED SBT ASSESSMENT MODEL RESULTS

The following graphical archive illustrates the estimation performance of a range of stock assessment models applied to simulated data from the SESAME VSM operating model. The operating model was parameterized to represent various plausible parameterizations of a fishery system roughly resembling Southern Bluefin Tuna. In the following figures, the assessment model is labeled at the top of each page (definitions in the body of the report, Table 2) and the performance indicator is labelled at the top of each panel (Table 8). The operating model to which the assessment model was applied is indicated on the x-axis of the boxplots or the sub-heading of the time series plots (definitions in Table 1).

The boxplots and time series of quantiles represent the frequency distributions of the ratio (assessment model estimate)/(operating model known value) for each performance indicator. The assessment model value corresponds to the estimate of the parameters at the Maximum Posterior Density (mode of the objective function). The frequency distributions result from applying the assessment model to 10 stochastic realizations from each operating model.

An asterisk (*) indicates a missing value, (e.g. because the Fox model does not estimate recruitment). An arrow (^) on the boxplots indicates that all values are off the scale. All time series plots are truncated at an upper value of 3.0. The Log10(max. gradient) boxplots provide a rough indication of convergence problems in the function minimizer. For the SCALIA and production models as implemented, we start to be concerned with values larger than –1. The "Penalty activation count" indicates the number of times that the ASPM model converged to a minimum that we consider to be implausible due to a numerical limitation in the model.

A 6.1 F_CALC

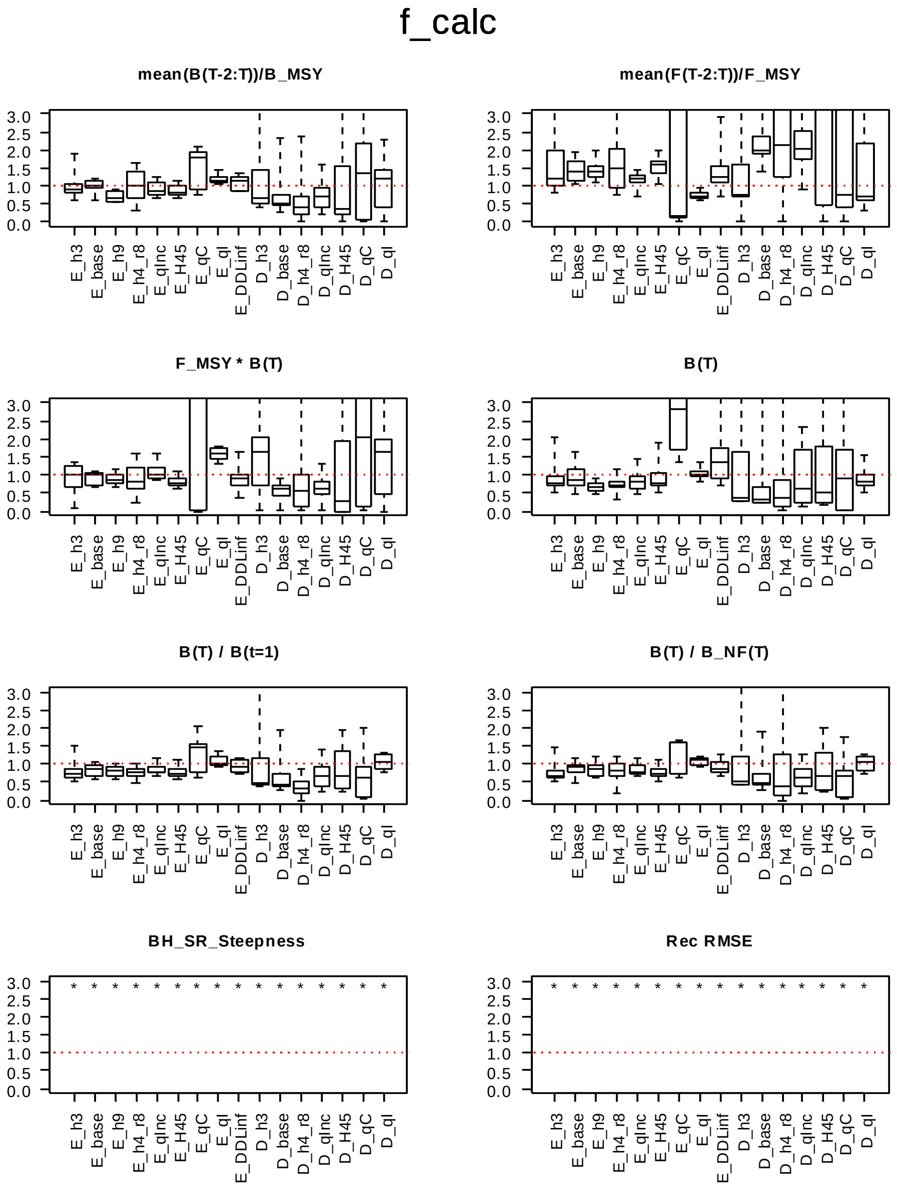


Figure A 6.1-a

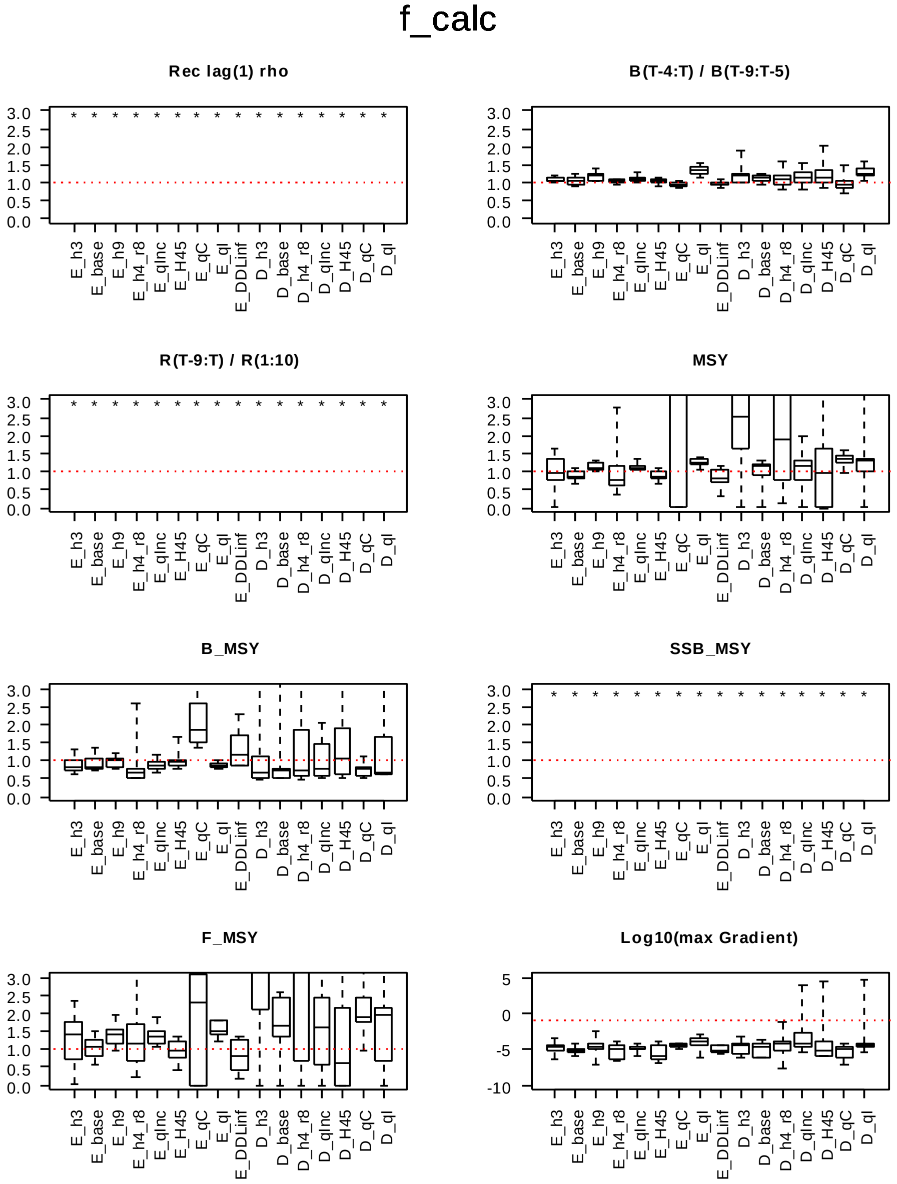
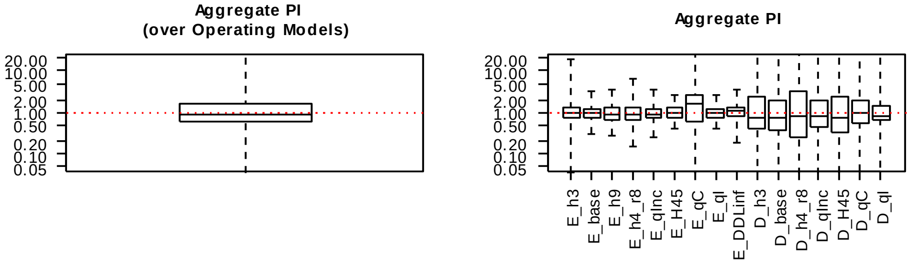

f_calc
Rec lag(1) rho
B(T-4:T) / B(T-9:T-5)
R(T-9:T) / R(1:10)
MSY
B_MSY
SSB_MSY
F_MSY
Log10(max Gradient)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5
0
-5
-10
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_ql
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_ql

Figure A 6.1-b

f_calc


Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_ql
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_ql

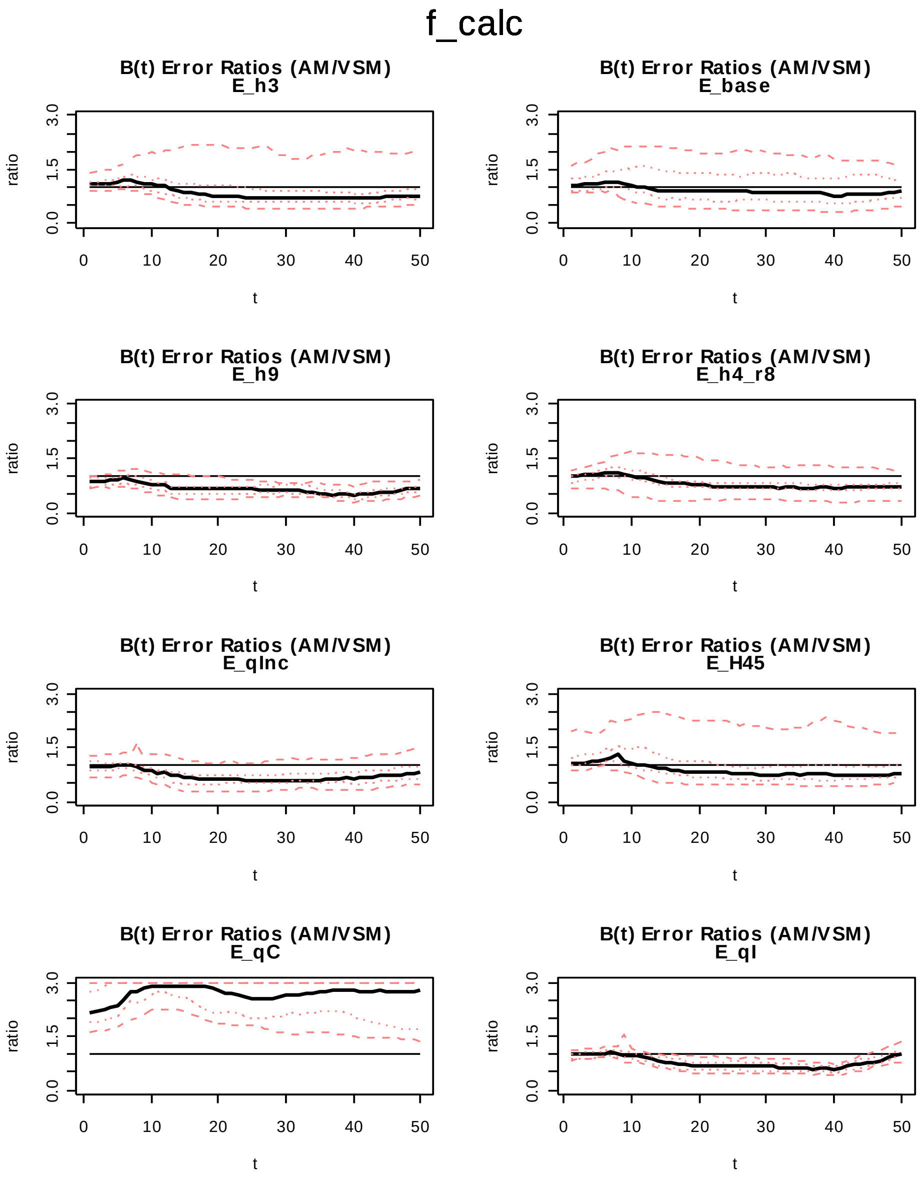
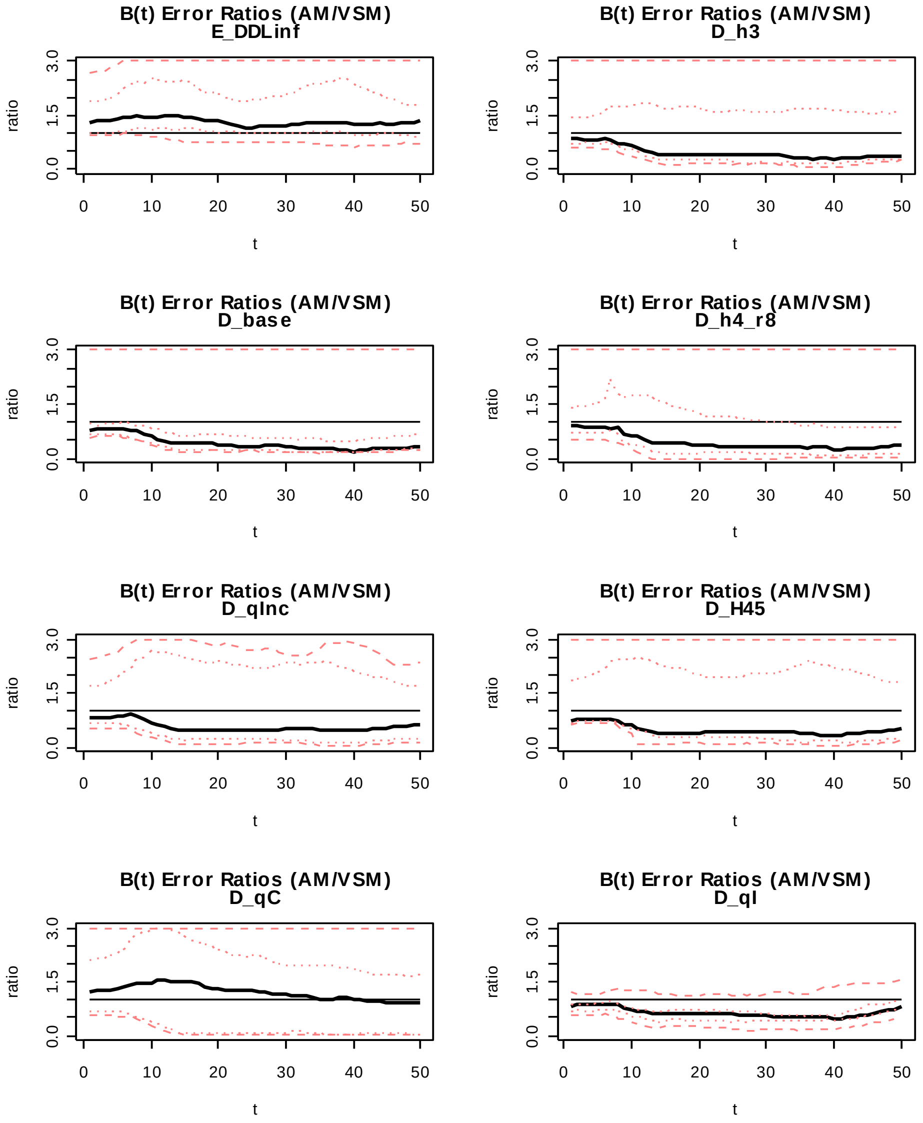

f_calc
B(t) Error Ratios (AM/VSM)
E_h3
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_h9
B(t) Error Ratios (AM/VSM)
E_h4_r8
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_H45
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.1-c

f_calc


B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t) Error Ratios (AM/VSM)
D_h3
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_h4_r8
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_H45
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

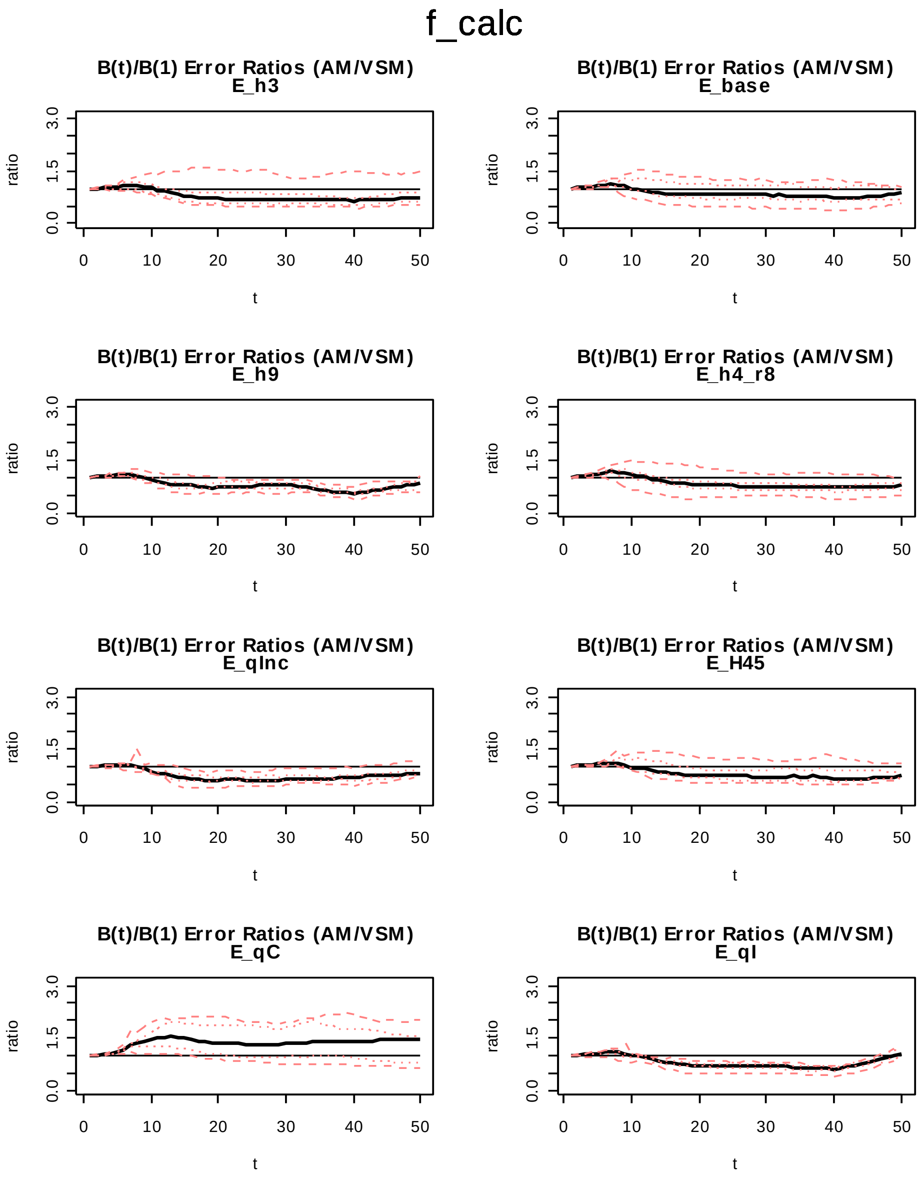
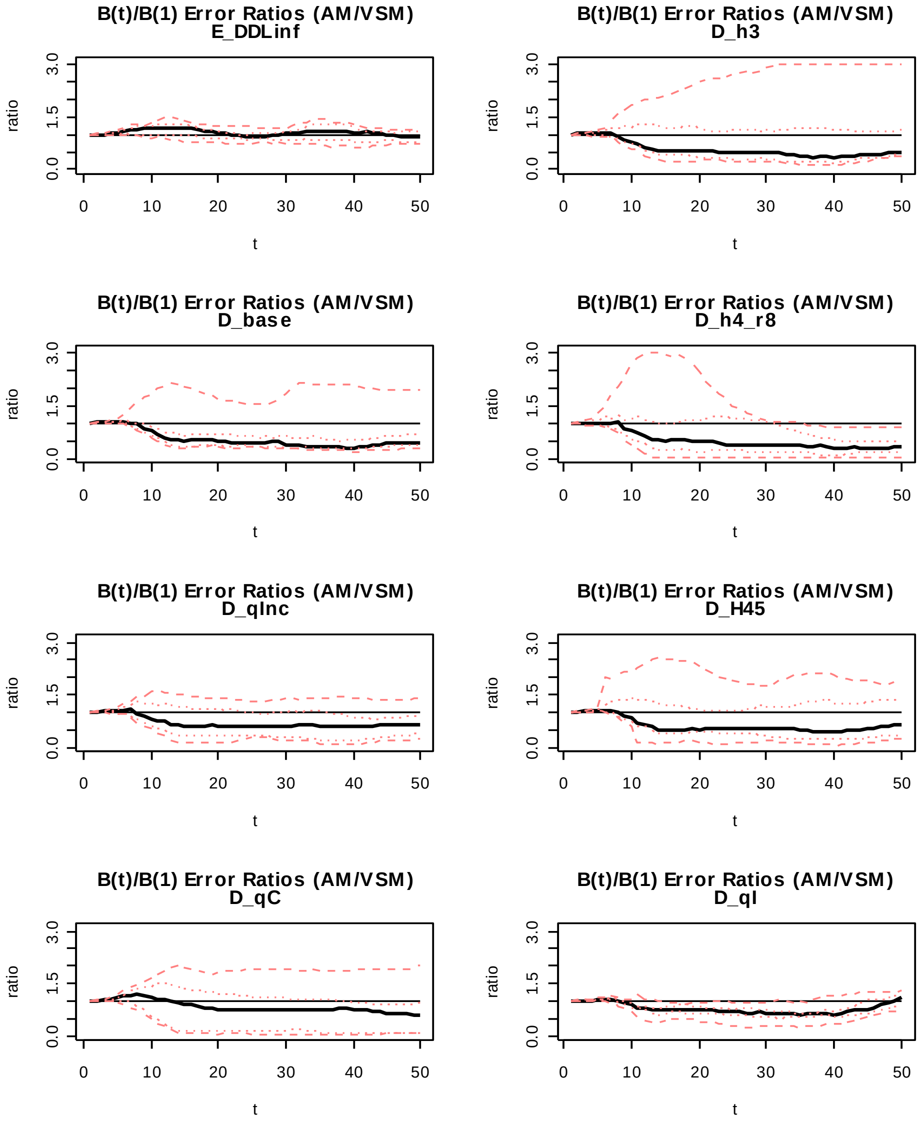

f_calc
B(t)/B(1) Error Ratios (AM/VSM)
E_h3
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_h9
B(t)/B(1) Error Ratios (AM/VSM)
E_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_qC
B(t)/B(1) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Figure A 6.1-d

f_calc



B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf

B(t)/B(1) Error Ratios (AM/VSM)
D_h3

B(t)/B(1) Error Ratios (AM/VSM)
D_base

B(t)/B(1) Error Ratios (AM/VSM)
D_h4_r8

B(t)/B(1) Error Ratios (AM/VSM)
D_qInc

B(t)/B(1) Error Ratios (AM/VSM)
D_H45

B(t)/B(1) Error Ratios (AM/VSM)
D_qC

B(t)/B(1) Error Ratios (AM/VSM)
D_ql

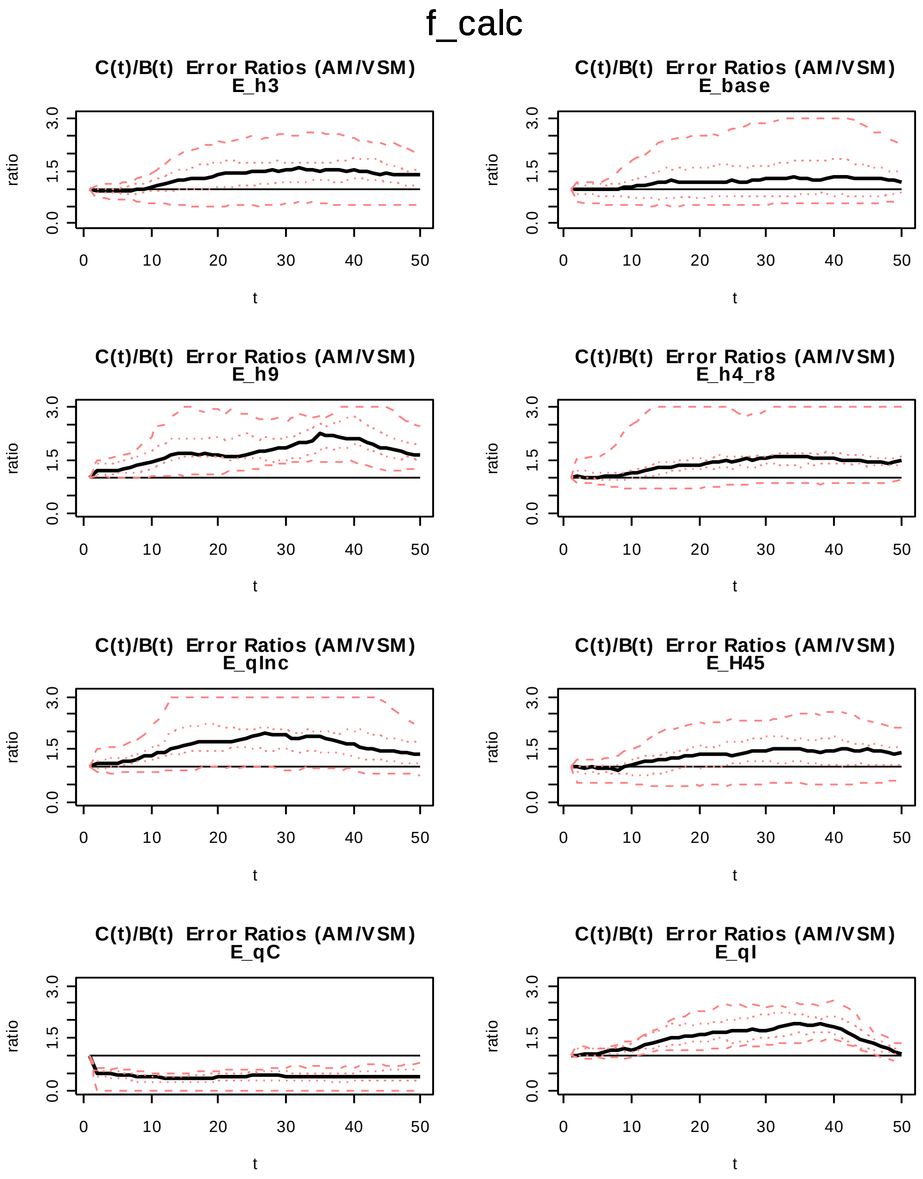

f_calc
C(t)/B(t) Error Ratios (AM/VSM)
E_h3
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h9
C(t)/B(t) Error Ratios (AM/VSM)
E_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_qC
C(t)/B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Figure A 6.1-e

A 6.2 ASPM_D2G

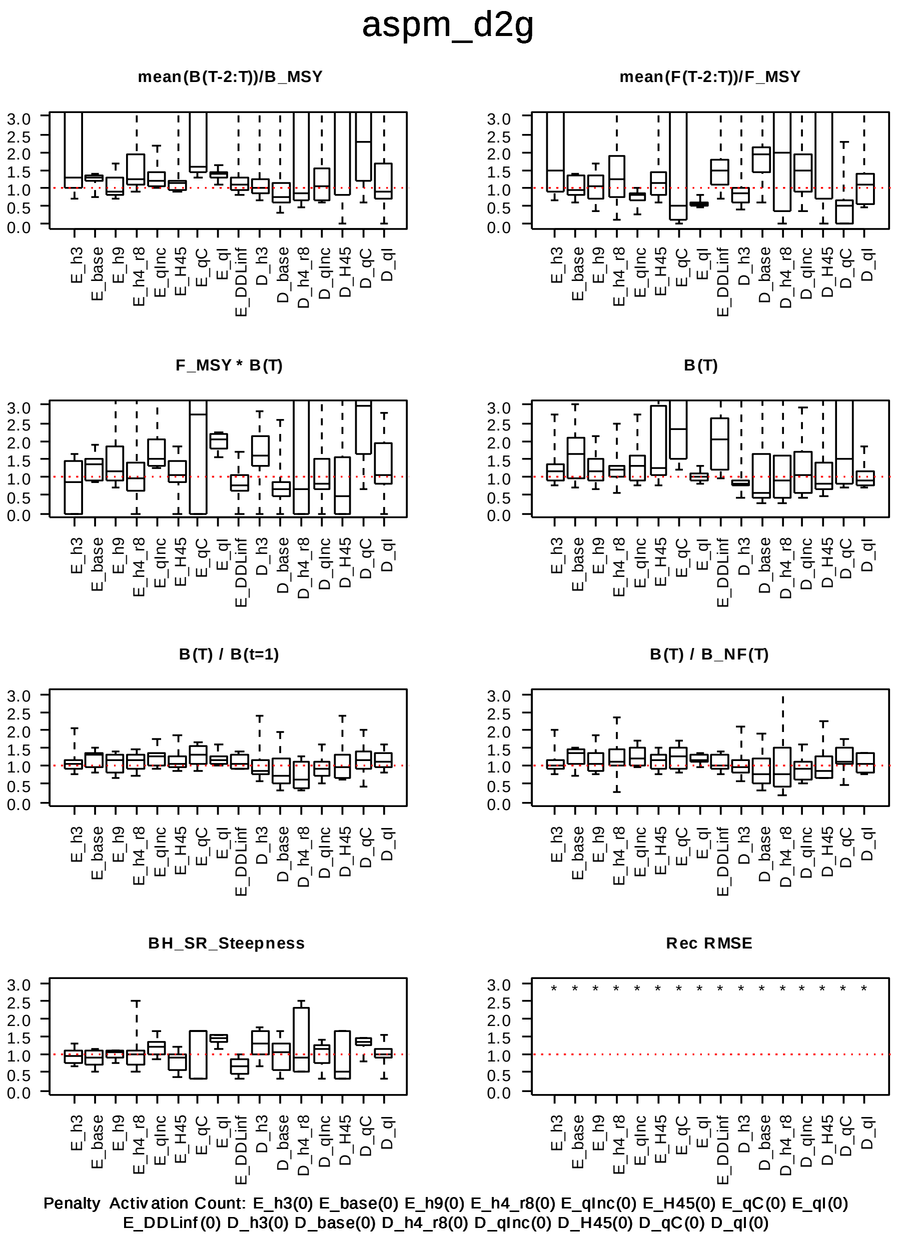
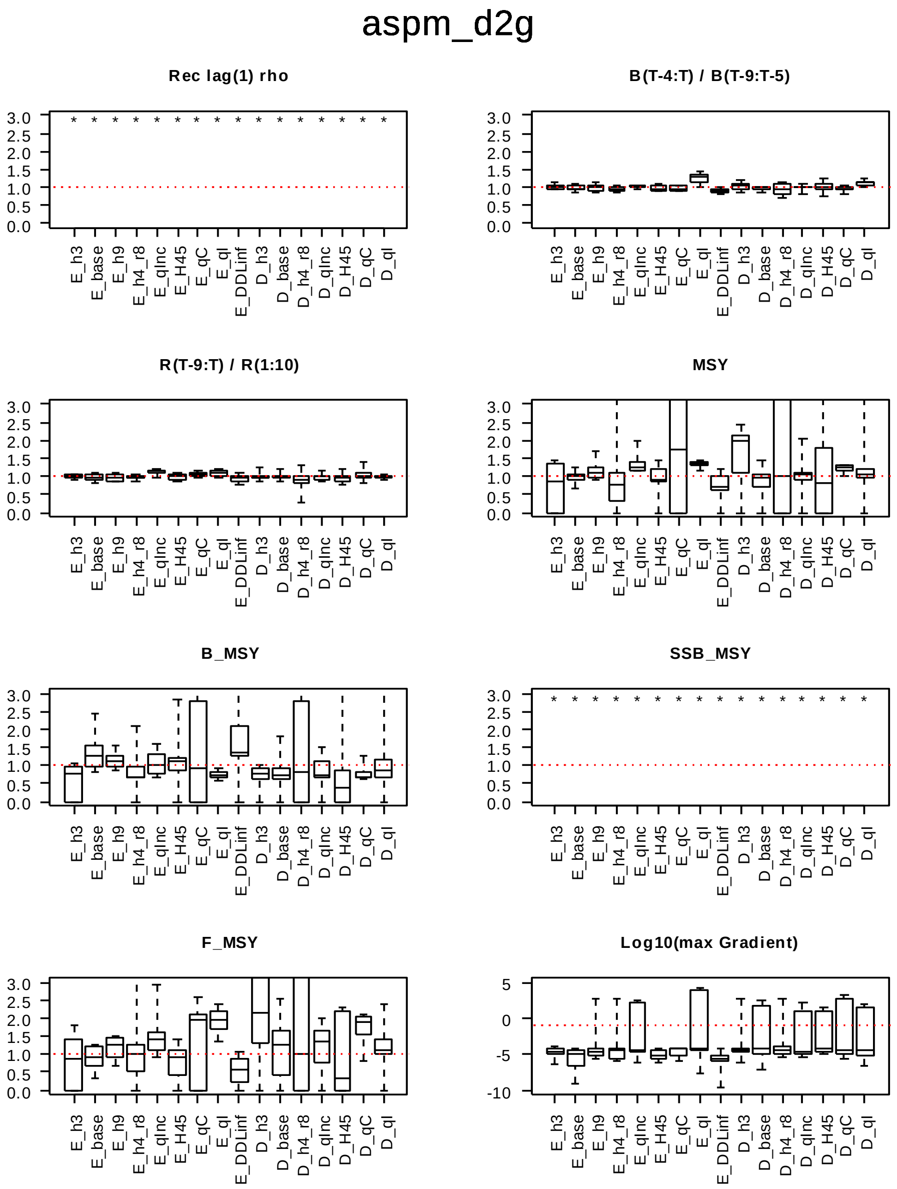


Figure A 6.2-a



Rec lag(1) rho

B(T-4:T) / B(T-9:T-5)

R(T-9:T) / R(1:10)

MSY

B_MSY

SSB_MSY

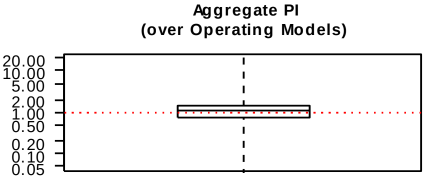
F_MSY

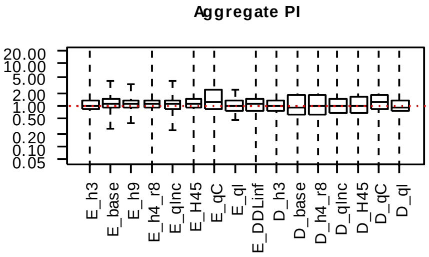
Log10(max Gradient)

Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.2-b

aspm_d2g





Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_ql(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_ql(0)

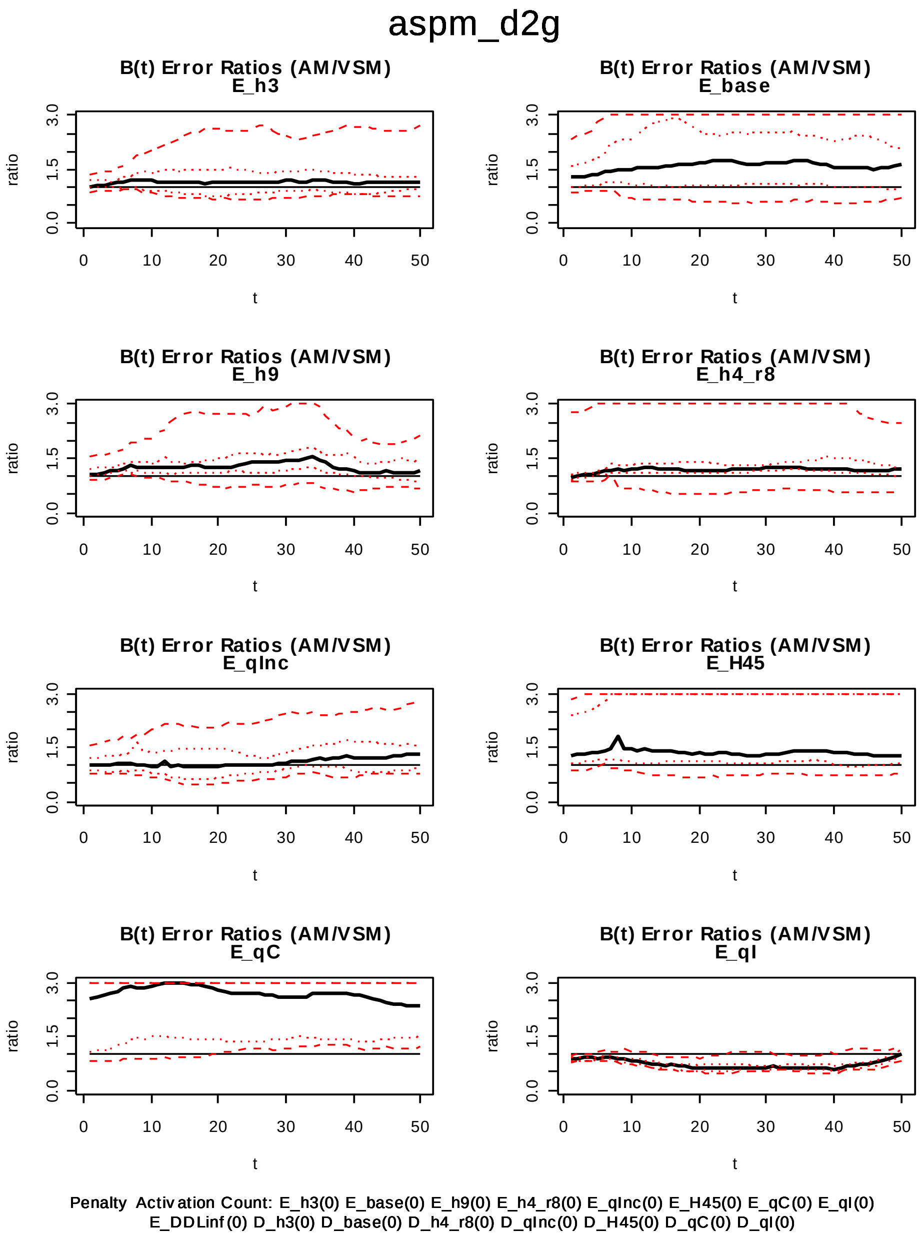
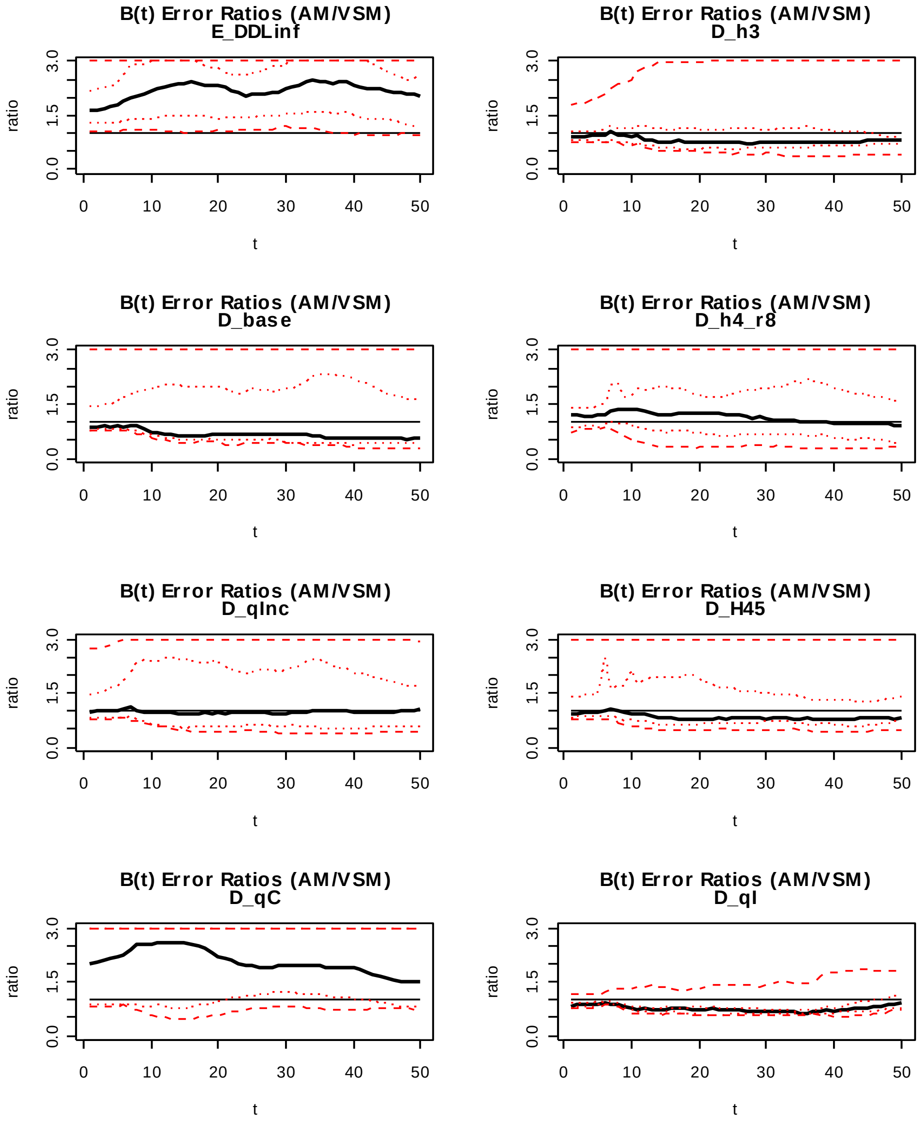

aspm_d2g
B(t) Error Ratios (AM/VSM)
E_h3
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_h9
B(t) Error Ratios (AM/VSM)
E_h4_r8
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_H45
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.2-c

aspm_d2g



B(t) Error Ratios (AM/VSM) E_DDLinf

B(t) Error Ratios (AM/VSM) D_h3

B(t) Error Ratios (AM/VSM) D_base

B(t) Error Ratios (AM/VSM) D_h4_r8

B(t) Error Ratios (AM/VSM) D_qInc

B(t) Error Ratios (AM/VSM) D_H45

B(t) Error Ratios (AM/VSM) D_qC

B(t) Error Ratios (AM/VSM) D_ql

Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_ql(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_ql(0)

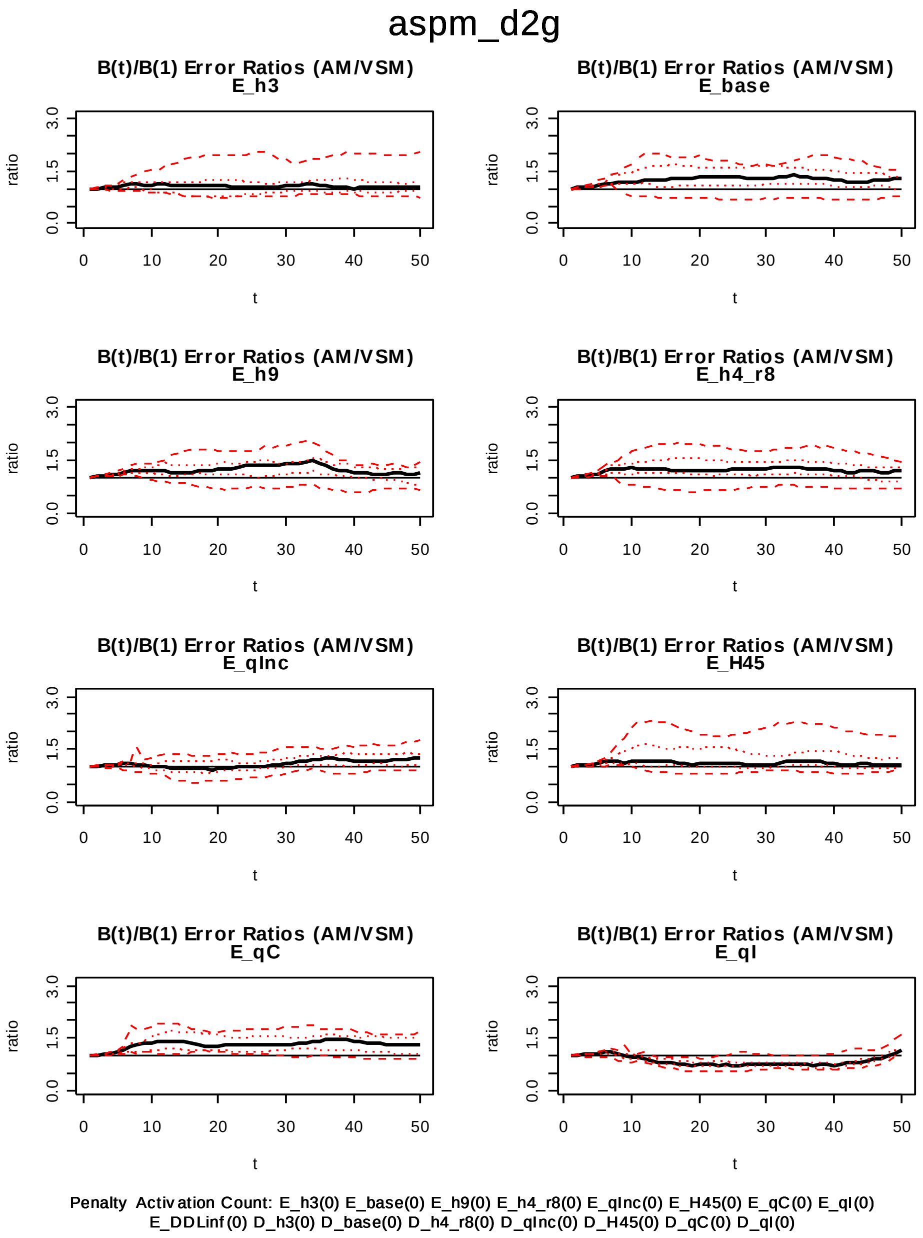
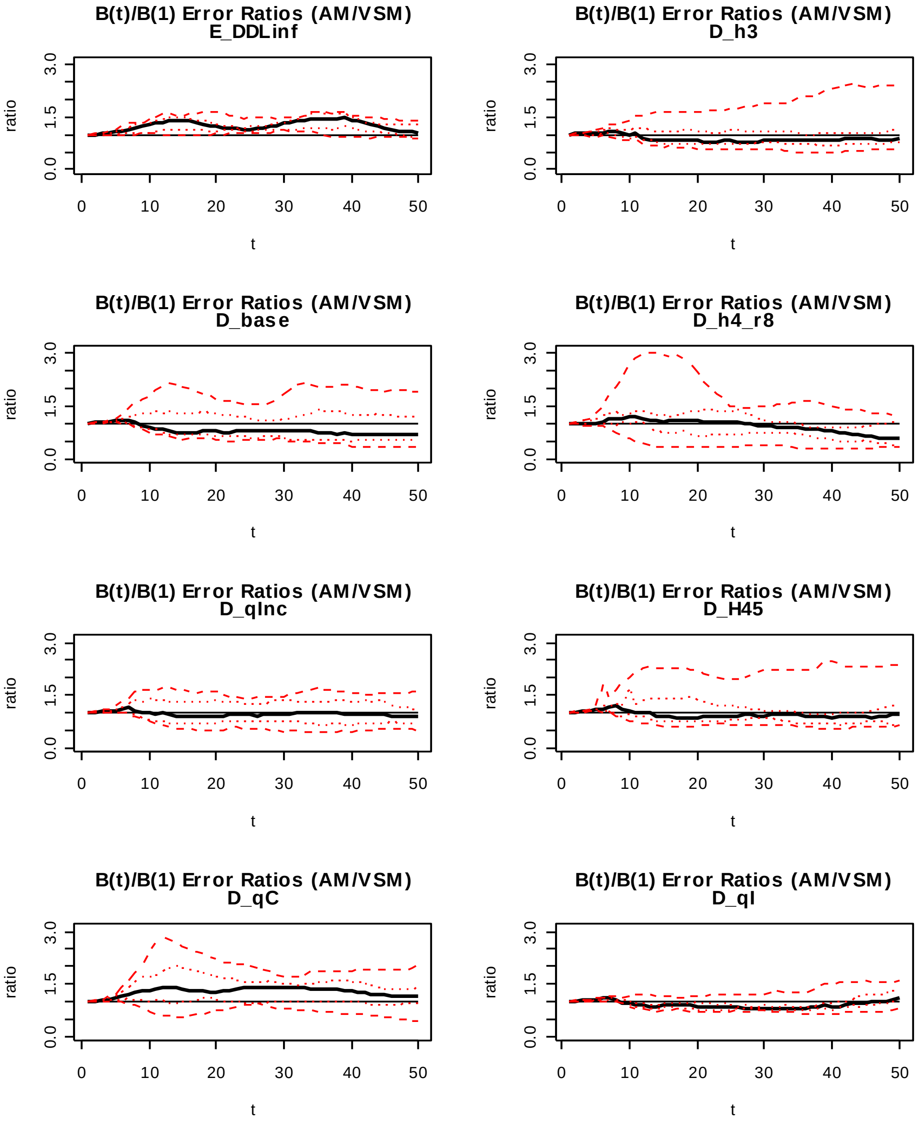

aspm_d2g
B(t)/B(1) Error Ratios (AM/VSM)
E_h3
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_h9
B(t)/B(1) Error Ratios (AM/VSM)
E_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_qC
B(t)/B(1) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.2-d

aspm_d2g



Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

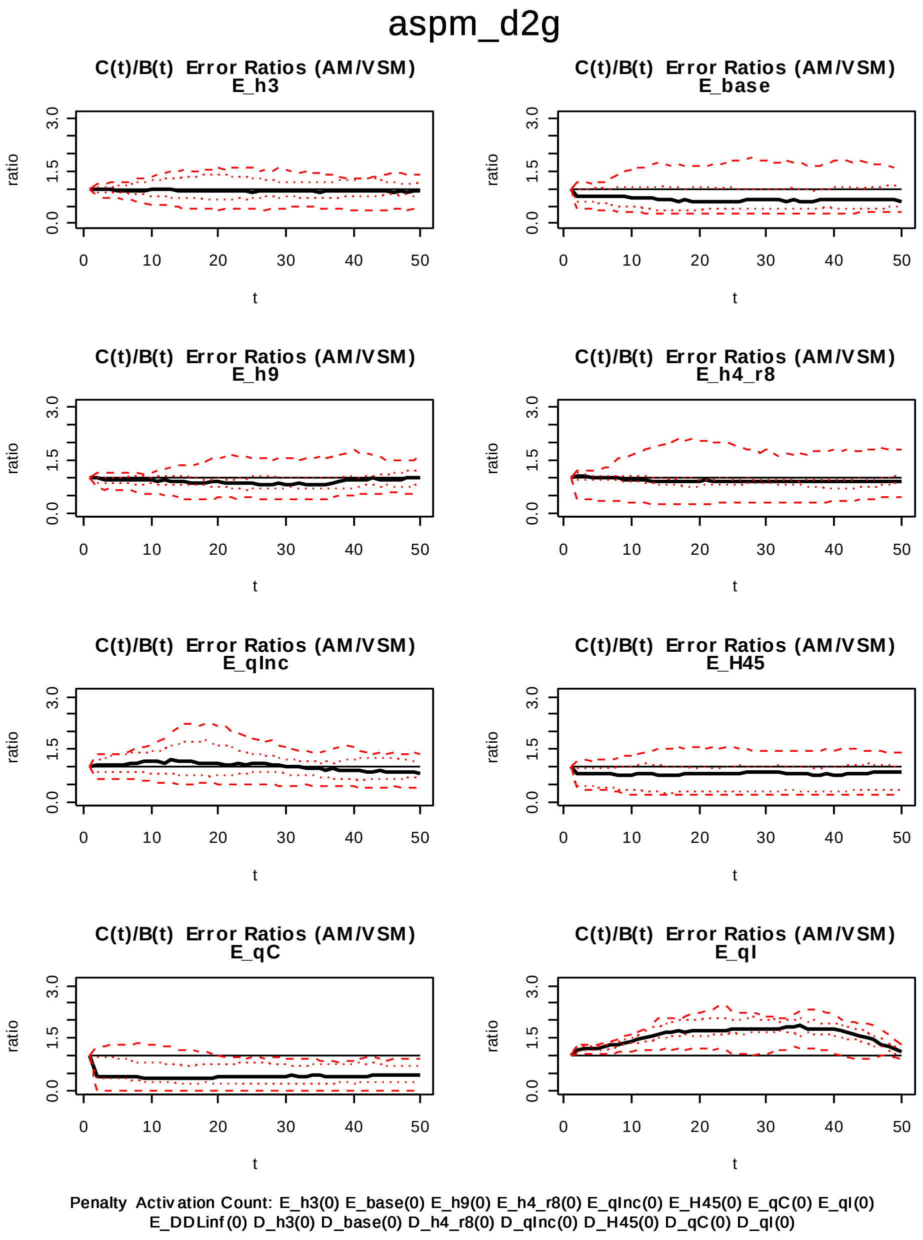
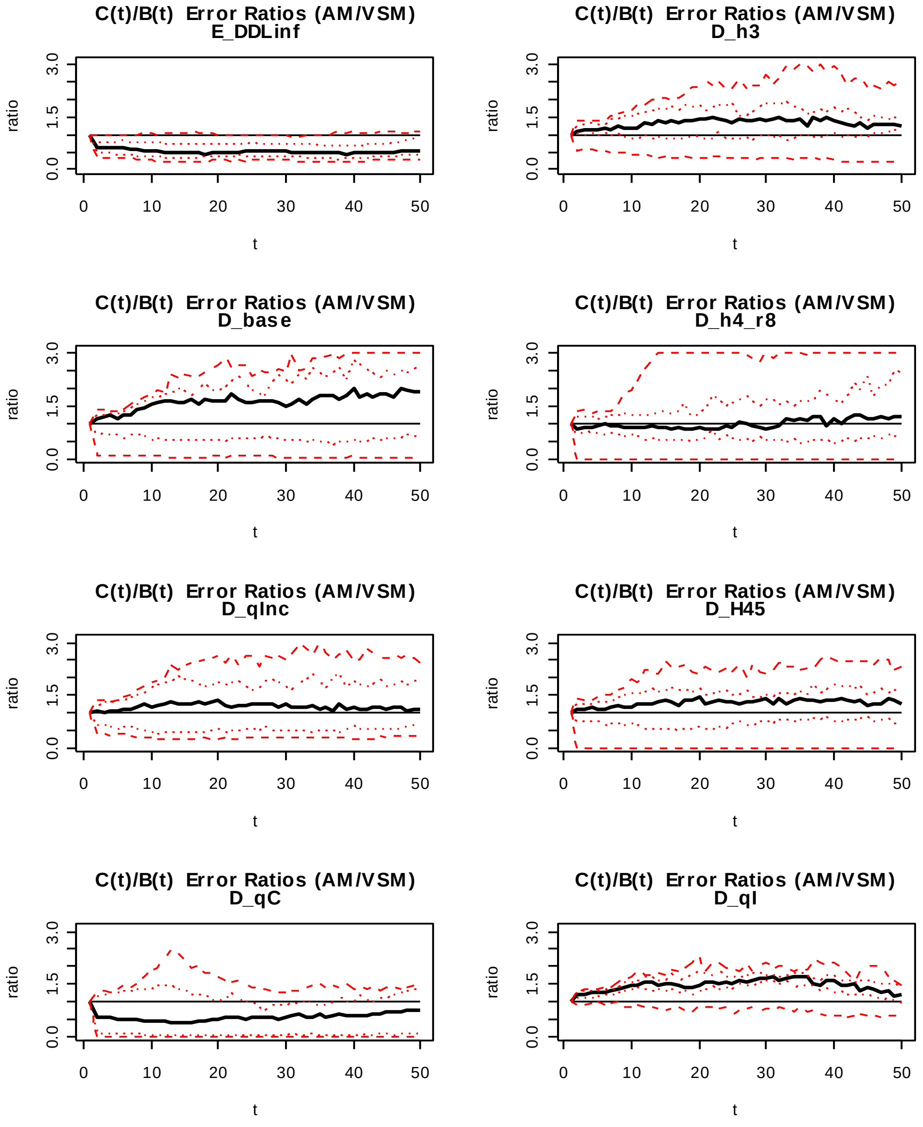
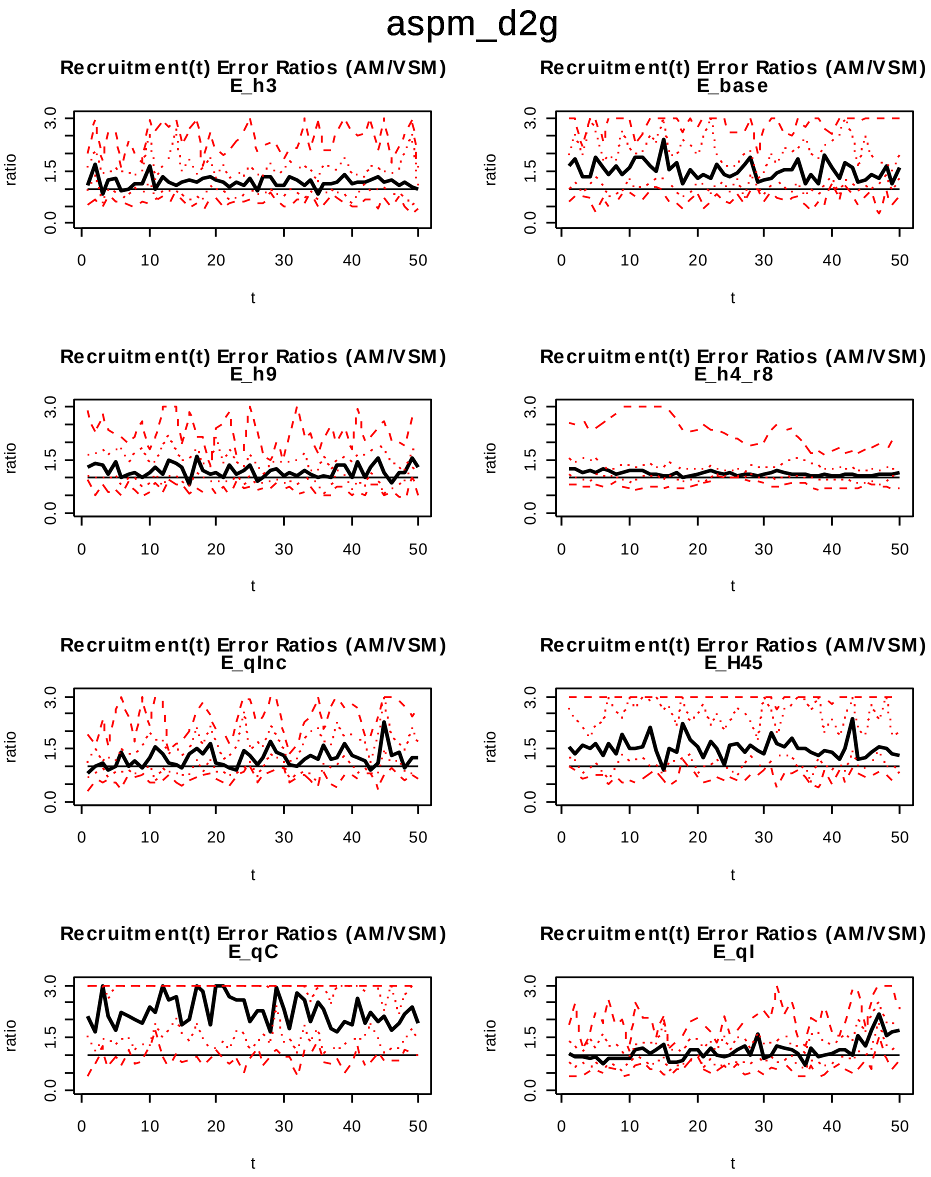

aspm_d2g
C(t)/B(t) Error Ratios (AM/VSM)
E_h3
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h9
C(t)/B(t) Error Ratios (AM/VSM)
E_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_qC
C(t)/B(t) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_ql(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_ql(0)

Figure A 6.2-e

aspm_d2g



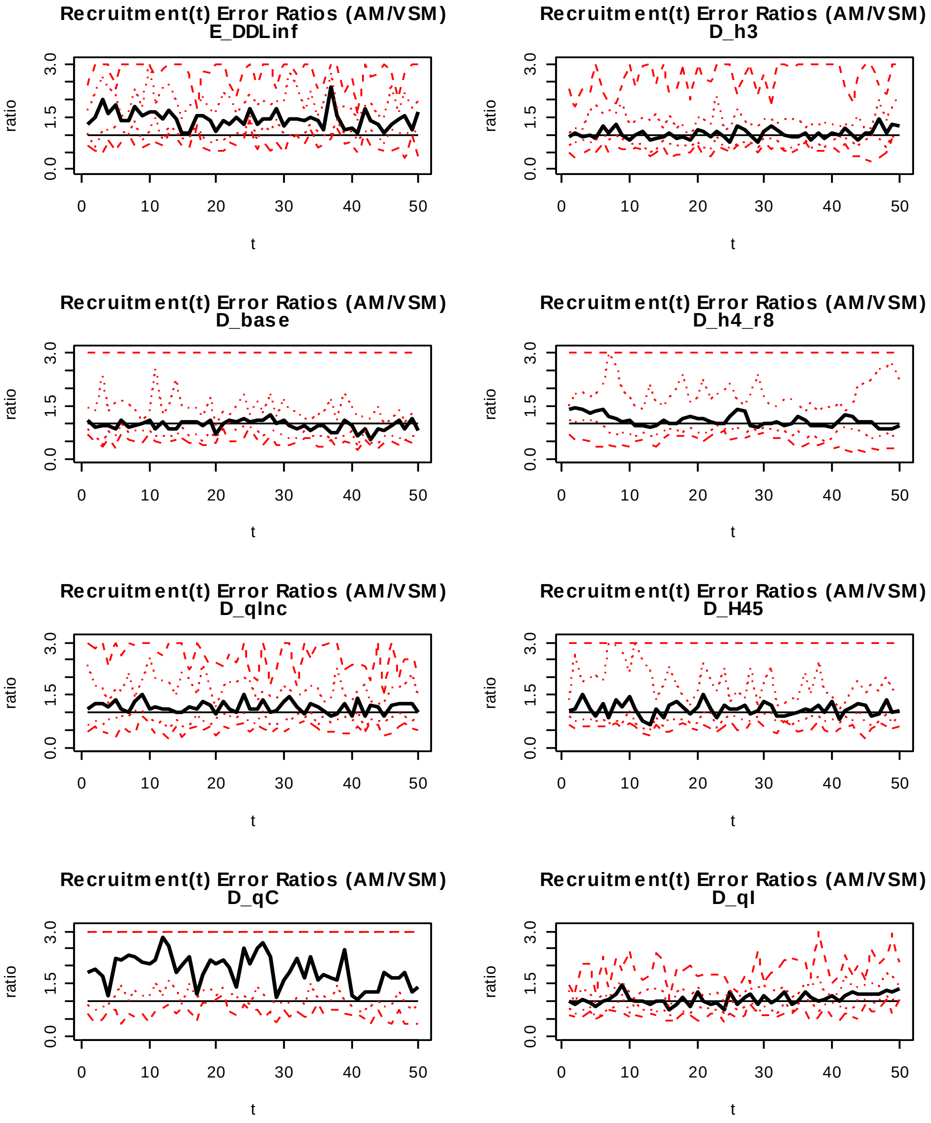
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)



Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.2-f

aspm_d2g



Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_ql(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_ql(0)

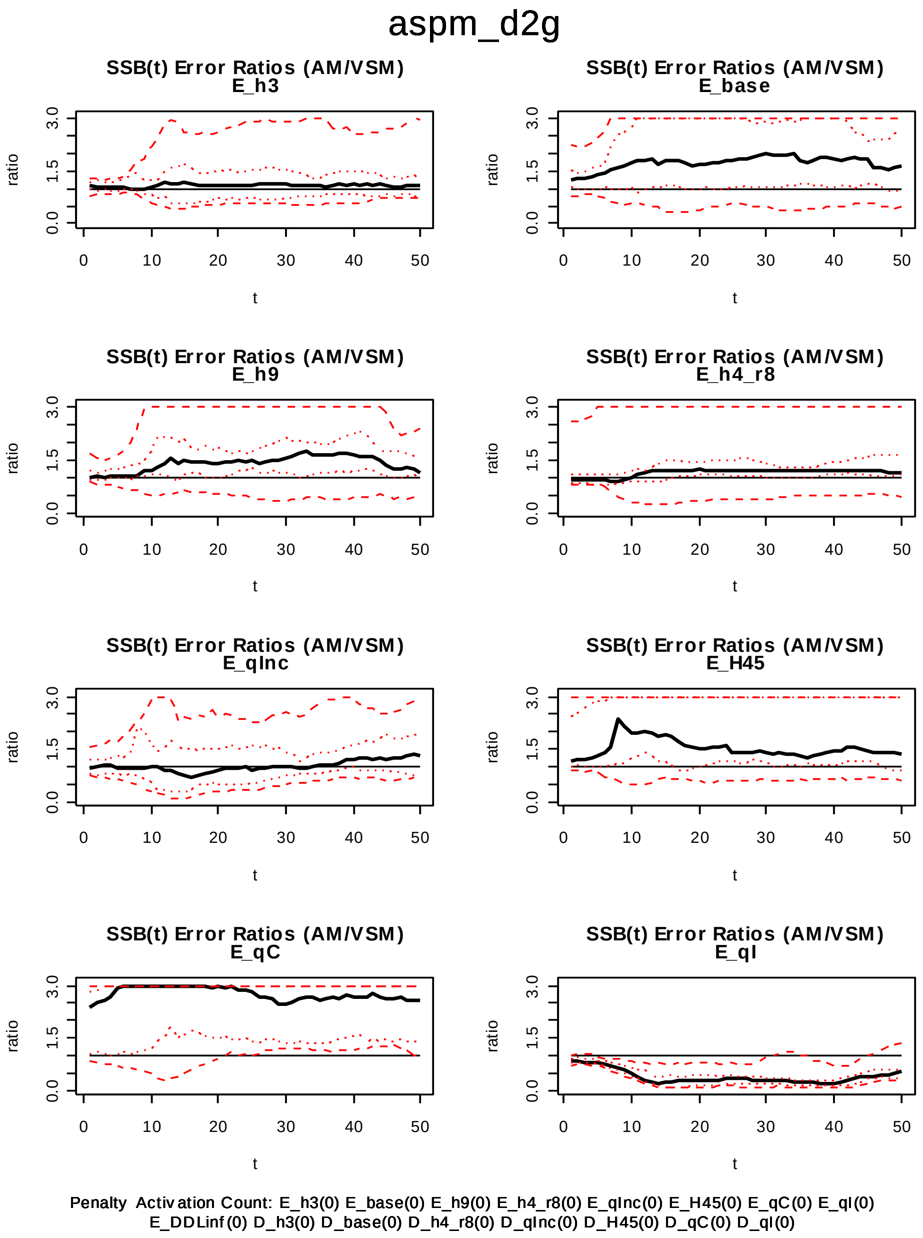

aspm_d2g
SSB(t) Error Ratios (AM/VSM)
E_h3
SSB(t) Error Ratios (AM/VSM)
E_base
SSB(t) Error Ratios (AM/VSM)
E_h9
SSB(t) Error Ratios (AM/VSM)
E_h4_r8
SSB(t) Error Ratios (AM/VSM)
E_qInc
SSB(t) Error Ratios (AM/VSM)
E_H45
SSB(t) Error Ratios (AM/VSM)
E_qC
SSB(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.2-g

A 6.3 ASPM_D6G

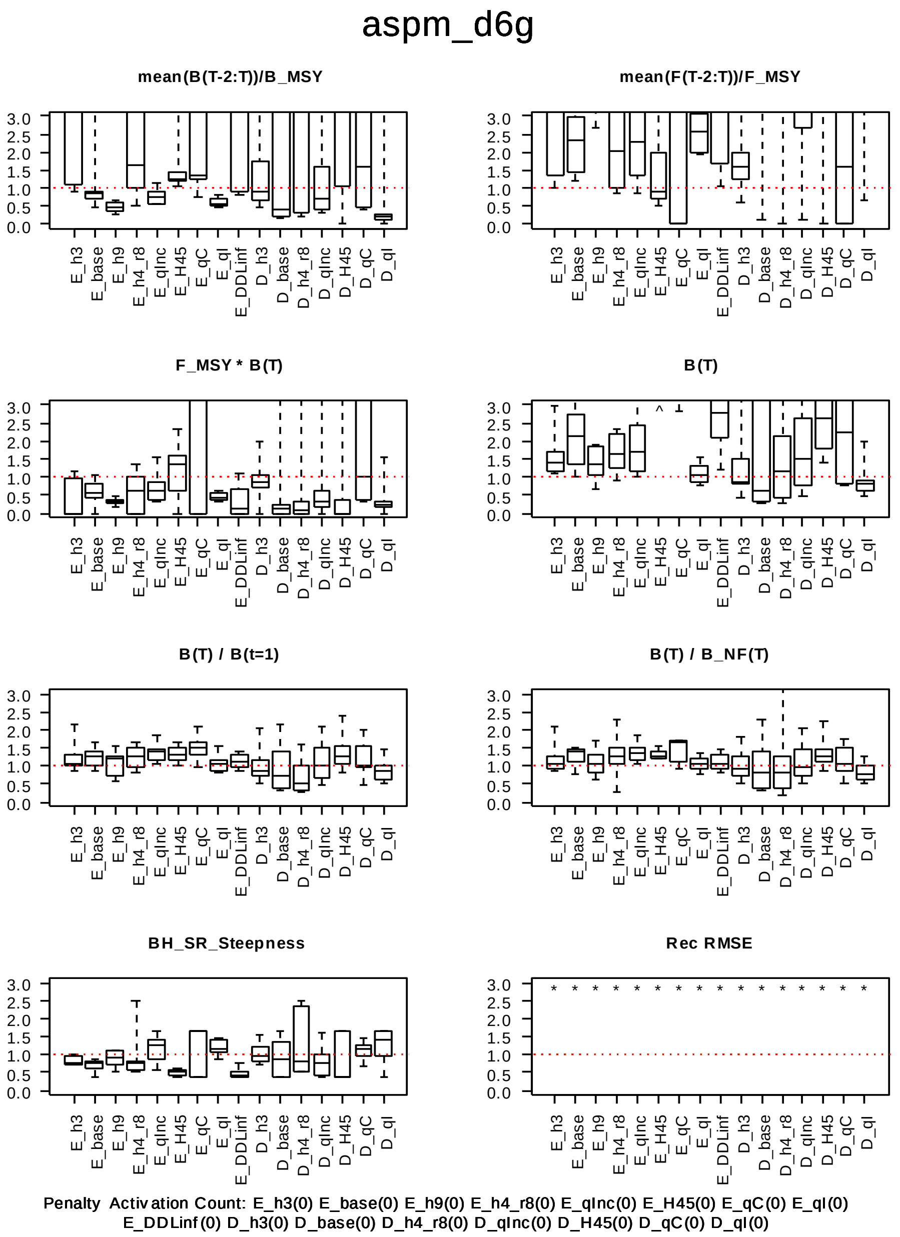
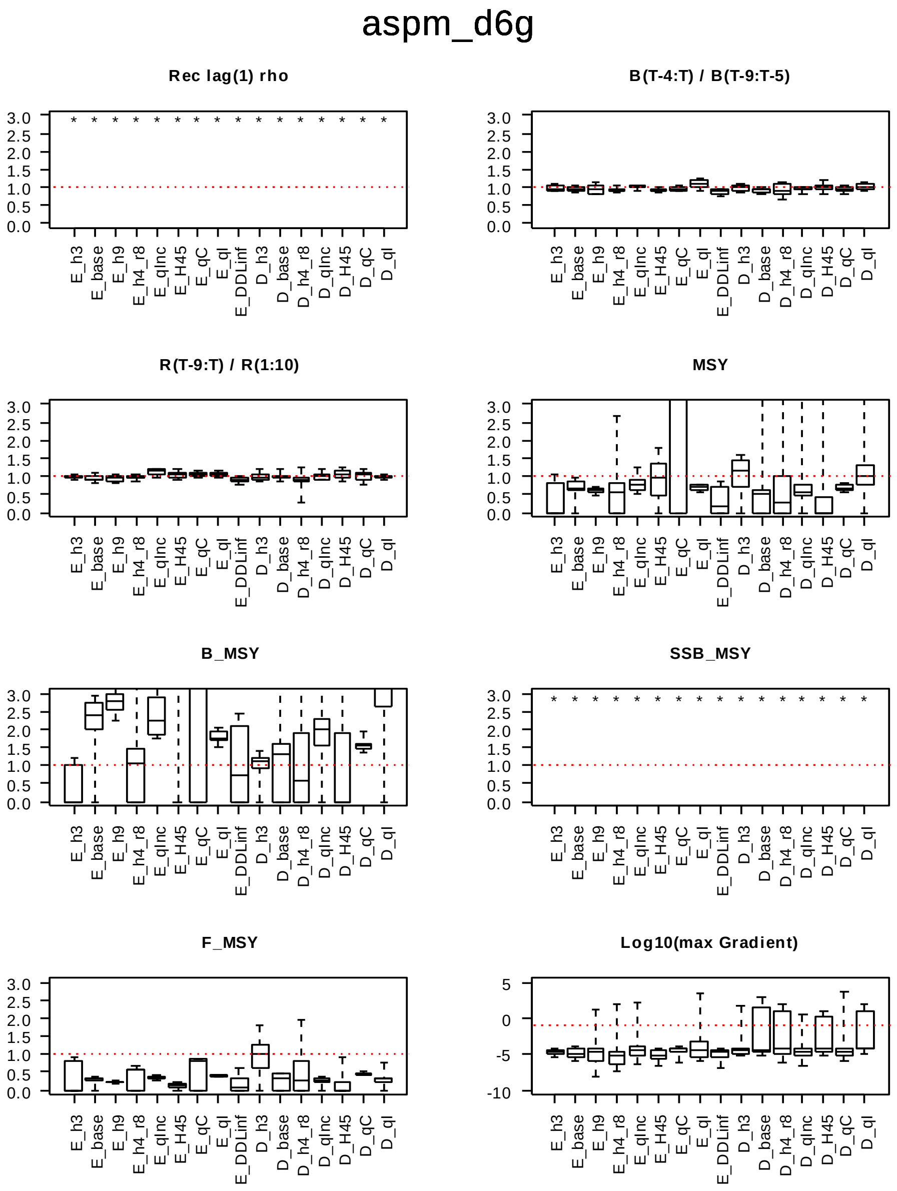


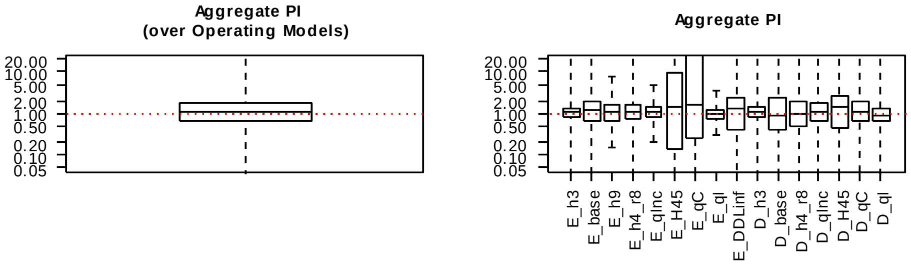
Figure A 6.3-a



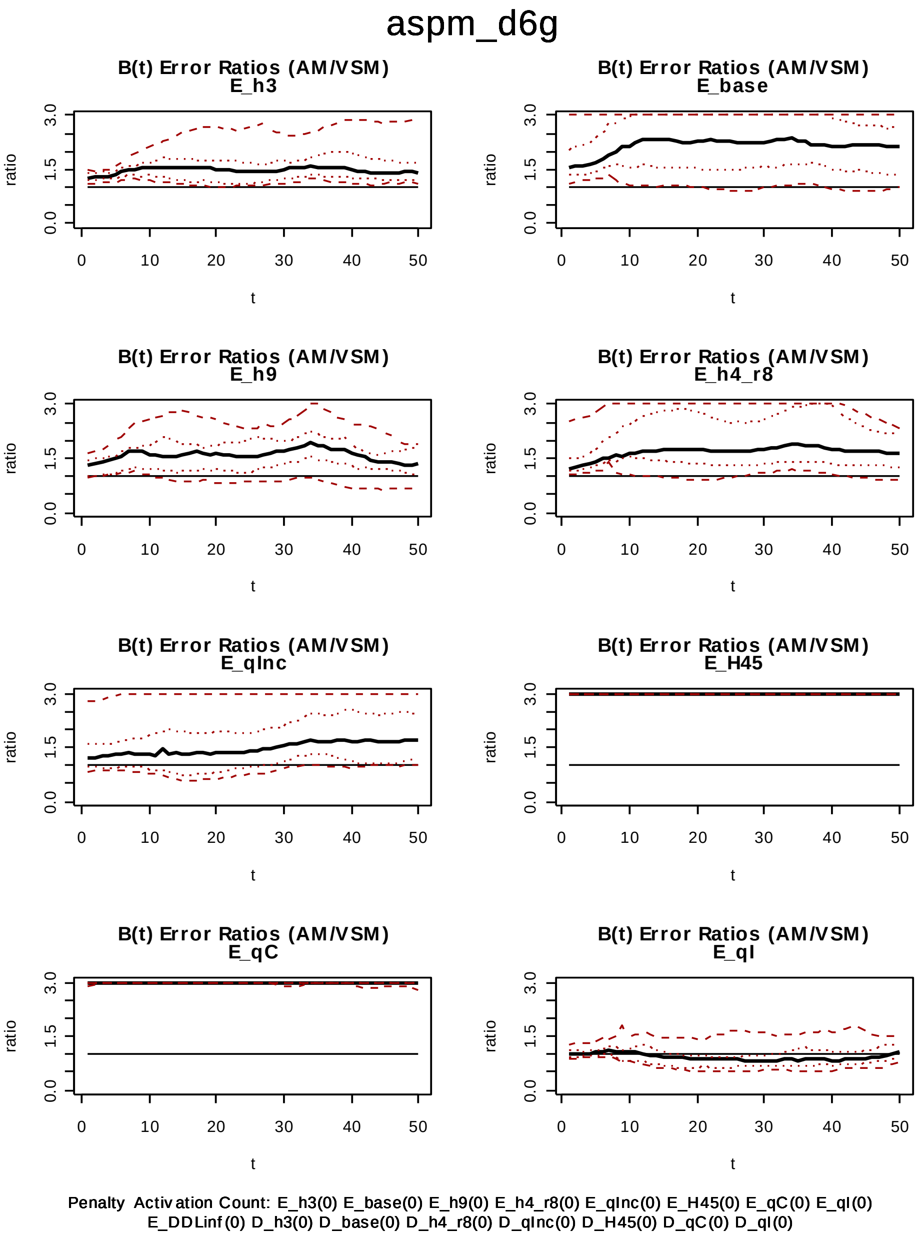
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_ql(0) E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_ql(0)

Figure A 6.3-b

aspm_d6g



Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)



aspm_d6g

B(t) Error Ratios (AM/VSM) E_h3

B(t) Error Ratios (AM/VSM) E_base

B(t) Error Ratios (AM/VSM) E_h9

B(t) Error Ratios (AM/VSM) E_h4_r8

B(t) Error Ratios (AM/VSM) E_qInc

B(t) Error Ratios (AM/VSM) E_H45

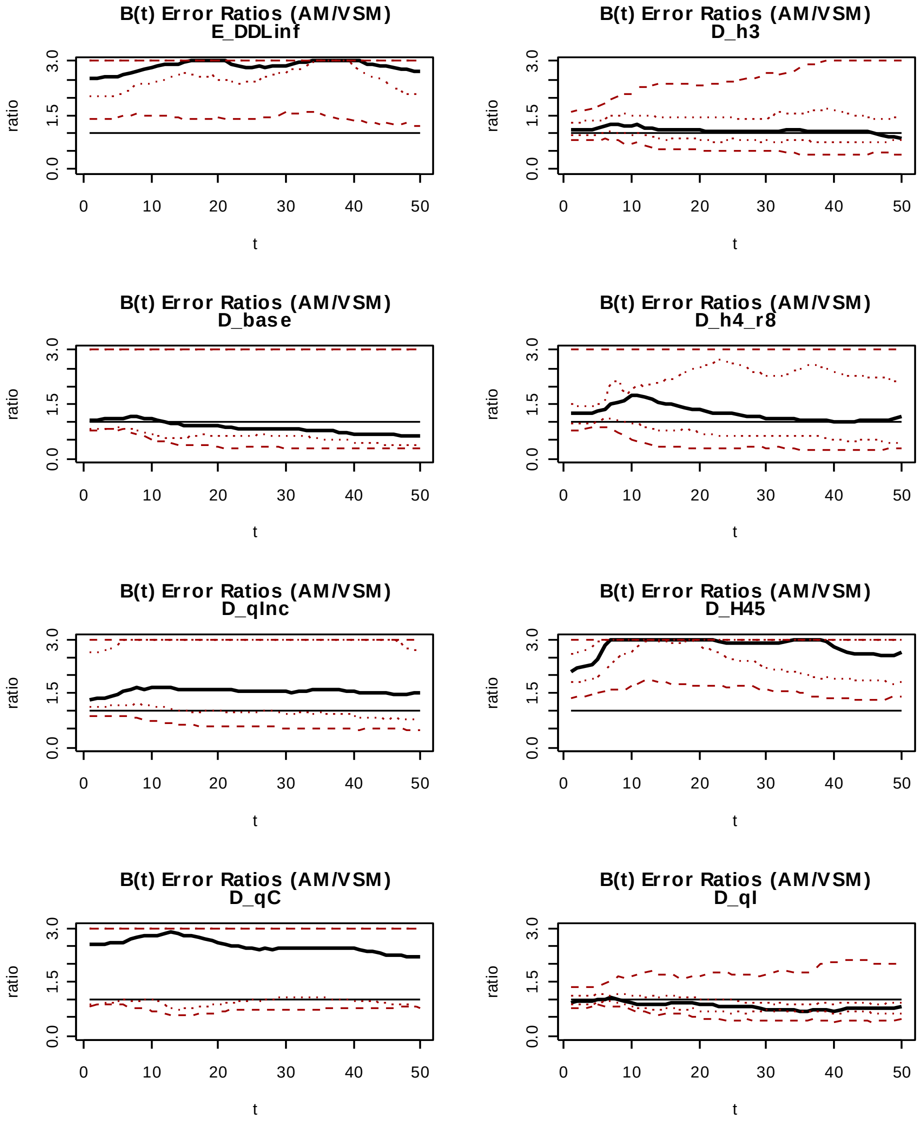
B(t) Error Ratios (AM/VSM) E_qC

B(t) Error Ratios (AM/VSM) E_qI

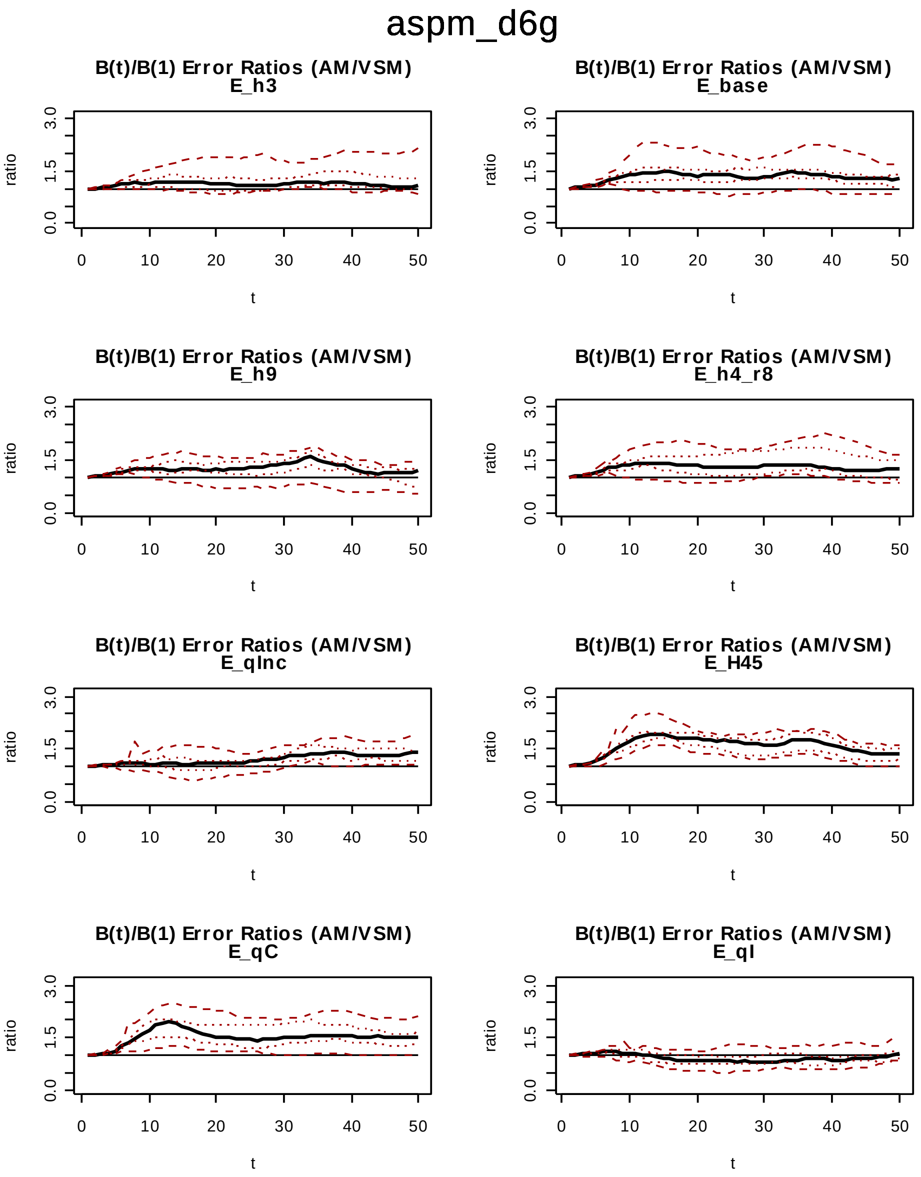
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0) E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.3-c

aspm_d6g



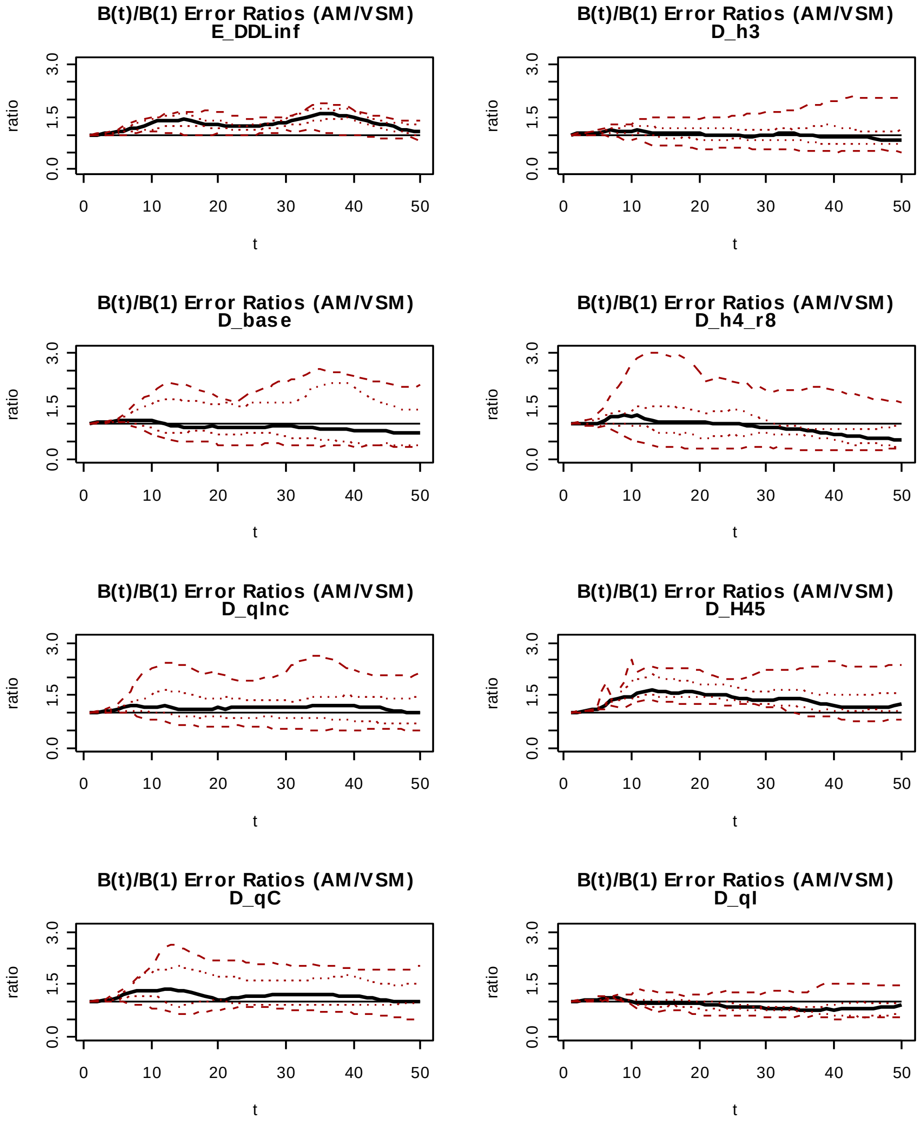
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)



Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.3-d

aspm_d6g



B(t)/B(1) Error Ratios (AM/VSM) E_DDLinf

B(t)/B(1) Error Ratios (AM/VSM) D_h3

B(t)/B(1) Error Ratios (AM/VSM) D_base

B(t)/B(1) Error Ratios (AM/VSM) D_h4_r8

B(t)/B(1) Error Ratios (AM/VSM) D_qInc

B(t)/B(1) Error Ratios (AM/VSM) D_H45

B(t)/B(1) Error Ratios (AM/VSM) D_qC

B(t)/B(1) Error Ratios (AM/VSM) D_qI

Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0) E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

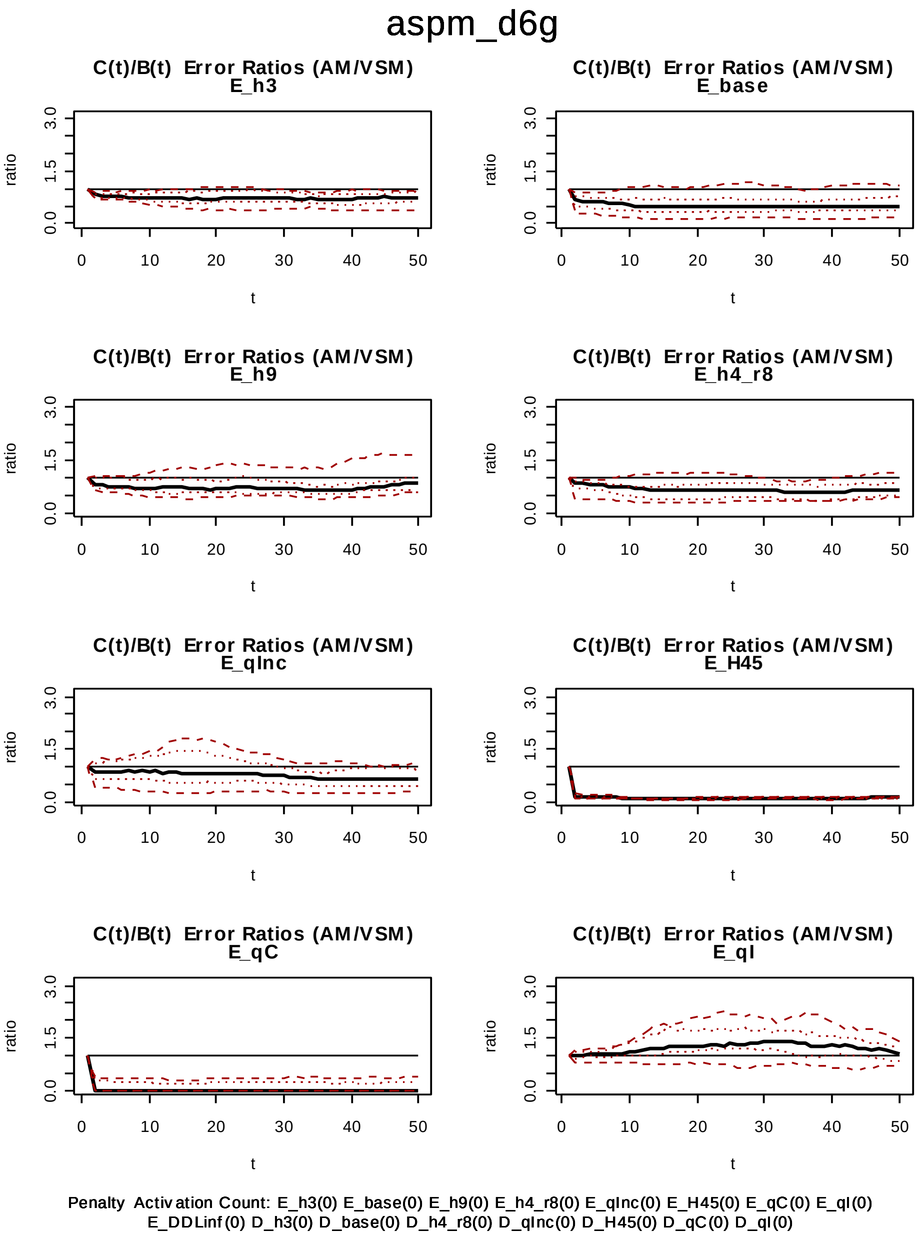
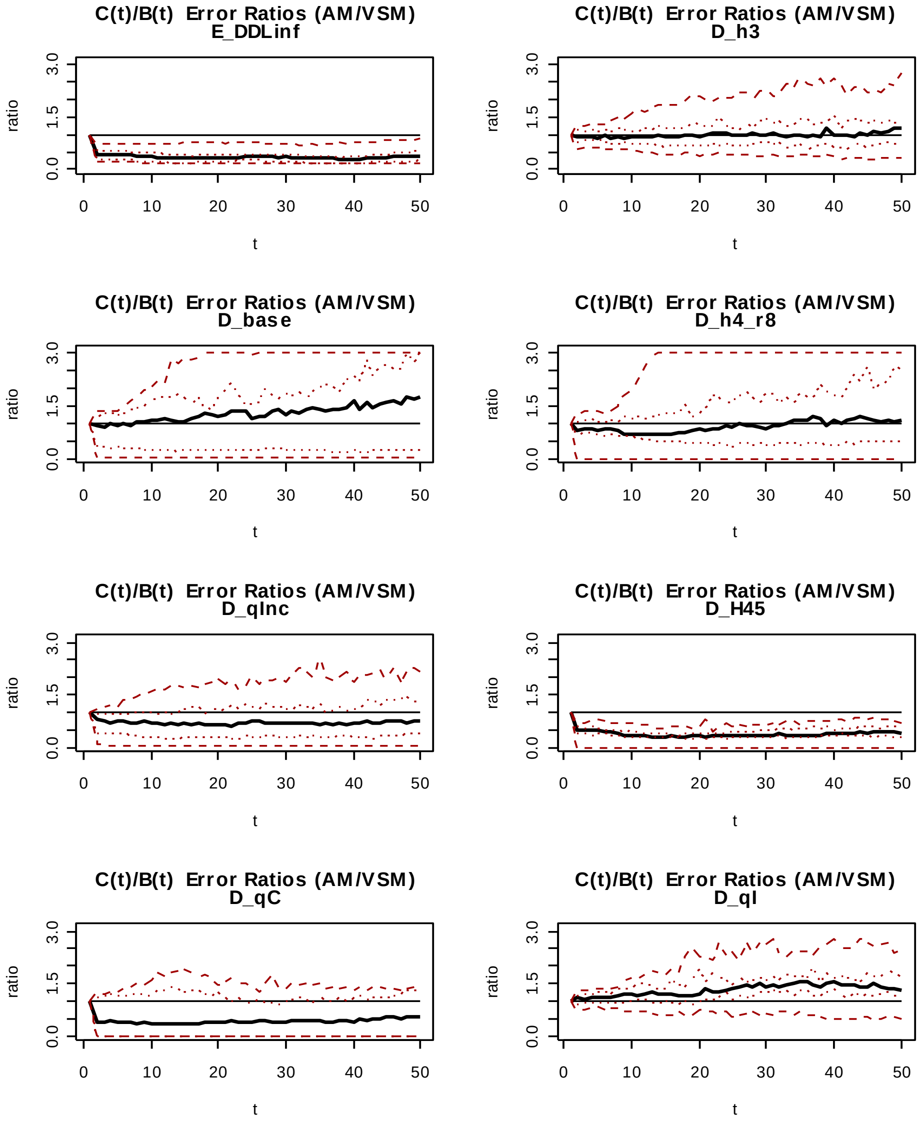
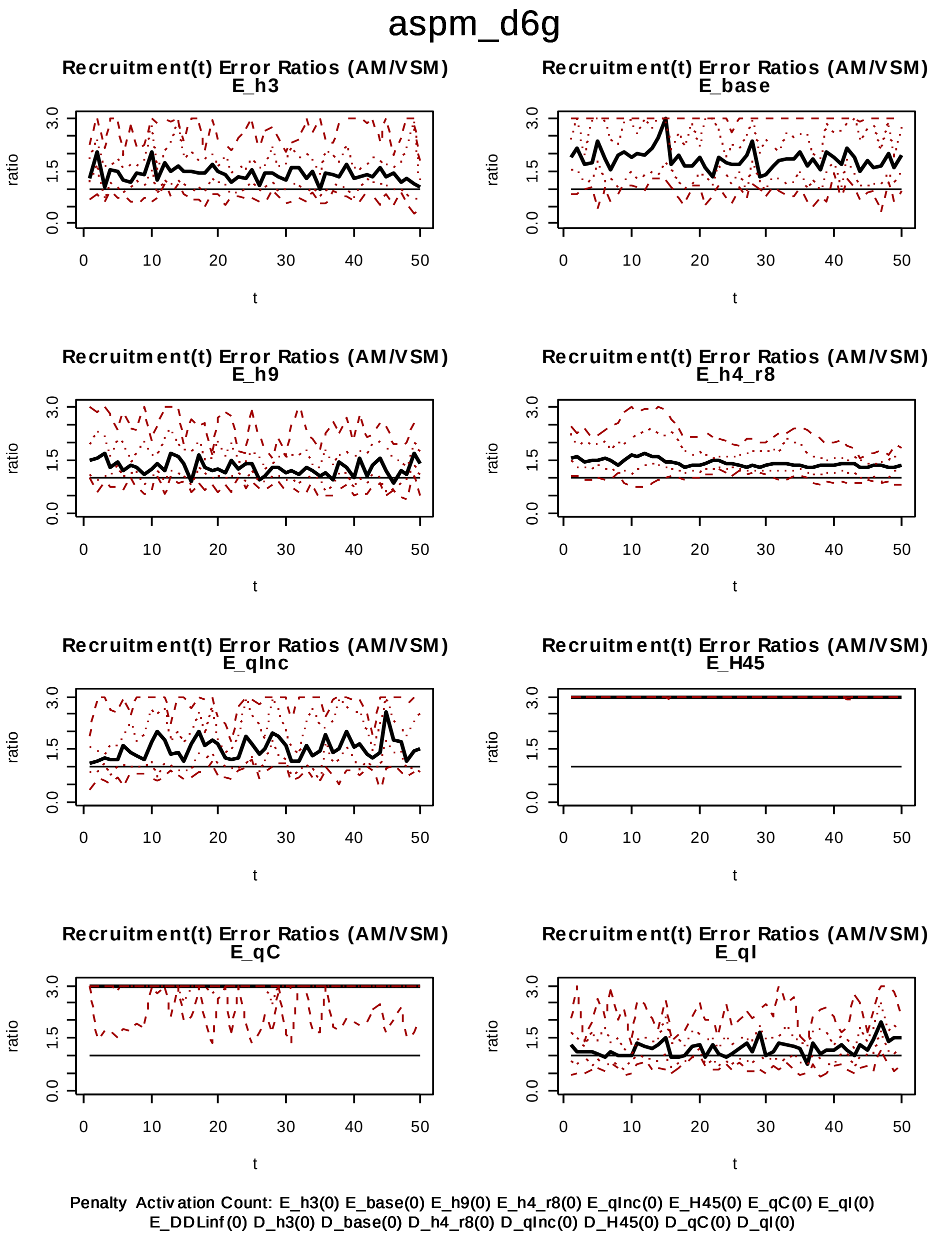

aspm_d6g
C(t)/B(t) Error Ratios (AM/VSM)
E_h3
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h9
C(t)/B(t) Error Ratios (AM/VSM)
E_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_qC
C(t)/B(t) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_ql(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_ql(0)

Figure A 6.3-e

aspm_d6g



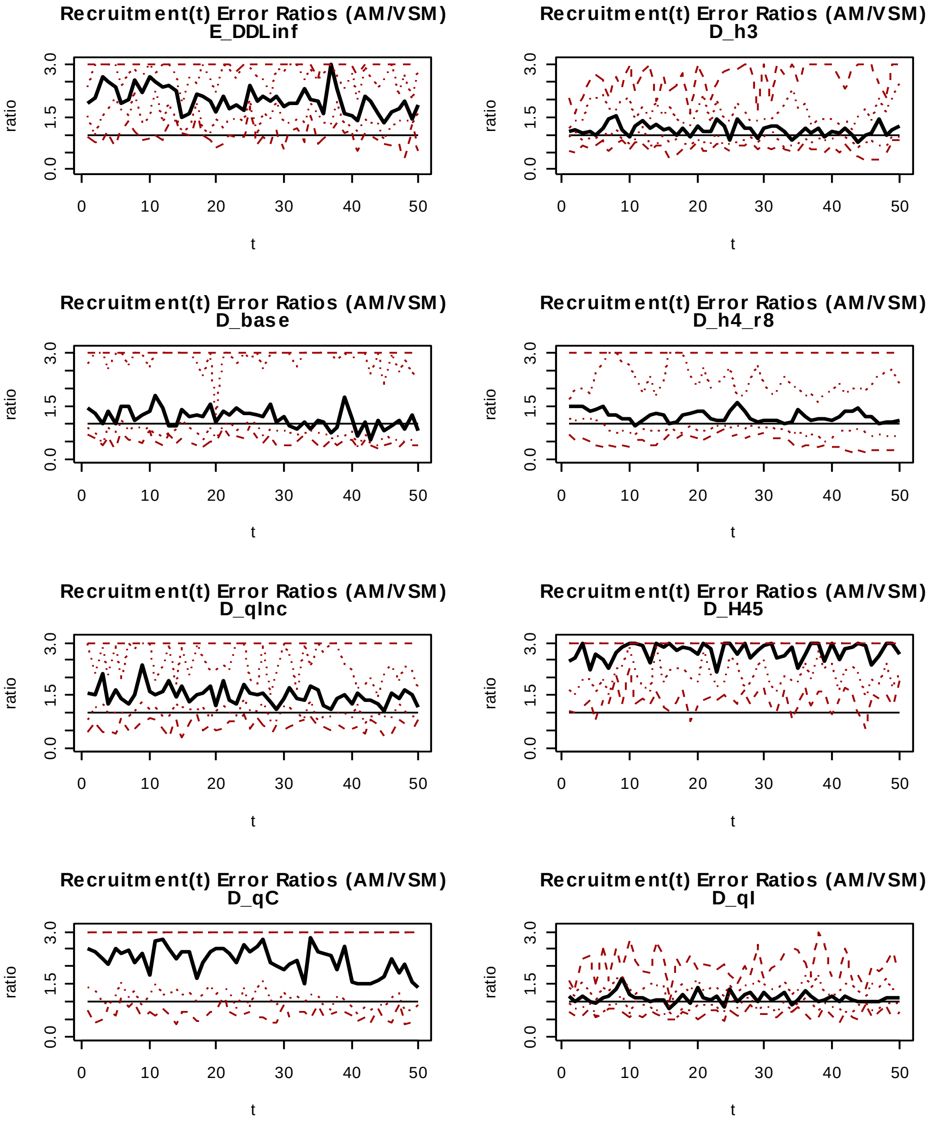
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)



Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.3-f

aspm_d6g



Recruitment(t) Error Ratios (AM/VSM) E_DDLinf

Recruitment(t) Error Ratios (AM/VSM) D_h3

Recruitment(t) Error Ratios (AM/VSM) D_base

Recruitment(t) Error Ratios (AM/VSM) D_h4_r8

Recruitment(t) Error Ratios (AM/VSM) D_qInc

Recruitment(t) Error Ratios (AM/VSM) D_H45

Recruitment(t) Error Ratios (AM/VSM) D_qC

Recruitment(t) Error Ratios (AM/VSM) D_qI

Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

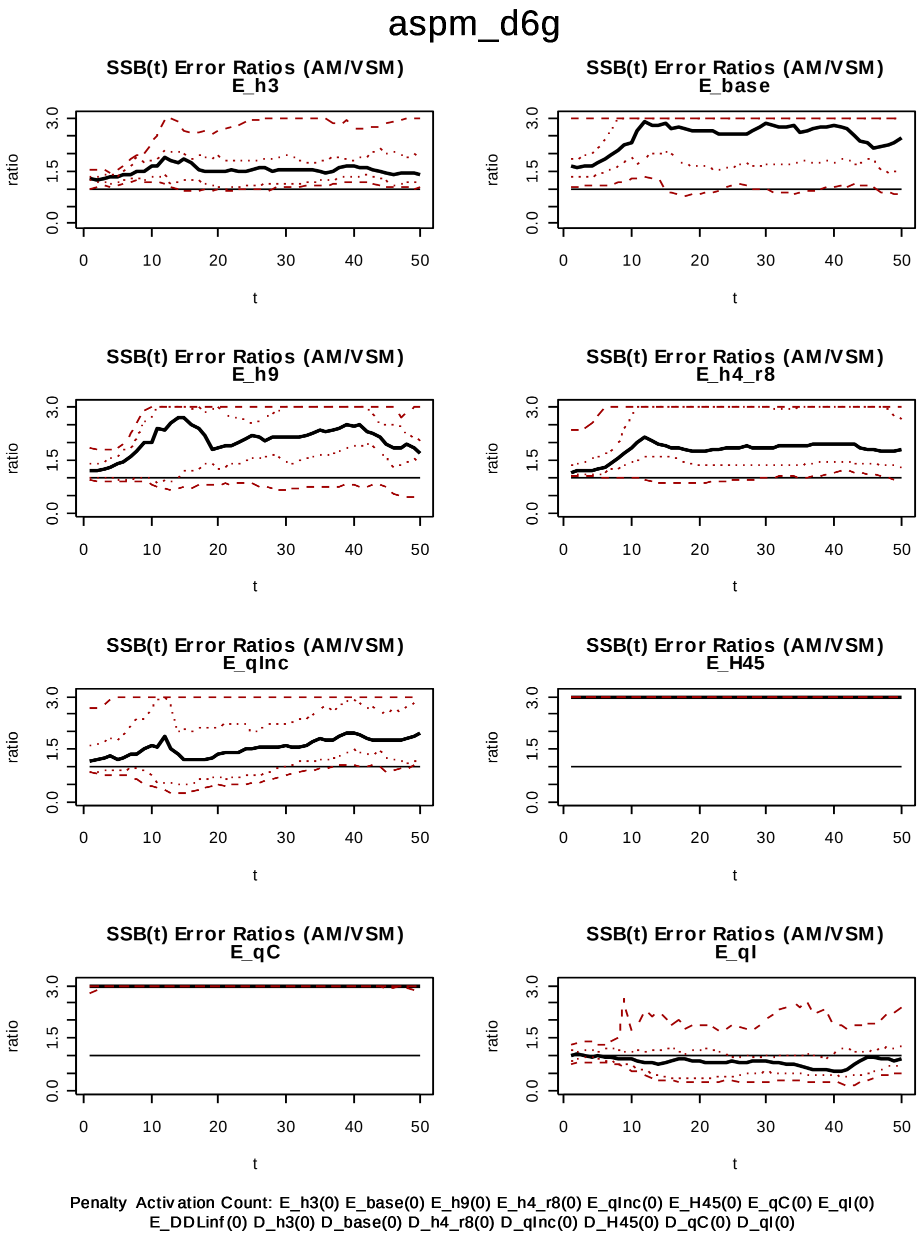

aspm_d6g
SSB(t) Error Ratios (AM/VSM)
E_h3
SSB(t) Error Ratios (AM/VSM)
E_base
SSB(t) Error Ratios (AM/VSM)
E_h9
SSB(t) Error Ratios (AM/VSM)
E_h4_r8
SSB(t) Error Ratios (AM/VSM)
E_qInc
SSB(t) Error Ratios (AM/VSM)
E_H45
SSB(t) Error Ratios (AM/VSM)
E_qC
SSB(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50
Penalty Activation Count: E_h3(0) E_base(0) E_h9(0) E_h4_r8(0) E_qInc(0) E_H45(0) E_qC(0) E_qI(0)
E_DDLinf(0) D_h3(0) D_base(0) D_h4_r8(0) D_qInc(0) D_H45(0) D_qC(0) D_qI(0)

Figure A 6.3-g

A 6.4 SC_BASE

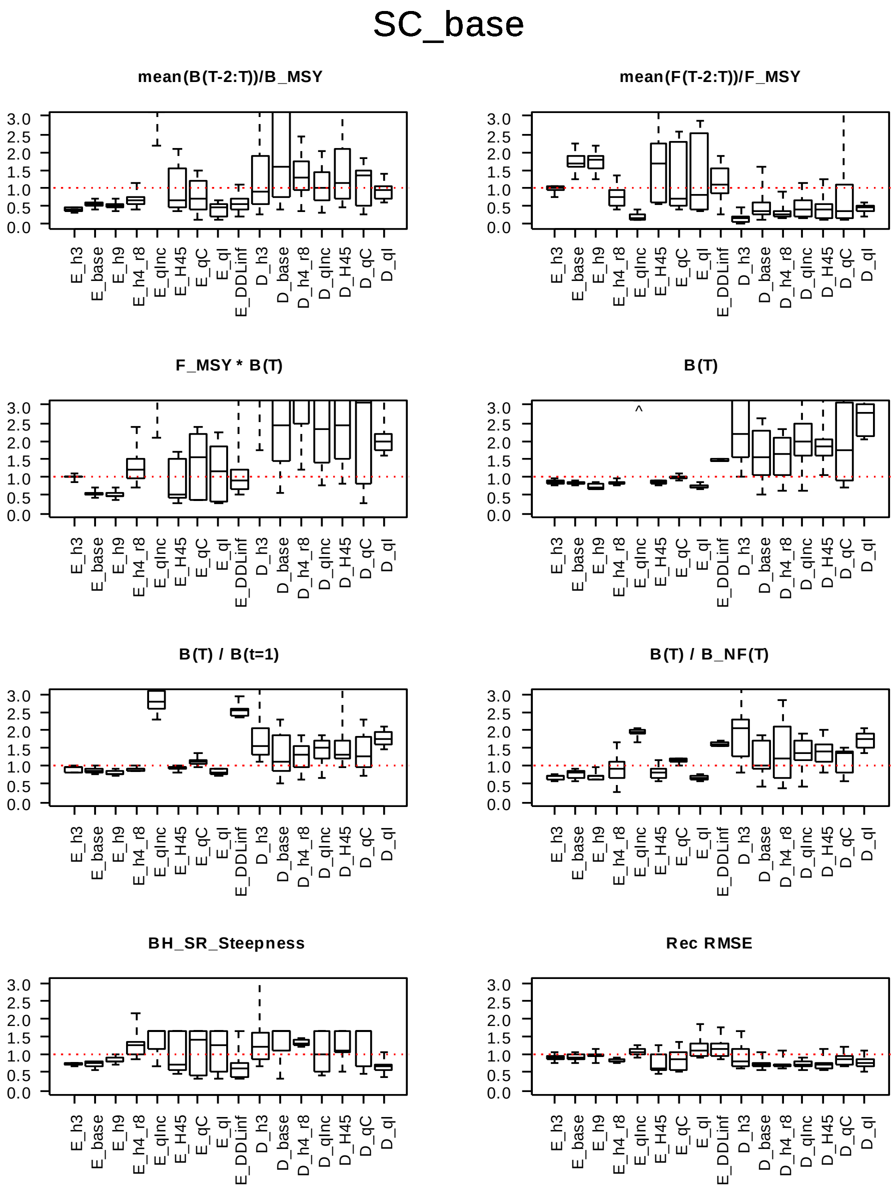


Figure A 6.4-a

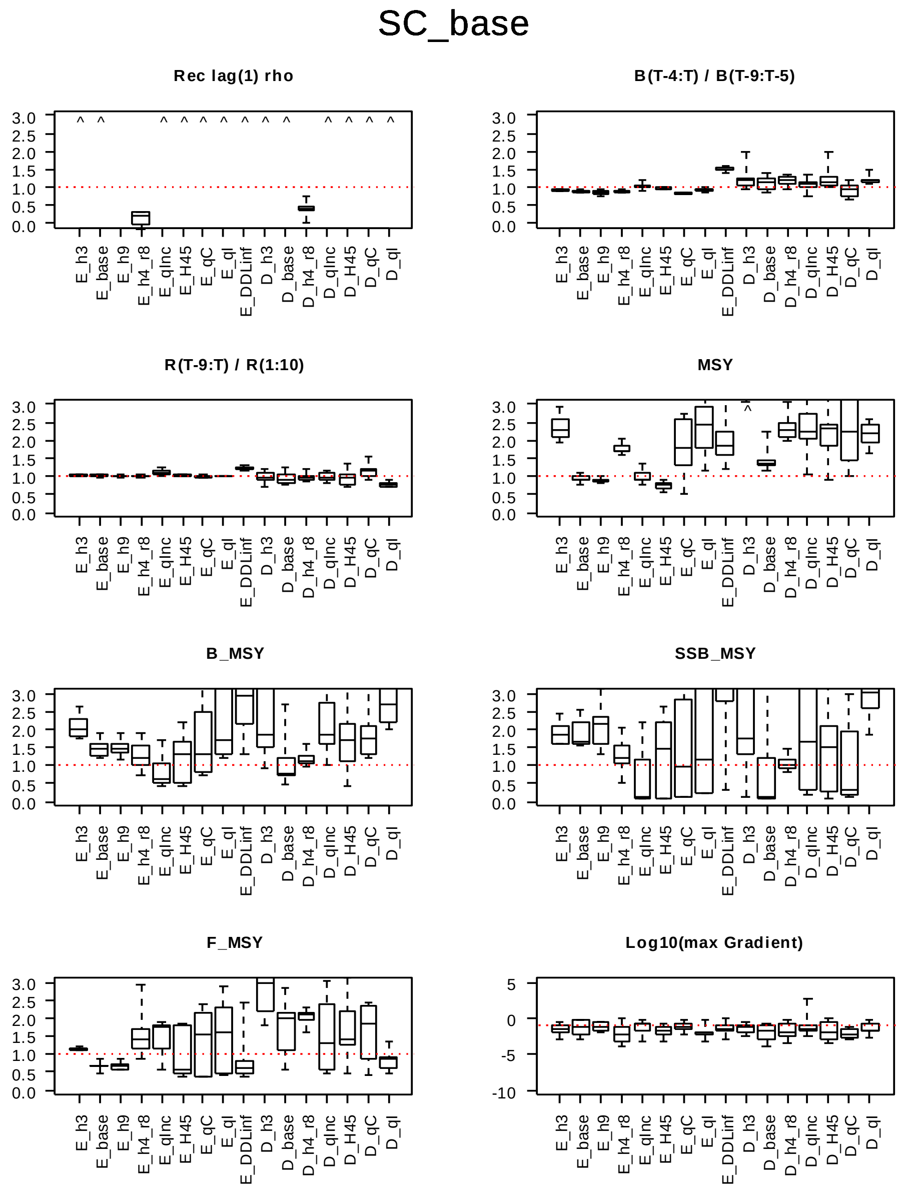
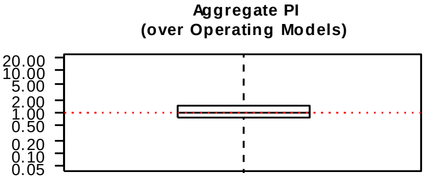
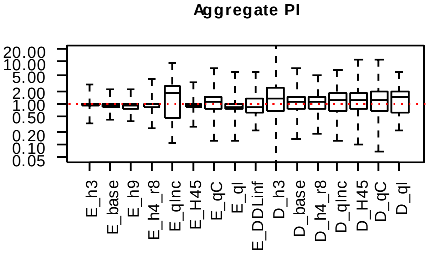

SC_base
Rec lag(1) rho
B(T-4:T) / B(T-9:T-5)
R(T-9:T) / R(1:10)
MSY
B_MSY
SSB_MSY
F_MSY
Log10(max Gradient)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5
0
-5
-10
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_qI
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_qI

Figure A 6.4-b

SC_base


Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05


Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_qI
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_qI

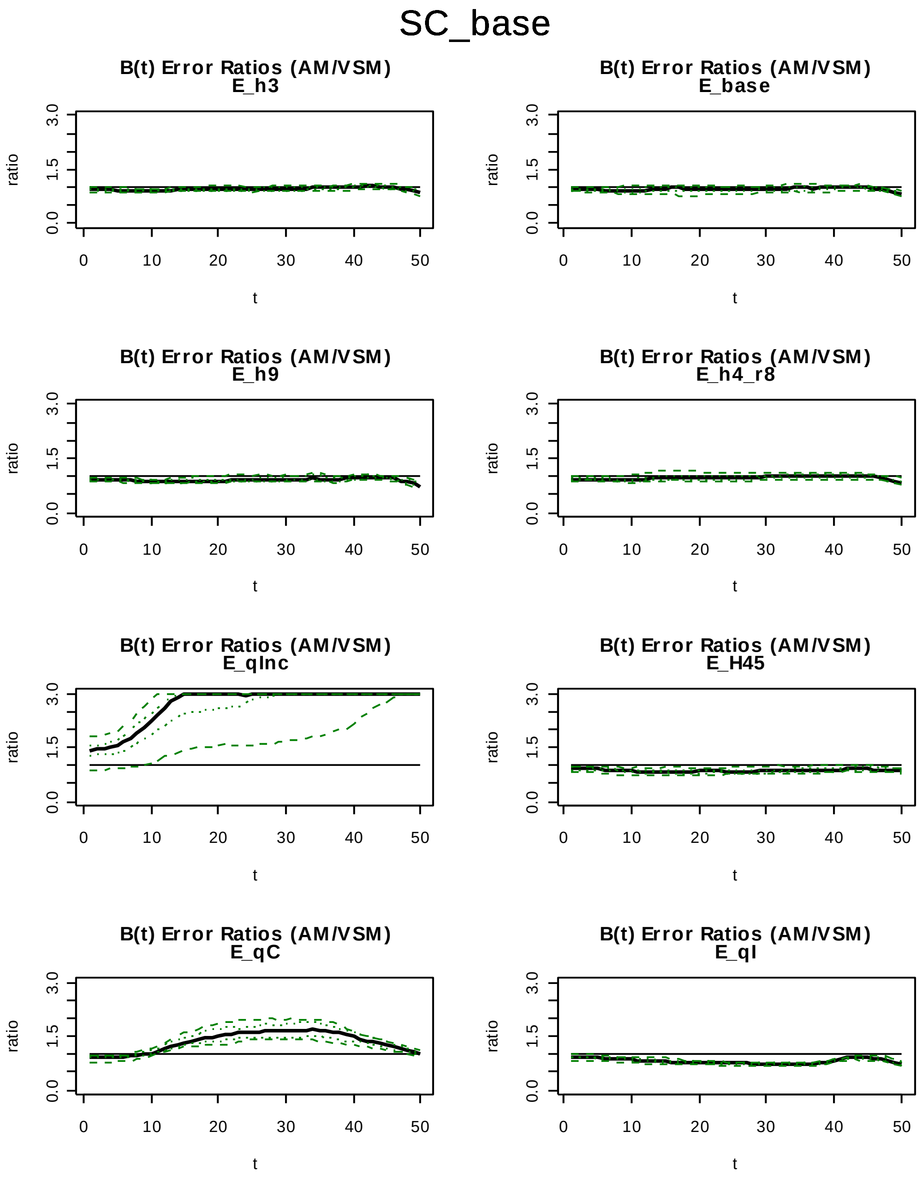
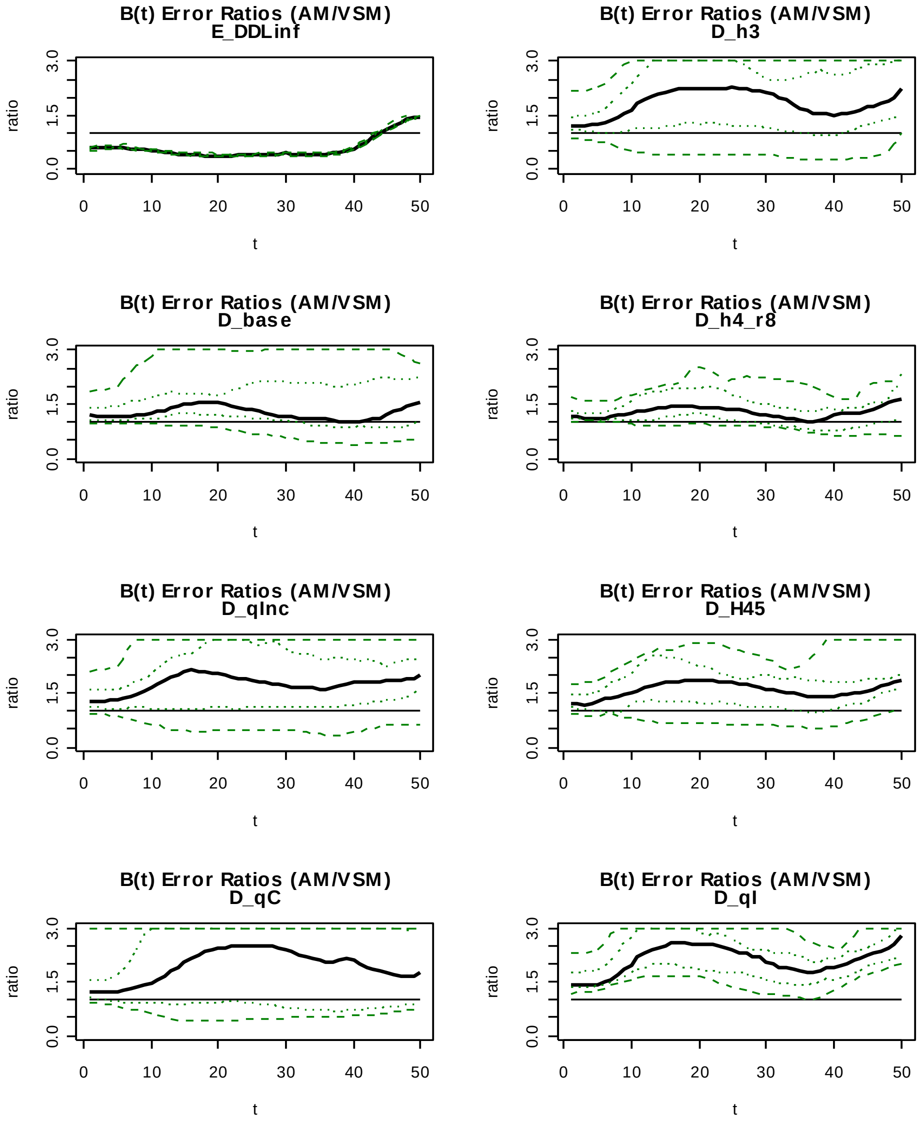

SC_base
B(t) Error Ratios (AM/VSM)
E_h3
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_h9
B(t) Error Ratios (AM/VSM)
E_h4_r8
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_H45
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Figure A 6.4-c

SC_base


B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t) Error Ratios (AM/VSM)
D_h3
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_h4_r8
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_H45
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

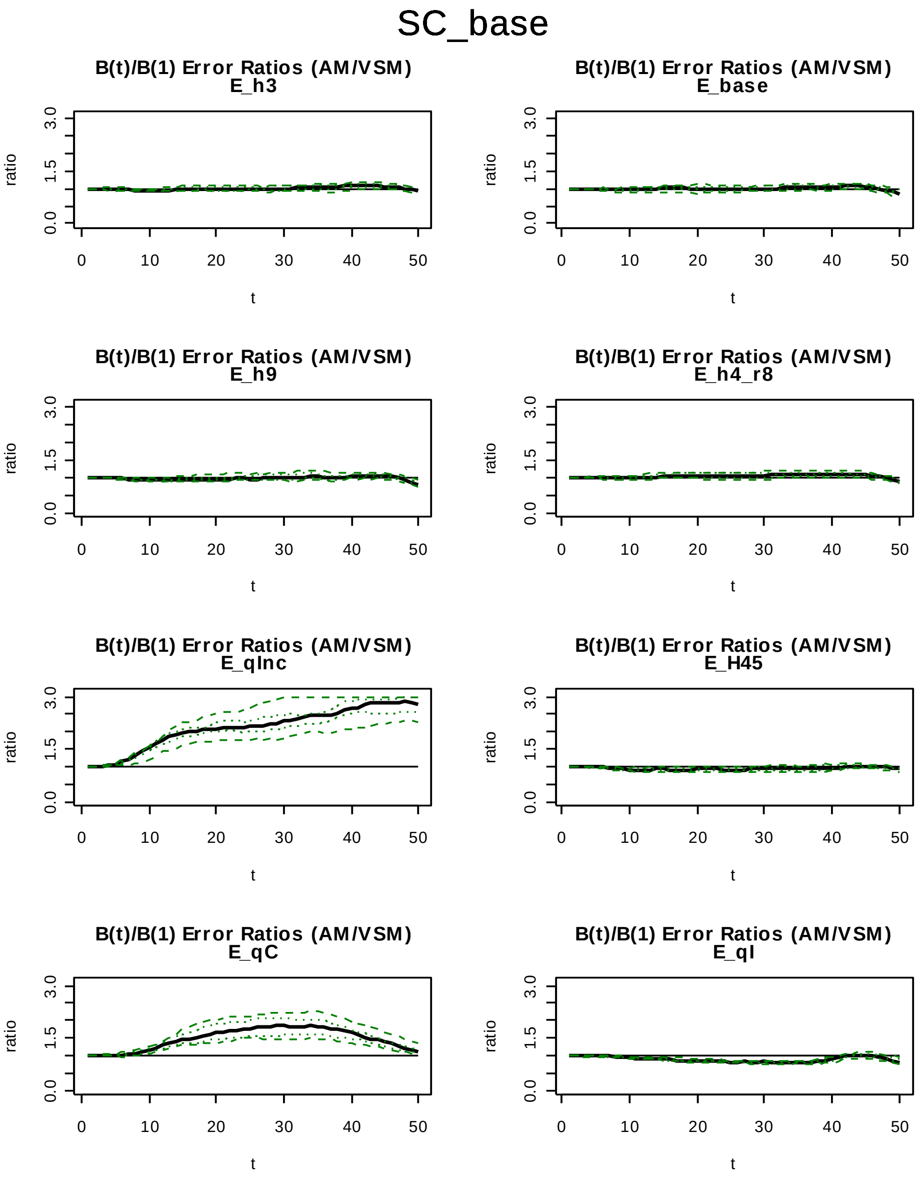
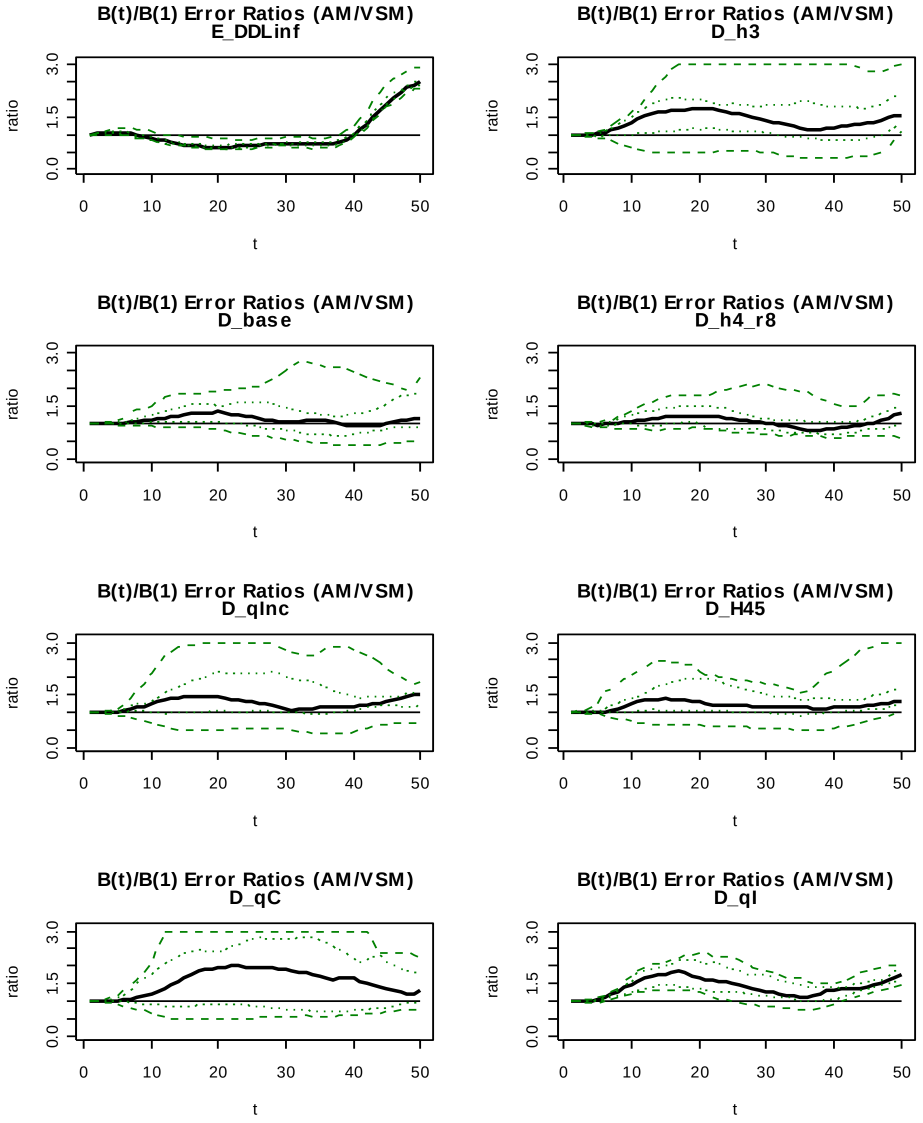

SC_base
B(t)/B(1) Error Ratios (AM/VSM)
E_h3
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_h9
B(t)/B(1) Error Ratios (AM/VSM)
E_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_qC
B(t)/B(1) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.4-d

SC_base


B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
D_h3
B(t)/B(1) Error Ratios (AM/VSM)
D_base
B(t)/B(1) Error Ratios (AM/VSM)
D_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
D_qInc
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
B(t)/B(1) Error Ratios (AM/VSM)
D_qC
B(t)/B(1) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

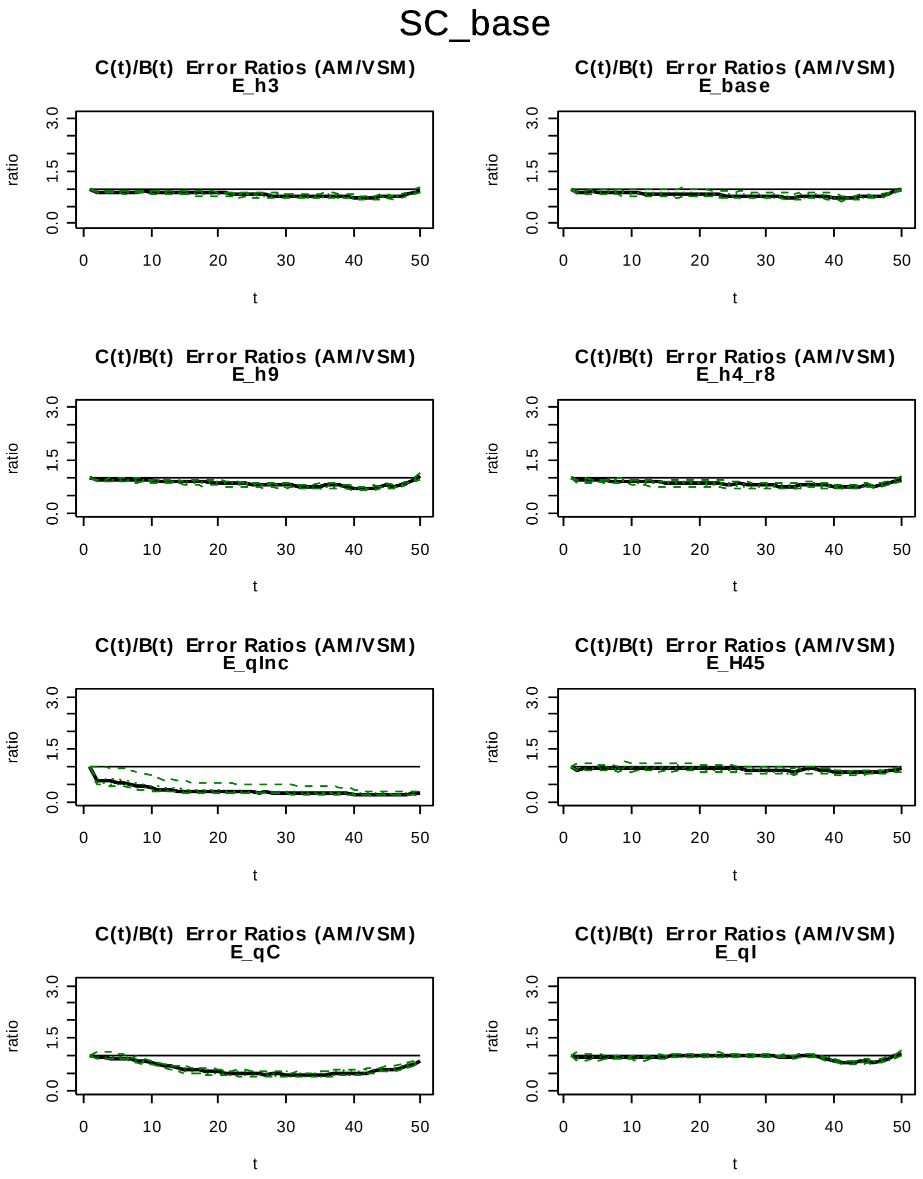
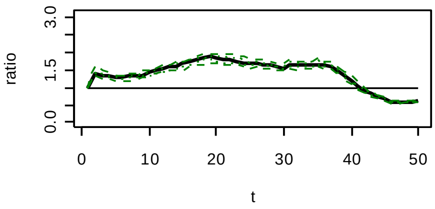

SC_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h3
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h9
C(t)/B(t) Error Ratios (AM/VSM)
E_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_qC
C(t)/B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

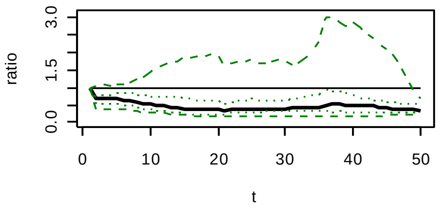
Figure A 6.4-e

SC_base

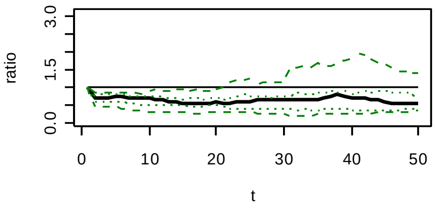
C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf



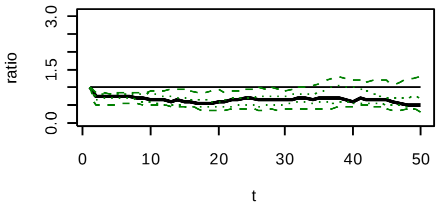
C(t)/B(t) Error Ratios (AM/VSM)
D_h3



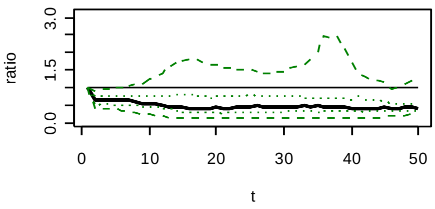
C(t)/B(t) Error Ratios (AM/VSM)
D_base



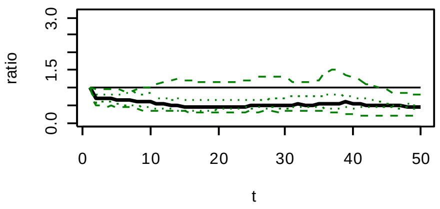
C(t)/B(t) Error Ratios (AM/VSM)
D_h4_r8



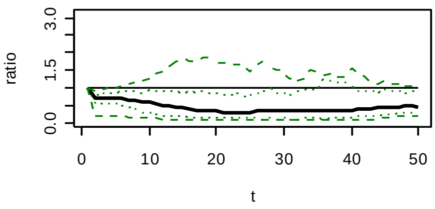
C(t)/B(t) Error Ratios (AM/VSM)
D_qInc



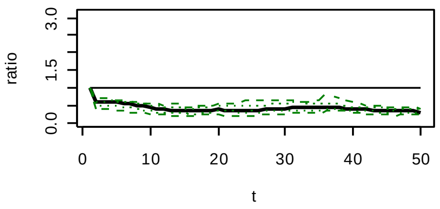
C(t)/B(t) Error Ratios (AM/VSM)
D_H45



C(t)/B(t) Error Ratios (AM/VSM)
D_qC



C(t)/B(t) Error Ratios (AM/VSM)
D_qI



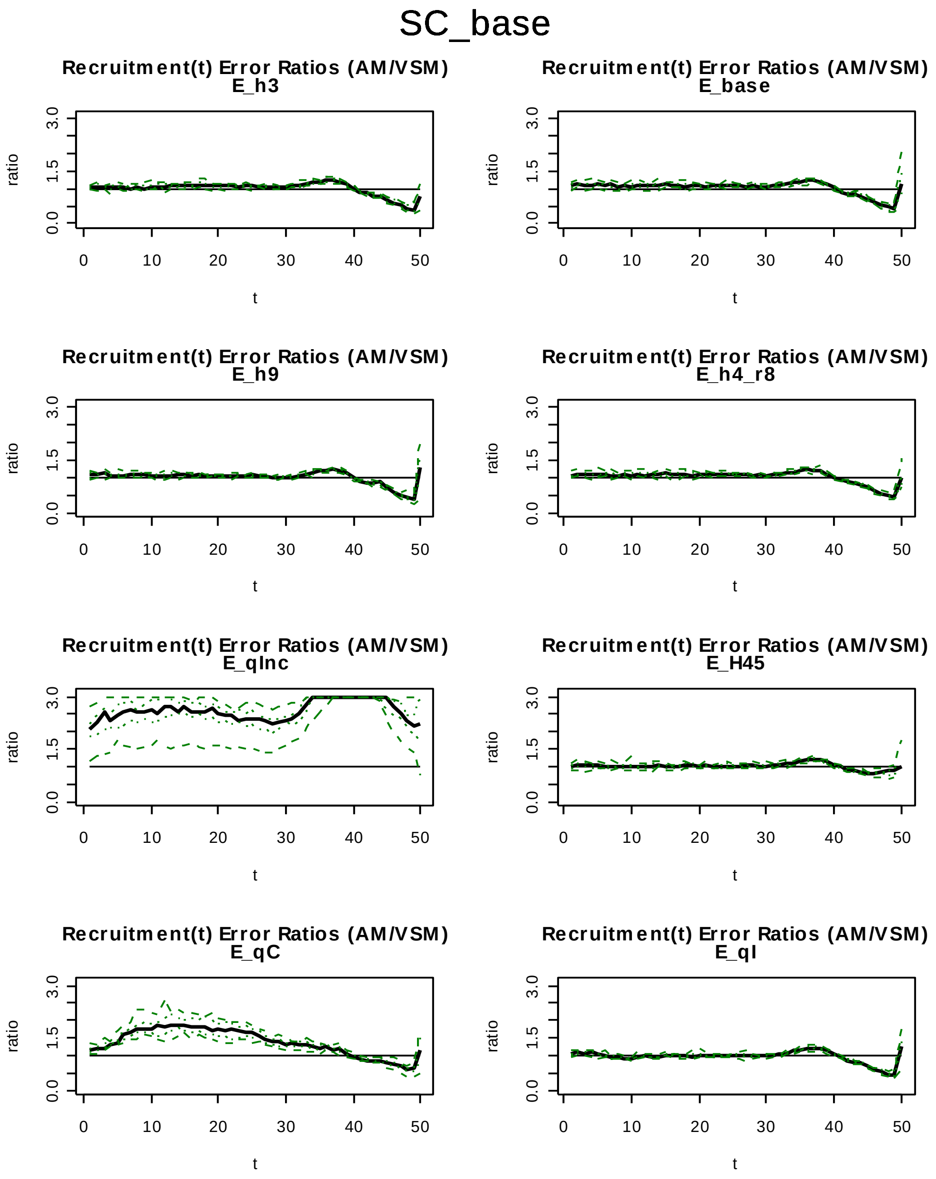
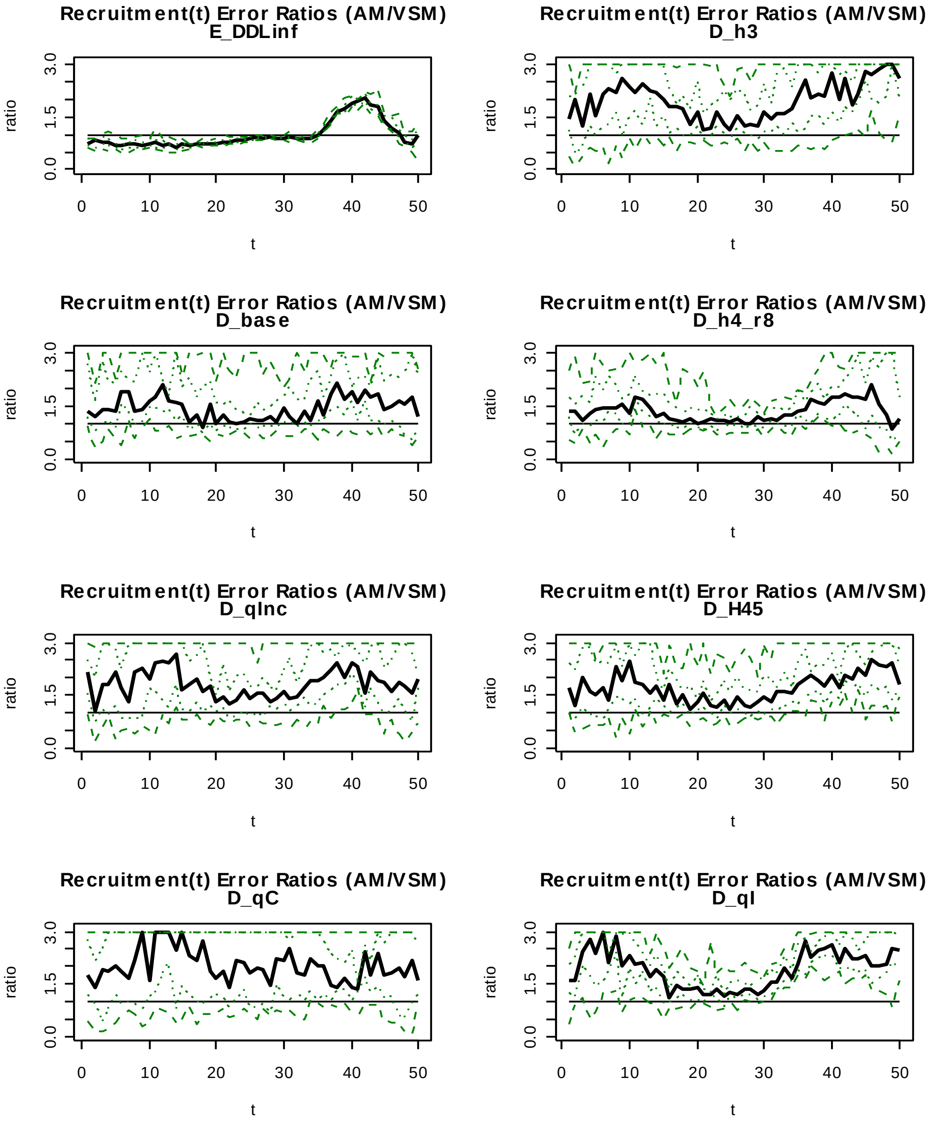

SC_base
Recruitment(t) Error Ratios (AM/VSM)
E_h3
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_h9
Recruitment(t) Error Ratios (AM/VSM)
E_h4_r8
Recruitment(t) Error Ratios (AM/VSM)
E_qInc
Recruitment(t) Error Ratios (AM/VSM)
E_H45
Recruitment(t) Error Ratios (AM/VSM)
E_qC
Recruitment(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.4-f

SC_base


Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf
Recruitment(t) Error Ratios (AM/VSM)
D_h3
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_h4_r8
Recruitment(t) Error Ratios (AM/VSM)
D_qInc
Recruitment(t) Error Ratios (AM/VSM)
D_H45
Recruitment(t) Error Ratios (AM/VSM)
D_qC
Recruitment(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

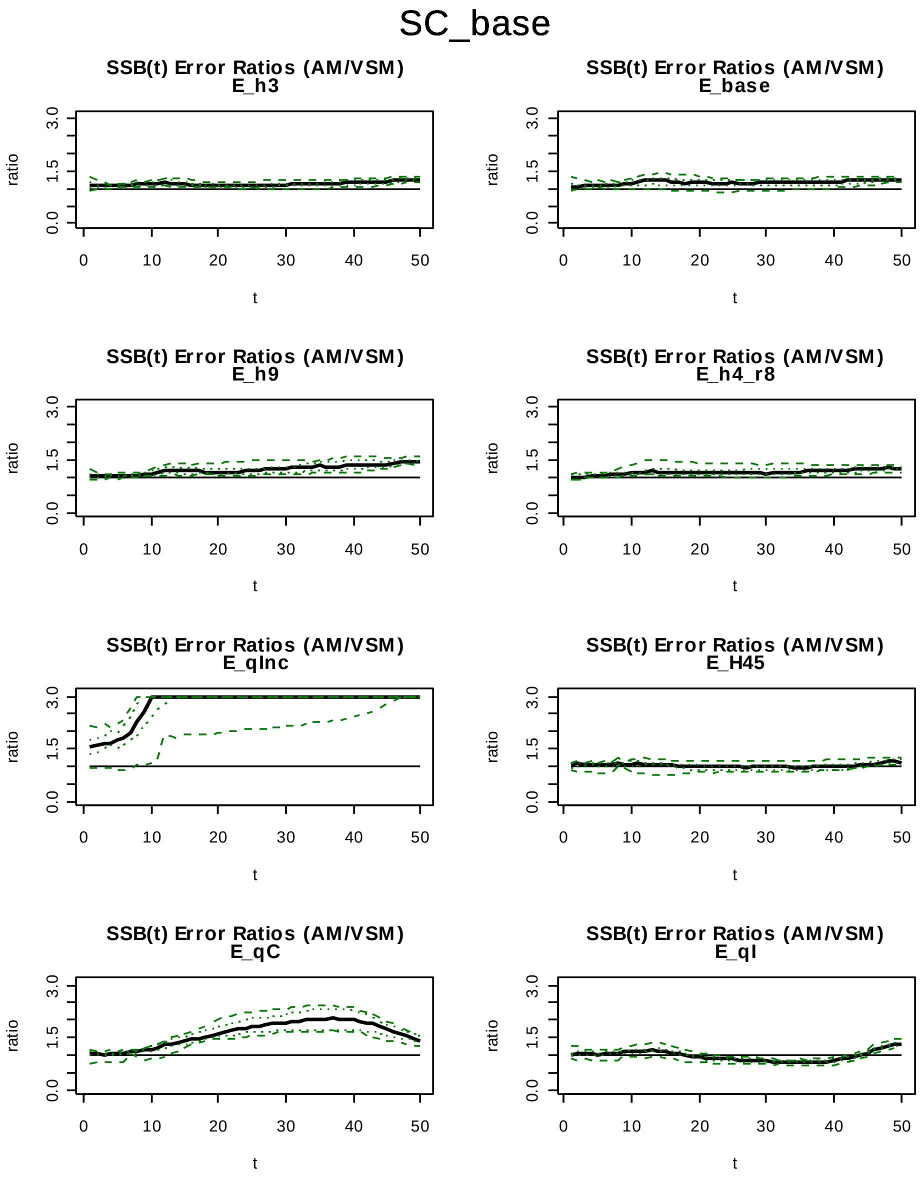

SC_base
SSB(t) Error Ratios (AM/VSM)
E_h3
SSB(t) Error Ratios (AM/VSM)
E_base
SSB(t) Error Ratios (AM/VSM)
E_h9
SSB(t) Error Ratios (AM/VSM)
E_h4_r8
SSB(t) Error Ratios (AM/VSM)
E_qInc
SSB(t) Error Ratios (AM/VSM)
E_H45
SSB(t) Error Ratios (AM/VSM)
E_qC
SSB(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.4-g

A 6.5 SC_Mest

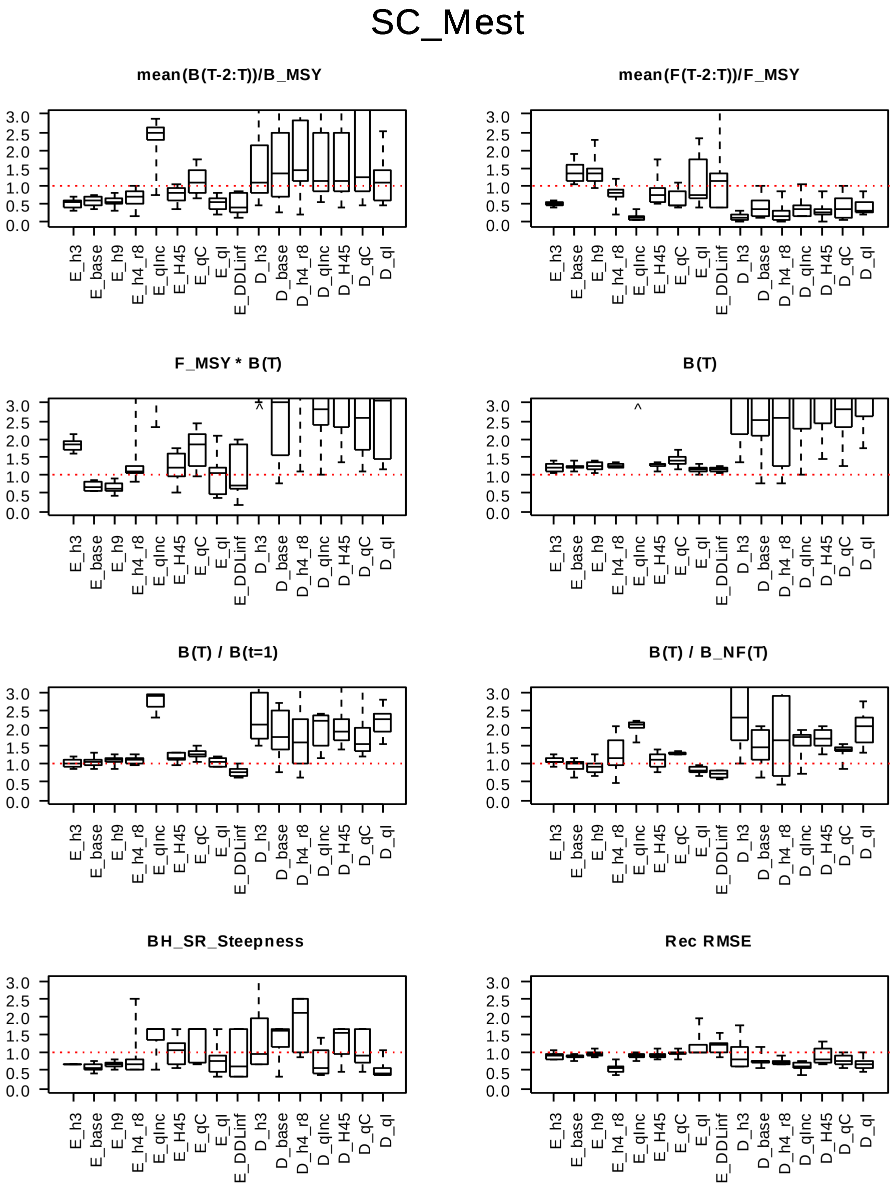


Figure A 6.5-a

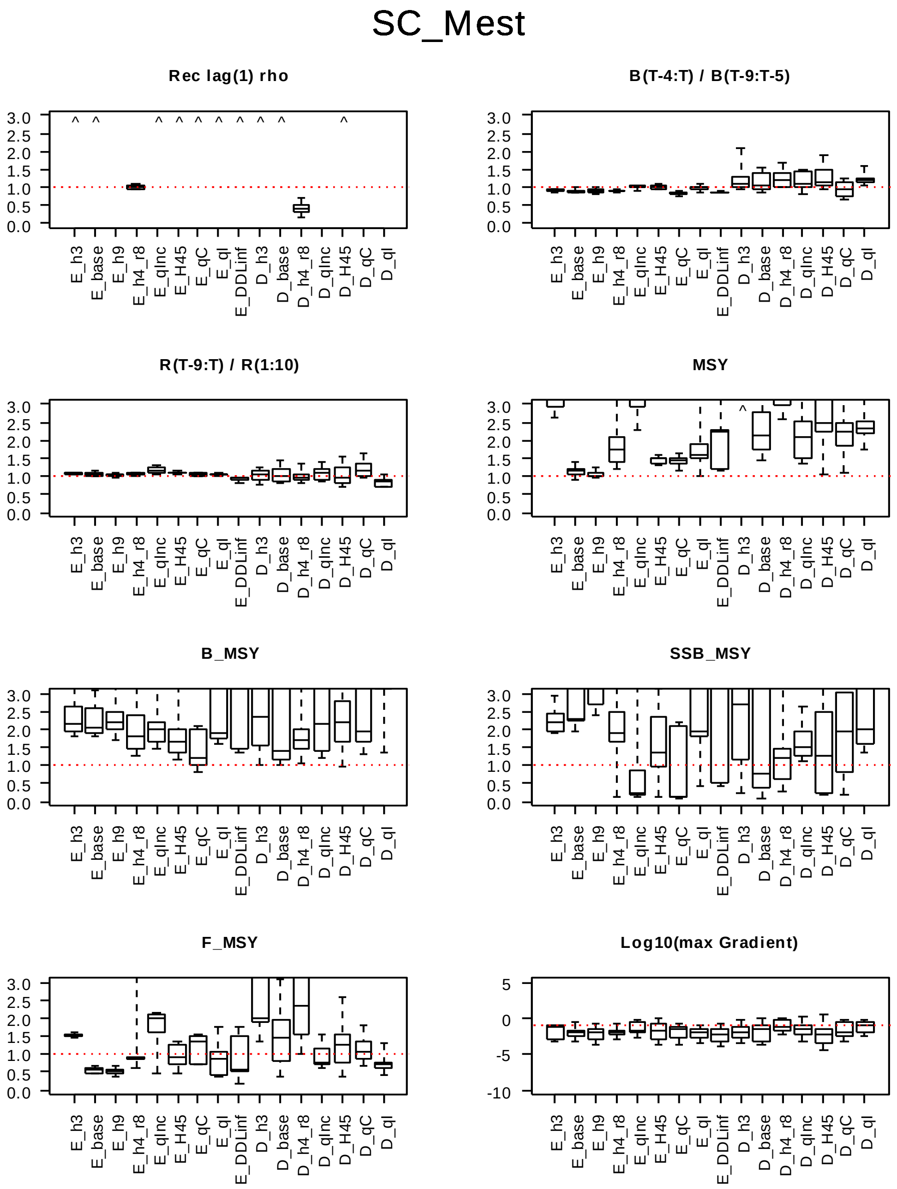
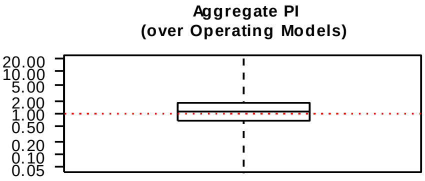
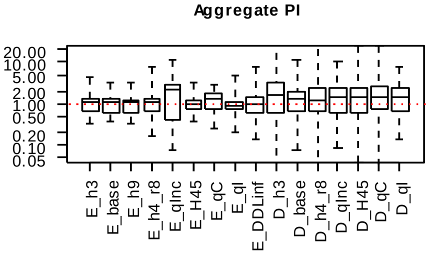

SC_Mest
Rec lag(1) rho
B(T-4:T) / B(T-9:T-5)
R(T-9:T) / R(1:10)
MSY
B_MSY
SSB_MSY
F_MSY
Log10(max Gradient)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5
0
-5
-10
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_ql
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_ql

Figure A 6.5-b

SC_Mest


Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05


Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_ql
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_ql

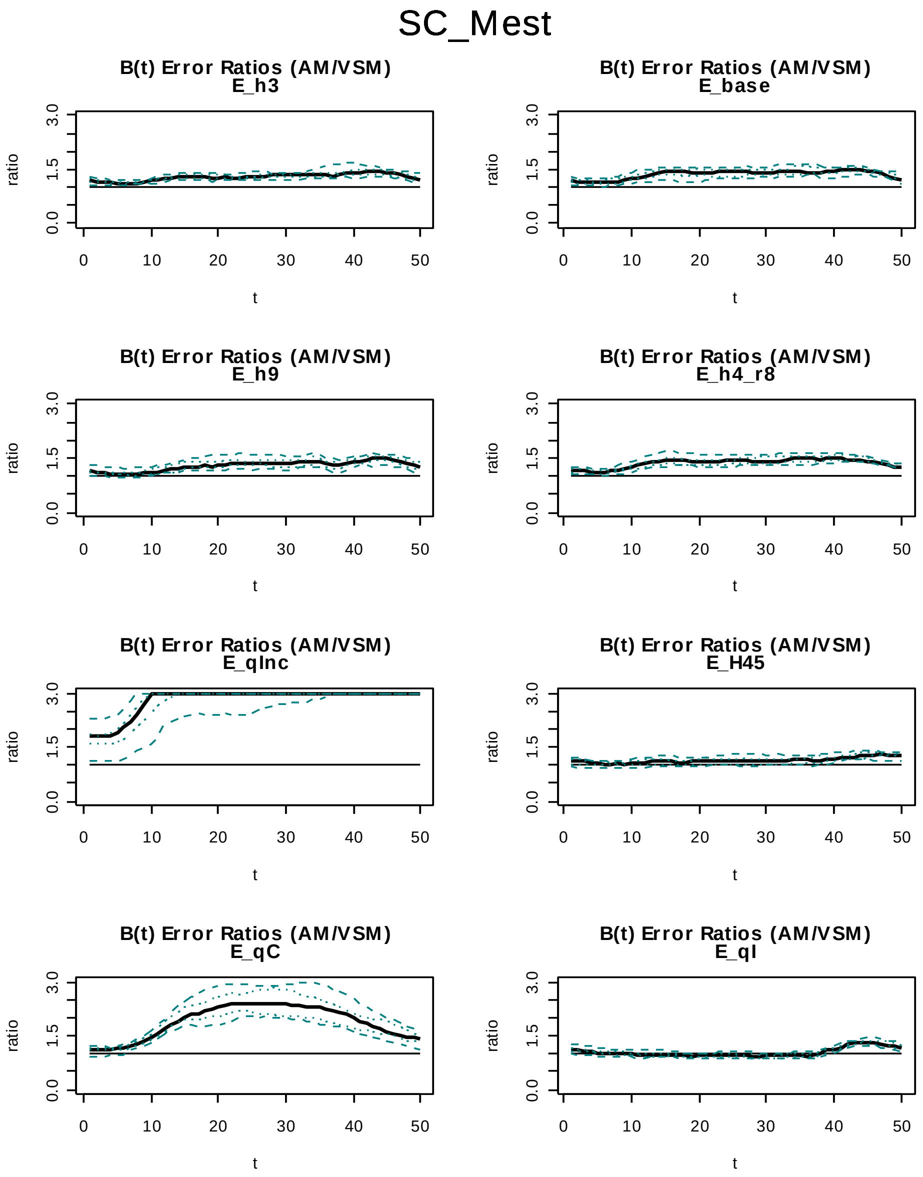
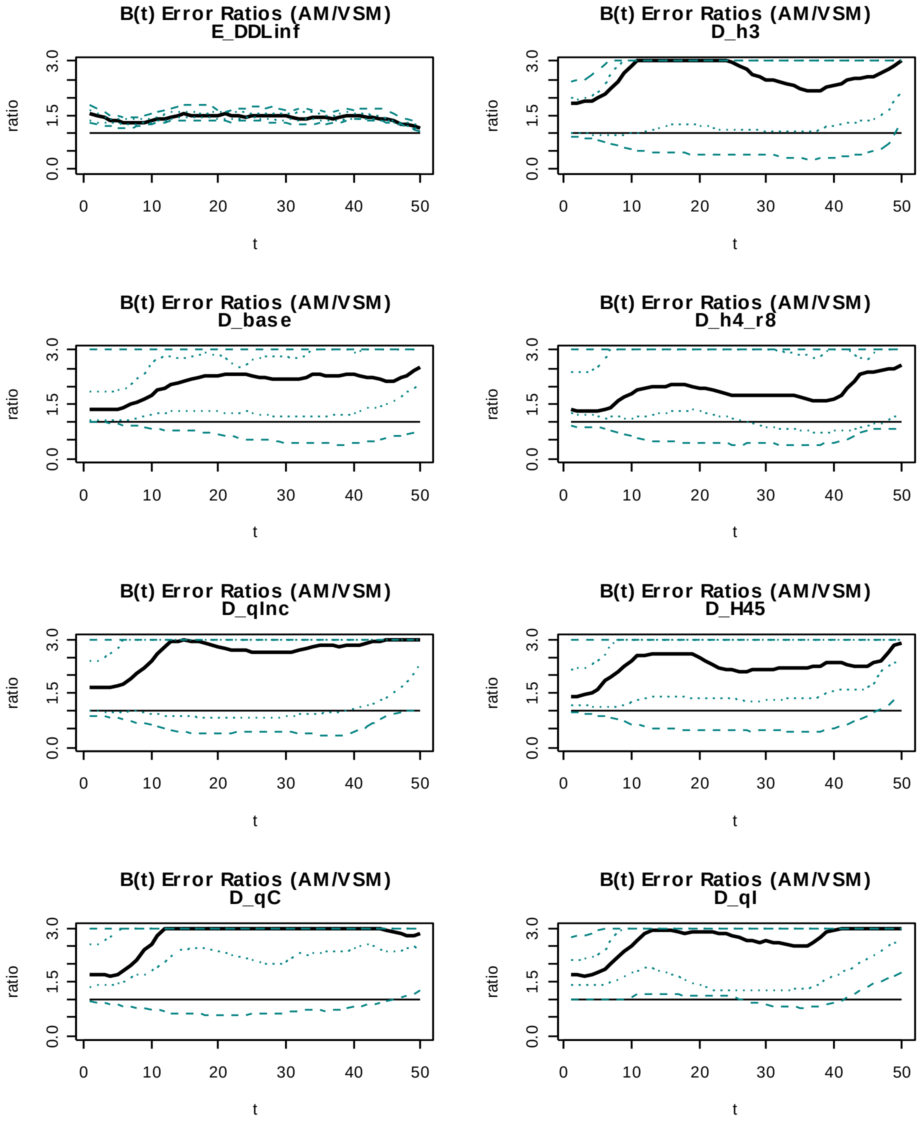

SC_Mest
B(t) Error Ratios (AM/VSM)
E_h3
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_h9
B(t) Error Ratios (AM/VSM)
E_h4_r8
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_H45
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.5-c

SC_Mest



B(t) Error Ratios (AM/VSM) E_DDLinf

B(t) Error Ratios (AM/VSM) D_h3

B(t) Error Ratios (AM/VSM) D_base

B(t) Error Ratios (AM/VSM) D_h4_r8

B(t) Error Ratios (AM/VSM) D_qInc

B(t) Error Ratios (AM/VSM) D_H45

B(t) Error Ratios (AM/VSM) D_qC

B(t) Error Ratios (AM/VSM) D_qI

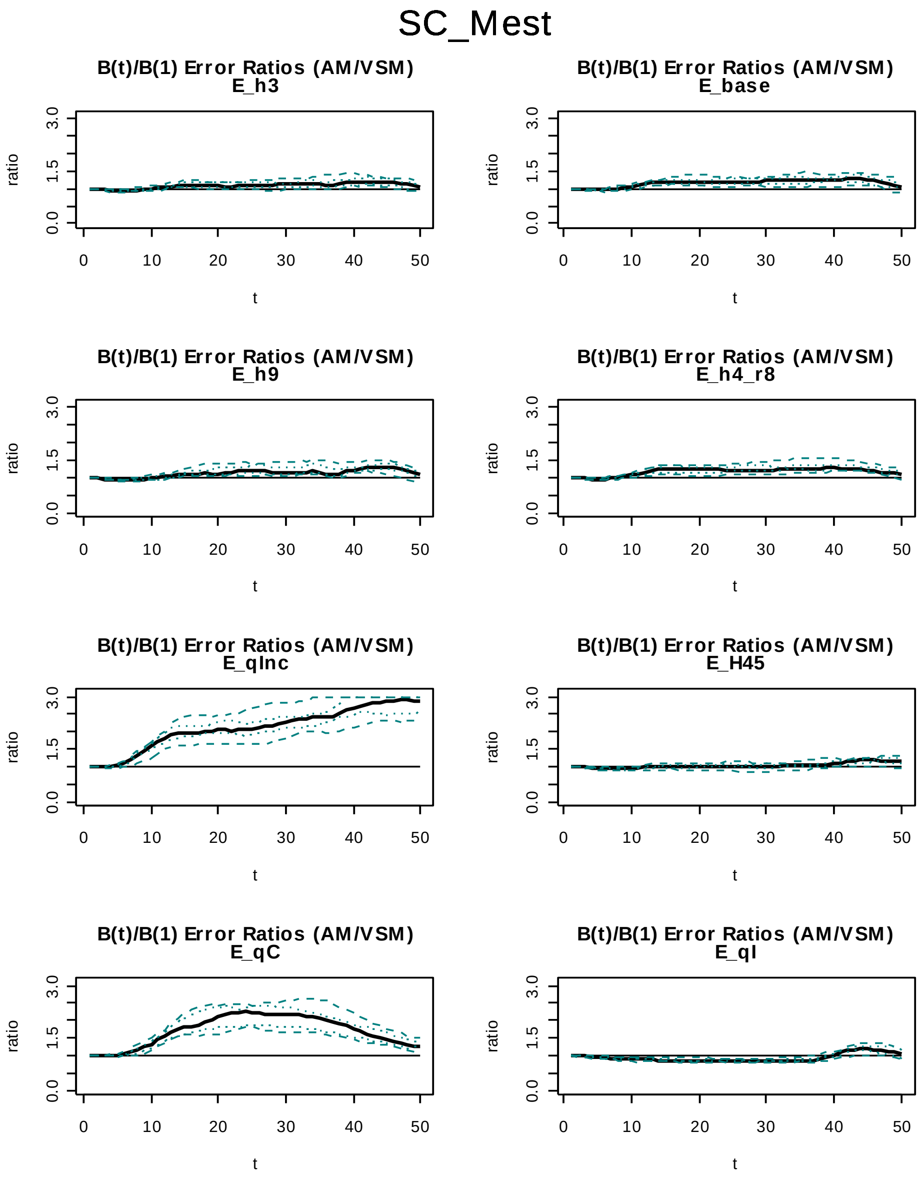
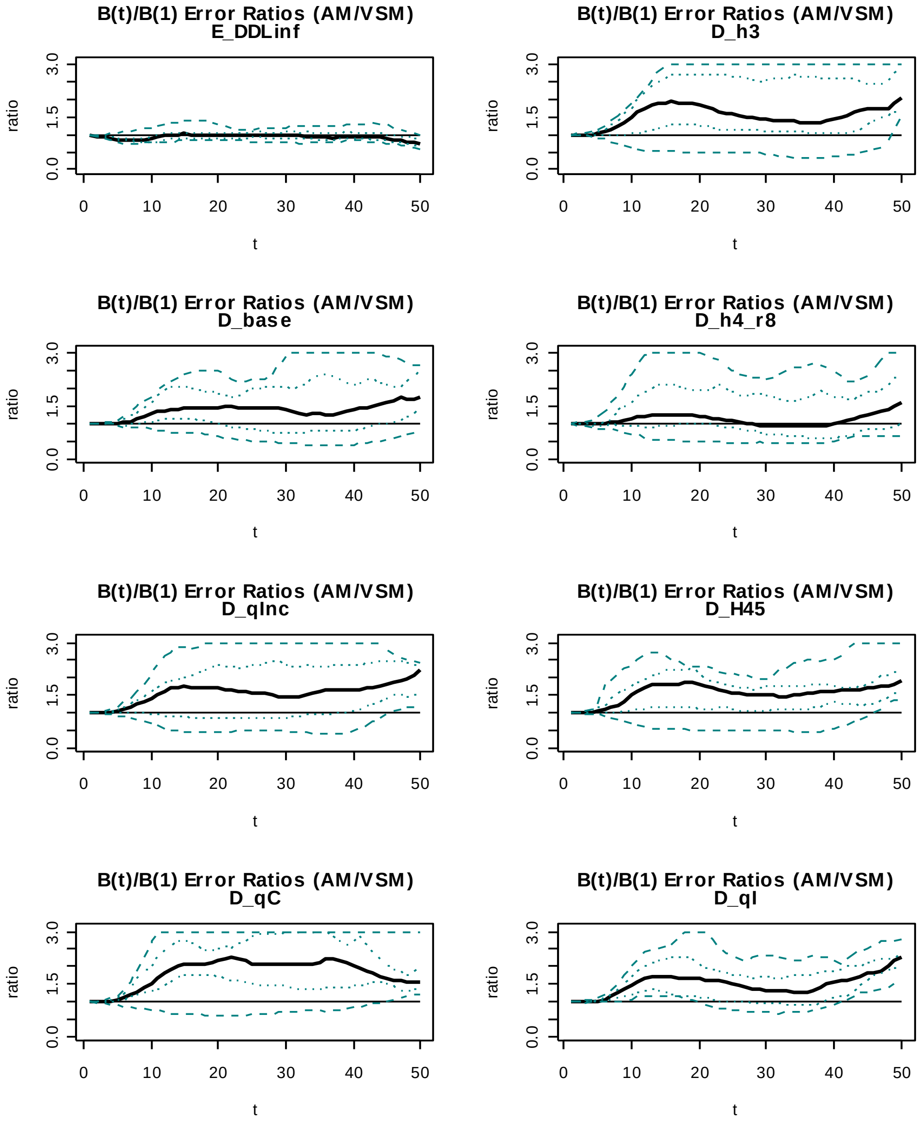

SC_Mest
B(t)/B(1) Error Ratios (AM/VSM)
E_h3
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_h9
B(t)/B(1) Error Ratios (AM/VSM)
E_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_qC
B(t)/B(1) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.5-d

SC_Mest


B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
D_h3
B(t)/B(1) Error Ratios (AM/VSM)
D_base
B(t)/B(1) Error Ratios (AM/VSM)
D_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
D_qInc
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
B(t)/B(1) Error Ratios (AM/VSM)
D_qC
B(t)/B(1) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

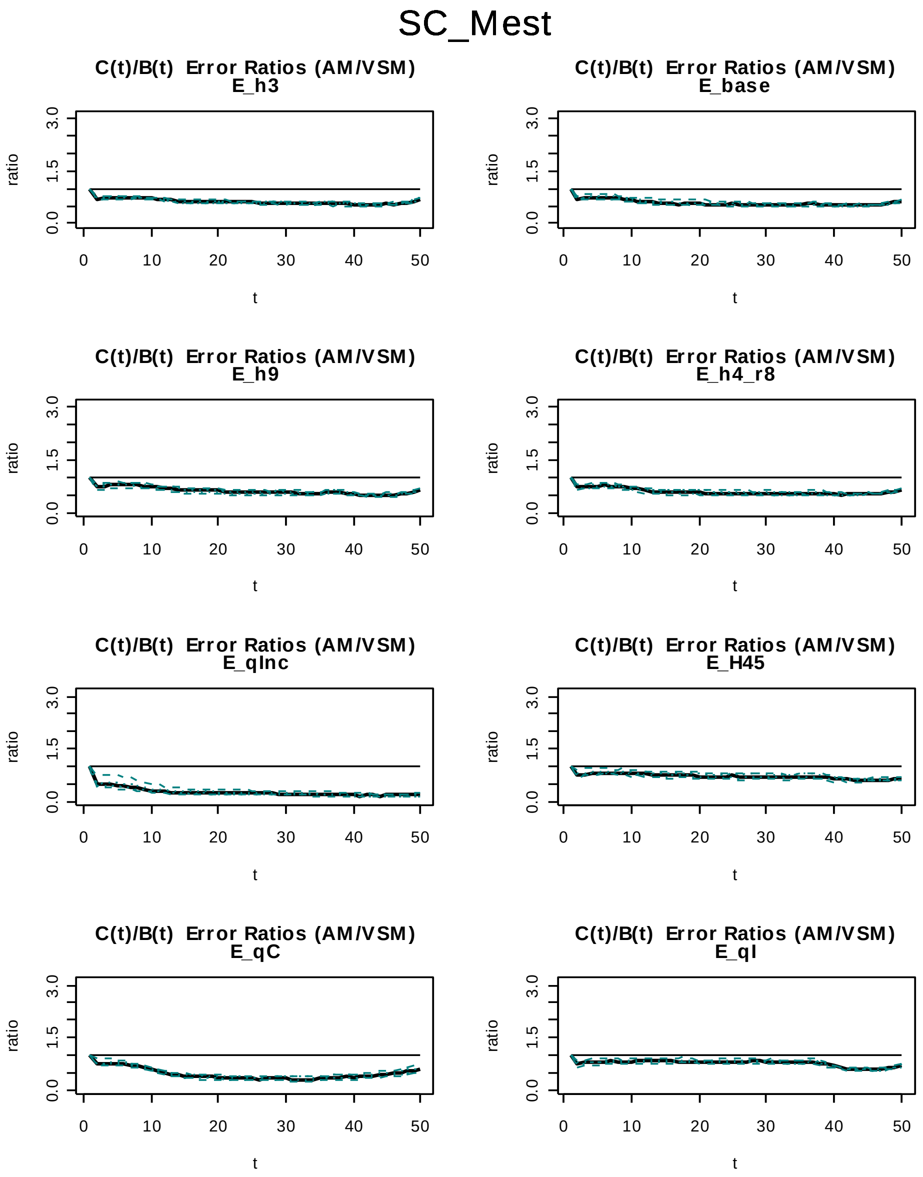
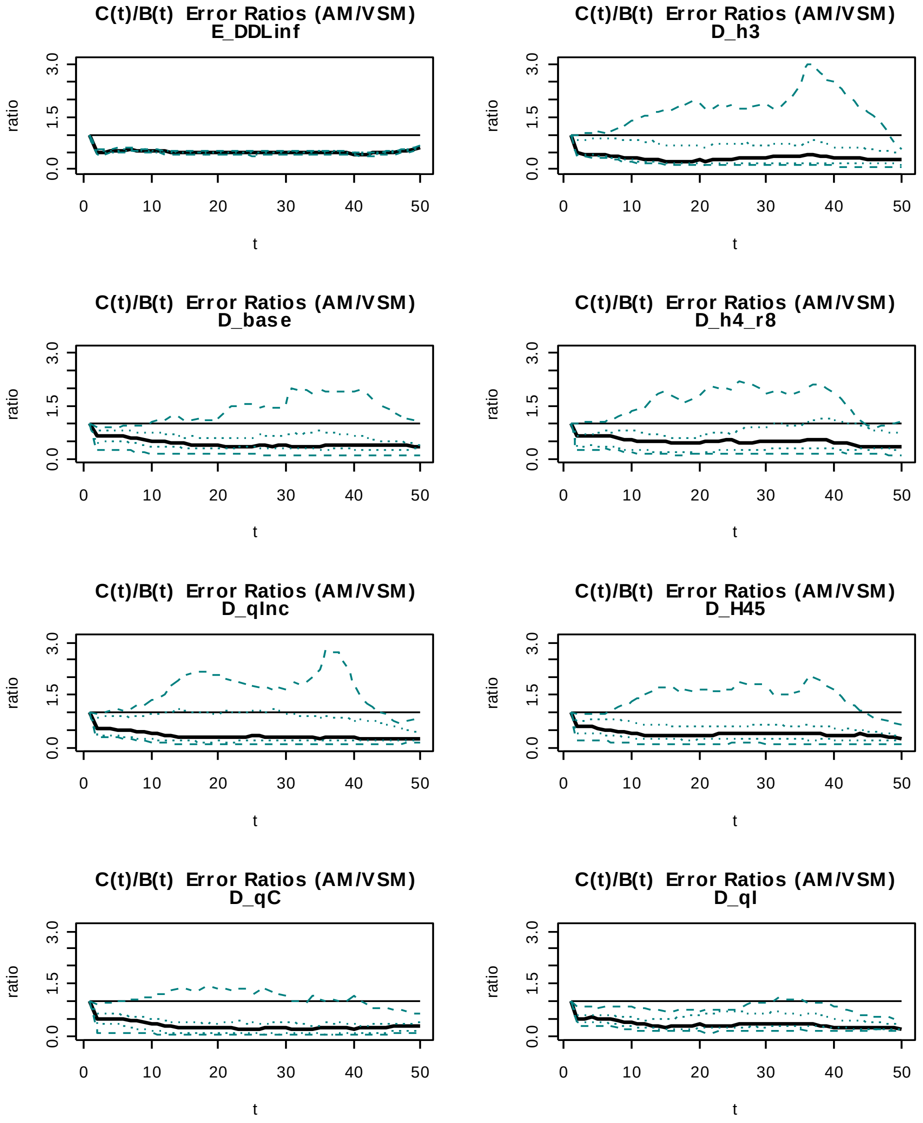

SC_Mest
C(t)/B(t) Error Ratios (AM/VSM)
E_h3
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h9
C(t)/B(t) Error Ratios (AM/VSM)
E_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
E_qlnc
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_qC
C(t)/B(t) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Figure A 6.5-e

SC_Mest


C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
C(t)/B(t) Error Ratios (AM/VSM)
D_h3
C(t)/B(t) Error Ratios (AM/VSM)
D_base
C(t)/B(t) Error Ratios (AM/VSM)
D_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
D_qInc
C(t)/B(t) Error Ratios (AM/VSM)
D_H45
C(t)/B(t) Error Ratios (AM/VSM)
D_qC
C(t)/B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

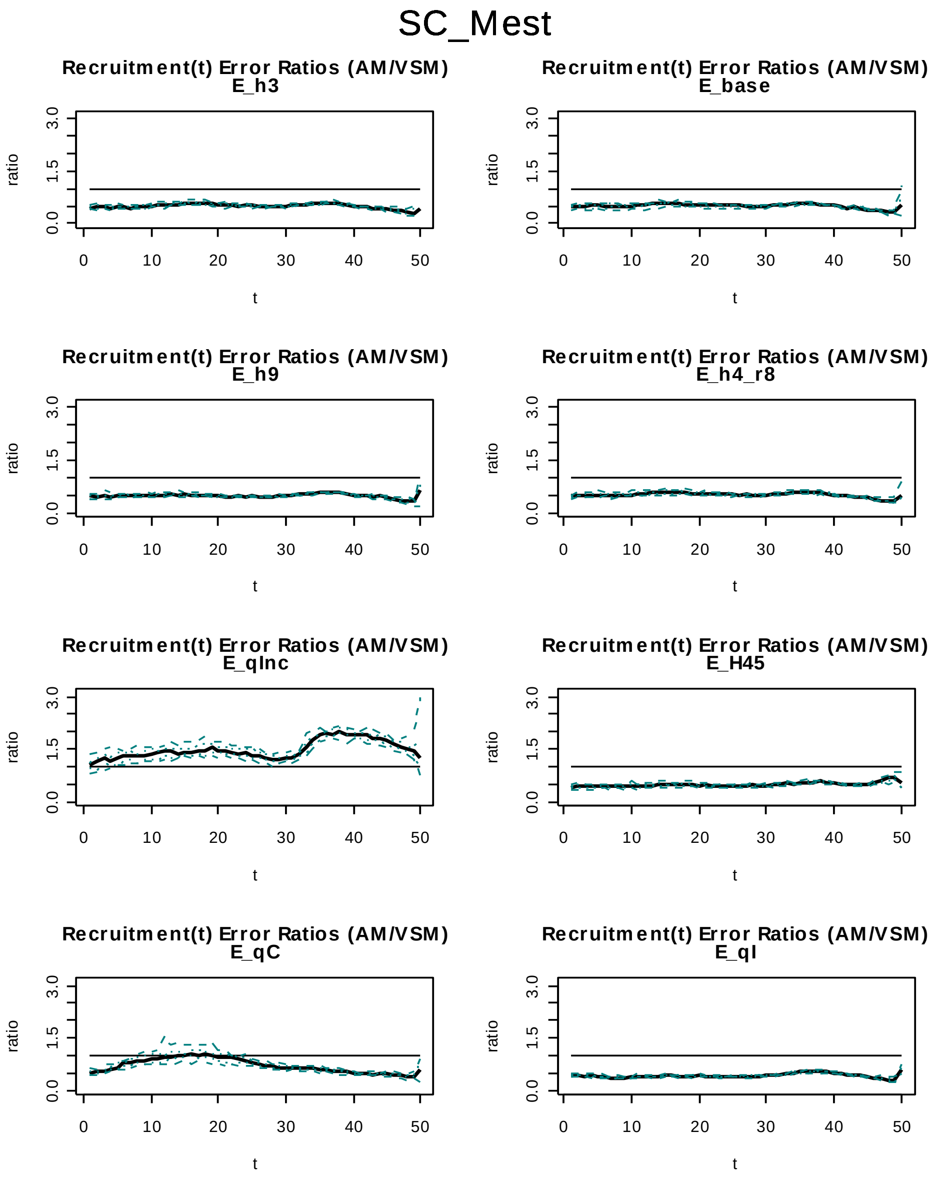
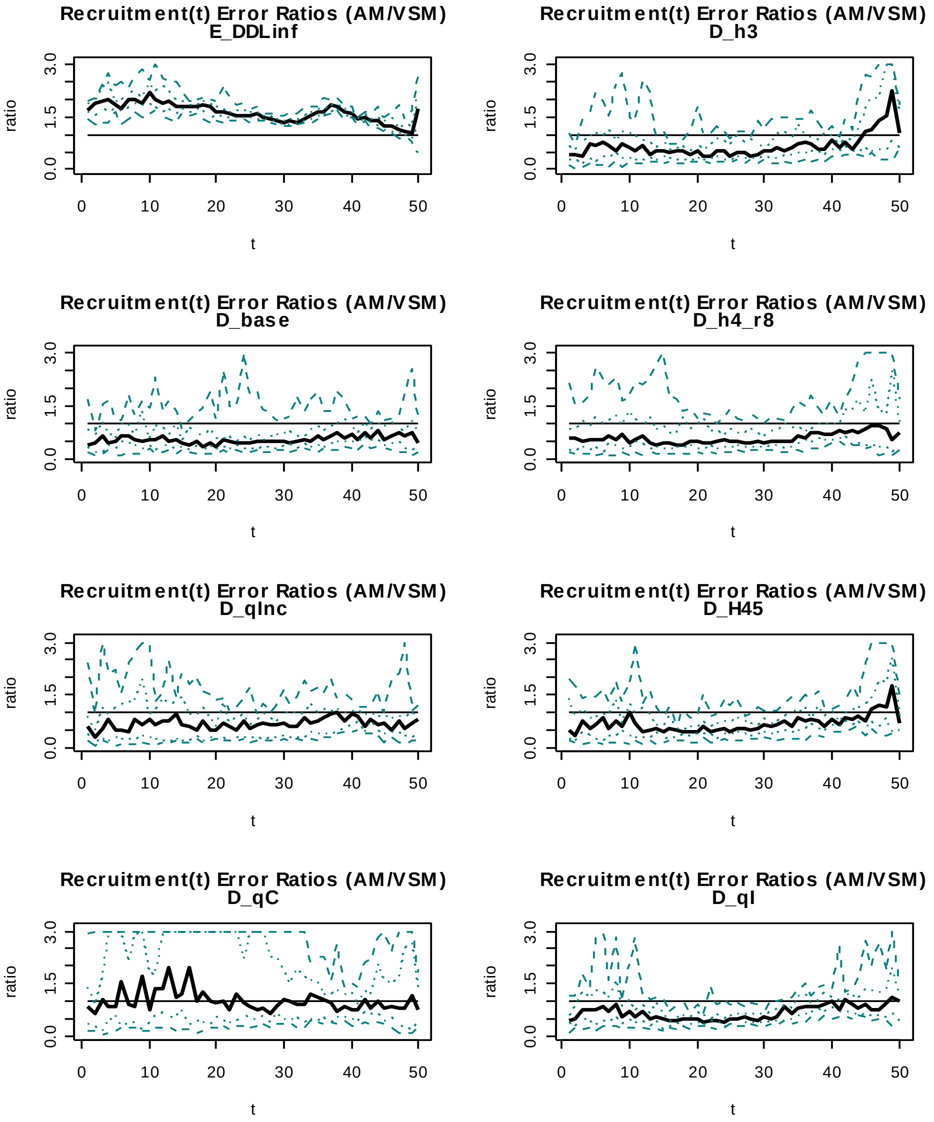

SC_Mest
Recruitment(t) Error Ratios (AM/VSM)
E_h3
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_h9
Recruitment(t) Error Ratios (AM/VSM)
E_h4_r8
Recruitment(t) Error Ratios (AM/VSM)
E_qInc
Recruitment(t) Error Ratios (AM/VSM)
E_H45
Recruitment(t) Error Ratios (AM/VSM)
E_qC
Recruitment(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.5-f

SC_Mest


Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf
Recruitment(t) Error Ratios (AM/VSM)
D_h3
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_h4_r8
Recruitment(t) Error Ratios (AM/VSM)
D_qInc
Recruitment(t) Error Ratios (AM/VSM)
D_H45
Recruitment(t) Error Ratios (AM/VSM)
D_qC
Recruitment(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

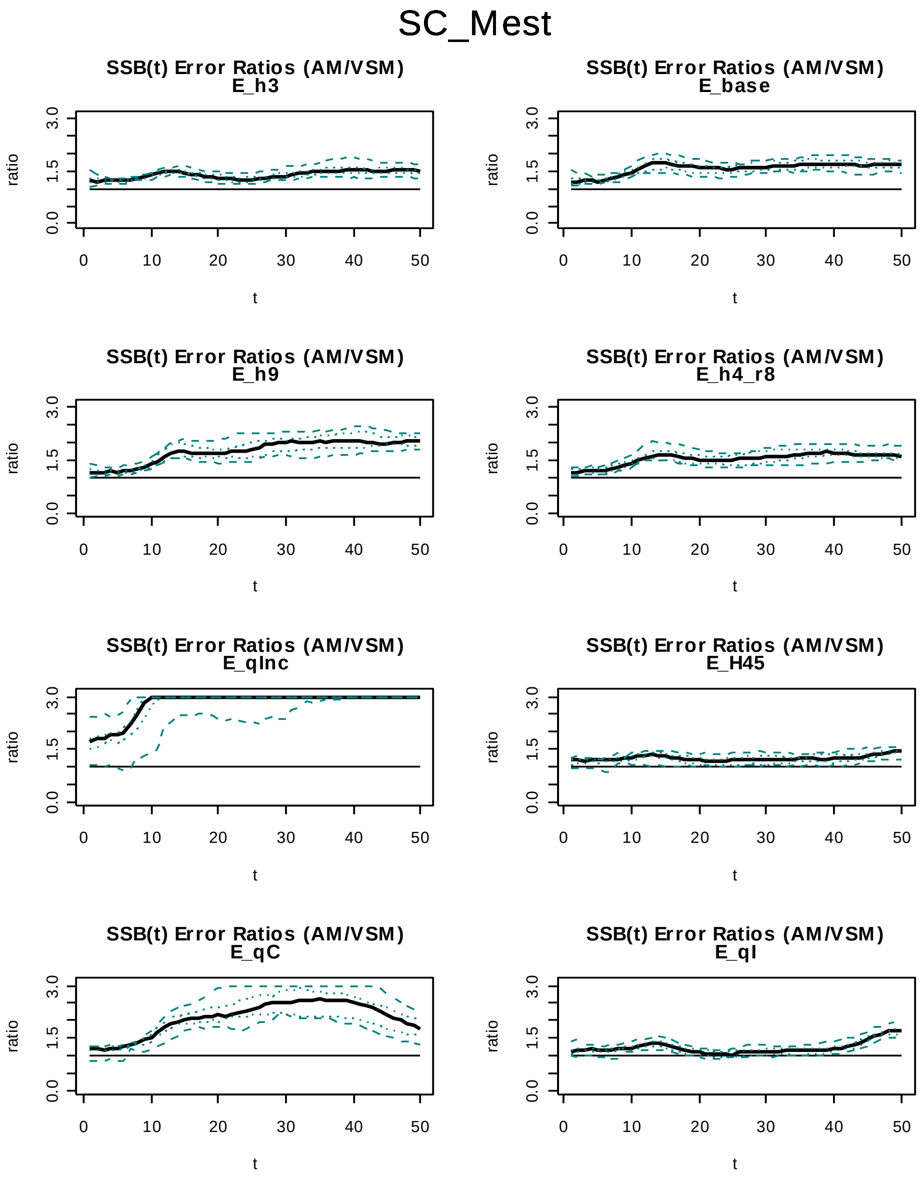

SC_Mest
SSB(t) Error Ratios (AM/VSM)
E_h3
SSB(t) Error Ratios (AM/VSM)
E_base
SSB(t) Error Ratios (AM/VSM)
E_h9
SSB(t) Error Ratios (AM/VSM)
E_h4_r8
SSB(t) Error Ratios (AM/VSM)
E_qInc
SSB(t) Error Ratios (AM/VSM)
E_H45
SSB(t) Error Ratios (AM/VSM)
E_qC
SSB(t) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.5-g

A 6.6 SC_noTag

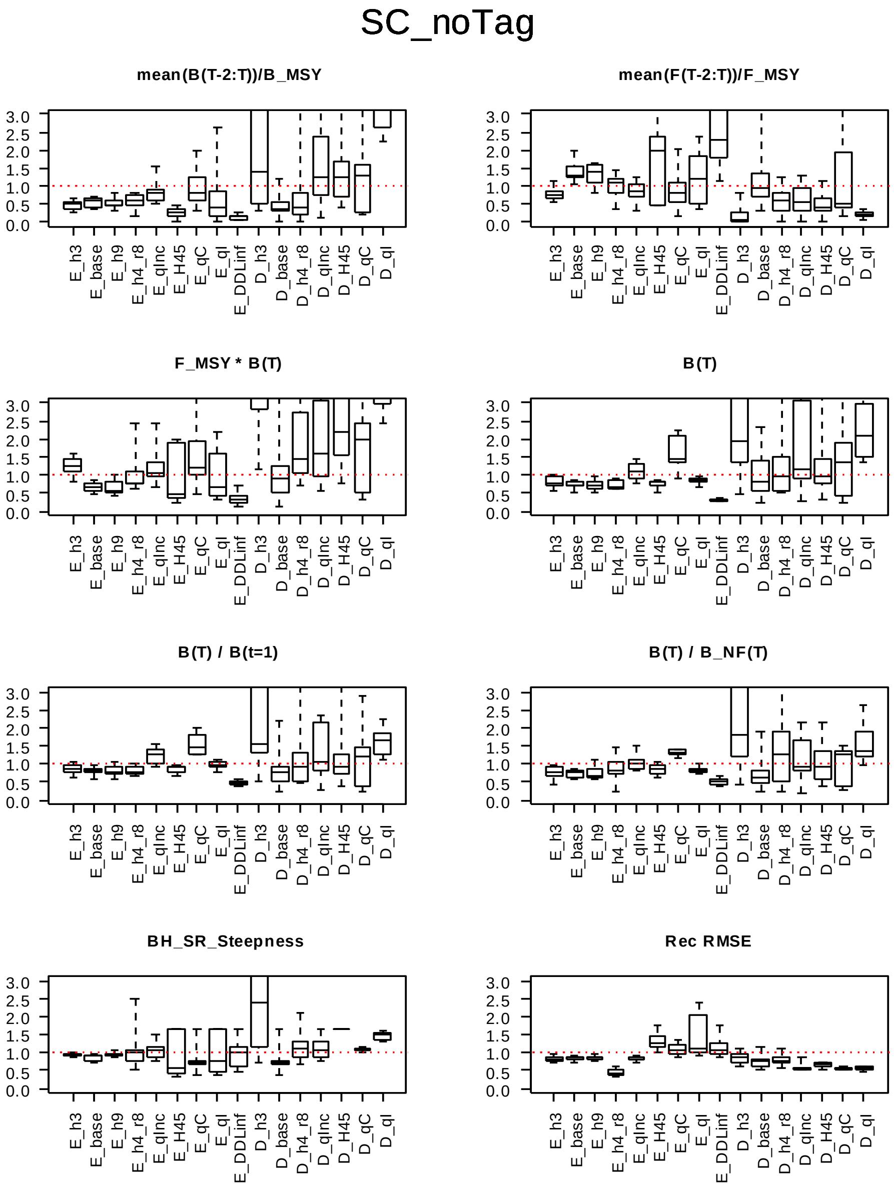


Figure A 6.6-a

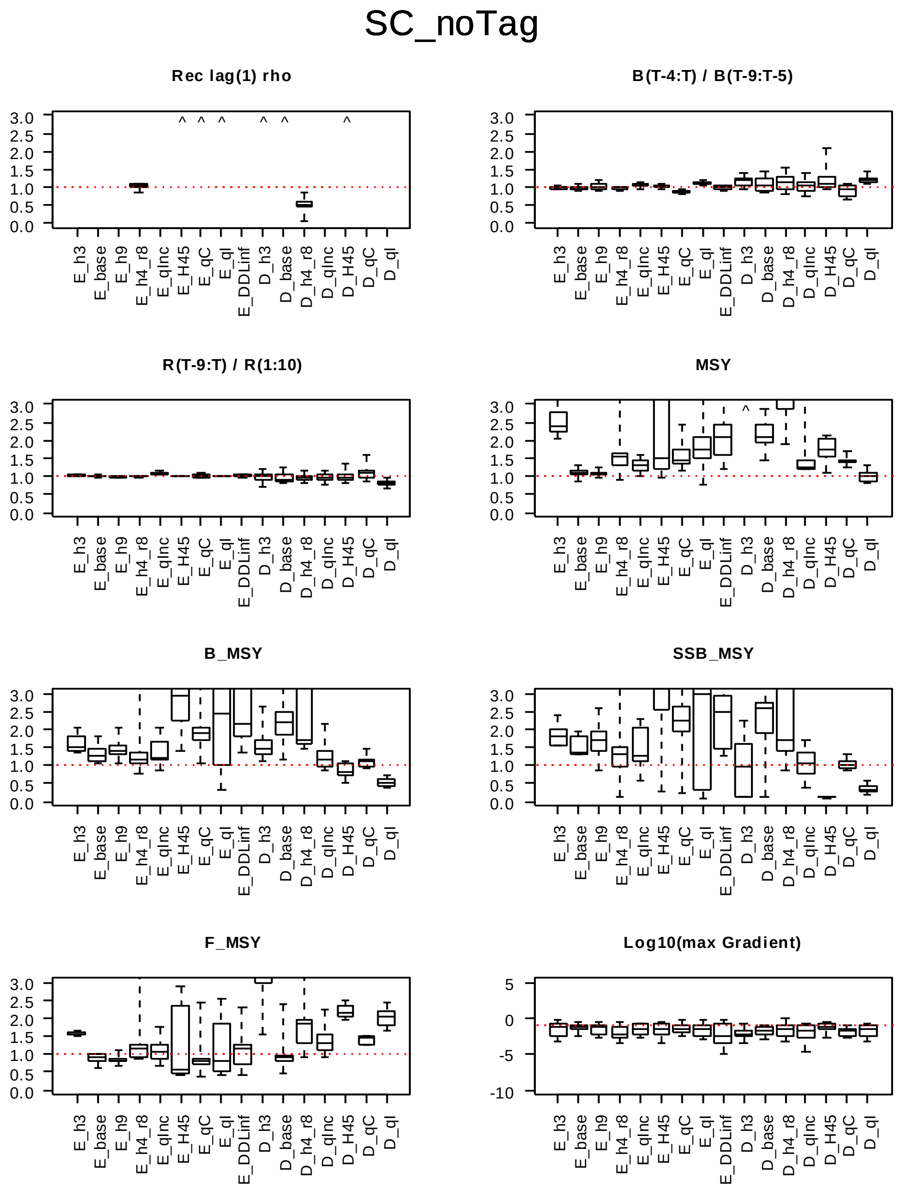
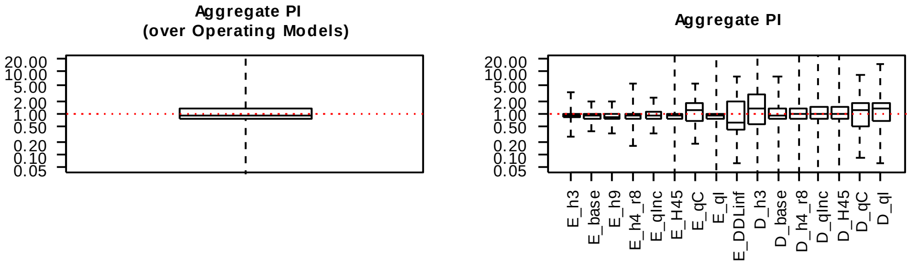

SC_noTag
Rec lag(1) rho
B(T-4:T) / B(T-9:T-5)
R(T-9:T) / R(1:10)
MSY
B_MSY
SSB_MSY
F_MSY
Log10(max Gradient)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
5
0
-5
-10
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_ql
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_ql

Figure A 6.6-b

SC_noTag


Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_qI
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_qI

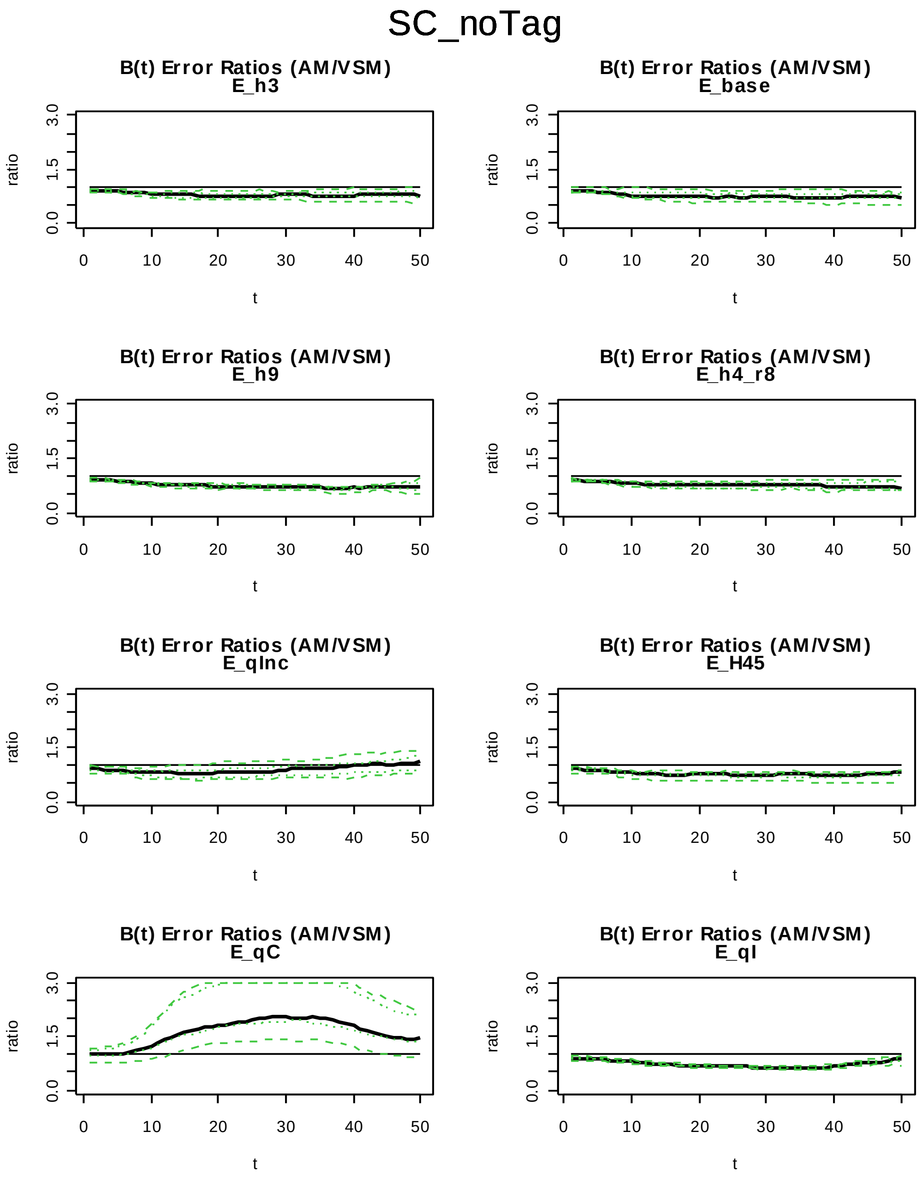
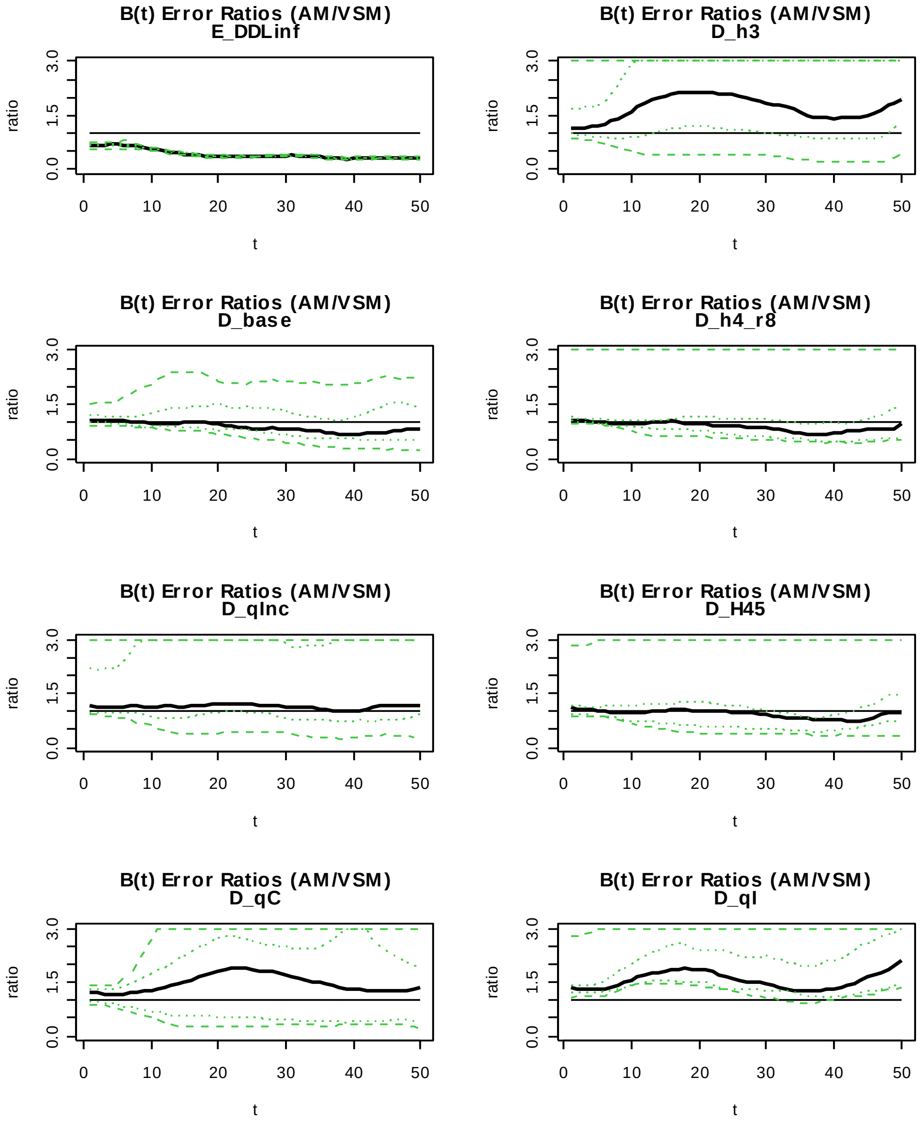

SC_noTag
B(t) Error Ratios (AM/VSM)
E_h3
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_h9
B(t) Error Ratios (AM/VSM)
E_h4_r8
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_H45
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.6-c

SC_noTag


B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t) Error Ratios (AM/VSM)
D_h3
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_h4_r8
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_H45
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

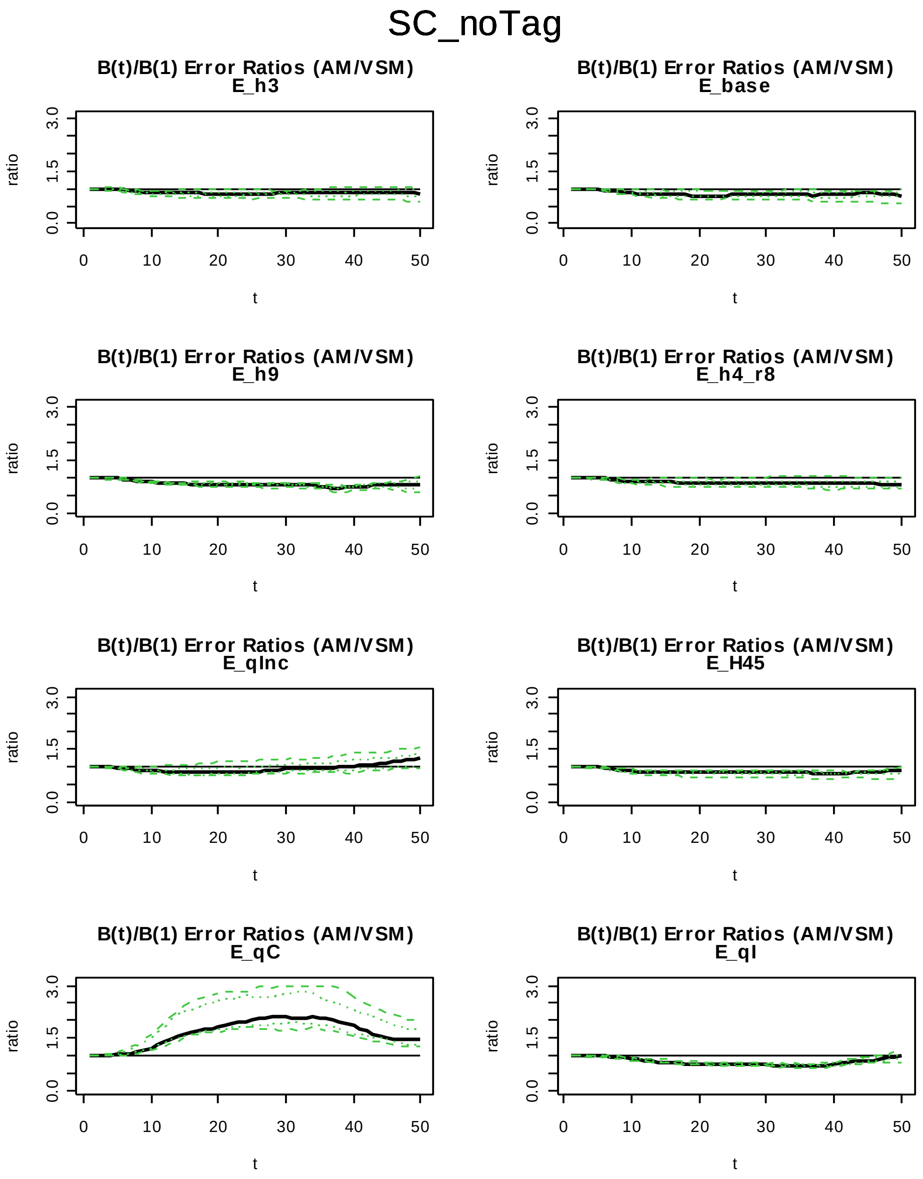
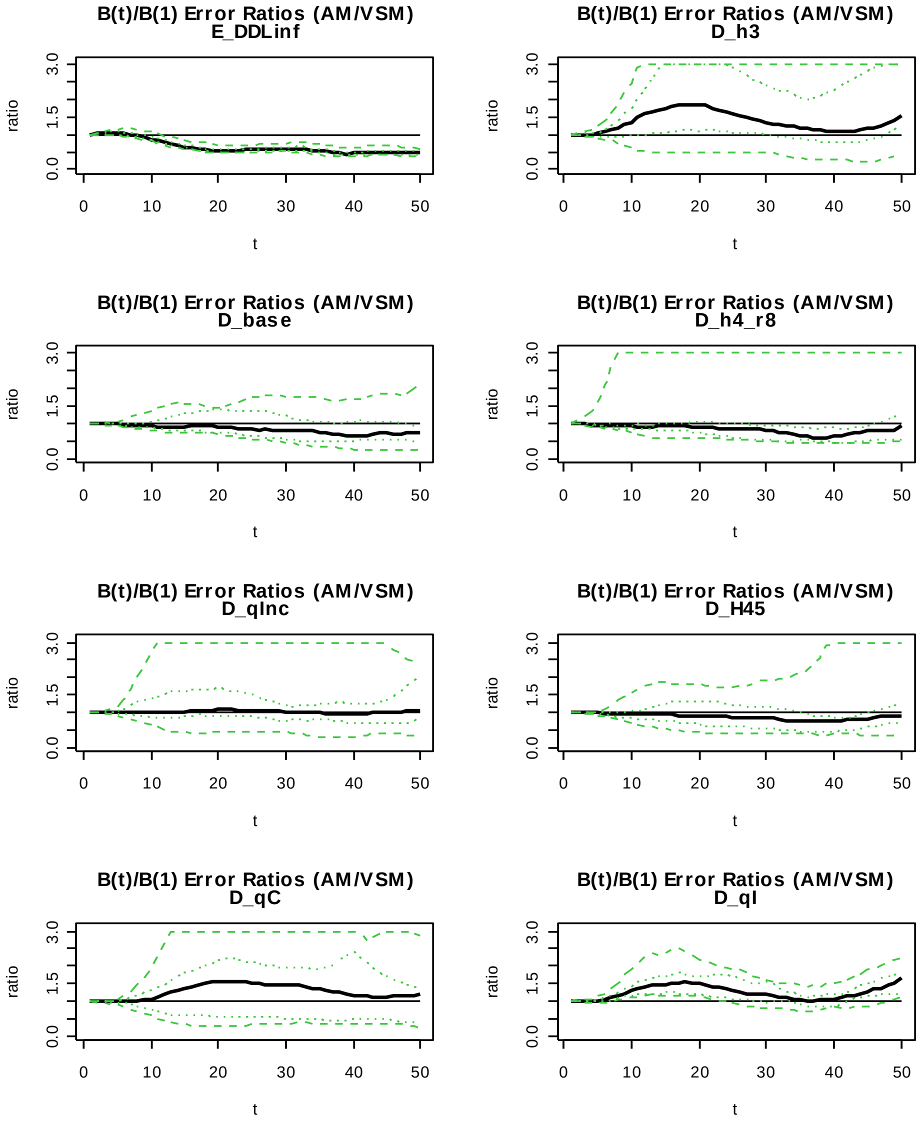

SC_noTag
B(t)/B(1) Error Ratios (AM/VSM)
E_h3
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_h9
B(t)/B(1) Error Ratios (AM/VSM)
E_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_qC
B(t)/B(1) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.6-d

SC_noTag



B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
B(t)/B(1) Error Ratios (AM/VSM)
D_h3
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
B(t)/B(1) Error Ratios (AM/VSM)
D_base
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
B(t)/B(1) Error Ratios (AM/VSM)
D_h4_r8
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
B(t)/B(1) Error Ratios (AM/VSM)
D_qInc
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
B(t)/B(1) Error Ratios (AM/VSM)
D_qC
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t
B(t)/B(1) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

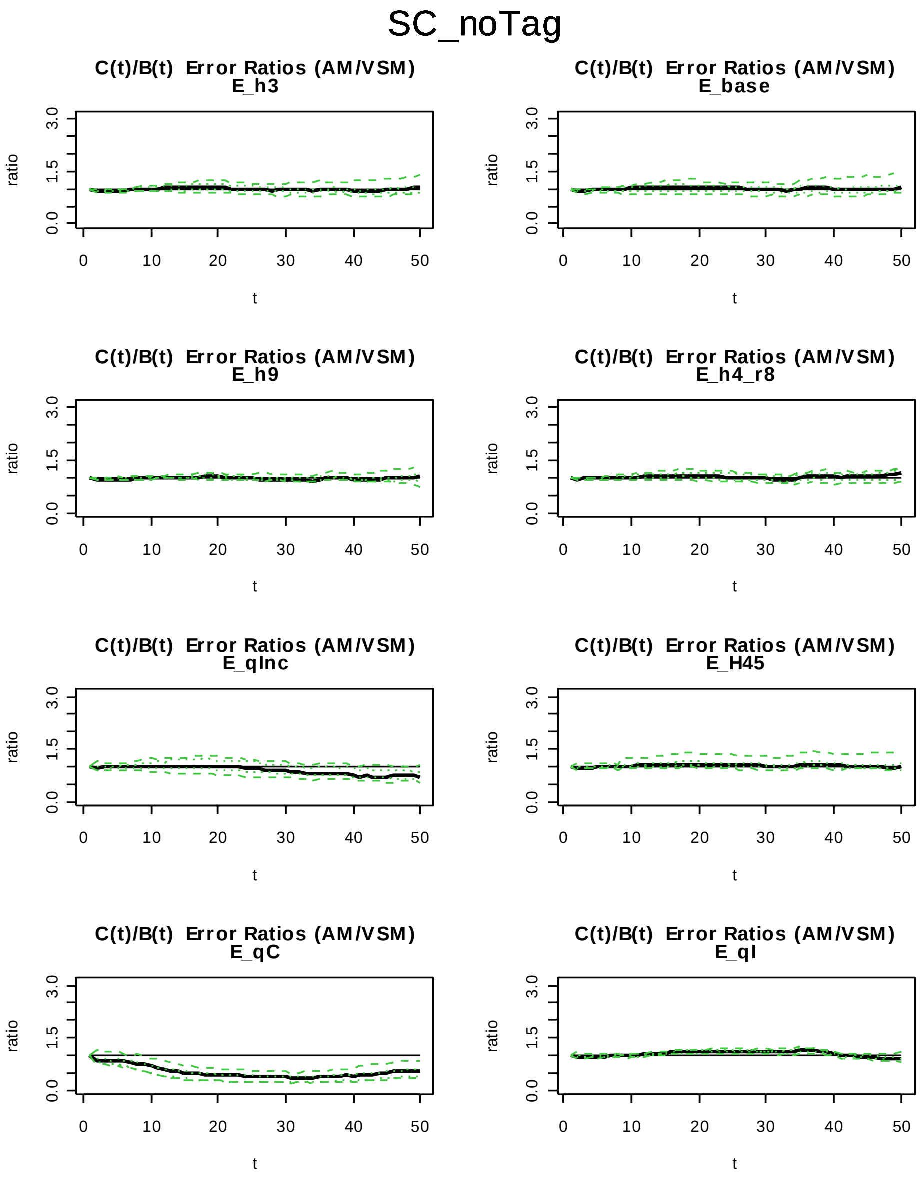
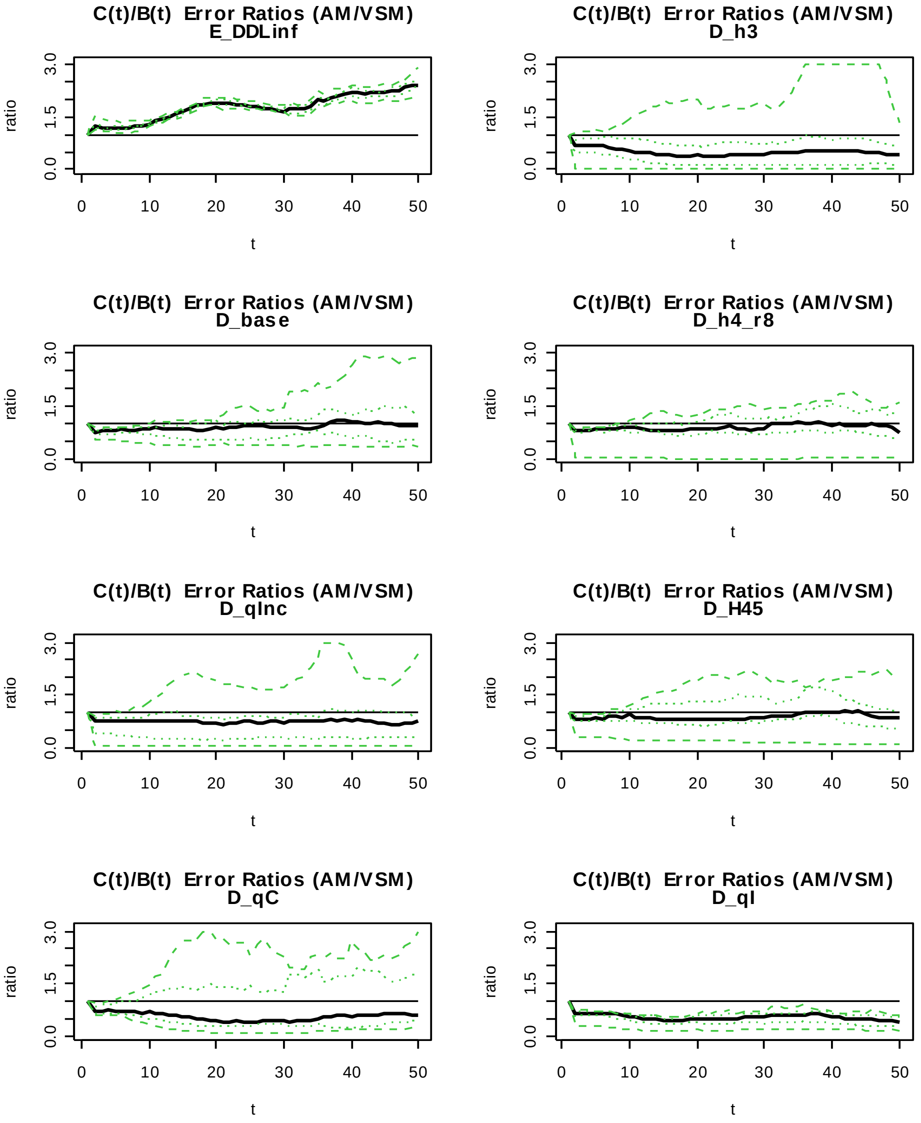

SC_noTag
C(t)/B(t) Error Ratios (AM/VSM)
E_h3
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h9
C(t)/B(t) Error Ratios (AM/VSM)
E_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_qC
C(t)/B(t) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.6-e

SC_noTag


C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

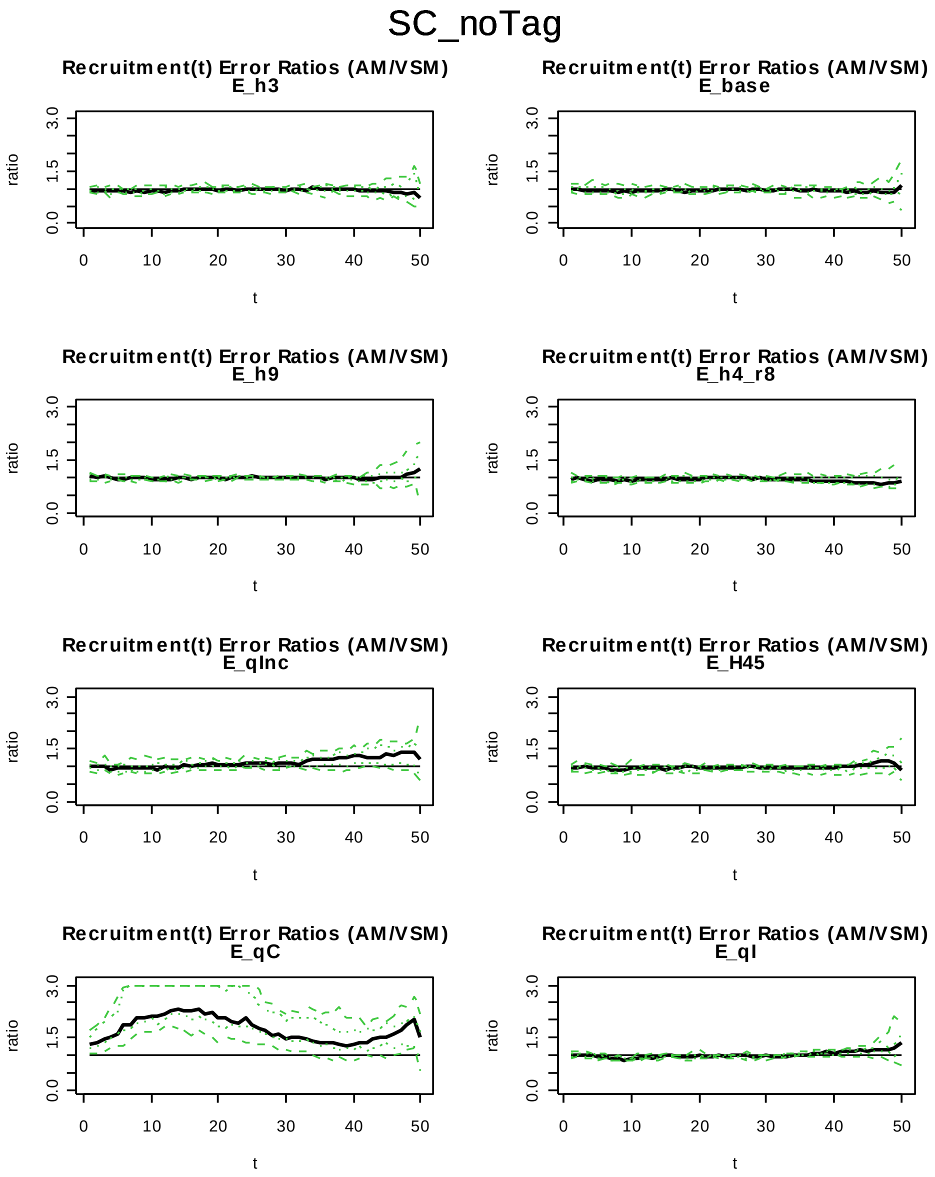
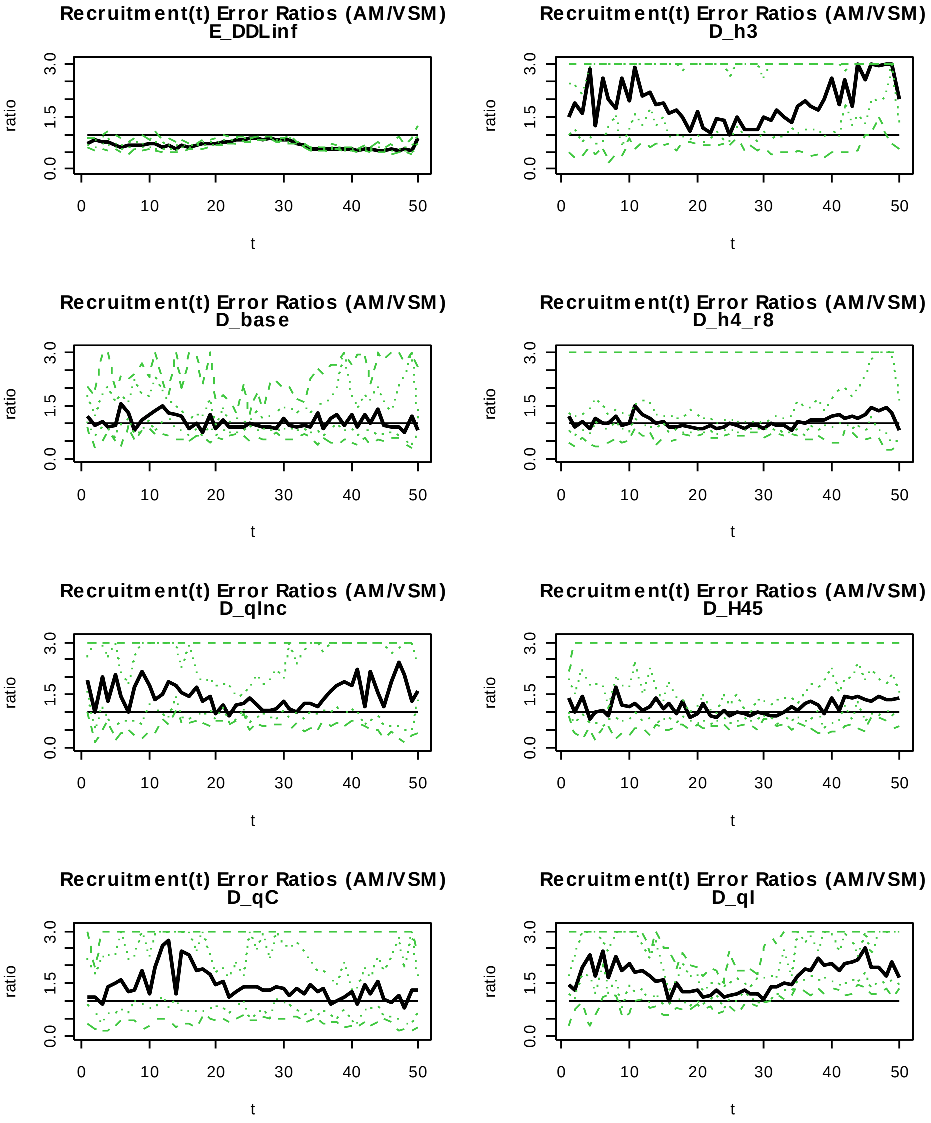

SC_noTag
Recruitment(t) Error Ratios (AM/VSM)
E_h3
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_h9
Recruitment(t) Error Ratios (AM/VSM)
E_h4_r8
Recruitment(t) Error Ratios (AM/VSM)
E_qInc
Recruitment(t) Error Ratios (AM/VSM)
E_H45
Recruitment(t) Error Ratios (AM/VSM)
E_qC
Recruitment(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.6-f

SC_noTag


Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf
Recruitment(t) Error Ratios (AM/VSM)
D_h3
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_h4_r8
Recruitment(t) Error Ratios (AM/VSM)
D_qInc
Recruitment(t) Error Ratios (AM/VSM)
D_H45
Recruitment(t) Error Ratios (AM/VSM)
D_qC
Recruitment(t) Error Ratios (AM/VSM)
D_ql
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

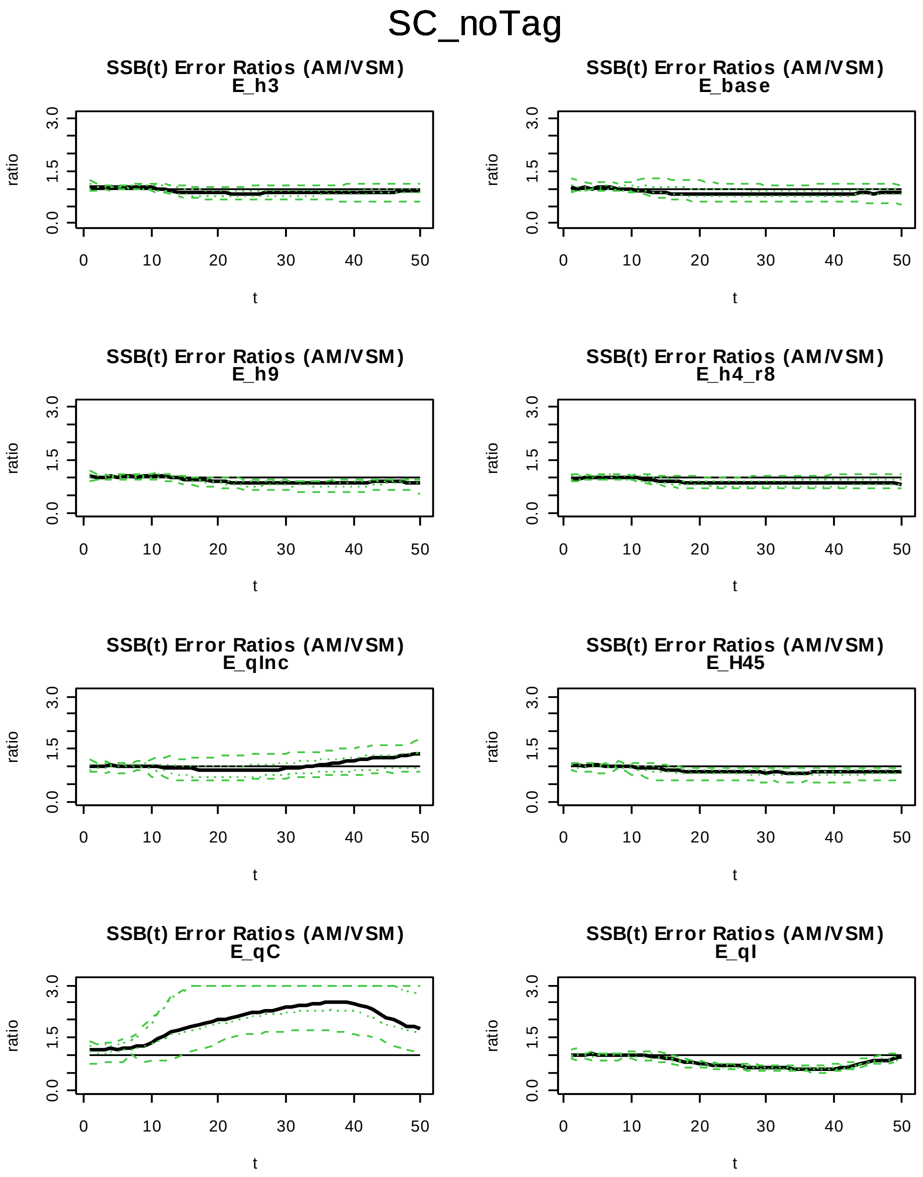

SC_noTag
SSB(t) Error Ratios (AM/VSM)
E_h3
SSB(t) Error Ratios (AM/VSM)
E_base
SSB(t) Error Ratios (AM/VSM)
E_h9
SSB(t) Error Ratios (AM/VSM)
E_h4_r8
SSB(t) Error Ratios (AM/VSM)
E_qInc
SSB(t) Error Ratios (AM/VSM)
E_H45
SSB(t) Error Ratios (AM/VSM)
E_qC
SSB(t) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.6-g

A 6.7 SC_2IDEAL

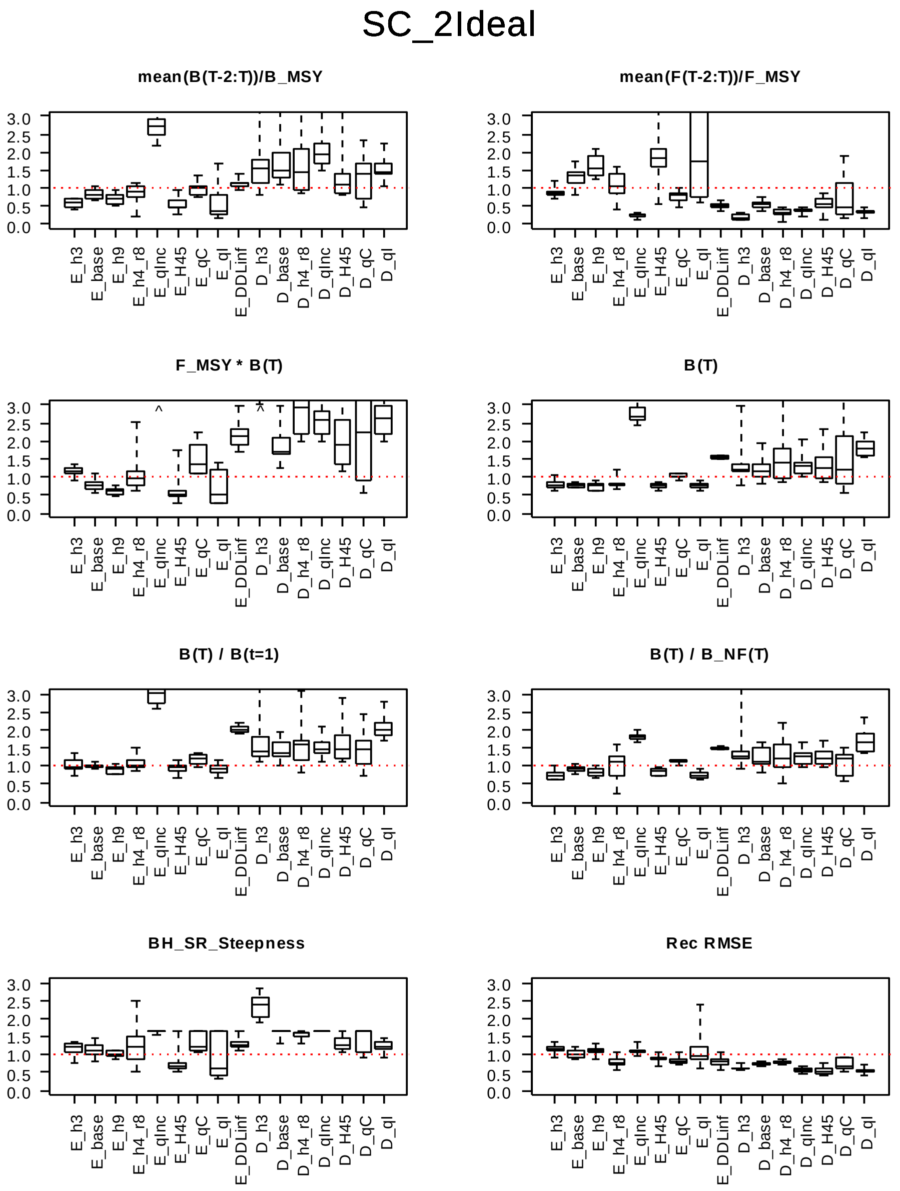


Figure A 6.7-a

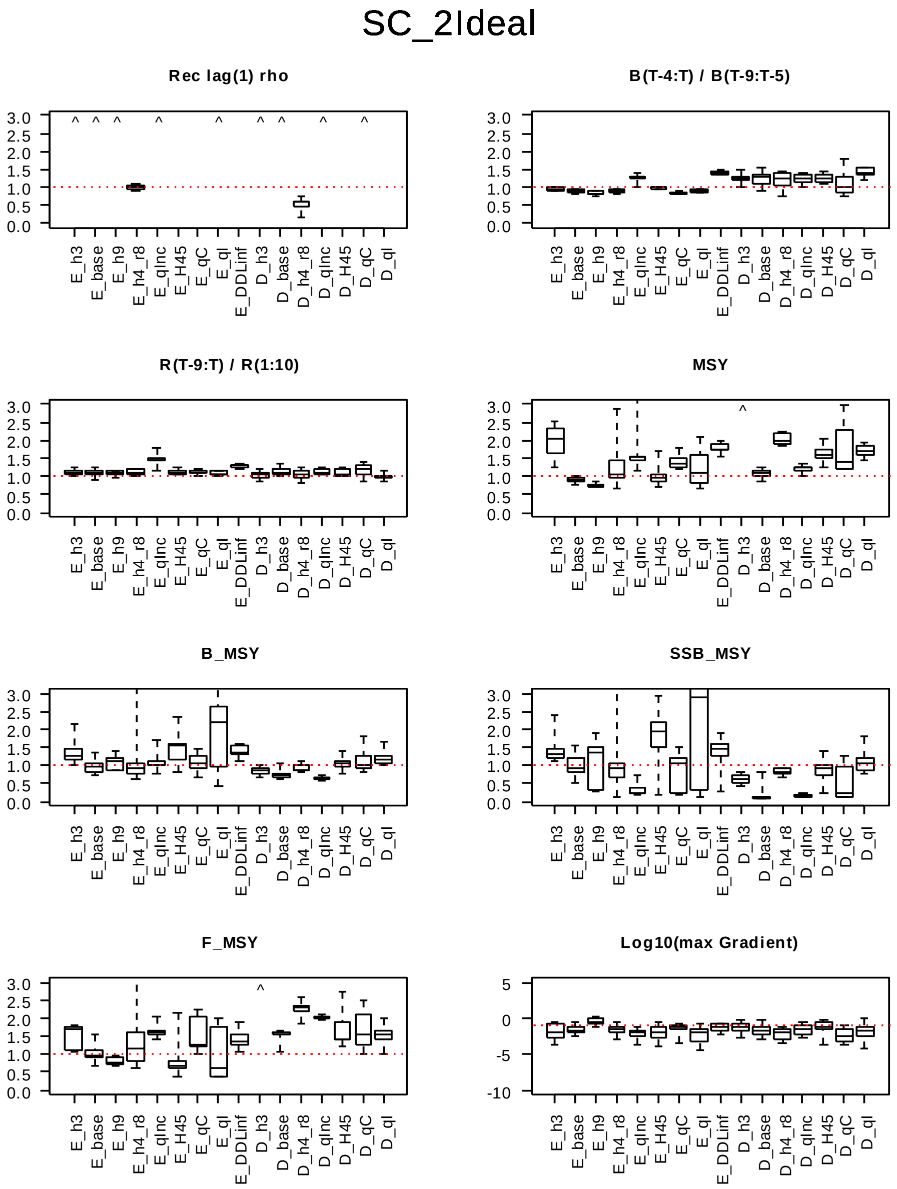
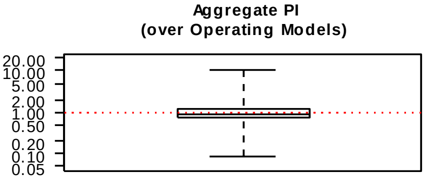
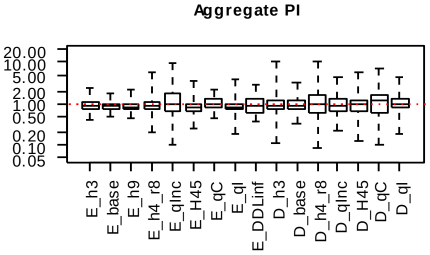
SC_2Ideal
Rec lag(1) rho
B(T-4:T) / B(T-9:T-5)
R(T-9:T) / R(1:10)
MSY
B_MSY
SSB_MSY
F_MSY
Log10(max Gradient)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
5
0
-5
-10
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_ql
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_ql

Figure A 6.7-b

SC_2Ideal



Aggregate PI
(over Operating Models)
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05



Aggregate PI
20.00
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
E_h3
E_base
E_h9
E_h4_r8
E_qInc
E_H45
E_qC
E_qI
E_DDLinf
D_h3
D_base
D_h4_r8
D_qInc
D_H45
D_qC
D_qI

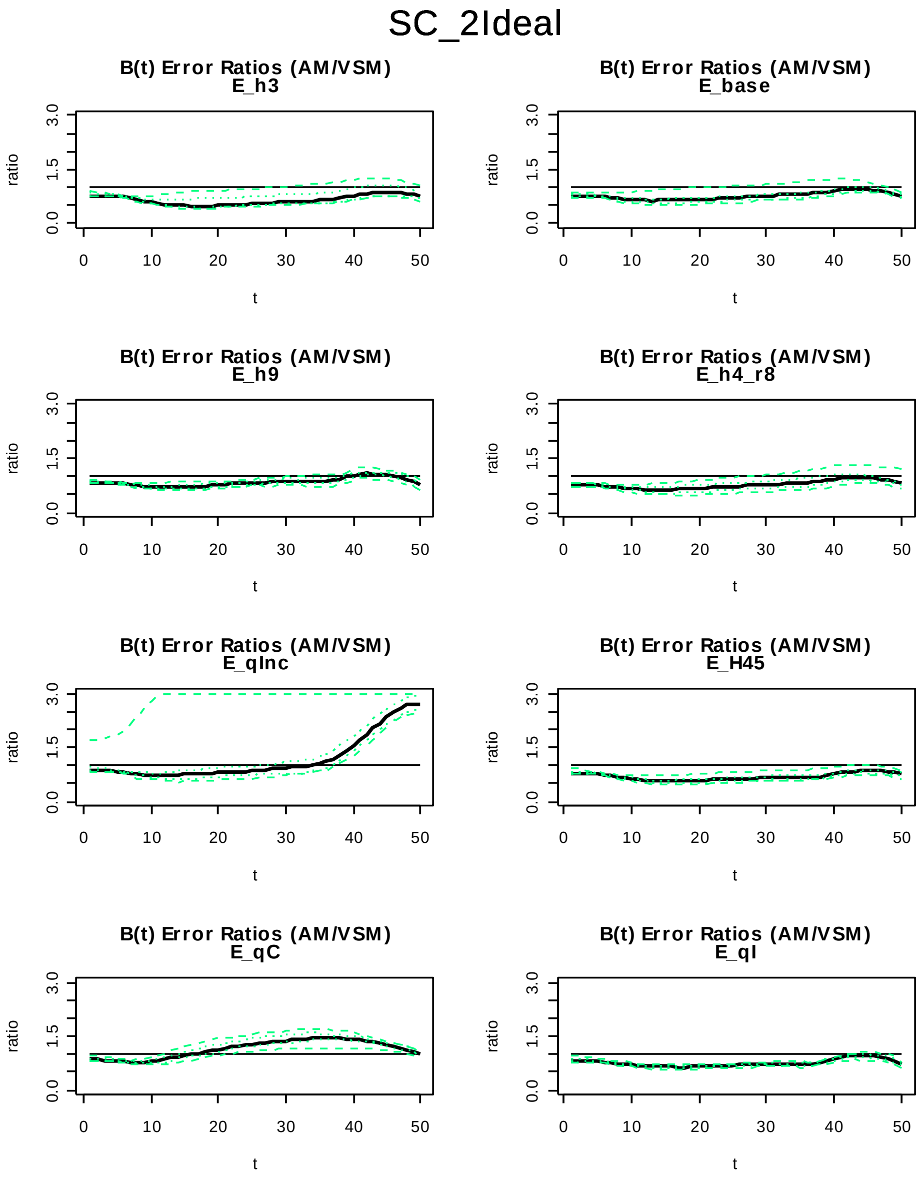
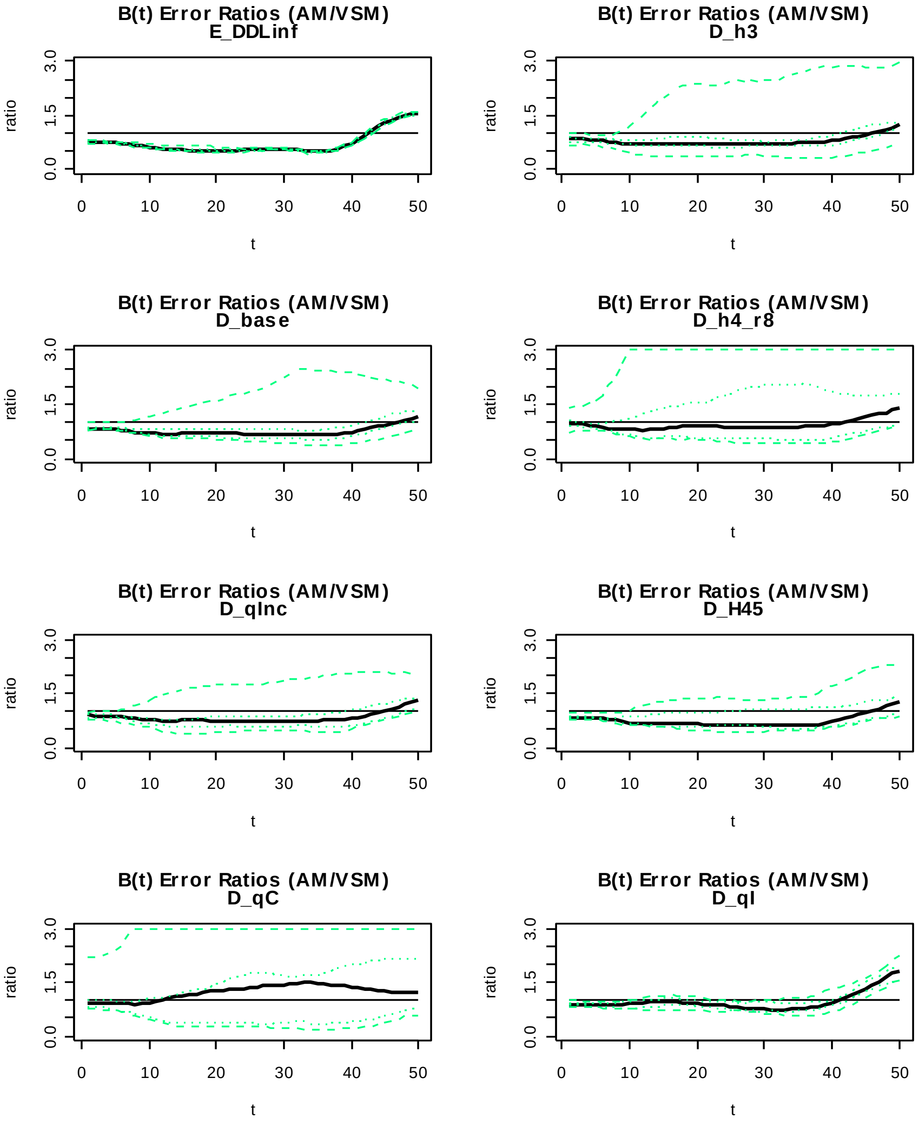

SC_2Ideal
B(t) Error Ratios (AM/VSM)
E_h3
B(t) Error Ratios (AM/VSM)
E_base
B(t) Error Ratios (AM/VSM)
E_h9
B(t) Error Ratios (AM/VSM)
E_h4_r8
B(t) Error Ratios (AM/VSM)
E_qInc
B(t) Error Ratios (AM/VSM)
E_H45
B(t) Error Ratios (AM/VSM)
E_qC
B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Figure A 6.7-c

SC_2Ideal


B(t) Error Ratios (AM/VSM)
E_DDLinf
B(t) Error Ratios (AM/VSM)
D_h3
B(t) Error Ratios (AM/VSM)
D_base
B(t) Error Ratios (AM/VSM)
D_h4_r8
B(t) Error Ratios (AM/VSM)
D_qInc
B(t) Error Ratios (AM/VSM)
D_H45
B(t) Error Ratios (AM/VSM)
D_qC
B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

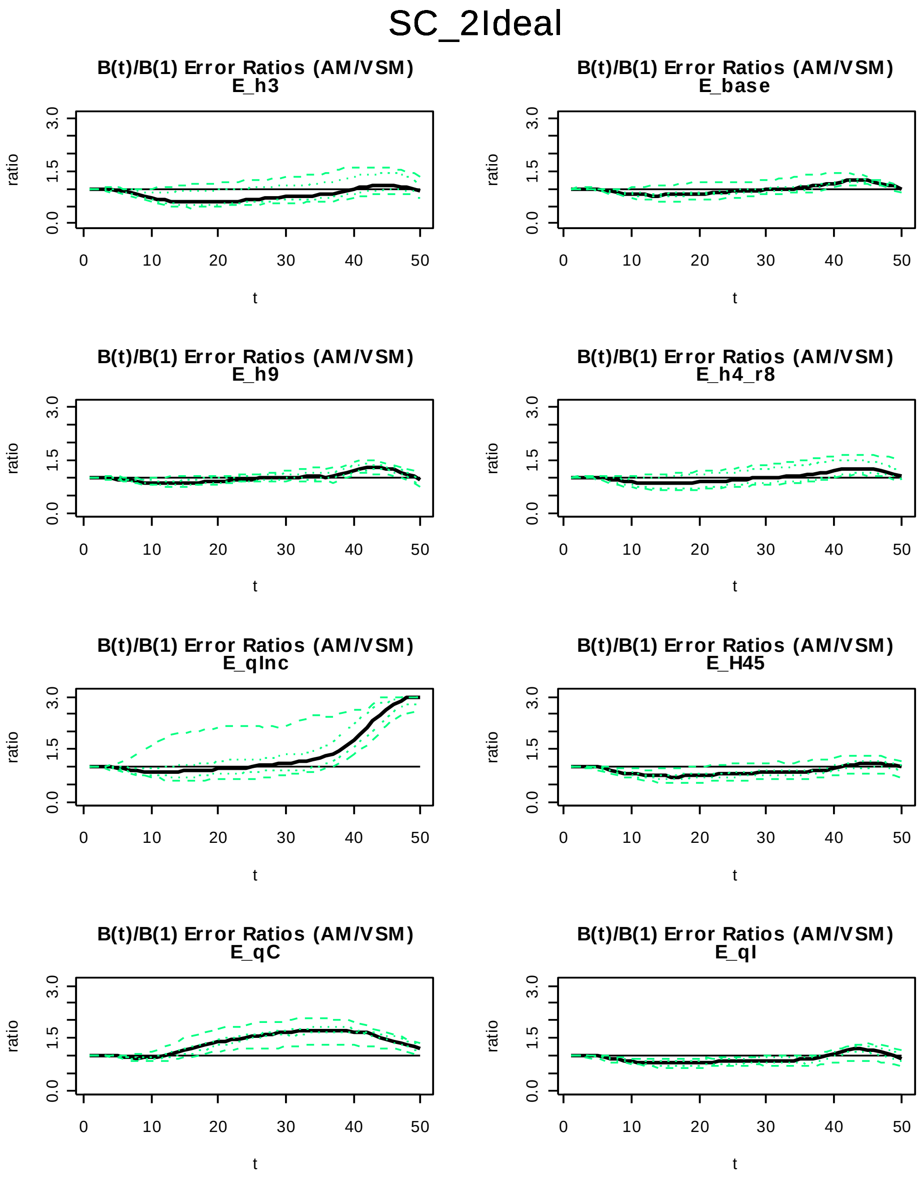
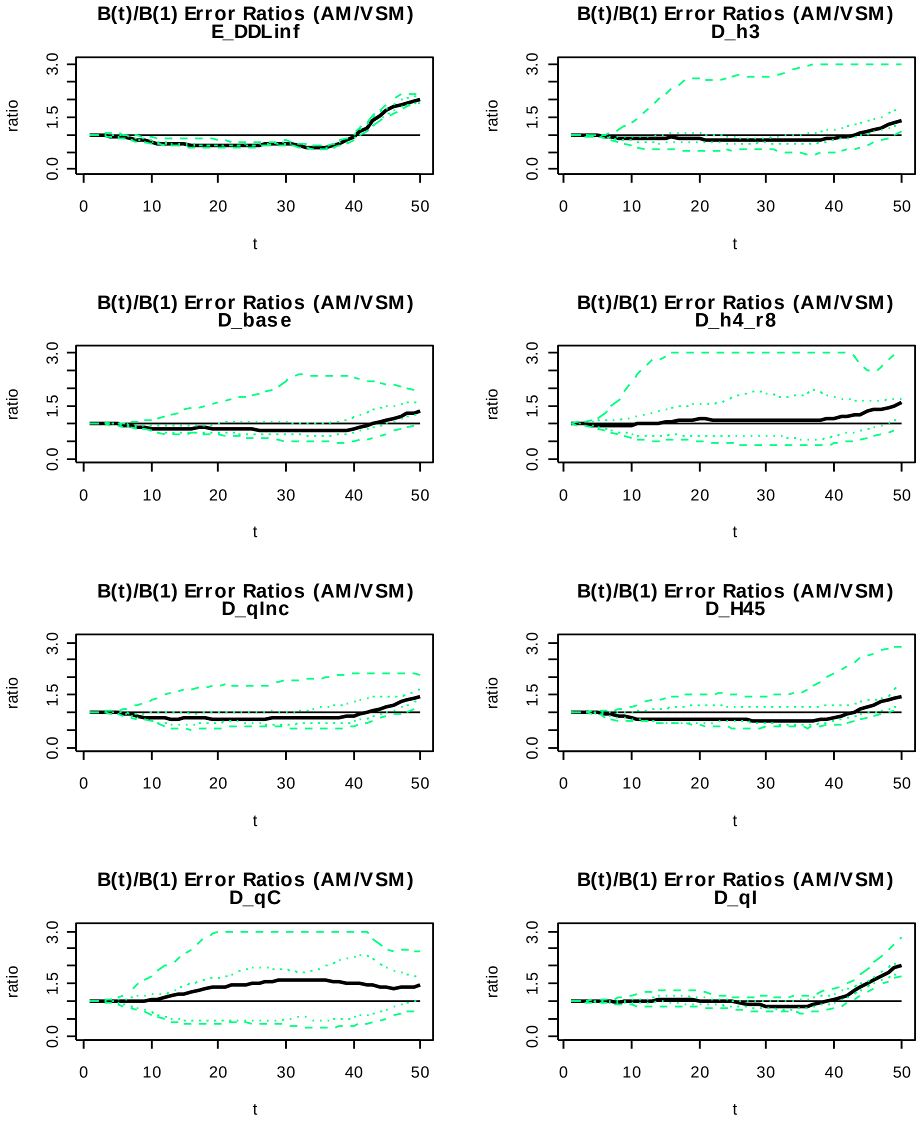

SC_2Ideal
B(t)/B(1) Error Ratios (AM/VSM)
E_h3
B(t)/B(1) Error Ratios (AM/VSM)
E_base
B(t)/B(1) Error Ratios (AM/VSM)
E_h9
B(t)/B(1) Error Ratios (AM/VSM)
E_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
E_qInc
B(t)/B(1) Error Ratios (AM/VSM)
E_H45
B(t)/B(1) Error Ratios (AM/VSM)
E_qC
B(t)/B(1) Error Ratios (AM/VSM)
E_ql
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Figure A 6.7-d

SC_2Ideal


B(t)/B(1) Error Ratios (AM/VSM)
E_DDLinf
B(t)/B(1) Error Ratios (AM/VSM)
D_h3
B(t)/B(1) Error Ratios (AM/VSM)
D_base
B(t)/B(1) Error Ratios (AM/VSM)
D_h4_r8
B(t)/B(1) Error Ratios (AM/VSM)
D_qInc
B(t)/B(1) Error Ratios (AM/VSM)
D_H45
B(t)/B(1) Error Ratios (AM/VSM)
D_qC
B(t)/B(1) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

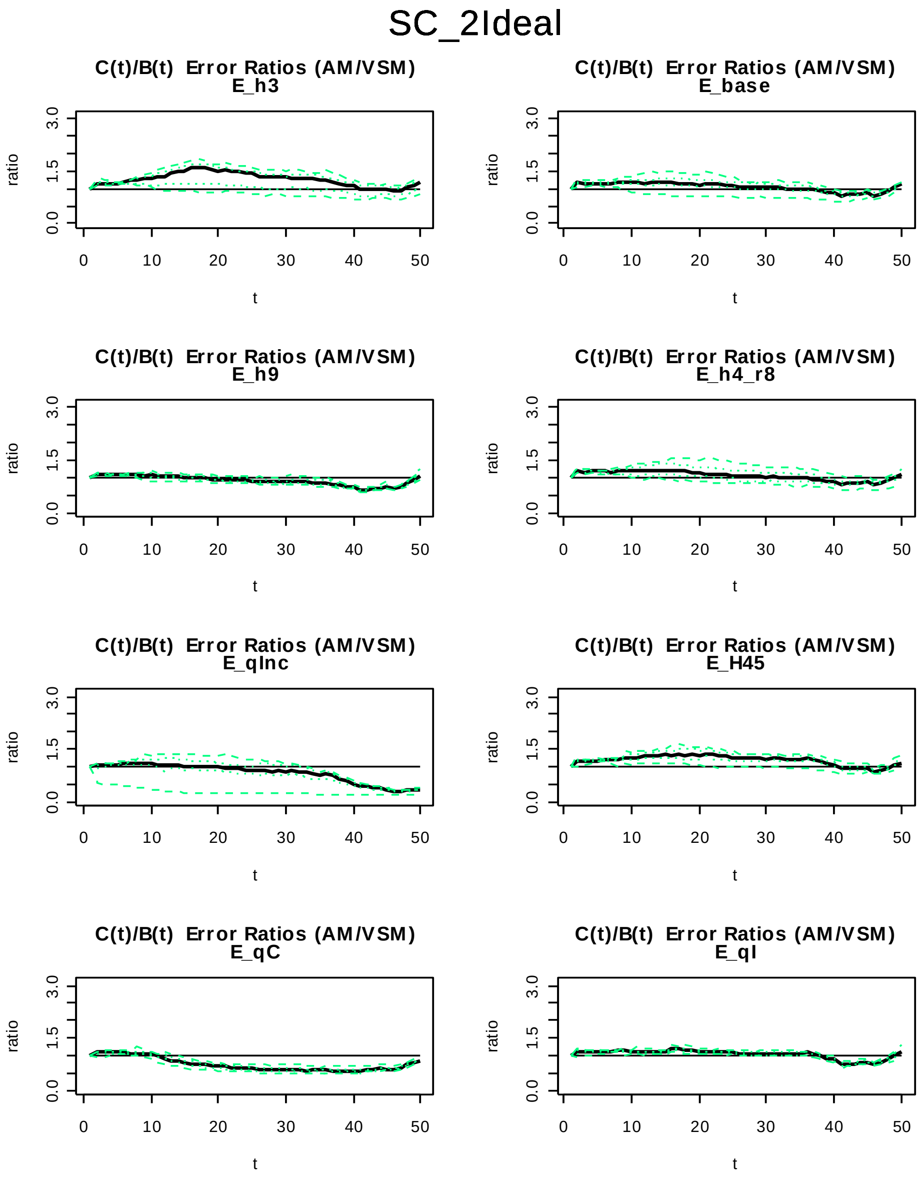
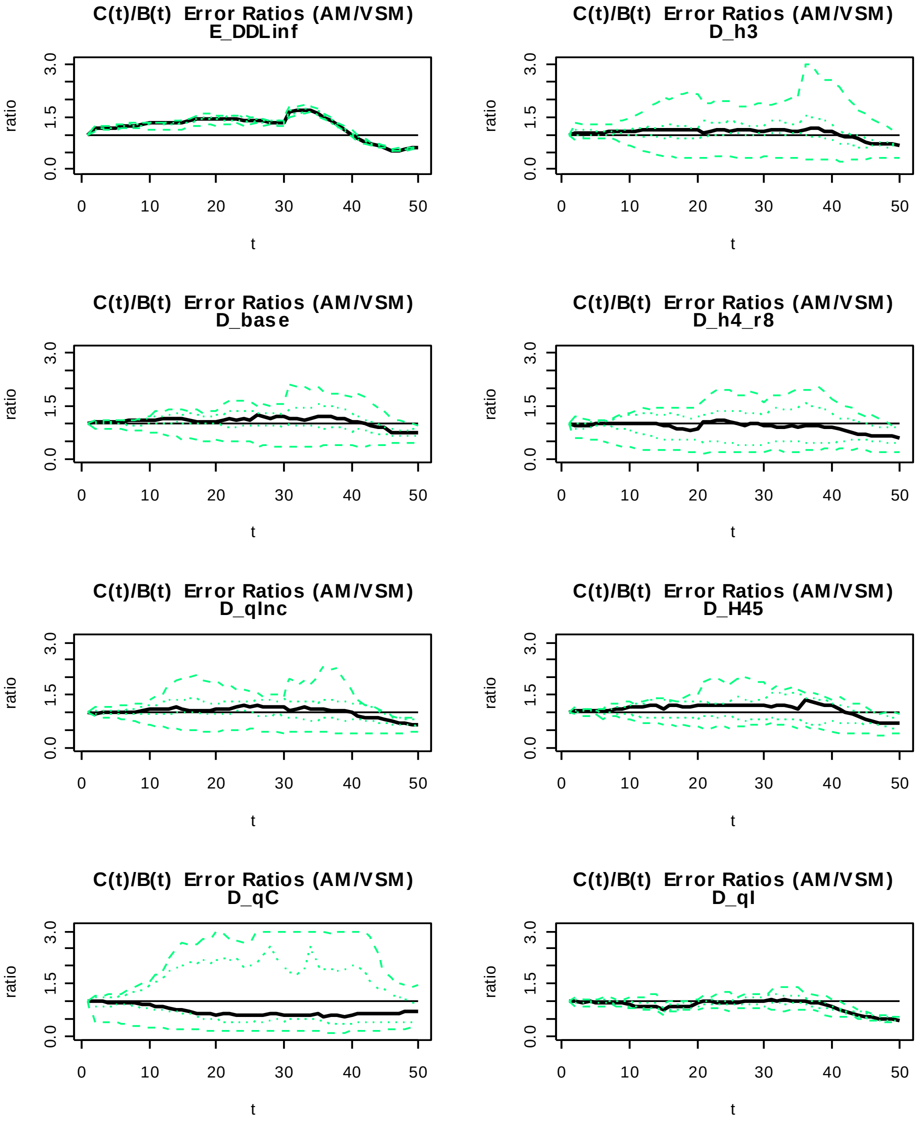

SC_2Ideal
C(t)/B(t) Error Ratios (AM/VSM)
E_h3
C(t)/B(t) Error Ratios (AM/VSM)
E_base
C(t)/B(t) Error Ratios (AM/VSM)
E_h9
C(t)/B(t) Error Ratios (AM/VSM)
E_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
E_qInc
C(t)/B(t) Error Ratios (AM/VSM)
E_H45
C(t)/B(t) Error Ratios (AM/VSM)
E_qC
C(t)/B(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

Figure A 6.7-e

SC_2Ideal


C(t)/B(t) Error Ratios (AM/VSM)
E_DDLinf
C(t)/B(t) Error Ratios (AM/VSM)
D_h3
C(t)/B(t) Error Ratios (AM/VSM)
D_base
C(t)/B(t) Error Ratios (AM/VSM)
D_h4_r8
C(t)/B(t) Error Ratios (AM/VSM)
D_qInc
C(t)/B(t) Error Ratios (AM/VSM)
D_H45
C(t)/B(t) Error Ratios (AM/VSM)
D_qC
C(t)/B(t) Error Ratios (AM/VSM)
D_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

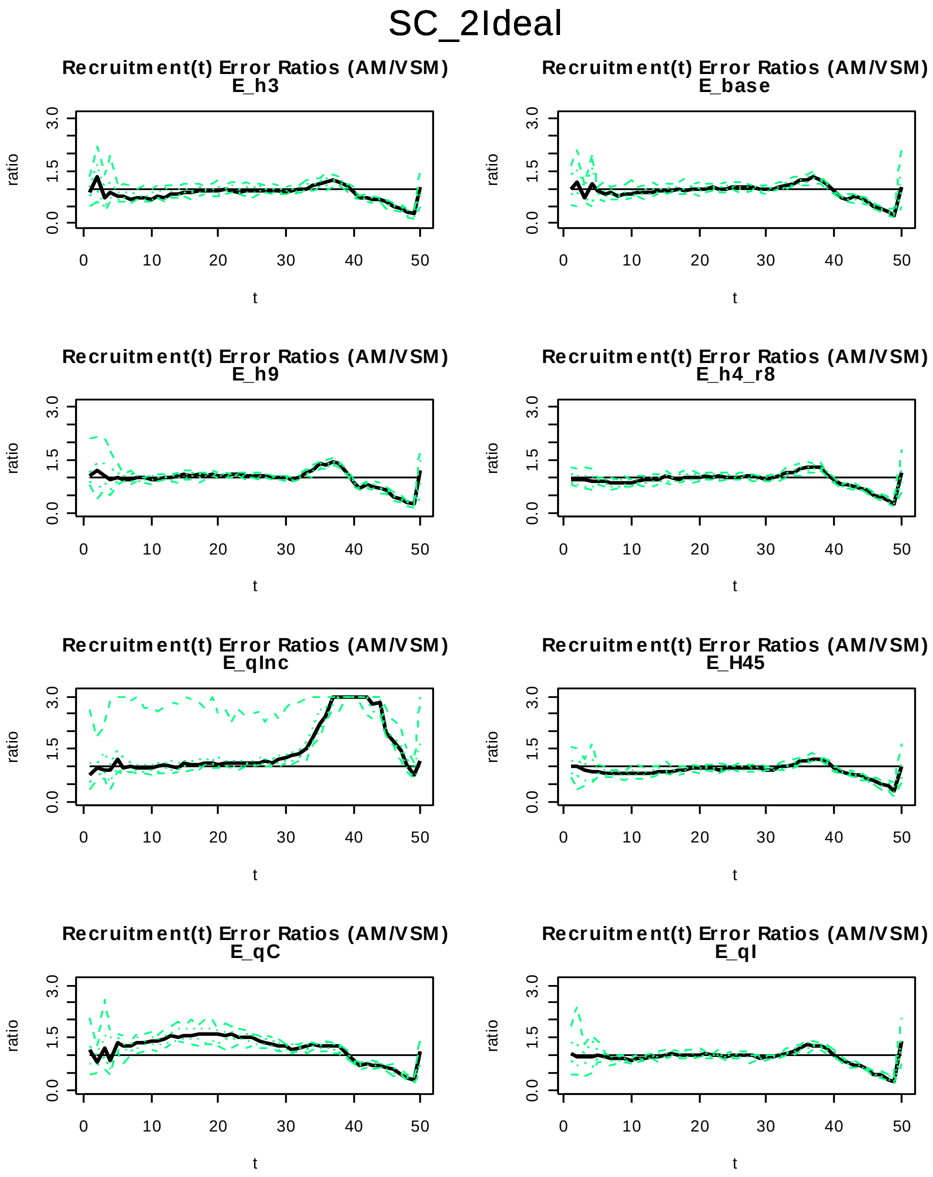
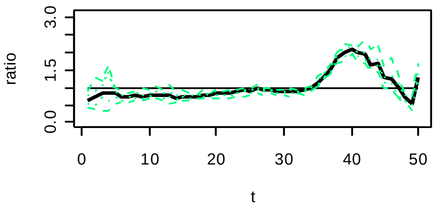

SC_2Ideal
Recruitment(t) Error Ratios (AM/VSM)
E_h3
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_h9
Recruitment(t) Error Ratios (AM/VSM)
E_h4_r8
Recruitment(t) Error Ratios (AM/VSM)
E_qInc
Recruitment(t) Error Ratios (AM/VSM)
E_H45
Recruitment(t) Error Ratios (AM/VSM)
E_qC
Recruitment(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
0
10
20
30
40
50
t

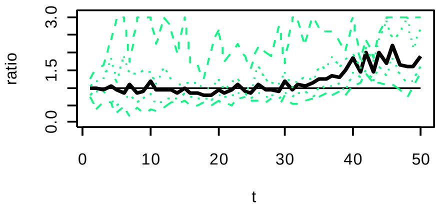
Figure A 6.7-f

SC_2Ideal

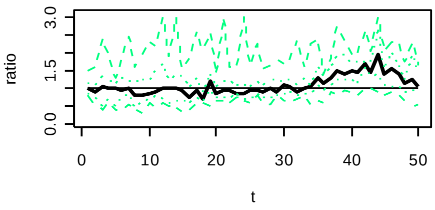
Recruitment(t) Error Ratios (AM/VSM)
E_DDLinf



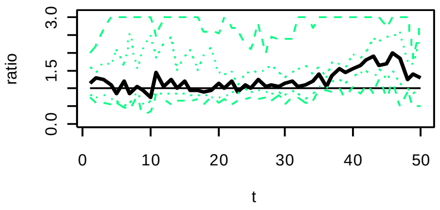
Recruitment(t) Error Ratios (AM/VSM)
D_h3



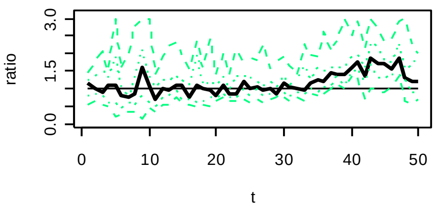
Recruitment(t) Error Ratios (AM/VSM)
D_base



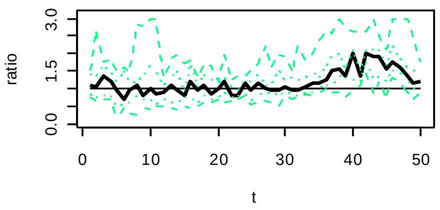
Recruitment(t) Error Ratios (AM/VSM)
D_h4_r8



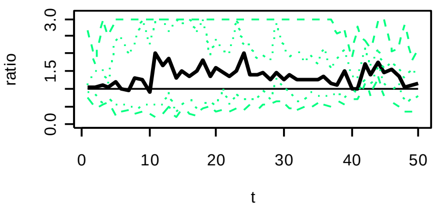
Recruitment(t) Error Ratios (AM/VSM)
D_qInc



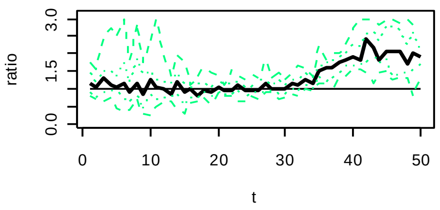
Recruitment(t) Error Ratios (AM/VSM)
D_H45



Recruitment(t) Error Ratios (AM/VSM)
D_qC



Recruitment(t) Error Ratios (AM/VSM)
D_qI



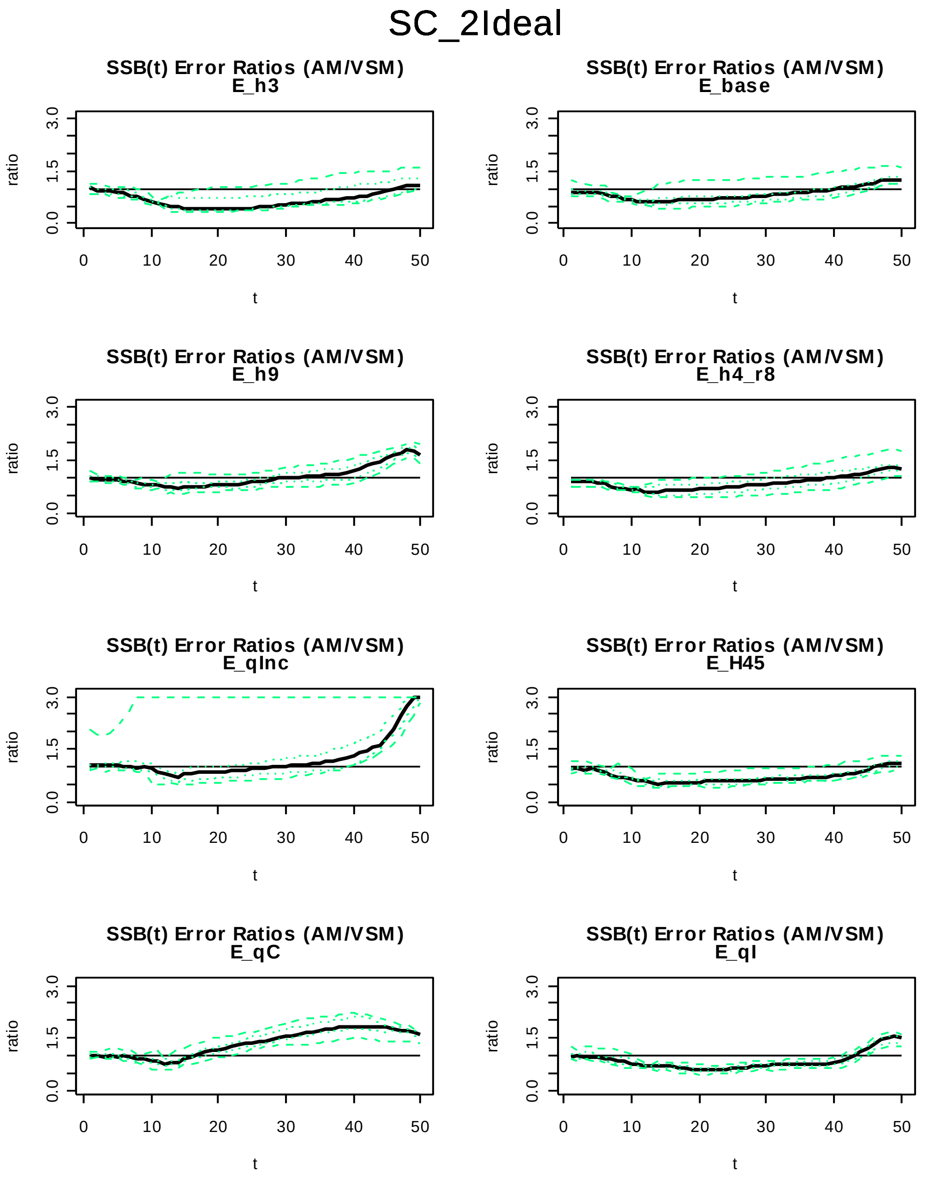

SC_2Ideal
SSB(t) Error Ratios (AM/VSM)
E_h3
SSB(t) Error Ratios (AM/VSM)
E_base
SSB(t) Error Ratios (AM/VSM)
E_h9
SSB(t) Error Ratios (AM/VSM)
E_h4_r8
SSB(t) Error Ratios (AM/VSM)
E_qInc
SSB(t) Error Ratios (AM/VSM)
E_H45
SSB(t) Error Ratios (AM/VSM)
E_qC
SSB(t) Error Ratios (AM/VSM)
E_qI
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.7-g

A 6.8 MF_YFT

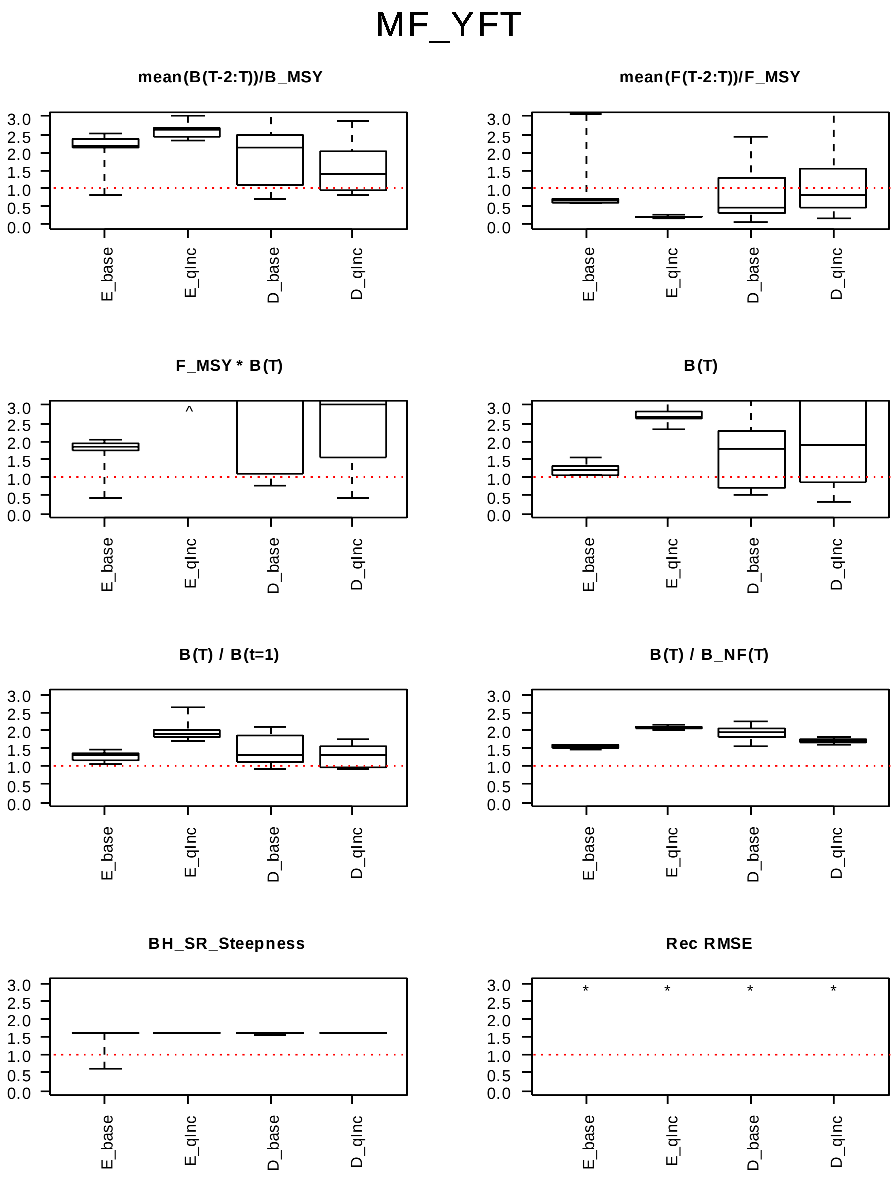


Figure A 6.8-a

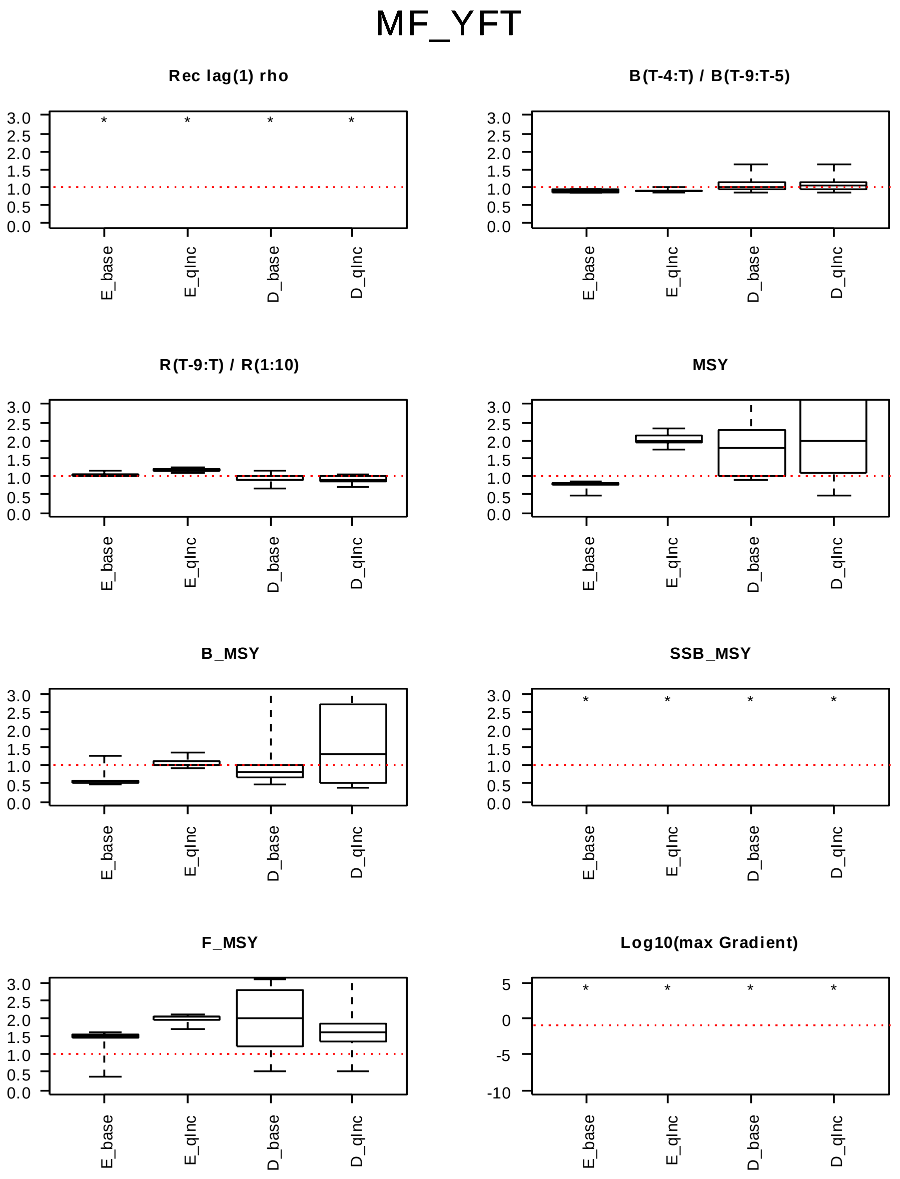

MF_YFT
Rec lag(1) rho
B(T-4:T) / B(T-9:T-5)
R(T-9:T) / R(1:10)
MSY
B_MSY
SSB_MSY
F_MSY
Log10(max Gradient)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5
0
-5
-10
*
E_base
E_qInc
D_base
D_qInc

Figure A 6.8-b

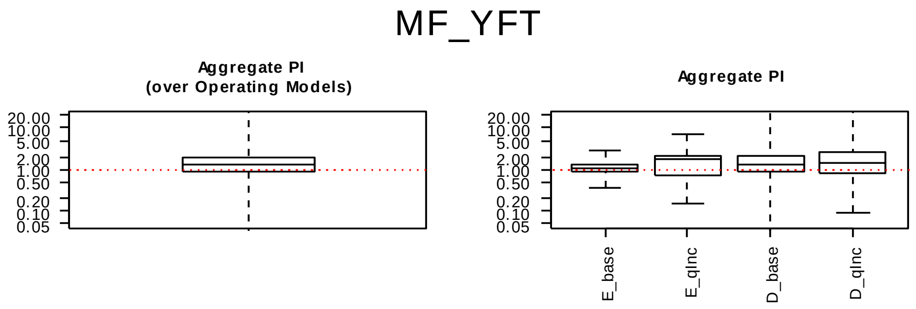


Figure A 6.8-c

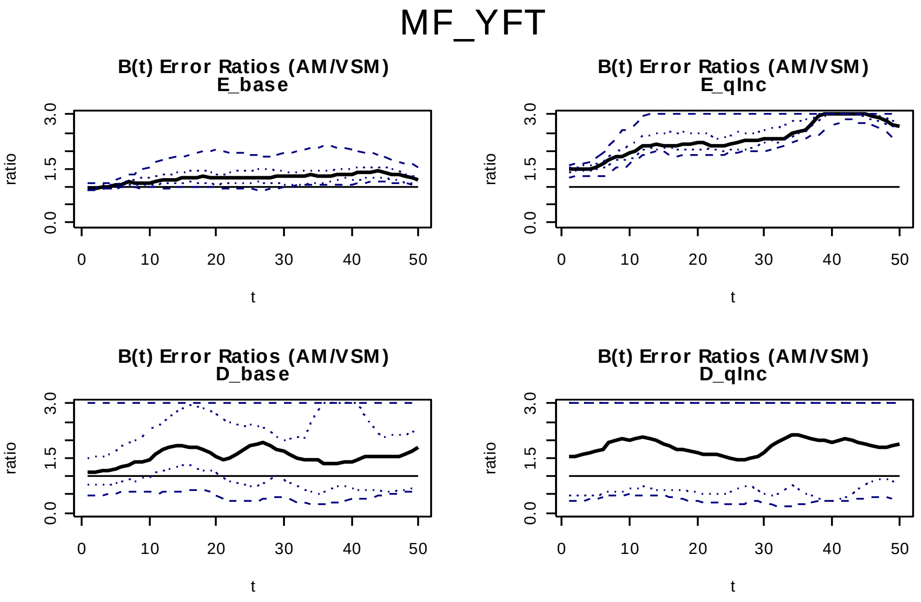


Figure A 6.8-d

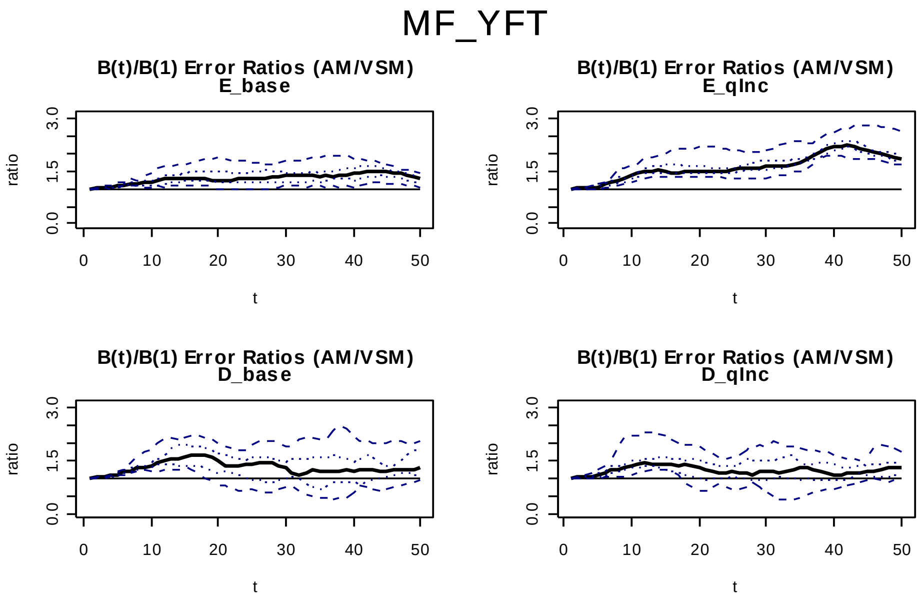


Figure A 6.8-e

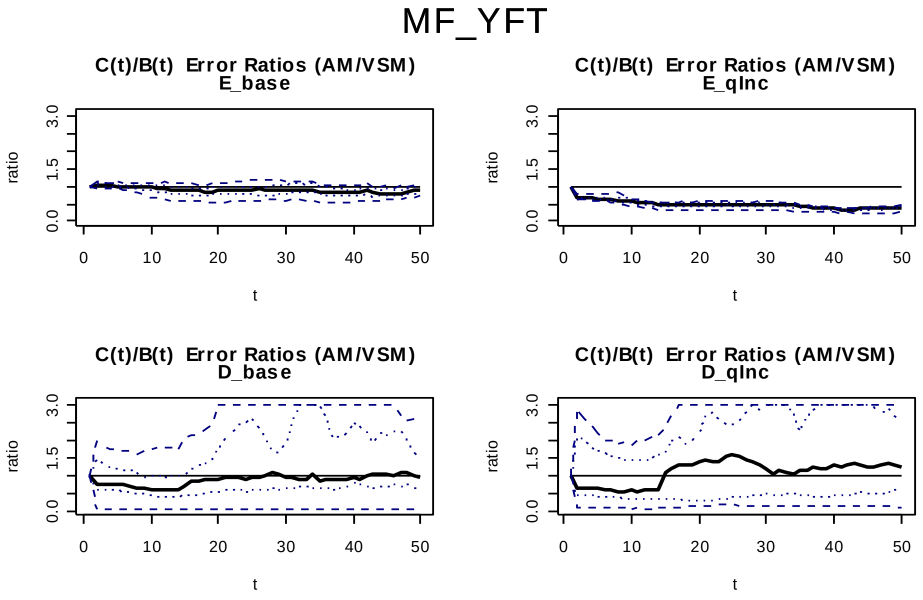


Figure A 6.8-f

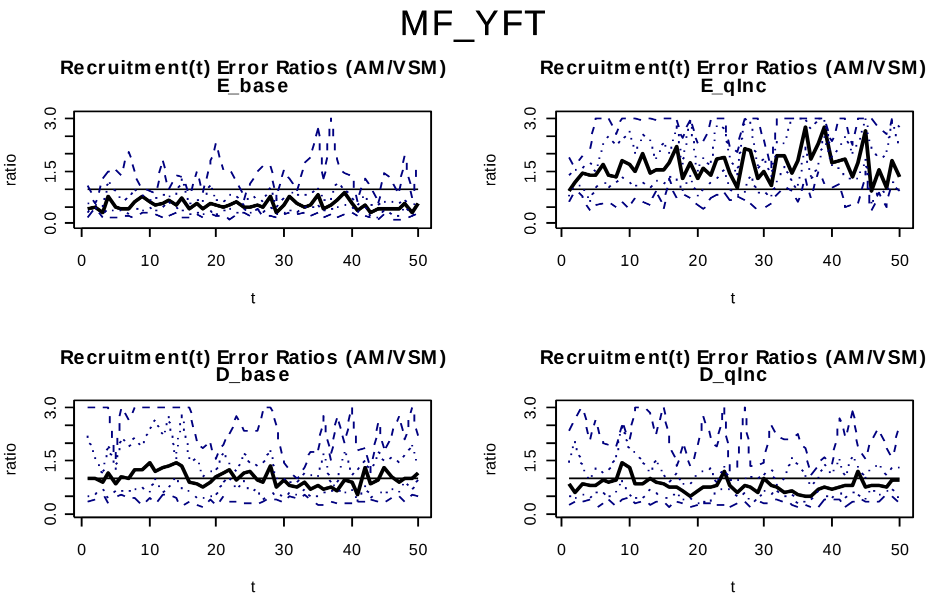

MF_YFT
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_qInc
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_qInc
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.8-g

A 6.9 MF_SCAN

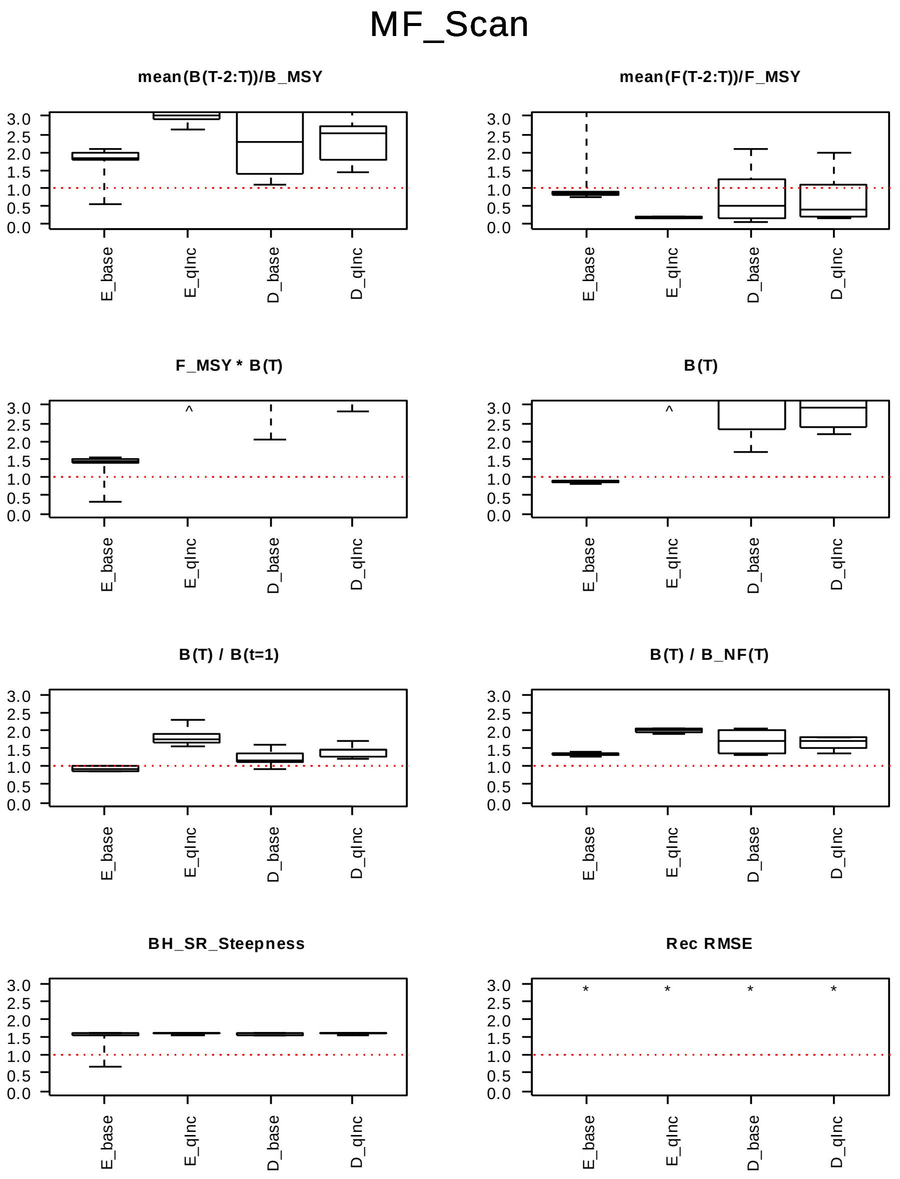


Figure A 6.9-a

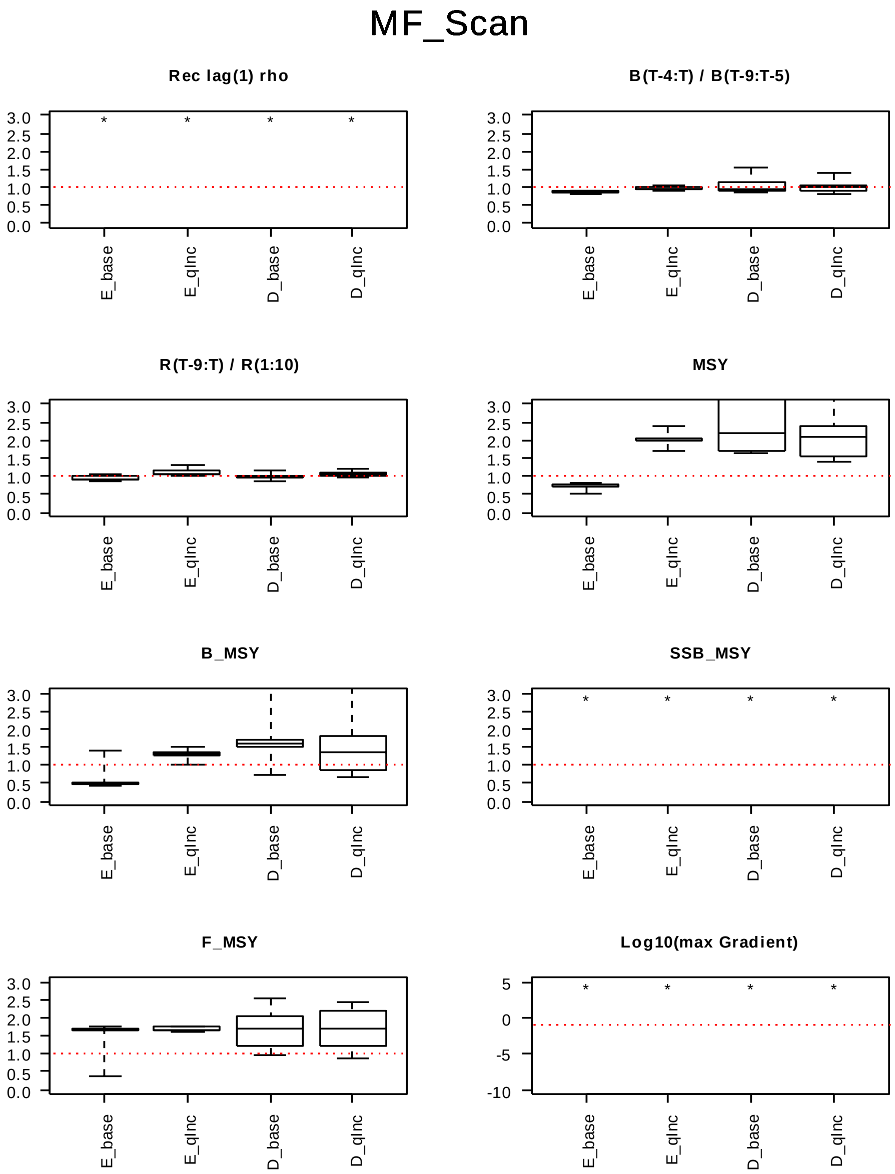

MF_Scan
Rec lag(1) rho
B(T-4:T) / B(T-9:T-5)
R(T-9:T) / R(1:10)
MSY
B_MSY
SSB_MSY
F_MSY
Log10(max Gradient)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5
0
-5
-10
E_base
E_qInc
D_base
D_qInc
*

Figure A 6.9-b

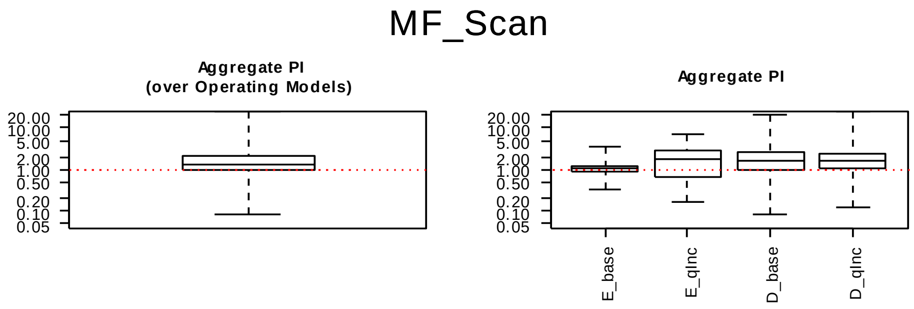


Figure A 6.9-c

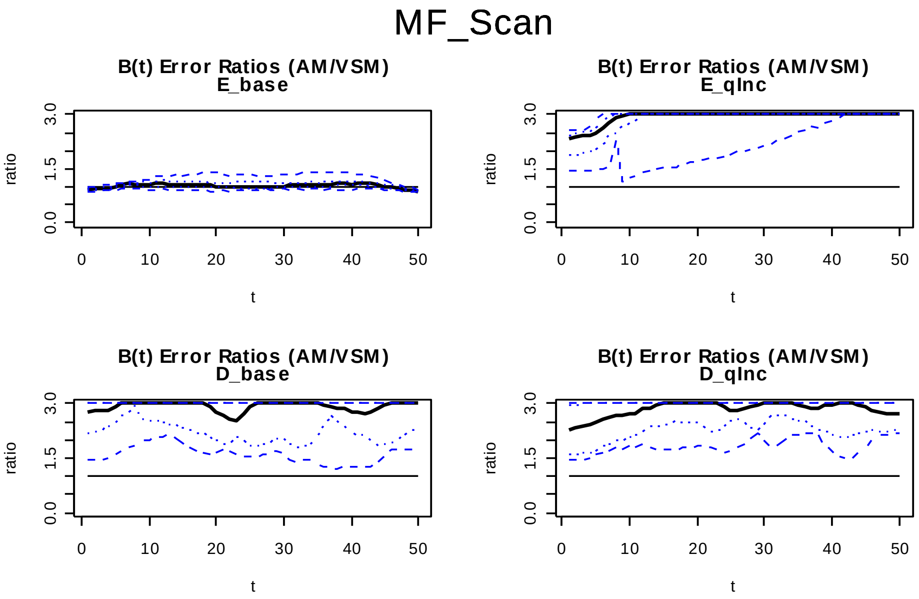


Figure A 6.9-d

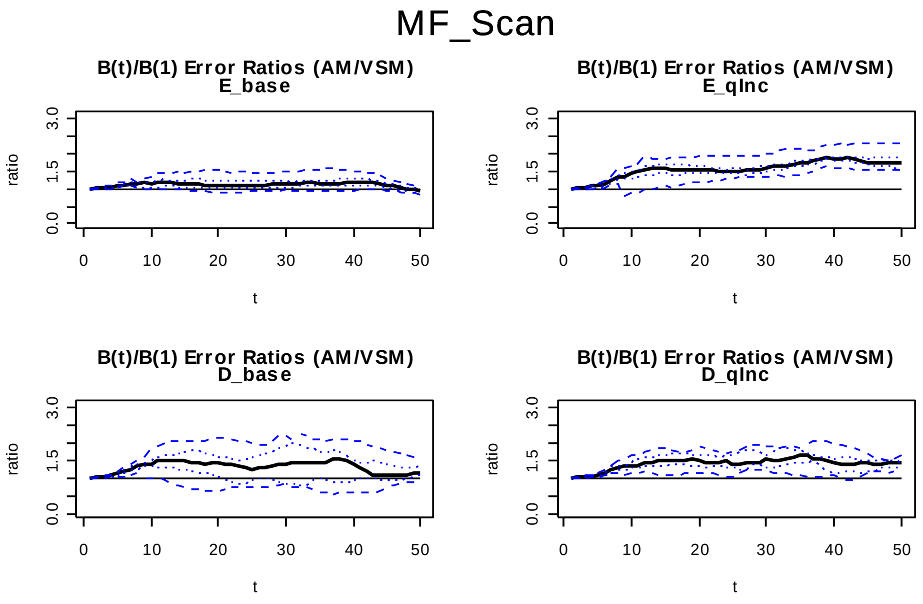


Figure A 6.9-e

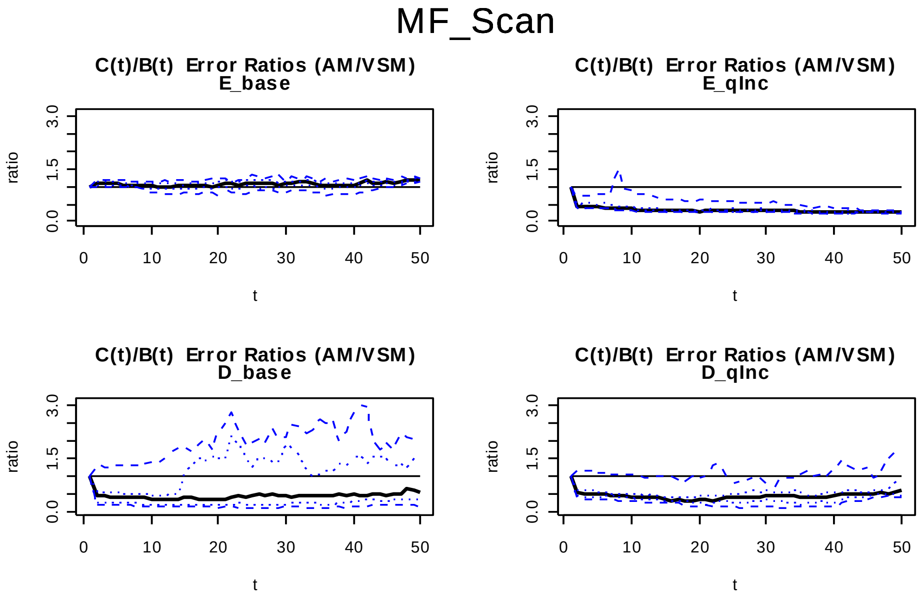


Figure A 6.9-f

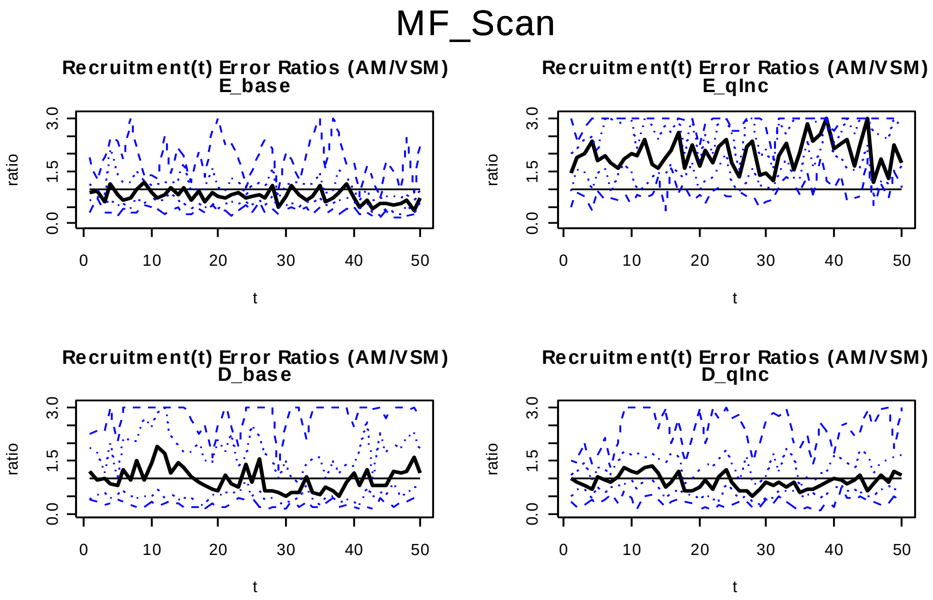

MF_Scan
Recruitment(t) Error Ratios (AM/VSM)
E_base
Recruitment(t) Error Ratios (AM/VSM)
E_qInc
Recruitment(t) Error Ratios (AM/VSM)
D_base
Recruitment(t) Error Ratios (AM/VSM)
D_qInc
ratio
0.0
1.5
3.0
t
0
10
20
30
40
50

Figure A 6.9-g

A 6.10 MF_qTS

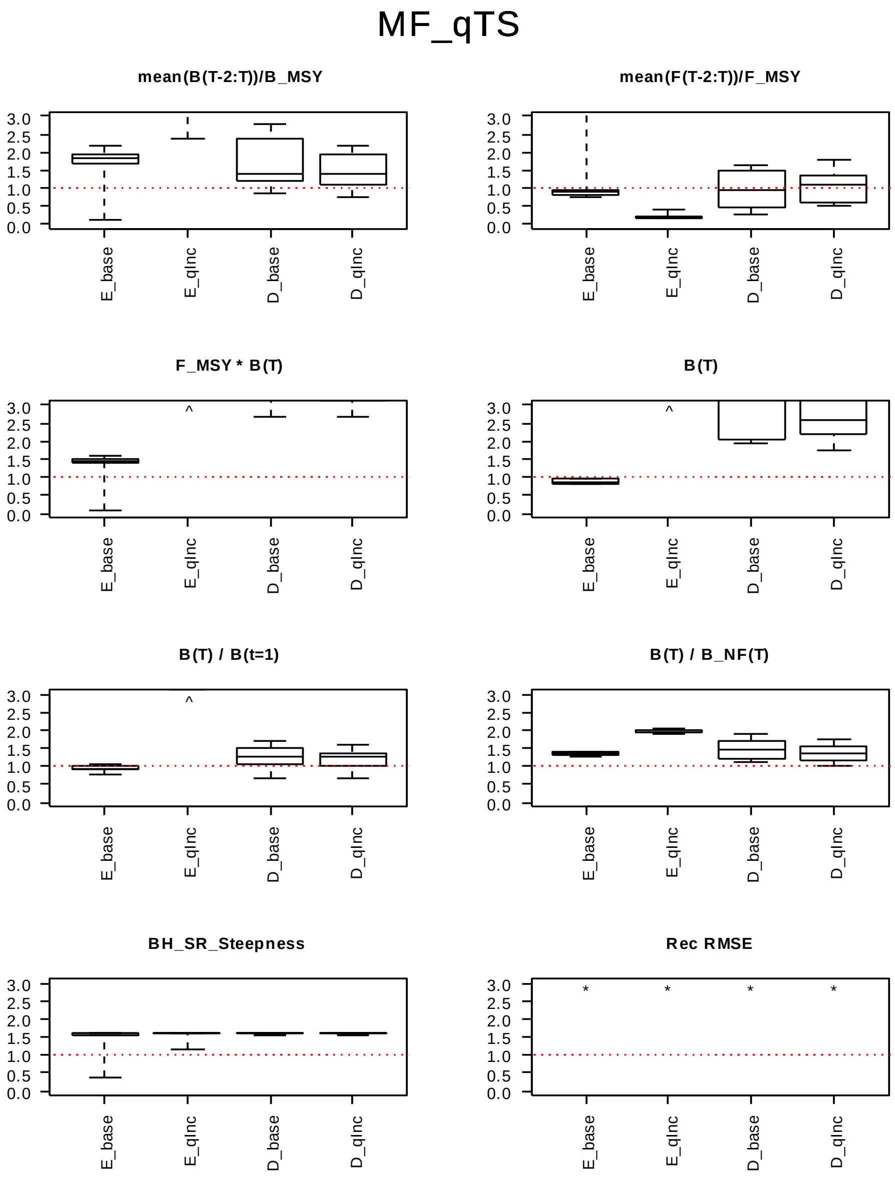


Figure A 6.10-a

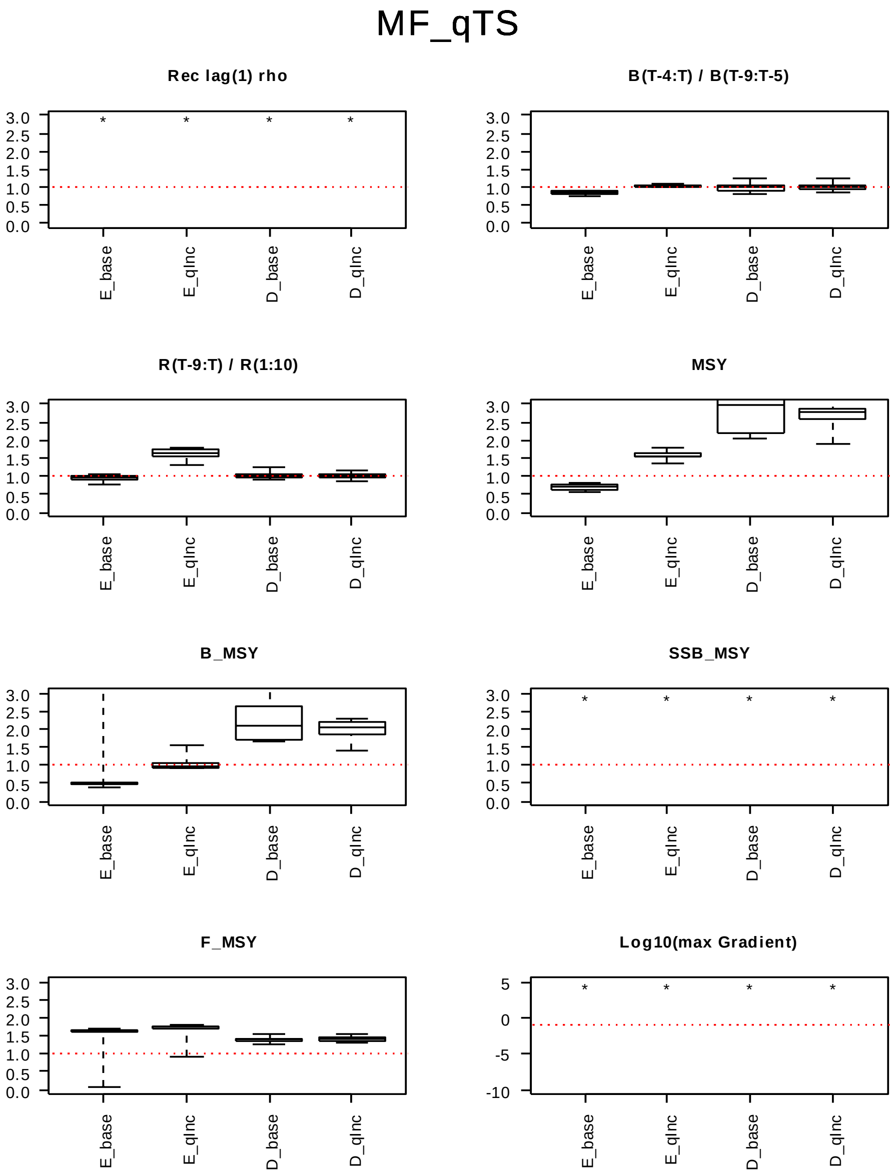


Figure A 6.10-b

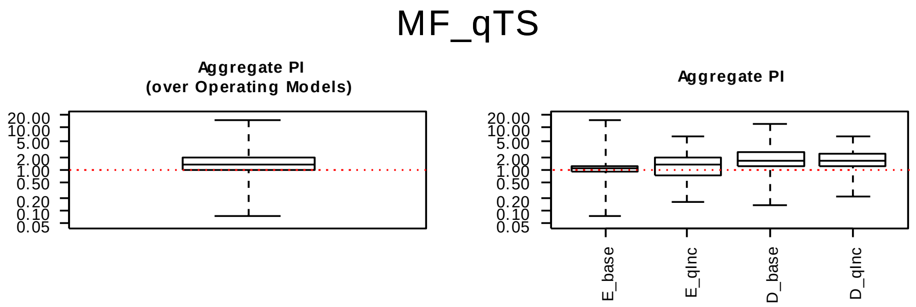


Figure A 6.10-c

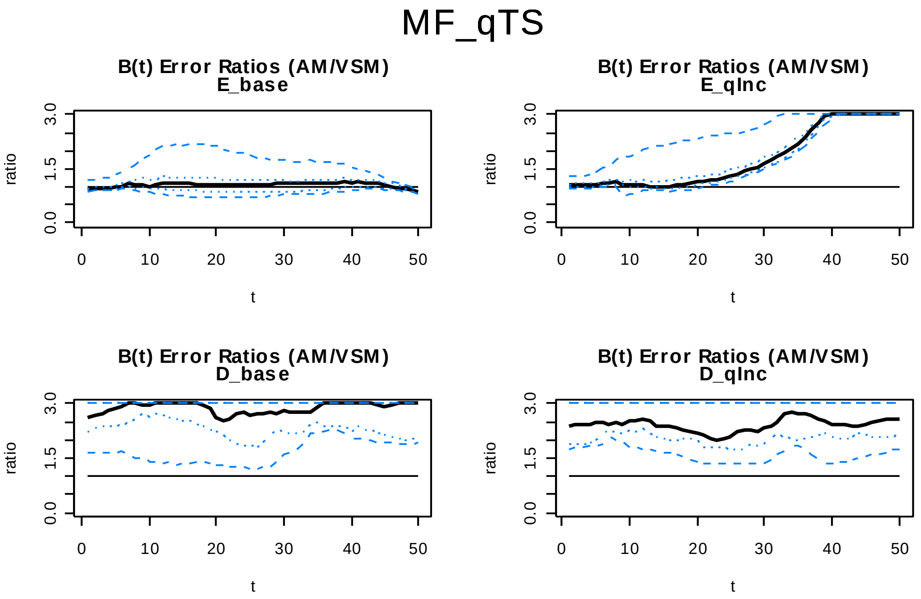


Figure A 6.10-d

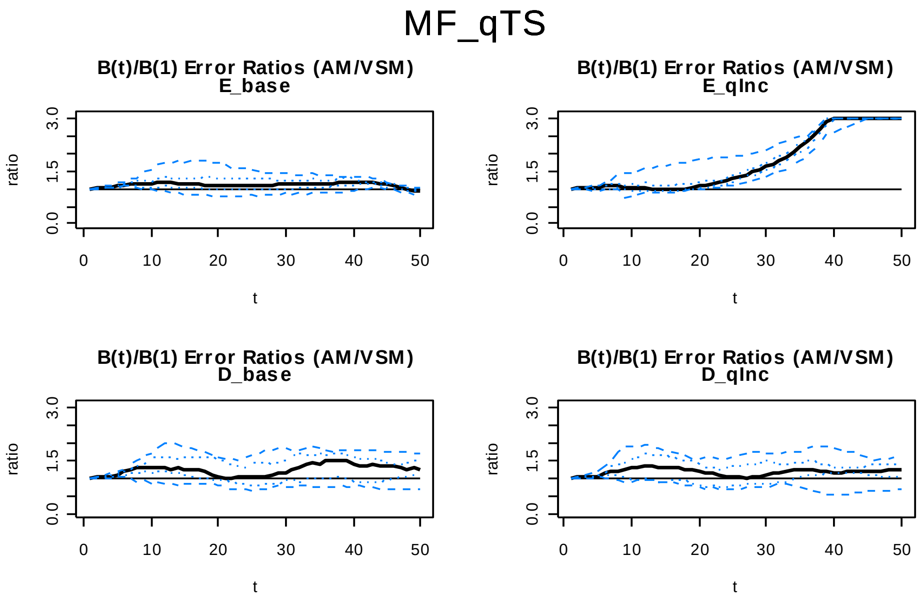


Figure A 6.10-e

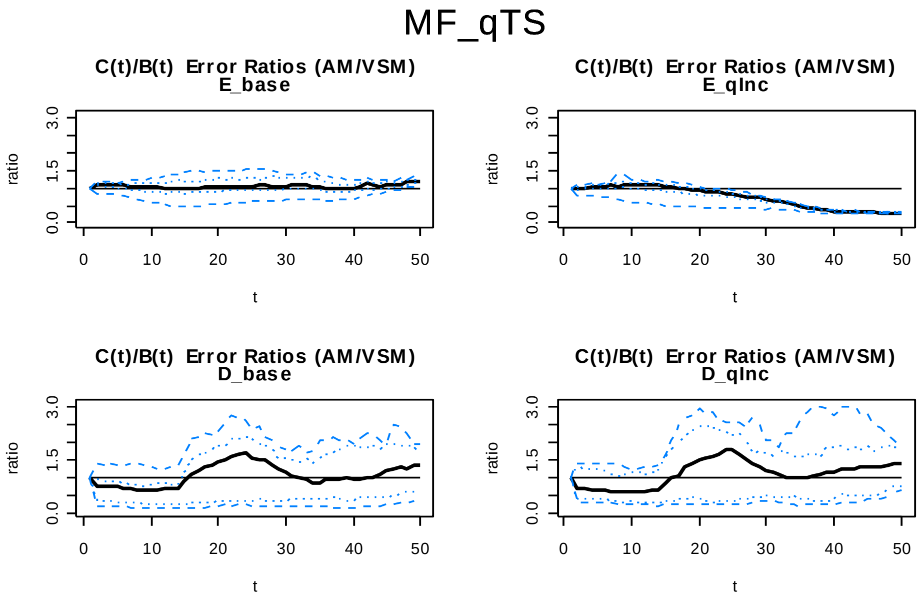


Figure A 6.10-f

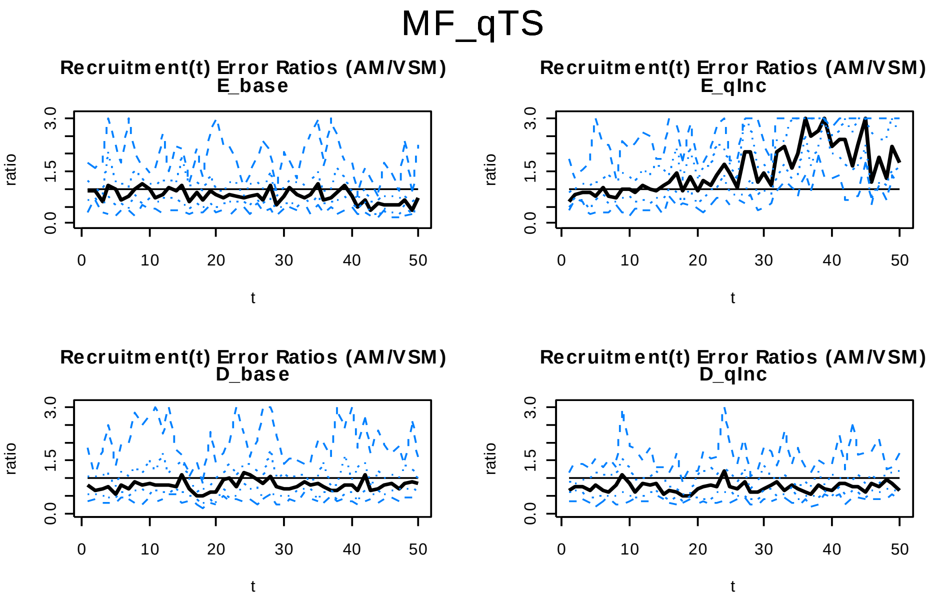


Figure A 6.10-g

A 6.11 BIH_2

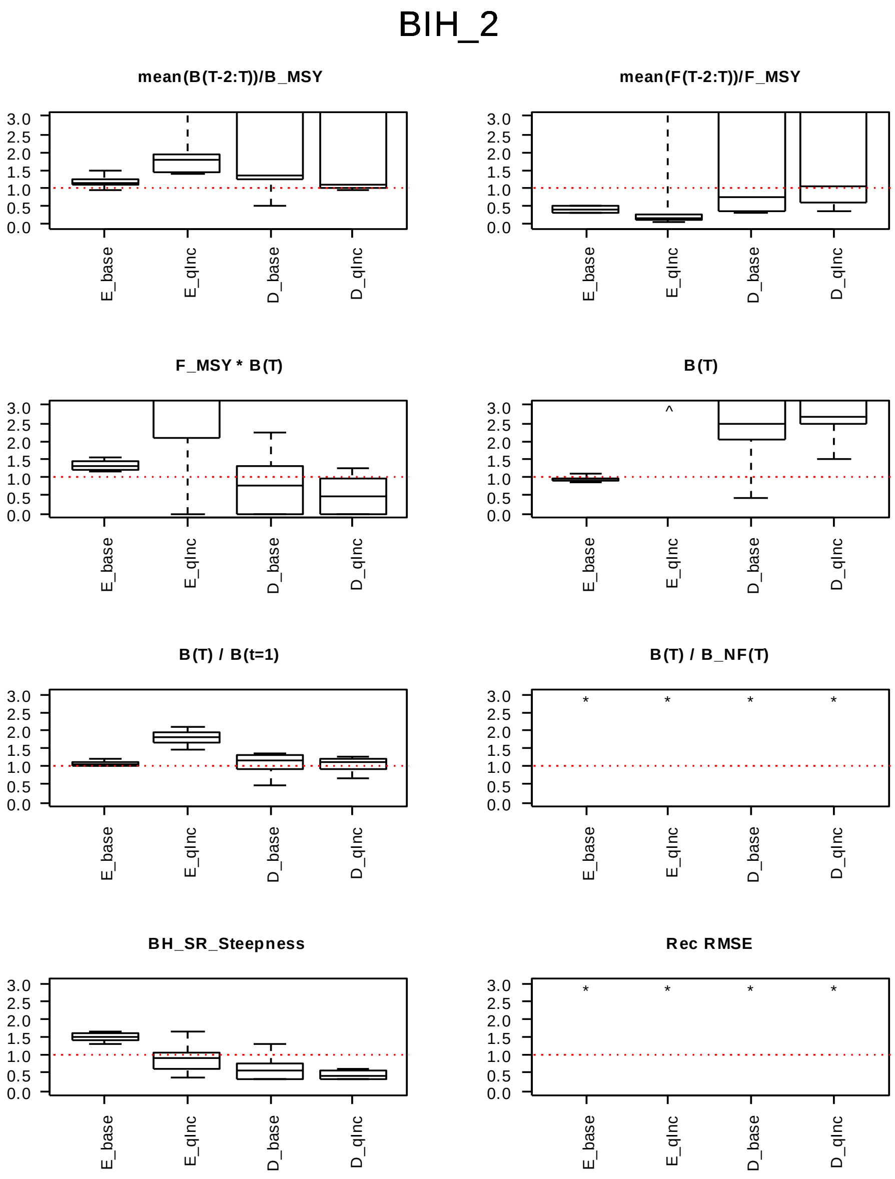


Figure A 6.11-a

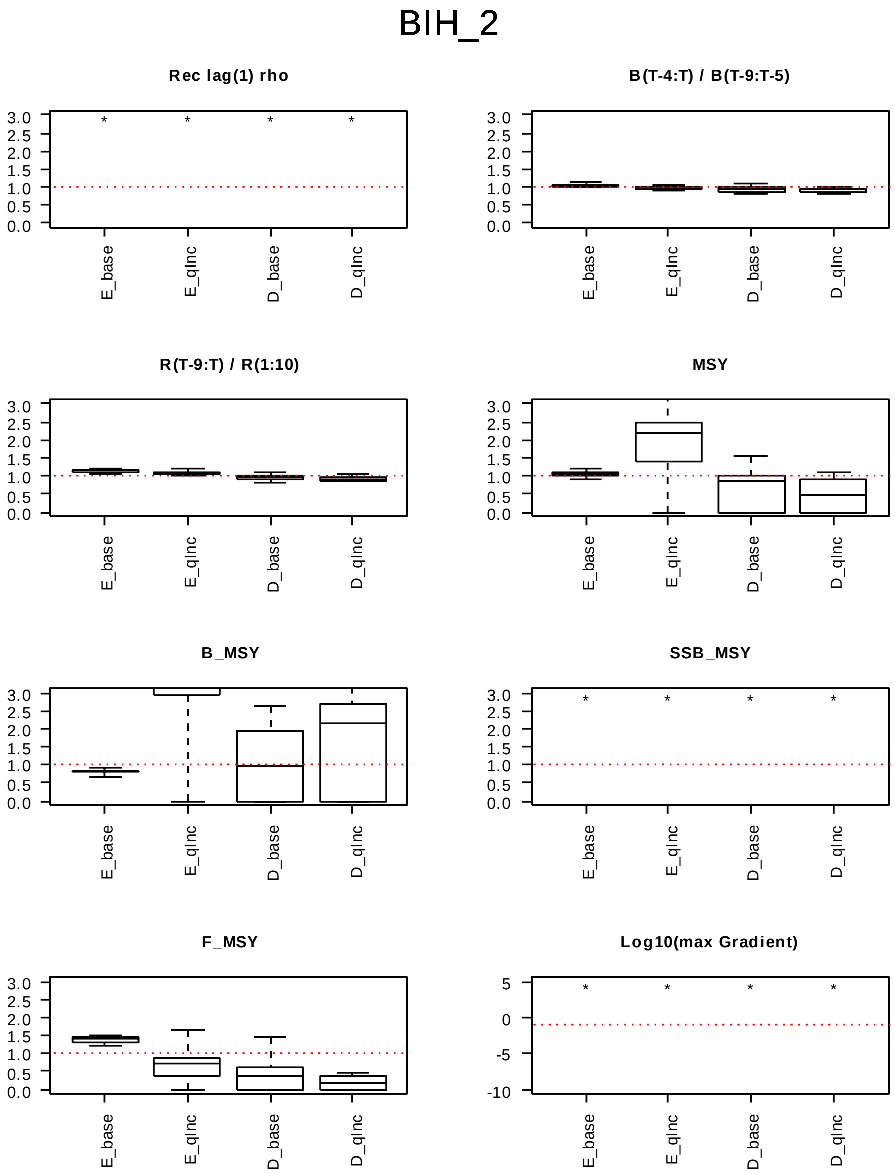


Figure A 6.11-b

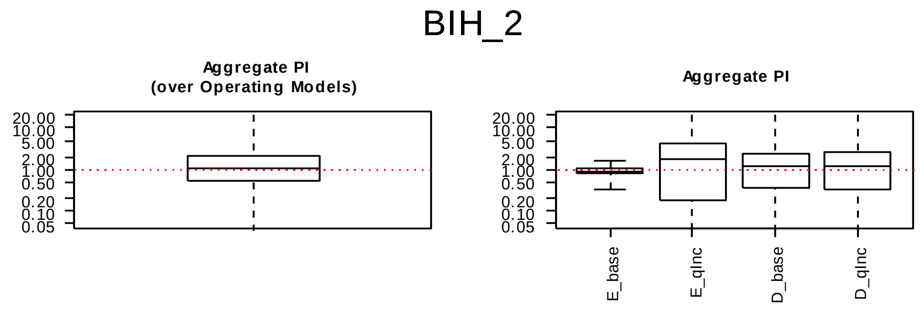


Figure A 6.11-c

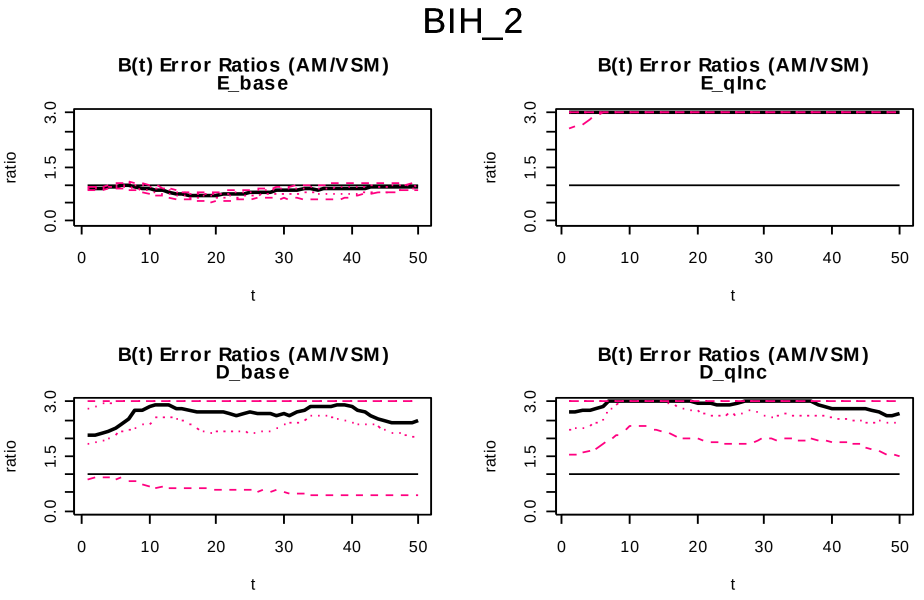


Figure A 6.11-d

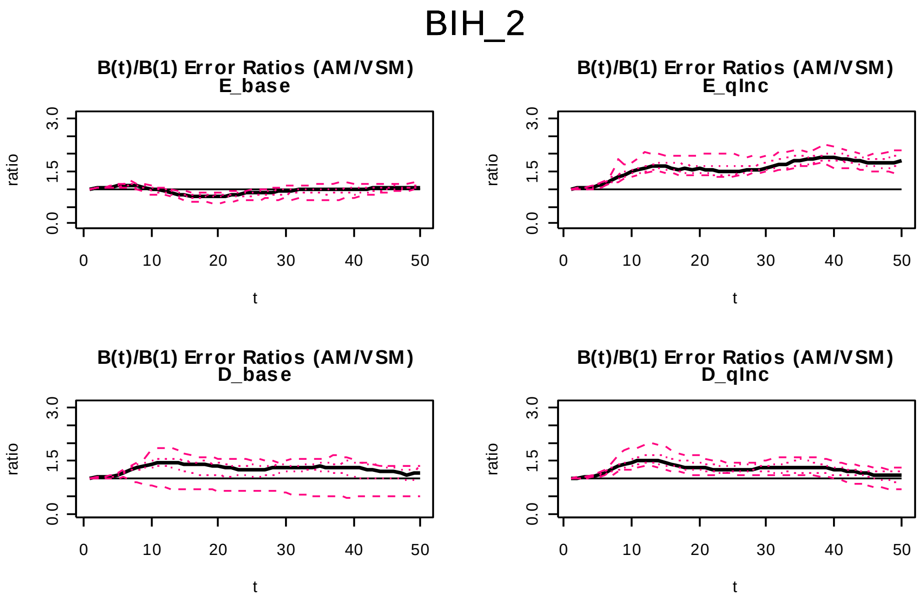


Figure A 6.11-e

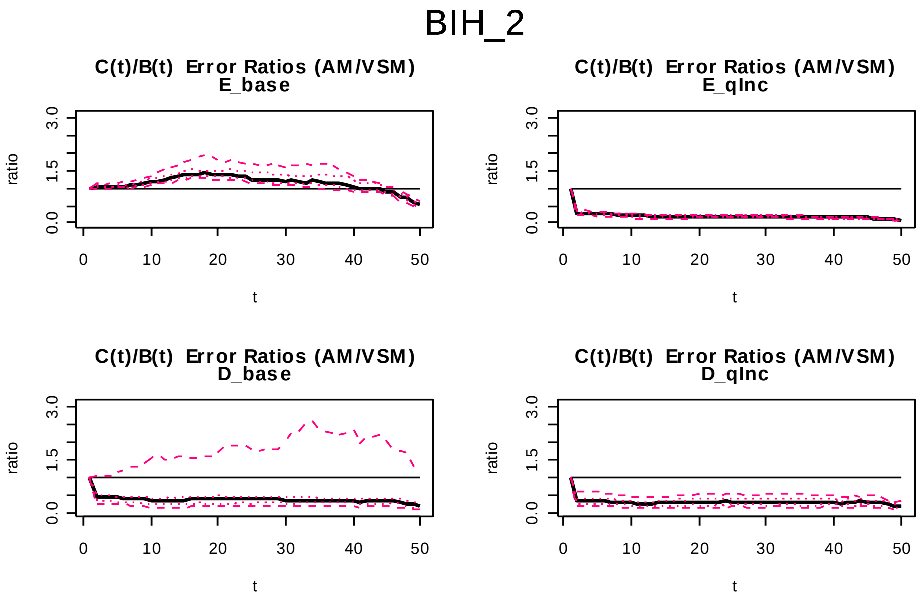


Figure A 6.11-f

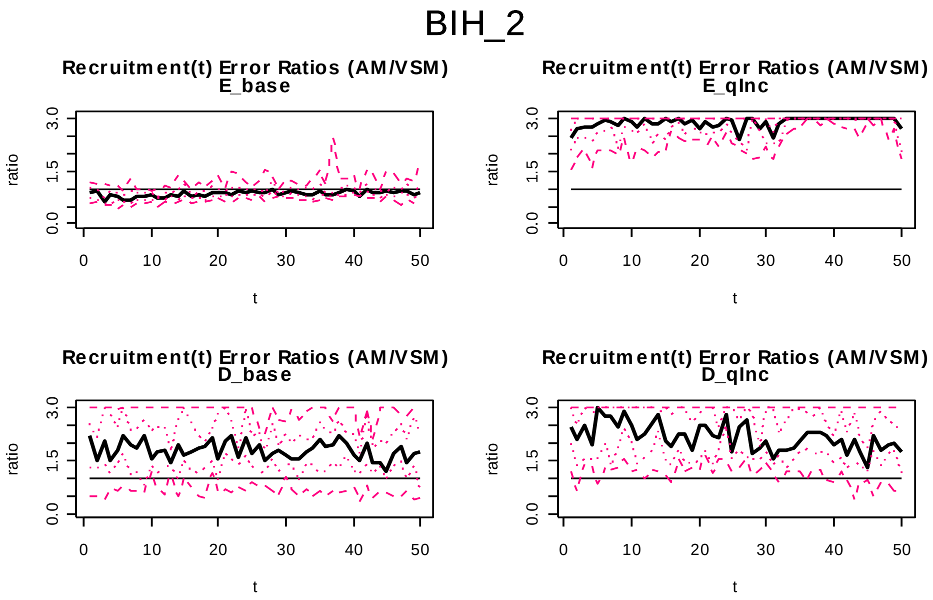


Figure A 6.11-g

APPENDIX 7 ACRONYMS USED IN THE SESAME REPORT

A-SCALA	Age-structured Statistical Catch-At-Length Analysis; a stock assessment model with many features of MULTIFAN-CL, independently implemented by the Inter-American Tropical Tuna Commission
ADAPT	the traditional stock assessment modelling approach used for the assessment of SBT
AAPM	Age-Aggregated Production Model; Fox or Schaefer in the SESAME study
AM	Assessment Model
ASPM	Age-Structured Production Model; an age-structured model that can be applied with minimal input data, but usually requires strong assumptions about natural mortality and selectivity that are derived from other analyses
CA	Catch-at-Age; pertaining to catch age frequency distributions
CCSBT	Commission for the Conservation of Southern Bluefin Tuna
CL	Catch-at-Length; pertaining to catch length frequency distributions
CPUE	Catch Per Unit Effort; catch rates generally assumed to have some relationship with fish abundance
MCMC	Markov Chain Monte Carlo – a stochastic method of approximating Bayesian posterior probability distributions
MPD	Maximum Posterior Density; pertaining to the parameter estimates at the mode of a likelihood-based objective function, which also includes additional constraints on the parameters that might be interpreted as Bayesian priors.
MP	Management Procedure; a procedure for making a management decision that is usually agreed before all the data on which it is dependent are available. The MP usually consists of data and a decision rule.
MSE	Management Strategy Evaluation; MP
MULTIFAN-CL	stock assessment modelling software designed especially for large highly migratory pelagic species and forming the basis for most WCPO tuna assessments in recent years
OM	Operating Model
PI	Performance Indicator; criteria used to compare assessment model estimates with actual values from the operating models
RFMO	Regional Fisheries Management Organization
SAG	Stock Assessment Group
SBT	Southen Bluefin Tuna
SC	Scientific Committee
SCALIA	Statistical Catch-at-Age/Length Integrated Analysis

SCTB-MWG	Standing Committee on Tuna and Billfish – Methods Working Group
SESAME	Simulation-Estimation Stock Assessment Model Evaluation
SPC-OFP	Secretariat of the Pacific Community – Oceanic Fisheries Programme
VPA	Virtual Population Analysis; a general term for population dynamics models
VSM	Virtual Stock Model; SESAME software developed to simulate fish and fishery dynamics, including data collection processes
WCPO	Western and Central Pacific Ocean
YFT	Yellowfin Tuna

APPENDIX 8 NON-TECHNICAL DESCRIPTION OF ASSESSMENT ISSUES FOR MANAGERS AND POLICY MAKERS

The proliferation of complicated stock assessment models (e.g. MULTIFAN-CL) has important implications for the manner in which stock assessment advice is provided to managers, and potentially changes the practical logistics of traditional stock assessment. The following non-technical summary provides an overview of major issues of concern with respect to the adoption of these models:

Integrated Assessment Models

Most modern stock assessment models attempt to describe the historical dynamics of the fish population, quantify the effects of fishing, and forecast the implications of future management actions. For the most part there have been few fundamental revolutions in the nature of stock assessment models or fisheries data over the past 20 years. But the proliferation of cheap computing power has led to continual incremental increases in model complication. The level of detail with which populations can be described has increased steadily, resulting in a more realistic representation of the fishery, potentially including:

- population age-structure
- spatial-structure and migration dynamics
- relationship between spawning stock size and recruitment
- temporal variability in fishery selectivity
- temporal variability in fishery catchability
- multi-species dynamics (predator-prey interactions)
- environmental effects on growth, mortality or migration

Statistical assessment models used for the pelagic fisheries (e.g. MULTIFAN-CL) now typically incorporate all the data related to:

- total catch
- catch length frequency samples
- catch age frequency samples
- standardized effort (or CPUE) as a relative abundance index
- tag releases and recaptures

These data can be dis-aggregated by fishing fleet and region if it is believed that this will provide a more appropriate analysis. Statistical sophistication is at the point that analyses attempt to:

- integrate all data into one framework rather than invoking several independent analyses, or excessively pre-processing the data (e.g. estimating lengths from ages)
- explicitly account for observation errors (e.g. fishery data sampling limitations)
- explicitly account for process errors (e.g. random variability in recruitment)

- estimate parameters based on the likelihood of observing the data given the model predictions
- estimate hundreds or even thousands of parameters simultaneously
- quantify the joint uncertainty in all parameter estimates (including reference points or other values derived from the parameters)
- quantitatively include auxiliary information (e.g. prior information based on experience in other systems)

However, worldwide there have been many important fishery assessment and management failures. This has focused attention on the limitations of mathematical models and the methods of communication of advice between scientists and managers. General issues of concern with the complicated models include:

- over-parameterization – there are limits to the type and number of parameters that can be reliably estimated for any given system. Too much freedom and the models might be fitting to noise. It is not trivial to identify what can be reliably estimated in many cases.

- model sensitivity – complicated assessment models inevitably include a number of arbitrary assumptions (e.g. related to the quality of the catch sampling, the relationship between effort and fishing mortality, etc.). The more complicated the model is, the greater the number of assumptions required. It is common for seemingly minor changes in model specifications to cause large changes in the estimated stock characteristics. With a large number of model components, it is difficult to identify which features should be examined in detail, particularly given that features interact in ways that are sometimes difficult to anticipate.

- uncertainty quantification – despite the statistical theory underpinning these models, there is often retrospective evidence that indicates we are less certain about the state of the stock than the uncertainty estimation suggests (e.g. our current assessment of the status of the stock 5 years ago is outside of the 95% confidence limits estimated using the same methodology 5 years earlier, and this happens much more than 5% of the time). It is less clear how well uncertainty quantification is likely to work when results are integrated over a range of plausible structural models, in part because methods for doing this are currently rather ad hoc. While desirable, this treatment of "model uncertainty" potentially opens the door for over-stating uncertainty.

- computational time – more complicated models take more time to develop and apply. This potentially limits the methods available with which to estimate parameter uncertainty conditional on the model being sufficiently "correct", and more importantly, limits the extent to which model specification uncertainty can be explored. The more complicated a model is, the more difficult it becomes to apply model fitting diagnostics, and understand the cause of problematic behaviour.

- technical expertise – assessment modellers need to be sufficiently experienced to recognize potential problems in model application. This is particularly a concern when employing complicated software developed by a third party.

Policy Implications for CCSBT and other RFMOs

1) For economically important fisheries with a rich data history, we expect that sophisticated integrative assessment models will play an important and increasing role in the provision of advice to managers in the major international pelagic fisheries worldwide (and other fora). However, as indicated below, these models do not represent a panacea for obtaining accurate stock status estimates, and will not eliminate the need for difficult management decisions.

2) Considerable uncertainty is inevitable in current methods of stock assessment. It is important that managers and assessment scientists 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). If management decisions are going to be based on the "best" point estimates of a stock assessment, it seems that relatively simple models often do provide inferences that are of comparable quality to the estimates provided by a single plausible specification of a complicated integrative model (unless perhaps if the data are exceptionally informative). However, with increasing recognition of the performance limitations and sensitivities of assessment models, and a shift in emphasis toward uncertainty quantification, it seems that the simple assessment models do not have sufficient structural flexibility for exploring model uncertainty.

3) Assessment scientists and managers should work together to identify methods for managing the fishery that are robust to the major foreseeable uncertainties. At the simplest level, this might involve decreasing the emphasis on advice pertaining to quantities that cannot be estimated reliably (e.g. MSY), to quantities that are generally better estimated (e.g. biomass relative to some historical point in time). At a more sophisticated level, 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 constant catch or effort in future projections). MPs are also evaluated using performance measures that should be readily defined from management objectives (whereas assessment model evaluation might include many estimators that are irrelevant, depending on the type of management decisions that are ultimately made). In an MP context, the complicated assessment models 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 should be no need for a comprehensive application of the complicated integrative models every time that a management decision is made. 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.

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 judgements into the process and potentially constrain the range of options under consideration inappropriately.

5) Managers 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).

6) The quest to achieve creative solutions that optimize management objectives and are robust to the major uncertainties about stock dynamics (using MP development or other sophisticated modelling methods) will usually 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.

7) 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 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 recognize that there is a risk of going too far and over-stating uncertainty. This could lead to over-pre-cautionary management and loss of reasonable economic opportunity. Identifying the appropriate balance in uncertainty quantification will be a major challenge.

8) There is likely to be continued pressure for scientists to provide increasingly complicated advice to managers in the future (e.g. with respect to ecosystem management and multiple stakeholder objectives, etc.). We expect that operating models will play an even more important role in the future as these models provide a means of evaluating assessment models and MPs. It is important to regularly conduct simulation testing tailored to the specific situations of interest, to maintain an understanding of our quantitative limitations as models and management objectives develop in new directions.

9) 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 could probably underpin many of 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).

10) We recommend that peer review should play an important part in stock assessment and management. This scrutiny provides an external view that can help to prevent a consistent group of stock assessment participants from becoming focused on a subset of issues at the expense of ignoring others. Similar review of fisheries managers should be employed to ensure that they are using the most effective methods for soliciting and acting on scientific advice.

11) As assessment models become more complicated, there will be an increasing tendency to rely on pre-existing software to decrease development time. This is advantageous in that the combined efforts of several developers can make software more widely applicable, and more robust to coding errors. However, we caution that this might occur at the expense of stifling development of specific features deemed to be of importance for any particular system by novice users. MULTIFAN-CL is currently poised to dominate pelagic fisheries stock assessment, and its rich flexible structure represents a good starting point. However, we would like to ensure access to the source code before embracing it for a major project.

APPENDIX 9 LIST OF WORKING PAPERS ARISING FROM THE SESAME PROJECT

Kolody, D. 2002. SCALIA: application of an integrated analysis stock assessment model to the 2002 SCTB methods working group simulated tuna fishery data. Meeting of the standing committee on tuna and billfish 15, Working Paper MWG-5.

Kolody, D., A. Preece, D. Ricard, P. Jumpannen, T. Jones, S. Cooper and T. Polacheck. 2003. Progress on a simulation study to evaluate stock assessment models for fisheries resembling southern bluefin tuna. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0209/29.

Kolody, D., and D. Ricard. 2003. Application of SCALIA and production models (Fox, Schaefer, and age-structured) to the SCTB MWG 2003 simulated tuna fishery data. meeting of the standing committee on tuna and billfish 16, Working Paper MWG-5.

Kolody, D. and P. Jumpannen. 2003. SCALIA simulation-estimation study results relevant to CCSBT management procedure development. Commission for the Conservation of Southern Bluefin Tuna doc. CCSBT-SC/0304/10.

Ricard, D. and D. Kolody. 2002. Application of production models to the assessment of the SCTB-MWG simulated tuna fishery data. Meeting of the standing committee on tuna and billfish 15, Working Paper MWG-8.

CSIRO Marine Laboratories

CSIRO Marine Research

Headquarters
Castray Esplanade, Hobart, Tasmania 7001
GPO Box 1538, Hobart, Tasmania 7001, Australia

Queensland Laboratory
133 Middle Street, Cleveland, Queensland 4163
PO Box 120, Cleveland, Queensland 4163, Australia

Western Australia Laboratory
Centre for Environment and Life Sciences Building
Underwood Avenue, Floreat, W.A. 6014, Australia

ISBN 0 876996 64 1
ISSN 0725–4598