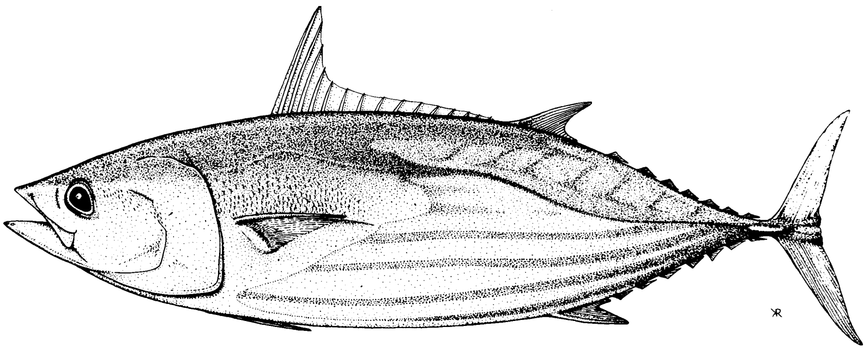
17th Meeting of the Standing Committee on Tuna and Billfish

SCTB 17
Majuro, Marshall Islands
9-18 August 2004

SCTB17 Working Paper

MWG–4(rev)

Comparison of stock assessment methods using an operational model



John Sibert

Pelagic Fisheries Research Program. JIMAR. University of Hawaii. USA.

July 2004

Comparison of stock assessment methods using an operational model

John R. Sibert
Pelagic Fisheries Research Program
Joint Institute for Marine and Atmospheric Research
University of Hawai'i at Manoa, Honolulu, HI 96822, U.S.A.
sibert@hawaii.edu

August 2004

Abstract

The results of the OFP simulation exercise to evaluate stock assessment models are reanalyzed. Standardized bias is used to evaluate the discrepancy between simulation and estimation and to provide a quantitative expression of estimation skill. Seven assessment models were applied to forty realizations of five fishery scenarios of increasing complexity. Production models performed well because of the exclusion of uninformative data in complex scenarios. MULTIFAN-CL performed well when data were informative and natural mortality correctly specified. Simulation exercises of this complexity do not definitively address the question of assessment model reliability.

1 Introduction

Stock assessment scientists are often asked the question "How do you know your model works?" The use of realistically complex stochastic fishery simulations, or "operational" models, to generate "data" on which to test assessment models is a commonly used approach to this question (NRC, 1998, Kolody *et al.*, 2004a). Model parameters, time series and reference points estimated by the assessment model are compared to corresponding operational model (OM) input parameters and output variables. This approach tacitly assumes that the output from the operational model is "truth" and that divergence between the operational and assessment models reveals a problem with the assessment model. Multiple stochastic simulations using the same suite of biological and fishery parameters, dubbed "realizations", are considered replicate samples of a universe of possible fishery outcomes. Data observed from a real fishery are considered to be a single sample from this universe.

Labelle (Labelle, 2003) conducted extensive tests of MULTIFAN-CL (MFCL) using a complex stochastic operational model for yellowfin tuna derived from a previous operational model for swordfish (Labelle, 2002). The SESAME project was undertaken by CSIRO Marine Research group to test assessment models for southern bluefin tuna using operational models (Kolody *et al.*, 2004a) and also included the OFP OM output. This paper attempts to synthesize the OFP results and the SESAME results and includes some results from applying the ASCALA model (Maunder and Watters, 2003) used at the Inter-American Tropical Tuna Commission. Rather than simply repeat the previous results a slightly different analytical approach is employed, and some aspects of the MULTIFAN-CL results are examined in greater detail.

The work described here is roughly equivalent conducting two hundred different stock assessments with each method. The following individuals contributed their time and expertize to this project: Marc Labelle, John Hampton (Secretariat of the Pacific Community), Mark Maunder (Inter-American Tropical Tuna Commission), Dale Kolody, Daniel Ricard, Jason Hartog (CSIRO Division of Marine Research), and Pierre Kleiber (NOAA Pacific Island Fisheries Science Center). Although this analysis could not have been accomplished without their hard work, the analytical approach and conclusions reached are the responsibility of the author.

2 Methods

2.1 Operational Model

The OFP (Labelle, 2003) simulations are intended to mimic fisheries for yellowfin tuna. The five model scenarios range in complexity from a single fleet or fishery operating in a single homogeneous region to a scenario with 16 different fleets operating in seven regions with movement of fish between regions. Forty different realizations of each scenario were computed and analyzed by each assessment model. All operational model output can be obtained by ftp from `ftp.soest.hawaii.edu/PFRP/MWG`.

2.2 Assessment Models

Seven different stock assessment models were tested. MULTIFAN-CL (Hampton and Fournier, 2001) was applied by Marc Labelle and John Hampton as presented at SCTB16 (Labelle, 2003). Some aditional analyses were conducted by the author using the version of MULTIFAN-CL obtained from the OFP web site `http://www.multifan-cl.org/`. ASCALA (Maunder and Watters, 2003) was applied by Mark Maunder. Dale Kolody and colleagues at CSIRO Marine Research applied SCALIA and four production model variants (Kolody *et al.*, 2004b).

2.3 Analytical Approach — Standardized Bias

Previous analysis of the OFP/MWG simulations used ratios between estimated and OM values to compare the results (Labelle, 2003, Kolody *et al.*, 2004b). In this paper, standardized bias is used to expresses the error in the estimate in relation to the between-realization standard deviation. It allows comparison between variables measured in different units and provides an intuitive means to evaluate the significance of the error. Standardized bias, however, tends to obscure the range of variables with intrinsically high inter-realization variance.

Let i index realizations; $i \quad i = 1, 2, \ldots, 40$. Let y_{ij} be the value of a variable from the OM and $\widehat{y}_{ij}$ be the corresponding estimate from the assessment model. Then bias, $b_{ij} = \widehat{y}_{ij} - y_{ij}$ so that negative values of b_{ij} are indicative of under estimates of the simulated variable. The average bias $\bar{b}_j = \frac{1}{40}\sum_i b_{ij}$ is the average bias over all scenarios; $\bar{b}_j = 0$ for an unbiased estimator. The variance of the bias is $s_j^2 = \frac{1}{40}\sum_i (b_{ij} - \bar{b}_j)^2$. The standardized bias is thus $\widetilde{b}_{ij} = \frac{b_{ij}}{s_j}$. Values of the subscript, $j = 1, 2, \ldots, N$, depend on the the variable under consideration. For time-series variables such as biomass, $N = 148$, four observations per year for 37 years. For age-dependent variables, such as mortality, $N = 20$ quarterly age classes. For scalar variables, such as B_{msy}, $N = 1$.

Standardized bias plots are generally presented in an array of seven rows by five columns. Each row is a different assessment model and each column is a fishery scenario. The range of the standardized bias plots is ± 5 standard deviations. Zero bias is indicated by a heavy dashed (green) line and ± 2 standard deviations are indicated by solid (red) lines. Individual realizations are indicated by solid black lines. For example see Fig. 1.

Standardized bias offers a convenient means of computing a simple measure of the "skill" of each assessment model to recover OM variables. The proportion of the total estimates with $|\widetilde{b}_{ij}| < 2$ can easily be computed for each combination $\times$ scenario combination. For an accurate, unbiased estimator approximately 95% of the estimates should fall within this interval.

A concise summary of the properties of the estimates of a suite of time-series variables and parameters from 200 stock assessments is nearly impossible. The results of the SESAME project (Kolody *et al.*, 2004a) runs to more than 400 pages. This presentation will emphasize the general skill of the assessment methods to recover biomass-related variables, explore the differences between scenarios, and examine aspects of the performance of MULTIFAN-CL in more detail.

3 Results

The results of the standardized bias analysis are generally consistent with the results presented by at SCTB16 (Labelle, 2003) and by the CSIRO group (Kolody *et al.*, 2004b). The results from this analysis are simply presented in the figures and tables. The reader is referred to (Labelle, 2003) and (Kolody *et al.*, 2004b) for more detailed comment on the general results.

The average skill level for estimating various measures of total biomass appeared to increase with scenario complexity, approaching 0.8 in some cases. Refer to the "Average" row in tables 1, 2, 3. The use of biomass ratios such as $B_{i,j}/B_{i,0}$ or $B_{i,j}/B_{i,msy}$ improves the skill level from some models but degrades the skill level in others.

There appears to be a systematic discrepancy between the operational model and MULTIFAN-CL with respect to natural mortality. M is underestimated in all scenarios for most ages (Fig. 7). Constraining M to be constant at the value used in the OM substantially improves the MULTIFAN-CL estimates of biological reference points and several measures of biomass for scenario 1 (1Fx1R), Fig. 8. In contrast, constraining M in scenario 2 (2Fx1R) produced only minor improvements in the MULTIFAN-CL estimates (Fig. 9).

Models are most instructive when they fail, and MULTIFAN-CL arguably failed in the 2Fx1R scenario. The MFCL skill level for total biomass estimates in this scenario is approximately 0.01. Estimates of total biomass, adult biomass and recruitment are lower than simulated values by approximately two standard deviations for most realizations over the entire time series (Figs. 1, 5, 6) Scaling total biomass by B_0 and B_{msy} increases the skill level to 0.2 and 0.3 respectively (Tables 2, 3).

Fishery 1 in both scenarios 1 and 2 is meant to simulate a longline fishery. However, the history of this fishery, Fig. 14, differs from observed longline fisheries by having high catch and high CPUE at the beginning of the time series. The commonly observed longline pattern is high CPUE at low catch at the beginning of the time series (Hampton and Kleiber, 2003). Fishery 2 in scenario 2 is meant to simulate a purse seine fishery. It starts ten years after the longline fishery with high CPUE and low catch (Fig. 15).

Figure 13 shows the relationship between operational model biomass and CPUE for scenario 2. The CPUE obviously reflects biomass for fishery 1, but not for fishery 2. Further, fishery 2 does not span the full range of biomass. There is thus little information about the stock and limited contrast in the fishery 2 data. The failure of MULTIFAN-CL to accurately assess the stock under scenario 2 is apparently due to uninformative data from one out of two the fisheries. The other two integrated models, ASCALA and SCALIA performed well on scenario 2.

The skill of MULTIFAN-CL to accurately estimate fishery parameters seems to increase with the complexity of the scenarios. This trend may be due to introduction of additional information in the data and to the structural assumptions of the model.

The apparent high skill of the Schaefer and Fox models is superficially surprising. However compromises were made with the data in order to utilize these models on scenarios with more than one fishery. Either (1) the nominal CPUE from one of the largest longline fisheries or (2) aggregated CPUE from all longline fisheries was used for these models (Kolody *et al.*, 2004b). In other words, uninformative data were either discarded or masked by averaging with more informative data.

4 Conclusions and Recommendations

The question "How do you know your model works?" cannot be definitively addressed though the use of complex simulations. When discrepancies are noted, it is difficult to determine whether the fault lies with the operational model or with the assessment model. Both types of models are abstractions of the natural world, and we do not know whether either is correct. True model verification depends on the testing of model prediction against empirical data.

In the clarity of hindsight, the important findings from the OFP simulation exercise could have be obtained by a much modest exercise. The discrepancy in natural mortality between the operational model and MULTIFAN-CL was evident in the simplest scenario, and could have been addressed earlier in the exercise. Similarly the problem of uninformative “purse seine” data became apparent in the second scenario. Since these two problems are both unresolved and not understood, it is impossible to understand anomalies from the more complex scenarios.

Finding: There is a discrepancy between the operational model and MULTIFAN-CL in the implementation of natural mortality. This discrepancy produces substantial bias in the estimates of biological reference points and biomass and recruitment time series with respect to the operational model. The nature of this discrepancy is unknown.
Recommendation: Develop a simplified simulator designed to elucidate the expression of natural mortality.

Finding: MULTIFAN-CL performs poorly when data from a substantial proportion of the fisheries are not informative.
Recommendation: Develop measures to diagnose the presence of uninformative data. Explore the possibility of introducing fishery-specific weights in the MULTIFAN-CL likelihood function.

Finding: Production models performed remarkably well. The reason for this good performance is the use of informative “longline” data and the exclusion of uninformative “purse seine” data
Recommendation: Routinely apply a production model to selected longline data and compare the estimated fishery parameters to those estimated by MULTIFAN-CL.

References

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

Hampton, J. and Kleiber, P. Stock assessment of yellowfin tuna in the western and central Pacific Ocean. SCTB16/YFT-1.

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

Kolody, D.S., Jumppanen P.C., Ricard, D.G., Hartog, J.R., Preece, A.L., and Polacheck, T.W. 2004. Extracts from SESAME: A Simulation-Estimation Stock Assessment Model Evaluation Project focused on Large Pelagic Species. SCTB17/MWG-3.

Labelle, M. 2002. An operational model to evaluate assessment and management procedures for the North Pacific swordfish fishery. U.S. Dept. Commerce. NOAA-TM-NMFS-SWFSC-341. 53 pp.

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. SCTB16/MWG-1. 32 pp.

Maunder, M. N. and Watters, G. W.. 2003. A-SCALA: an age-structured statistical catch-at-length analysis for assessing tuna stocks in the eastern Pacific Ocean. IATTC Bull. 22(5):435-582.

National Research Council (NRC). 1998. Improving fish stock assessments. Committee on Fish Stock Assessment Methods. National Research Council. National Academy Press, Washington D.C. 1998. 176 pp.

5 Figures and Tables

List of Figures

List of Tables

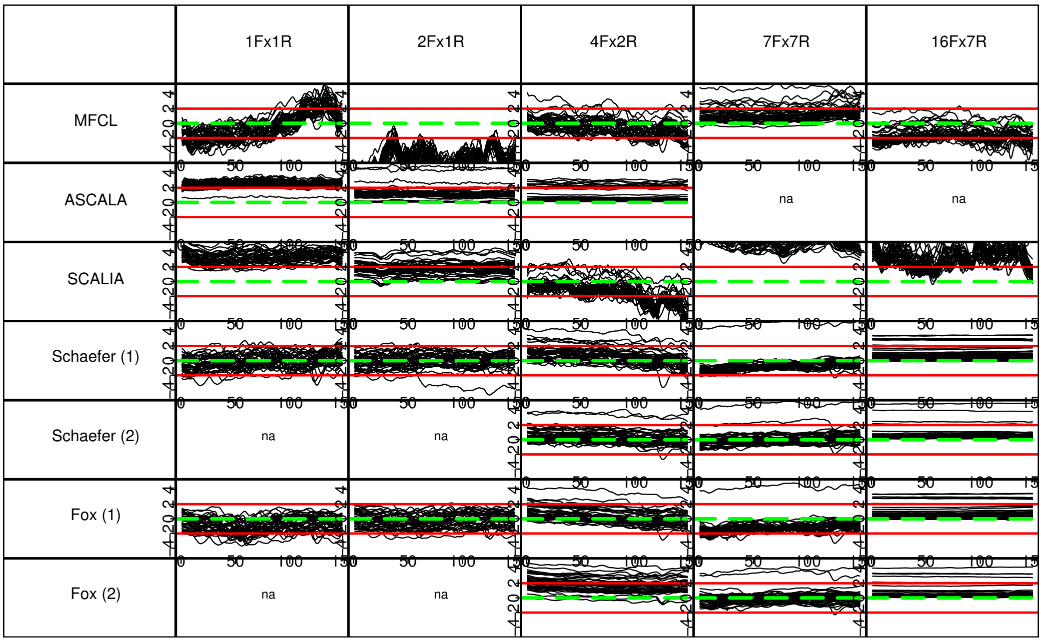


Figure 1: Standardized bias for estimated total biomass, B_{ij}, time series.

Table 1: Proportion of total biomass bias estimates with $|\widetilde{b}_{ij}| < 2$.

Model	Scenario					
	1Fx1R	2Fx1R	4Fx2R	7Fx7R	16Fx7R	Average
MFCL	0.653	0.014	0.878	0.822	0.612	0.596
ASCALA	0.104	0.882	0.716	NA	NA	0.567
SCALIA	0.045	0.636	0.662	0.000	0.176	0.304
Schaefer (1)	0.930	0.968	0.874	0.966	0.859	0.919
Schaefer (2)	NA	NA	0.916	0.956	0.876	0.916
Fox (1)	0.862	0.968	0.902	0.893	0.861	0.897
Fox (2)	NA	NA	0.592	0.949	0.875	0.805
Average	0.519	0.694	0.792	0.764	0.710	0.705

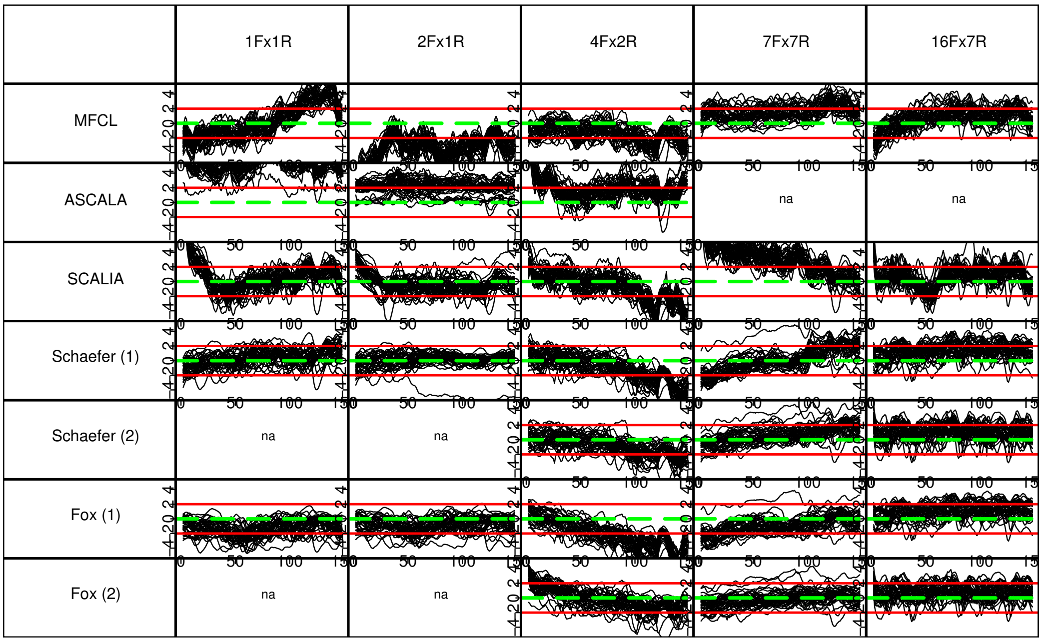


Figure 2: Standardized bias for estimated $B_{i,j}/B_{i,0}$ time series.

Table 2: Proportion of $B_{i,j}/B_{i,0}$ bias estimates with $|\widetilde{b}_{ij}| < 2$.

Model	Scenario					
	1Fx1R	2Fx1R	4Fx2R	7Fx7R	16Fx7R	Average
MFCL	0.579	0.200	0.670	0.658	0.782	0.578
ASCALA	0.017	0.381	0.544	NA	NA	0.314
SCALIA	0.762	0.875	0.692	0.312	0.799	0.688
Schaefer (1)	0.877	0.955	0.689	0.759	0.717	0.800
Schaefer (2)	NA	NA	0.764	0.840	0.804	0.802
Fox (1)	0.725	0.851	0.470	0.800	0.791	0.728
Fox (2)	NA	NA	0.796	0.926	0.851	0.858
Average	0.592	0.652	0.661	0.716	0.791	0.686

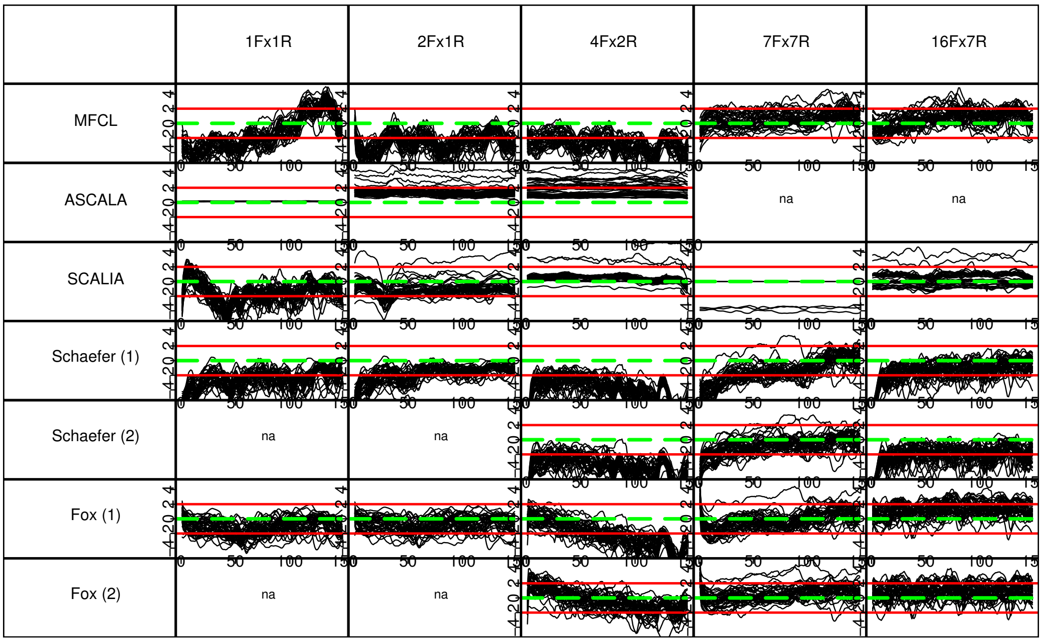


Figure 3: Standardized bias for estimated $B_{i,j}/B_{i,msy}$ time series.

Table 3: Proportion of $B_{i,j}/B_{i,msy}$ bias estimates with $|\widetilde{b}_{ij}| < 2$.

Model	Scenario					
	1Fx1R	2Fx1R	4Fx2R	7Fx7R	16Fx7R	Average
MFCL	0.403	0.307	0.237	0.898	0.852	0.539
ASCALA	0.975	0.839	0.592	NA	NA	0.802
SCALIA	0.635	0.876	0.901	0.925	0.949	0.857
Schaefer (1)	0.304	0.634	0.110	0.651	0.727	0.486
Schaefer (2)	NA	NA	0.123	0.839	0.488	0.483
Fox (1)	0.766	0.889	0.439	0.888	0.761	0.749
Fox (2)	NA	NA	0.796	0.907	0.844	0.849
Average	0.617	0.709	0.457	0.851	0.770	0.674

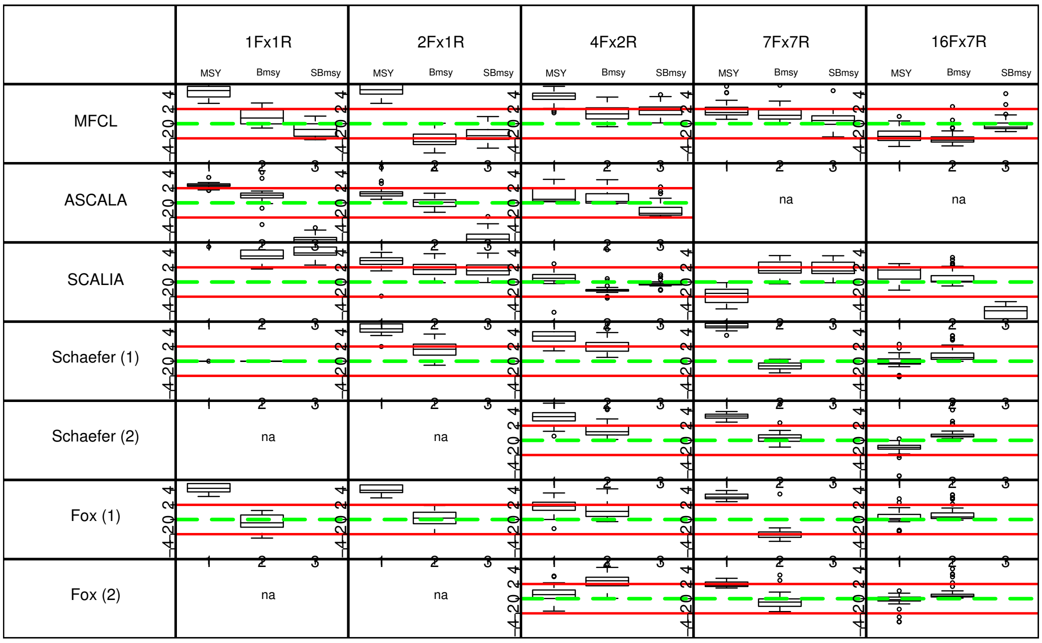


Figure 4: Standardized bias of selected reference points, MSY, total biomass at MSY (Bmsy) and adult biomass at MSY (SBmsy). SBmsy is not estimated by the Schaefer and Fox models.

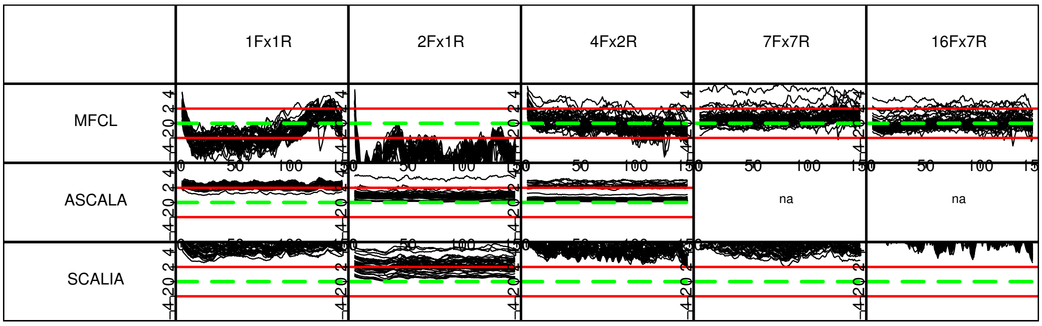


Figure 5: Standardized bias for estimated adult biomass time series.

Table 4: Proportion of adult biomass bias estimates with $|\widetilde{b}_{ij}| < 2$.

Model	Scenario					
	1Fx1R	2Fx1R	4Fx2R	7Fx7R	16Fx7R	Average
MFCL	0.574	0.073	0.900	0.899	0.927	0.675
ASCALA	0.193	0.917	0.704	NA	NA	0.604
SCALIA	0.000	0.446	0.000	0.000	0.000	0.089
Average	0.256	0.479	0.535	0.449	0.463	0.433

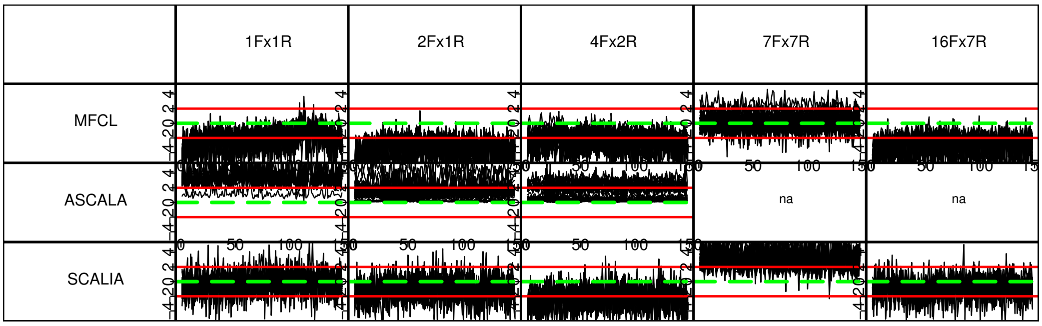


Figure 6: Standardized bias for estimated recruitment time series.

Table 5: Proportion of recruitment bias estimates with $|\widetilde{b}_{ij}| < 2$.

Model	Scenario					
	1Fx1R	2Fx1R	4Fx2R	7Fx7R	16Fx7R	Average
MFCL	0.618	0.196	0.495	0.939	0.312	0.512
ASCALA	0.073	0.807	0.790	NA	NA	0.556
SCALIA	0.930	0.807	0.527	0.208	0.859	0.666
Average	0.540	0.603	0.604	0.573	0.586	0.582

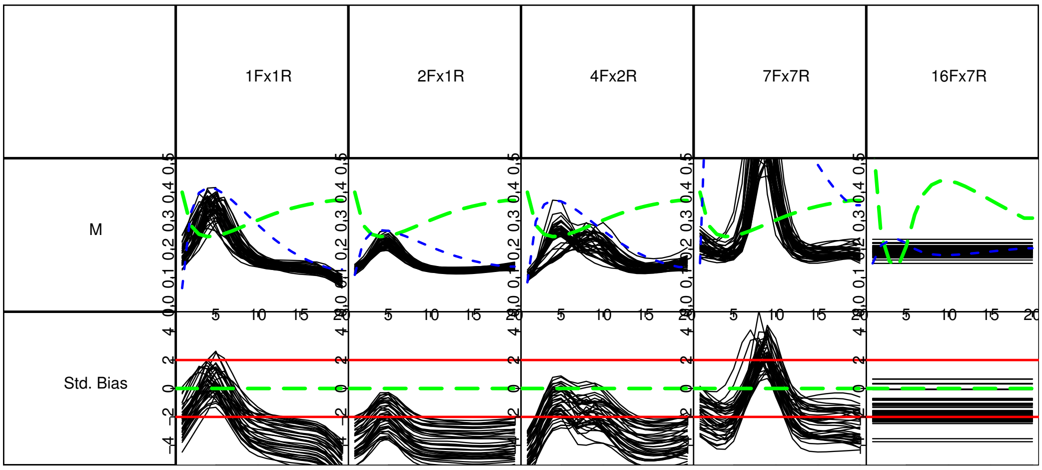


Figure 7: MULTIFAN-CL estimates of natural mortality (M) and standardized bias of the estimates (Std. Bias.) for each scenario. Solid black lines are the MFCL estimates of natural mortality, $\widehat{M}_{ij}; j = 1, \ldots, 20$. Heavy (green) dashed line is value of M used in operational model, identical for all realizations of a scenario. Light (blue) dashed line is a $1 - M$ scaled to the same range as $\widehat{M}_{ij}$.

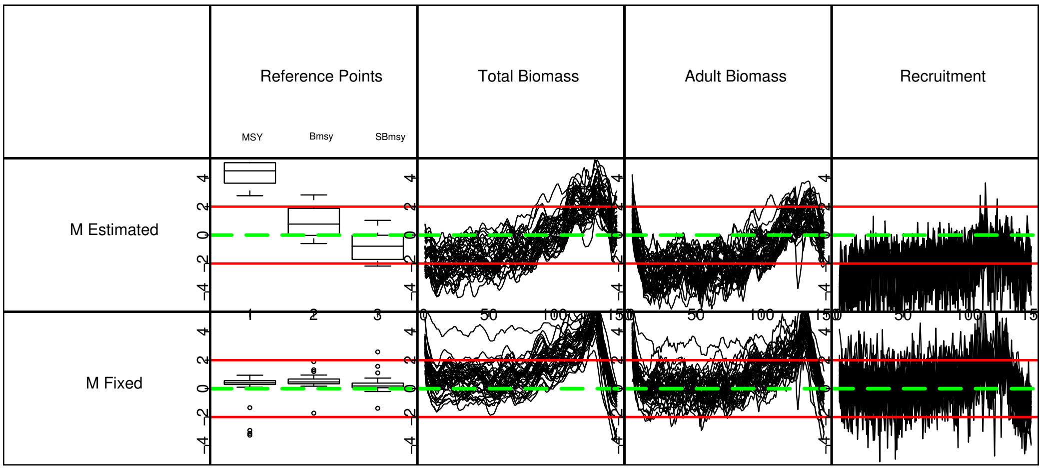


Figure 8: Comparision of MULTIFAN-CL estimates for scenario 1 (1 fishery, 1 region) with M fixed at the values used in the simulation.

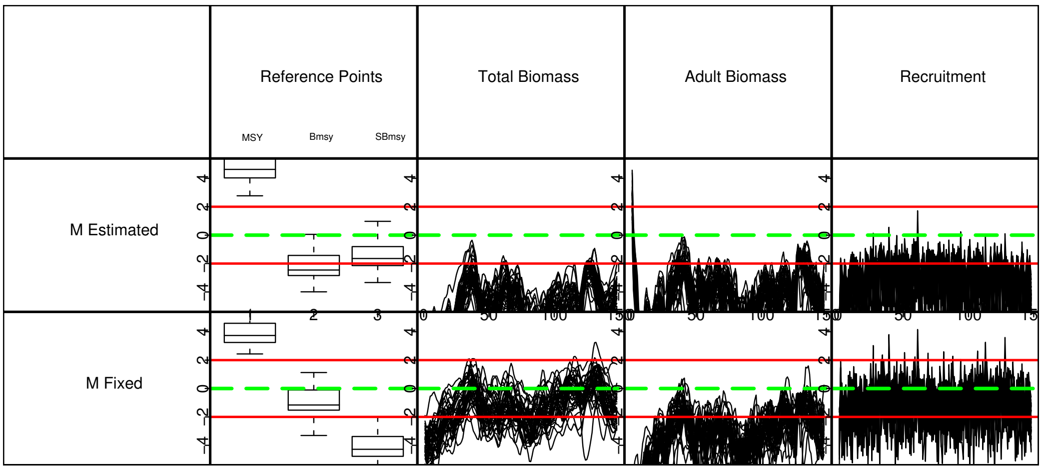


Figure 9: Comparision of MULTIFAN-CL estimates for scenario 2 (2 fishery, 1 region) with M fixed at the values used in the simulation.

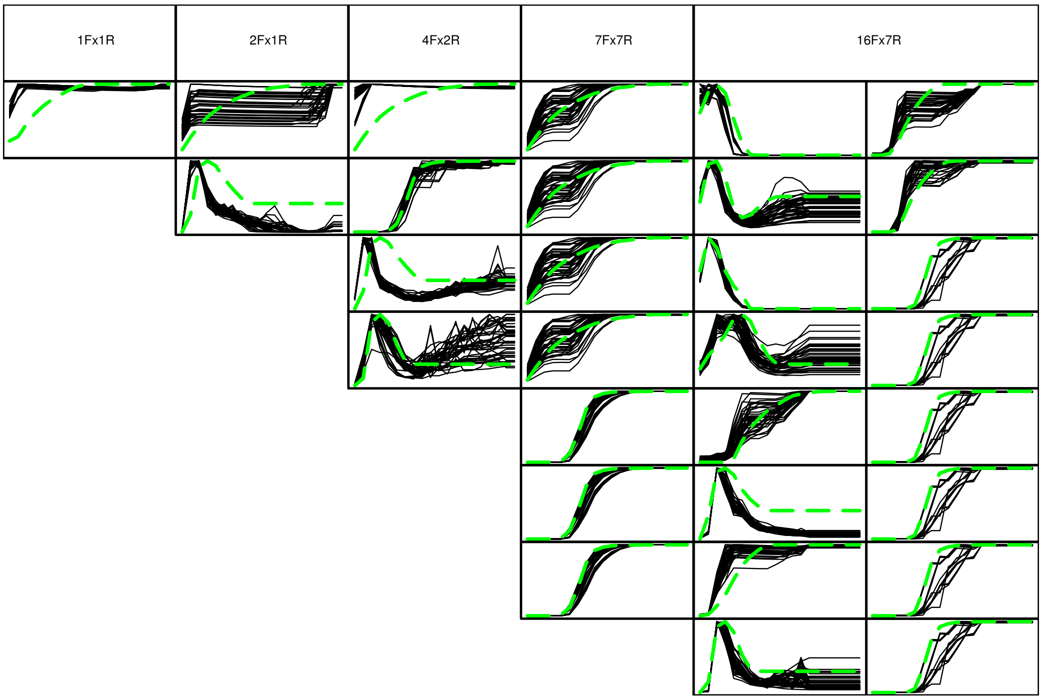


Figure 10: MULTIFAN-CL estimates of slectivity by age for each scenario. Heavy (green) dashed line indiates value used in operational model.

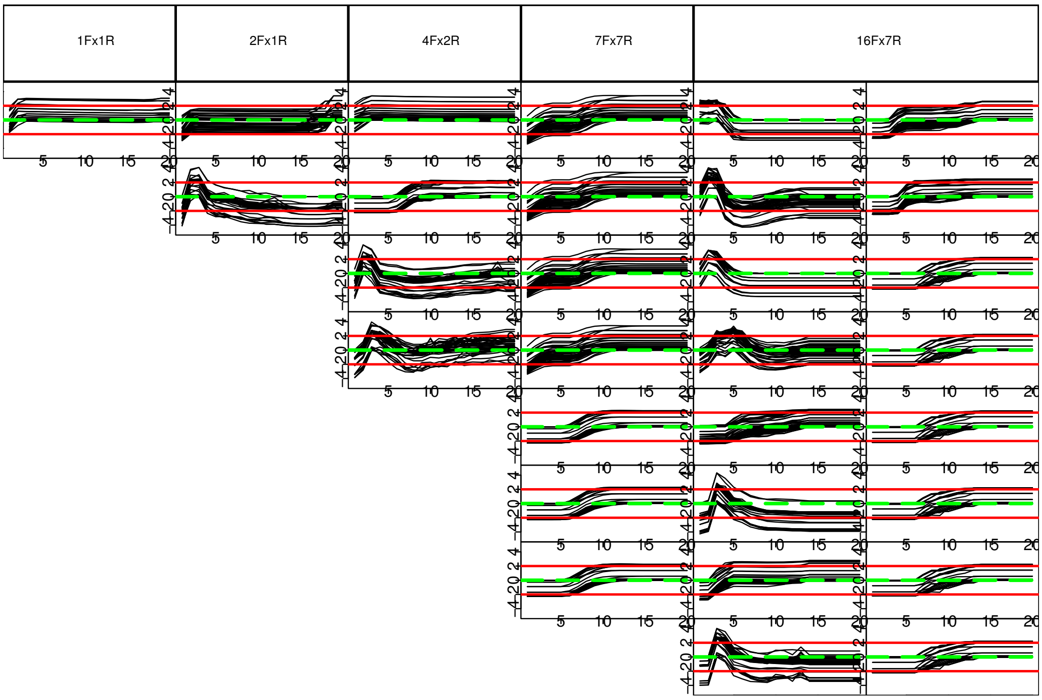


Figure 11: Standardized bias of MULTIFAN-CL estimates of slectivity by age for each scenario.

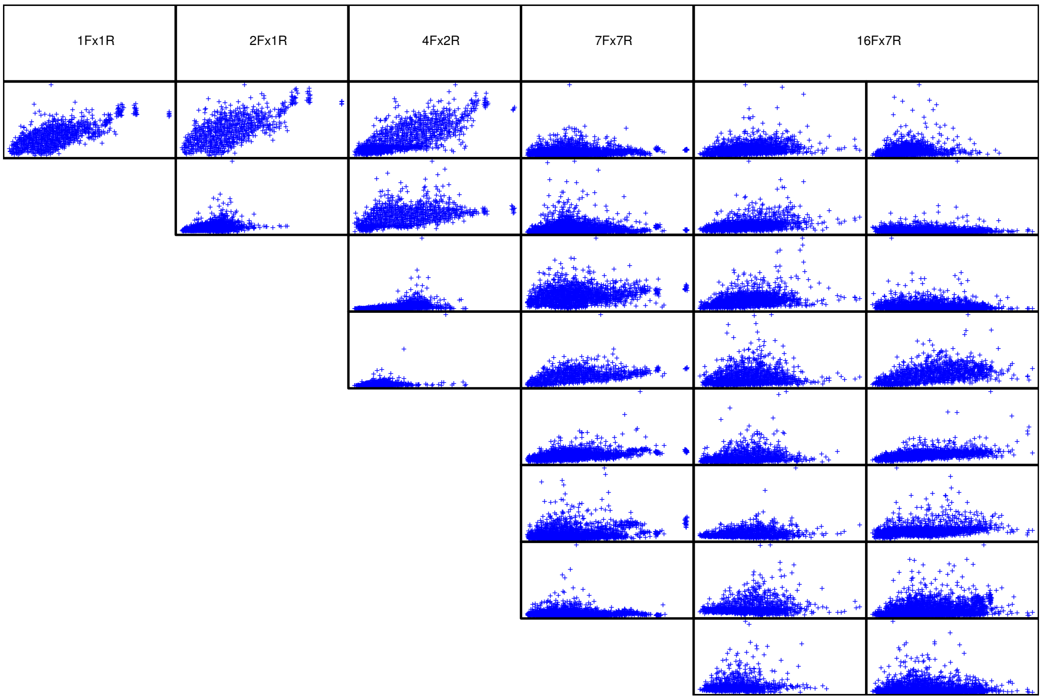


Figure 12: CPUE and biomass scatter plots for all scenarios, fisheries and realizations from operational model. Biomass is on the abscissa and CPUE is on the ordinate.

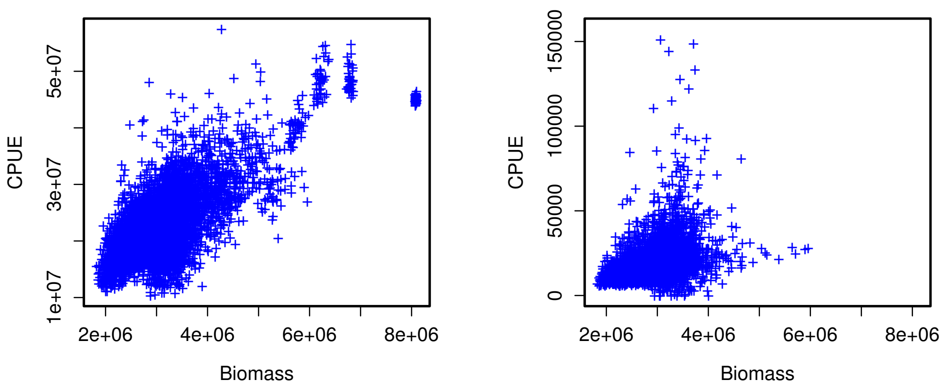


Figure 13: CPUE and biomass scatter plots of fisheries one (left) and two (right) for scenario 2 (2Fx1R) for all realizations from the operational model.

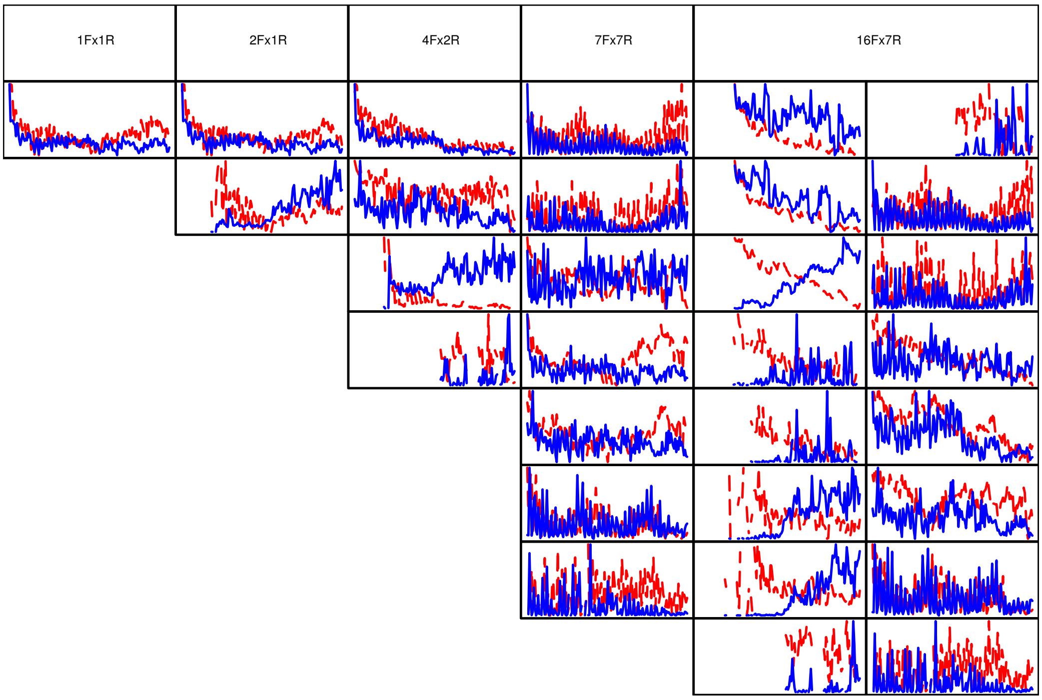


Figure 14: Histories of fisheries for all scenarios averaged over realizations from the operational model. Solid (blue) line is catch; dashed (red) line is CPUE.

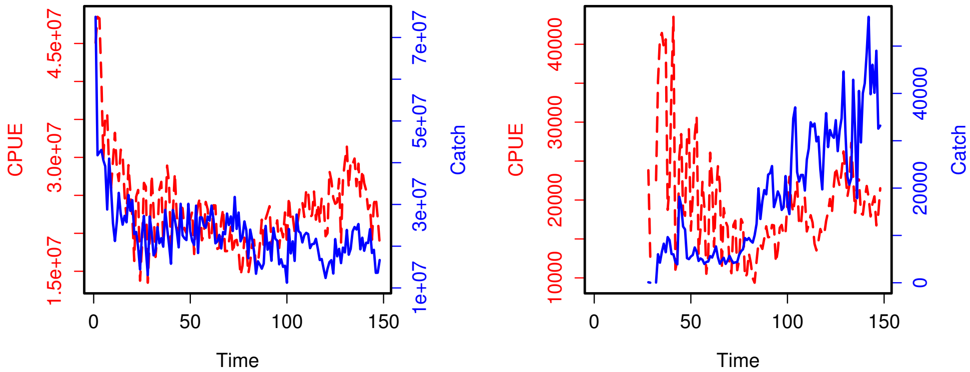


Figure 15: Histories of fisheries one (left) and two (right) for scenario 2 (2Fx1R) averaged over all realizations from the operational model. Solid (blue) line is catch; dashed (red) line is CPUE.