Sensitivity analysis of the 2021 WPTT Indian Ocean yellowfin tuna stock assessment within Stock Synthesis 3

Prepared for Europêche Tuna Group

May 19, 2023

Landmark Fisheries Research

211-2414 St. Johns Street

Port Moody, BC

1 Introduction

Landmark Fisheries Research was contracted by Europeche to conduct a sensitivity analysis of the current Indian Ocean Yellowfin Tuna (IOYT) stock assessment. The current assessment fits Stock Synthesis v3.30.17 (Methot and Wetzel 2013; Methot et al. 2020) to long-line fishery CPUE indices, catch-at-length observations, and mark-recapture tagging over four spatial areas on a quarterly time-step. The resulting fitted model extends estimates of life-history parameters, fishery selectivity, movement rates, and biological reference points.

This document describes the modifications that were made under the sensitivity analysis, and the effects those modifications had on model estimates. The set of sensitivities was based on a review of the IOYT assessment model by Landmark Fisheries Research, Ltd (LFR) in September 2022 (LFR 2022).

Note: residual plots have been removed from the figures in this document to save file size. The complete set of figures and tables is also provided in the "LFR-IOYT-SS-Jul13-SuppFigs.docx" file.

2 Methods

IOYT SS3 model sensitivity analyses were broken into three batteries of tests, described below. The first battery (A) tests the effect of alternative assumptions about spatial structure of the IOYT resource, an aspect of the current model which is not conclusively supported by the available data. The second battery (B) modifies the weighting on length composition data by fleet to reflect the relative catch levels of each fleet, as opposed to the current approach where every fleet is equally weighted. Finally, the last battery (C) tests alternative tag latency periods for a 2-area model as tested in A; the latency period is the number of quarters required to pass before a released tag is assumed to be uniformly mixed with the general population, and is set at 4 quarters for the IOYT assessment (Fu et al, 2021). All modifications were based on the input, and parameter, and data files for the 2021 stock assessment

'base model' (Table 3, Fu et al, 2021). We compare estimates from each sensitivity run to estimates from the "basic model" described in the current IOYT assessment (hereafter referred to as the "Assessment" model).

A. Alternative spatial structures

One problem identified with the current IOYT assessment model is that the spatial structure is overly complex and not justified by the data (Methot 2019; LFR 2022). The current IOYT assessment model is spatially stratified into 4 regions (Figure 1): two equatorial/tropical regions (R1 and R4) and two temperate regions (R2 and R3). Bi-directional movement is estimated between R1/R2, R1/R4, and R3/R4 by fitting a mark-recapture model to tagging data, but very few tags (27 in total out of 39,433 released in R1/R2) are recovered in R4. As such, SS models estimate (practically) zero longitudinal movement between R1 and R4 in recent assessments (Fu et al. 2018; Urtizbera et al. 2019; Fu et al. 2021).

Two simplified spatial structures were tested to reduce spatial complexity, including (i) a single-area model (A-1area) combining R1, R2, R3, and R4 and (ii) a two-area model (A-2area) in which R1 and R2 are combined into one area and R3 and R4 are combined into another (all models summarized in Table 1). For both models, tagging data was omitted and quarterly longline fishery CPUE indices were recalculated as the mean of the CPUE indices for the areas being aggregated. In the two-area model we assumed no movement between the areas, as there does not appear to be any significant movement of IOYT from the western spatial strata (R1R2) to the eastern strata (R3R4). We also estimated survey-specific catchability in the two-area model.

B. Length composition weighting

Another problem identified with the IOYT assessment model is a poor fit to length composition data in individual quarters, despite acceptable fits to the time-averaged composition. This issue may be related to the choice of a multinomial likelihood for the compositional data, which dominates the total objective function. IOYT assessment authors re-weight the total length composition

likelihood function by scaling fleet length compositions for each quarter to a sample size of 5, with each length bin having the same proportions as the raw data (Fu et al 2021). We attempted to resolve this issue by catch-weighting the sample size so that length composition influence better matched the relative catch influence of each fleet. We tested two alternative weighting schemes using time-averaged catches by fleet (B-byFleet) and also using fleet-/quarter-specific (B-byFleetQuarter). For B-byFleet, we calculated the fleet-specific mean catch across quarters ($\overline{C_f}$), then calculated sample size (n_f) as

$$n_f = \frac{\bar{C}_f}{\text{mean}(\bar{C}_f)} \times 5. \qquad \text{Eq 1.}$$

This weighting produces a mean sample size across fisheries of 5 but gives fisheries with relatively large catches sample sizes greater than 5 (max: 22.00), while fisheries with relatively small catches have sample sizes smaller than 5 (min: 0.03; Table 2). Similarly, we calculated fleet-/quarter-specific sample sizes as

$$n_{f,t} = \frac{C_{f,t}}{\text{mean}(C_{f,t})} \times 5 \qquad \text{Eq 2}$$

where $C_{f,t}$ are fleet-/quarter-specific catches. Fleet-/quarter-specific sample sizes were much more variable than fleet-specific sample sizes, ranging from 0.00 to 139.58 (Figure 2).

C. Tag latency period

The current IOYT assessment model assumes a 4-quarter (or 1 year) mixing period for tagged fish, during which the overall influence of tagged fish on parameter estimates is reduced as tags recovered before the end of the mixing period are excluded. A 1 to 4 quarter mixing period is commonly assumed in tuna assessments; however, this assumption is difficult to validate and analyses

of tag mixing for western Pacific Ocean skipjack tuna (*Katsuwonis pelamis*) have demonstrated that four quarters may not be sufficient time for tagged fish to have adequately mixed with the untagged population (Kolody and Hoyle, 2014), especially for larger spatial areas such as in the A-2area model.

To test the sensitivity of IOYT assessments to the latency period, three alternative mixing periods were tested for the A-2area model: 2 quarters / 0.5 years (C-mix2), 8 quarters / 2 years (C-mix8), and 12 quarters / 3 years (C-mix12). Furthermore, exploration of the tagging data indicated that tags have been primarily recovered from purse seine fisheries (specifically, fleets 6, 8, 16, and 17) and relatively small numbers of tags have been recovered from other fisheries. Therefore, the models defined for this part only included tags recovered by the four purse seine fisheries with appreciable catch of tagged fish to test whether fitting to tags from smaller fisheries with low sample sizes could cause model bias.

3 Results

We separately present results for the single-area models (A-1area and the "B" models) and two-area models (A-2area and the "C" models) given the similar structure of models within these two groups.

3.1 Effects of aggregating catches and longline CPUE indices

The CPUE index series averaged across all areas (R1-R4) increased from the mid-1970s to the late 1980s and subsequently declined, showing a 67% decrease in the average decadal survey index from the 1980s to the 2010s (Figure 3b).

When averaging CPUE for the 2-area models, indices for the combined R1R2 area followed a similar trajectory of increasing until the mid-1980s then declining thereafter; the average survey index declined by 57% from the 1980s to 2010s (Figure 3c). Survey indices for the combined R3R4 area were reasonably stable in the 1970s and 1980s but rapidly declined to very low values

thereafter, with spawner indices declining by 86% between the 1980s and 2010s (Figure 3c). Survey indices in the combined R3R4 area were on average 45% smaller than for the R1R2 area.

Annual landings in the combined R1R2 area accounted for an average of 70% of total landings (Figure 4). The proportion of annual landings taken from R1R2 increased from an average of 55% in the 1950s to 78% in the 2010s.

3.2 One-area model fits and estimates

Each of the three one-area models approximated the mean survey index reasonably well, though none of these models replicated the high seasonal variability of the observed index (Figure 5Figure 5). There did not appear to be any long-term trend in the survey index residuals (Figure 6), though there were seasonal trends in the residuals (Figure 7) similar to those produced by the assessment model.

A-1area fits to the time-aggregated length composition data were very similar to assessment model fits; specifically, fits to the time-aggregated compositions were reasonably close (**Erreur ! Source du renvoi introuvable.**), but fits to individual quarters were often poor. For instance, both models underestimate the proportion of fish between 100-140 cm in longline length-composition and overestimate the proportion of fish larger than 140 cm until the early 2000s, after which this behaviour is reversed (**Erreur ! Source du renvoi introuvable.**-SuupFigs). The purse seine length-composition residuals from A-1area also had large blocks of positive and negative residuals that were similar to the those produced by the assessment models (Figure 10-SuppFigs). While A-1area produced residual patterns that were similar to the assessment model, the overall magnitude of the residuals was smaller, and the sum of the negative log likelihood for the length composition data was 142.8 units smaller for A-1area than the assessment model, indicating a better fit to the data (Table 3).

B-byFleet and B-byFleetQuarter fits to the time-aggregated length composition data were similar to those of A-1area and the assessment model (Figure 11-Figure 12). In general, weighting the likelihood tended to affect residual magnitude but had little impact on residual patterns (Figure 13-SuppFigs through Figure 16-SuppFigs).

Spawning biomass estimates from A-1area were about twice as high as from the assessment model, while spawning biomass estimates from B-byFleet and B-byFleetQuarter were 58% and 82% higher than assessment model estimates, respectively (Figure 17). Estimated unfished recruitment under A-1area was 29% and 15% higher than under B-byFleet and B-byFleetQuarter, respectively, while estimated catchability under A-1area was 26% and 12% lower than under B-byFleet and B-byFleetQuarter, respectively (Table 4).

Estimated F_{MSY} from the assessment model (0.167 yr^{-1}) was similar to estimates from A-1area (0.182 yr^{-1}) and the "B" models (0.171-0.177 yr^{-1}). However, estimated SB_{MSY}, and, by extension, MSY, were higher under A-1area and the "B" models than under the assessment model (Table 5). The assessment model and the single-area models we tested also give a very different perception of stock status; under the assessment model, terminal *F* exceeds F_{MSY} by 16% and terminal spawning biomass is 63% of SB_{MSY}, while, under the single-area models we tested, terminal *F* was at most 70% of F_{MSY} while terminal spawning biomass was 23-54% greater than SB_{MSY}. A-1area and the "B" models also allocate proportionally more MSY among fleets in area R1R2 than area R3R4 (Table 6). All fleets operating in R1R2 received larger MSY allocations under A-1area or the "B" models than under the assessment model. All fleets operating in R3R4 also received larger MSY allocations under A-1area than under the assessment model (albeit to a lesser degree than R1R2 fleets); however, R3R4 fleets received less allocation under the "B" models than under the assessment model. For the "B" models, this reduced allocation in R3R4 was outweighed by increased allocations in R1R2.

3.3 Two-area model fits and estimates

The two-area models appeared to be less reliable than the one-area models, as issues arose in fitting each model to the survey indices. In particular, in the R1R2 area, model-predicted survey indices tended to be higher than observed indices in early years, and lower in later years; in R3R4, this trend was reversed (Figure 18-Figure 19).

Fits of the two-area models to the length composition were similar to those of A-1area in that fits to the time-averaged compositions were acceptable but fits to quarter-specific compositions exhibited large blocks of positive or negative residuals (Figure 20-SuppFigs-Figure 22-SuppFigs).

Fits to the tagging data were generally poor (Figure 23). The model over-predicted returns for tag groups 1-9 and 128-131, which was not surprising given the relatively small amounts of returns and releases from these groups (Figure 24-Figure 26); however, the models chronically under-estimated tag returns overall (Figure 24-SuppFigs through Figure 26-SuppFigs). Increasing the mixing latency period for tags slightly improved fits to the tag recaptures, as evidenced by the slightly smaller residuals for C-mix8 than C-mix4, but the chronic under-estimation persisted so fits were not acceptable.

To attempt to alleviate issues of model fit, additional two-area models were tested that (i) allowed movement between areas, and (ii) included tag recoveries from all fleets. However, neither of these changes reduced the residual pattern in the survey index fits nor did they improve fits to the tagging data.

Each model in each area estimated similar trends in spawning biomass, with biomass maintaining relatively stable levels until the mid-1980s and subsequently declining (Figure 27). Spawning biomass estimates from A-2area were higher than the current assessment model estimates in R1R2 until the early 2010s, but biomass estimates have since been similar between the two models. Each of the C-models estimated levels of spawning biomass that were slightly above assessment model levels until the mid-1980s, but were subsequently nearly always below, as these models estimated an earlier and stronger decline than the assessment model. In R3R4, A-2area estimated slightly lower levels of spawning biomass than the assessment model, but has estimated relatively higher

levels since 1980. In contrast, the C-models estimated spawning biomass levels in R3R4 about 50% below assessment levels until the 1980s, but has estimated relatively higher levels since the early 2000s. After the 2010s, the two-area models further diverged from the current assessment, which estimated a slightly declining spawning biomass over 2010 – 2020 in contrast to a 30-55% rebound for the two-area models.

Estimated unfished recruitment scaled with spawning biomass; specifically, unfished recruitment from A-2area was smaller than from each of the one-area models, and unfished recruitment from the "C" models was, in turn, smaller than from A-2area (Table 4). Catchability represented an important divergence between A-2area and the "C" models, as A-2area estimated 26% lower catchability in R3R4 than R1R2 and the "C" models estimated 2-7% higher catchability in R3R4 than R1R2. Additionally, catchability estimates were 66-86% higher under the "C" models than A-2area for R1R2, and 130-166% higher for R3R4. The two-area models allocated 67-69% of recruitment to R1R2 (Table 4).

The overall scale of the stock appeared to be highly influenced by the tagging data, with lower biomass estimates associated with higher numbers of tagging data. As the influence of tagging data increased (i.e., as it was introduced in the C-models, and as more tags are included in the likelihood as the mixing period decreases) spawning biomass estimates become smaller. This inverse relationship between the magnitude of spawning biomass and the degree of tagging data influence may be related to the persistent underestimation of recovered tags in each of the "C" models. Specifically, smaller biomass leads to larger harvest rates, and thus a larger proportion of tags being recovered.

F_{MSY} estimates from A-2area (0.170 yr^{-1}) and the "C" models (0.153-0.158 yr^{-1}) were similar to estimated F_{MSY} from the assessment model (0.167 yr^{-1}). Estimated MSY and SB_{MSY} under A-2area were higher than under the assessment model, but estimated MSY and SB_{MSY} under the "C" models were lower (Table 5). The two-area model estimates of stock status were more optimistic than assessment model estimates but more pessimistic than one-area model estimates (Table 5). In particular, A-2area estimated that terminal *F* was approximately equal to F_{MSY} while terminal SSB was between 76-91% of SB_{MSY}. Compared to the

assessment model, each two-area model allocated more MSY to fleets operating in R1R2 and less to fleets operating in R3R4, though total MSY was higher under the two-area models (Table 6).

4 Discussion and recommendations

In our previous review of the IOYT assessment, LFR (2022) identified several sources of uncertainty that appeared to be leading to biased results from the current IOYT assessment model, specified in the SS3 stock assessment package. This paper addressed three of those via a sensitivity analysis of SS3 to spatial structure of the stock, weighting of length composition data, and the tagging data mixing period in response to changes in spatial complexity.

Previous reviews of the IOYT stock assessment have noted that the assumption of 4 spatial regions for IOYT appears to be overly complex, particularly as tagging data indicates very little exchange of individuals between regions. We demonstrate above that a model with no spatial structure (A-1area) fits the survey indices about as well as the assessment model and fits the observed length-composition better than the assessment model. Similarly, two-area models with no movement (A-2area and the C models) fit the observed data relatively well, providing a basis for a simplified spatial structure for IOYT without the need to partition the stock latitudinally and/or account for exchange of fish between areas. Biases in predicting the survey index by the 2-area models suggest that single-area models may be a more defensible basis for future work, however these model fit issues may be resolved via further tuning of the models themselves or under alternative derivations of the survey indices. In either the single-area or two-area cases, fleets could continue to represent different spatial area and gear combinations via an areas-as-fleets approach like Pacific Halibut and Atlantic Halibut stock assessments (Stewart and Hicks 2022; Johnson et al. 2022), as well as others.

If a decision is made to reduce the spatial complexity of the assessment model then future work should focus on deriving new aggregate survey indices for single-area and 2-area models. When reducing the spatial complexity for our analyses, aggregated survey indices were calculated as the mean of existing survey indices for the component areas, which are themselves the output

from generalized linear models fit to longline catch and effort data (Kitakado et al., 2021). This differs from the current assessment, where the relative scale of each area's CPUE index is fixed via regional scaling factors based on region size and the relative catch rate, allowing a common catchability coefficient to be estimated over the entire stock. This is partially responsible for the scale mismatch for R3R4 in our 2-area model and the sum of those regions in the current assessment. Ideally, for the combined survey indices in our analysis, regional scaling factors would be derived from a similar analysis starting with raw data, but this was impossible given the scope of this analysis. We alternatively tested models that were fitted to catch-weighted survey indices; however, this had little impact on model estimates.

Another issue raised in LFR's previous review of the IOYT assessment was the lack of the fit to the length composition data. One issue was that assigning equal effective sample sizes to each length composition observation essentially overweighs data from smaller fleets, causing these fleets to have undue influence over the likelihood. Another, potentially related problem was that the multinomial compositional likelihood dominated the total likelihood. We attempted to address these issues by catch-weighting the effective sample sizes used in the multinomial likelihood for the length composition data. While this approach was effective at reducing the influence of length composition data from smaller fleets, it also increased the effective sample size of larger fleets, leading the length composition likelihood to account for even greater proportion of the total likelihood value. Additionally, increasing the effective sample size for larger fleets did not alleviate the troubling residual patterns that were present in model fits to these data. Therefore, alternative approaches to fitting the compositional likelihood should be considered. One option is to scale the catch-weighted sample sizes used in the B model so that their maximum values are smaller (e.g., closer to the value of 5 used in the assessment) but differences in the sample sizes between small and large fleets are retained. Another option is to consider an alternative compositional likelihood. Stock synthesis includes an option to fit compositional data using a Dirichlet multinomial error distribution; however, this distribution also requires a tuning stage for setting the effective sample size (Thorson et al., 2017). A better choice may be to fit IOYT length composition data using a logistic-normal likelihood function (Schnute and Haigh 2007; Francis

2014), for which the likelihood function is a sum of squared logit-residuals and is thus self-weighting by a variance parameter. There is therefore no need to down-weight sample sizes as done for IOYT, and the likelihood function value is often on a similar scale to likelihoods for other data sources. Relative sizes of individual quarterly samples can be included as annual weights on the residual sum of squares to represent changes in sampling effort.

The “C” models that we tested did not significantly improve the chronic underestimation of the tag recoveries that was exhibited by the assessment model. We hypothesized that attempting to fit the assessment model to tags that were recovered by fleets with negligible overall levels of tag recoveries could cause issues in fitting to tags recovered by the large purse seine fisheries, though this did not appear to be the case, as the models underestimated tag recoveries whether these recoveries from smaller fleets were included or not. Increasing the tag mixing period resulted in only marginal improvements to model predictions. One potential issue was that significant amounts of tagging data were discarded as the mixing period increased, leaving few tags remaining for inference (

Table 7-Table 8). Adding likelihood weights to the tagging data likelihood is one possible route for achieving better fits to the tagging data, particularly if the likelihood is being dominated by compositional data, though the influence of these weights on both the tagging fits and other model components would be need to be scrutinized.

Another issue with the tagging model is that SS3 requires ages to be assigned to fish in each tag group so that they can be tracked through time, however, the tag group age is not directly observed but is instead inferred based on the mean length-at-age relationship. Significant overlap in length distributions among year class leads to age misspecification, and thus to the wrong fishing selectivity and mortality being applied to tagged fish. Two alternative approaches may reduce biases in the tagging model associated with age assignments to tag release groups. First, uncertainty in the length-at-age relationship could be accounted for using an age-length key (Fu 2022). Second, and perhaps ideally, the age-based approach could be entirely replaced by a length-based approach with size-class transition matrices (e.g., Hillary and Eveson 2015).

Our previous review recommended the construction of a bespoke model for IOYT assessments. The results presented here reinforce that recommendation, as assessment model biases were not significantly reduced across the range of sensitivity analyses performed. Further incremental tweaks that do not require major model modifications, such as deriving new survey indices or re-weighting the lenth composition effective sample sizes could be performed in an attempt to reduce these biases; however it is unlikely that these incremental changes will affect fits the length composition data, which appears to be the more proximate problem. On the other hand, major modifications such as fitting the length composition data using a logistic-normal likelihood or fitting a length-based tag model address model deficiencies more directly and may therfore be more effective at reducing model bias. These features are not included in SS3 and so neither of these modifications can be accommodated by simply altering SS data or control files. Instead, incorporating these changes into the current assessment framework requires modifying the SS3 code, which is arduous for anyone who does not have previous experience developing stock synthesis code, and would likely require contracting an SS3 developer. Moreover, there would need to be extensive model testing to

ensure that no new biases or software bugs are introduced. A more feasible approach to including these features in an IOYT assessment model is to construct a bespoke model, in which population dynamics processes and data likelihoods are defined to incorporate unique features of the IOYT fishery and address the issues described above. Moreover, extending a custom model into an operating model for closed loop simulation and management strategy evaluation (MSE), which is already being explored for IOYT (Kolody and Jumppanen 2021; IOTC 2018), would be relatively straightforward.

References

Fu, D., Langley, A., Merino, G., Ijurco, A.U. 2018. Preliminary Indian Ocean Yellowfin Tuna Stock Assessment 1950-2017 (Stock Synthesis). IOTC–2018–WPTT20–33.

Fu, D., Urtizberea, A., Cardinale, M., Methot, R. D., Hoyle, S., and Merino, G. 2021. Preliminary Indian Ocean yellowfin tuna stock assessment 1950-2020 (Stock Synthesis). IOTC-WPTT23, 12.

Fu, D. 2022. A length-based extension to the Brownie-Peterson model and its performance. Fisheries Research, 249:106248.

Hillary, R. M. and Eveson, J. P. 2015. Length-based brownie mark-recapture models: Derivation and application to Indian ocean skipjack tuna. Fisheries Research, 163:141–151.

IOTC. 2018. Report of the 21st Session of the IOTC Scientific Committee. Seychelles, 3 – 7 December 2018. IOTC–2018–SC21–R. 250 pp.

Johnson, S.D.N., Hubley, B., Cox, S. P., and den Heyer, C. E. 2022. Framework Assessment of Atlantic Halibut on the Scotian Shelf and Southern Grand Banks (NAFO Divs 3NOPs4VWX5Zc) Model Update (In Press). Can. Sci. Adv. Sec. Res. Doc.

Kolody, D. and Hoyle, S. 2015. Evaluation of tag mixing assumptions in western Pacific ocean Skipjack tuna stock assessment models. Fisheries Research, 163:127–140.

Landmark Fisheries Research. 2022. Review of 2021 WPTT Indian Ocean yellowfin tuna stock assessment and feasibility of alternative assessment. Prepared for Europêche Tuna Group.

Langley A., and Million, J. 2012. Determining an appropriate tag mixing period for the Indian Ocean yellowing tuna stock assessment. IOTC-2012-WPTT14-31.

Methot, R.D. 2019. Recommendations on the configuration of the Indian Ocean yellowfin tuna stock assessment model.

Methot, R.D., Wetzel, C.R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research 142 (2013) 86–99.

Methot, R.D., Wetzel, C.R., Taylor, I.G., Doering, K. 2020. Stock Synthesis User Manual Version 3.30.16.

Stewart, I., and Hicks, A. 2022. Assessment of the Pacific halibut (*Hippoglossus stenolepis*) stock at the end of 2021. IPHC-2022-SA-01. 36 p.

Thorson, J.T., Johnson, K.F., Methot, R.D., and Taylor, I.G. 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fisheries Research 192: 84–93. doi:10.1016/j.fishres.2016.06.005.

Urtizberea, A., Fu, D., Merino, Gorka., Methot, R., Cardinale, M., Winker, H., Walter, J., Murua, H. 2019. Preliminary Assessment of Indian Ocean Yellowfin Tuna 1950-2018 (Stock Synthesis, V3.30). IOTC-2019-WPTT21-50.

Tables

Table 1. Model names and configurations.

Model name	Description
A-1area	Single stock with an implicit uniform distribution across the Indian Ocean
A-2area	Two populations with shared recruitment for the western spatial strata (R1R2) and eastern spatial strata (R3R4) and no movement between areas. Tagging data omitted. No movement between areas.
B-byFleet	Based on A-1area, but length composition sample sizes are weighted by the catch in each area, averaged across years.
B-byFleetQuarter	Based on A-1area, but length composition sample sizes are weighted by the annual catch in each area.
C-mix4	Based on A-2area but including tagging data. The mixing period for tags is 4 seasons (1 year).
C-mix6	Based on C-mix4 but the mixing period for tags is 6 seasons (1.5 years).
C-mix8	Based on C-mix4 but the mixing period for tags is 8 seasons (2 years).

Table 2. Summary of the catch data for each of the 21 fleets in the IOYT assessment.

Fleet Number	Name	Mean Catch (t)	Total length samples	Adjusted length samples	Time-averaged catch-weighted sample size
1	GI1a	5901.03	115953	520	9.48
2	HD1a	5249.75	34600	250	8.43
3	LL1a	1698.03	7716	75	2.73
4	OT1a	21.56	24794	120	0.03
5	BB1b	2346.3	308116	675	3.77
6	PSFS1b	13698.23	32335175	760	22.00
7	LL1b	4533.1	183856.4	1005	7.28
8	PSLS1b	13394.97	170425447	765	21.52
9	TR1b	577.74	0	0	0.93
10	LL2	1541.72	131042	945	2.48
11	LL3	424.43	224413	920	0.68
12	GI4	1405.28	452003	195	2.26
13	LL4	2028.61	189532	945	3.26
14	OT4	878.65	8300	100	1.41
15	TR4	1052.45	34064	115	1.69
16	PSFS2	1211.03	3283561	425	1.95
17	PSLS2	1586.11	14398028	520	2.55
18	TR2	391.32	0	0	0.63
19	PSFS4	435.94	642546	145	0.70
20	PSLS4	304.65	1486702	250	0.49
21	LF4	6686.15	111133	180	10.74

Table 3. Negative log likelihood of the length composition data by fleet and model. All models assumed a multinomial distribution for length composition, with the Assessment, A, and C models used an effective sample size of 5. The B models used catch-weighted effective sample sizes so the likelihood values for thee models are not directly comparable to likelihood value from the other models as a measure of fit.

Fleet	Assessment	A-1area	A-2area	B-byFleet	B-byFleetQ	C-mix4	C-mix6	C-mix8
1) GI1a	211.6	200.1	199.7	379.5	780.5	203.4	201.4	199.9
2) HD1a	62.0	59.0	60.6	97.7	389.7	67.5	66.4	64.8
3) LL1a	33.3	31.5	31.7	17.6	22.6	32.5	32.8	32.4
4) OT1a	69.5	70.4	71.8	15.5	15.4	69.8	71.2	70.0
5) BB1b	177.3	175.9	178.6	137.6	241.9	175.5	178.7	183.3
6) PSFS1b	446.3	427.9	438.1	1811.5	1900.6	442.3	438.5	439.6
7) LL1b	398.4	379.2	393.4	557.1	696.5	415.0	412.6	409.1
8) PSLS1b	297.5	294.7	290.2	1197.7	1254.8	298.6	287.0	287.1
10) LL2	509.2	482.9	516.5	244.2	347.3	571.1	570.0	563.7
11) LL3	323.6	323.7	320.9	65.6	82.2	318.3	318.9	318.9
12) GI4	86.4	76.8	75.0	36.6	63.4	74.7	74.8	74.7
13) LL4	334.5	297.3	299.7	197.6	223.0	304.0	303.3	302.6
14) OT4	63.3	75.0	77.5	24.2	66.9	70.6	78.9	78.8
15) TR4	44.4	38.7	38.9	14.5	20.4	39.5	39.1	38.9
16) PSFS2	315.3	326.0	330.9	127.8	221.7	331.3	330.4	331.2
17) PSLS2	319.2	320.5	317.9	165.8	294.2	324.6	317.7	316.9
19) PSFS4	101.0	98.3	96.3	19.2	53.3	92.8	93.5	93.4
20) PSLS4	115.2	112.9	116.4	22.6	37.4	121.4	119.0	117.8
21) LF4	89.3	63.4	64.2	134.7	353.3	65.3	65.0	64.9
Total	3997.2	3854.4	3918.1	5257.0	7065.2	4018.2	3999.1	3987.9

Table 4. Transformations of leading model parameters including unfished recruitment (R0; 1000s), catchability in area 1 (q1; area 1 is R1R2R3R4 in the one-area models and R1R2 in the two-area models), catchability in area 2 (q2; R3R4 in the two-area models), and the proportion of recruitment allocated to area R1R2 (p; only estimated in two-area models).

Model	R0	q1	q2	p
A-1area	202.4	9.72E-06	-	-
A-2area	147.4	2.96E-05	2.18E-05	0.690
B-byFleet	156.3	1.32E-05	-	-
B-byFleetQuarter	176.2	1.11E-05	-	-
C-mix4	105.0	5.50E-05	5.79E-05	0.677
C-mix6	108.1	5.34E-05	5.63E-05	0.673
C-mix8	112.4	4.91E-05	5.02E-05	0.678

Table 5. Reference points and estimated stock status by model.

Model	MSY (kt)	F_{MSY}	SB_{MSY} (kt)	F_{2020} / F_{MSY}	SB_{2020} / SB_{MSY}
Assessment	427	0.167	1228	1.155	0.630
A-1area	702	0.182	1790	0.485	1.541
A-2area	513	0.170	1423	0.697	0.977
B-byFleet	572	0.171	1431	0.681	1.233
B-byFleetQuarter	633	0.177	1624	0.523	1.278
C-mix4	390	0.153	1095	1.021	0.858
C-mix8	392	0.154	1132	0.983	0.757
C-mix12	403	0.158	1162	0.967	0.908

Table 6. Fleet allocation of MSY by Stock Synthesis reference point estimation routine. Strata refers to the longitudinal strata in which each fleet operates, i.e., W for the western strata (R1R2) and E for the eastern strata (R3R4). Total allocations by strata are listed at the bottom.

Fleet	Strata	Assess.	A-1area	A-2area	B-byFleet	B-byFleetQ	C-mix4	C-mix6	C-mix8
1) GI1a	W	57.9	126.5	88.5	96.1	111.1	65.5	66.7	68.6
2) HD1a	W	97.7	137.5	98.1	118.7	136.0	88.0	87.8	88.3
3) LL1a	W	0.2	0.4	0.3	0.3	0.4	0.2	0.2	0.2
4) OT1a	W	0.2	0.5	0.3	0.3	0.4	0.2	0.2	0.2
5) BB1b	W	14.1	36.2	24.9	28.1	26.5	17.1	17.4	18.1
6) PSFS1b	W	29.1	45.5	32.9	39.7	47.4	27.5	27.6	28.0
7) LL1b	W	11.0	17.0	12.3	14.8	18.1	10.3	10.3	10.5
8) PSLS1b	W	65.5	150.2	103.9	117.9	121.5	76.0	76.3	79.5
9) TR1b	W	5.8	14.6	10.0	11.4	10.9	7.0	7.1	7.5
10) LL2	W	4.0	8.2	5.6	7.1	8.5	4.9	4.9	5.0
11) LL3	E	0.3	0.5	0.4	0.4	0.5	0.3	0.3	0.3
12) GI4	E	9.1	10.5	7.9	8.5	9.5	5.1	5.2	5.4
13) LL4	E	0.6	0.7	0.6	0.6	0.7	0.4	0.4	0.4
14) OT4	E	38.9	42.7	31.6	33.0	34.4	20.5	20.9	21.9
15) TR4	E	8.1	8.5	6.0	6.6	6.3	3.9	4.0	4.2
16) PSFS2	W	0.5	1.1	0.8	1.0	1.2	0.6	0.6	0.6
17) PSLS2	W	4.6	13.6	9.3	10.5	11.5	6.5	6.7	6.9
18) TR2	W	3.4	10.8	7.4	8.5	8.1	5.2	5.3	5.6
19) PSFS4	E	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1
20) PSLS4	E	0.5	0.5	0.3	0.4	0.4	0.2	0.2	0.2
21) LF4	E	74.9	76.7	71.3	67.5	80.0	49.6	49.8	51.8
TOTAL									
	W	294.0	562.0	394.3	454.4	501.4	309.3	311.3	319.1
	E	132.6	140.2	118.2	117.1	131.8	80.2	80.9	84.4

Table 7. Summary of tag recapture timing by tag group, aggregated across age-at-release.

Tag Groups	Release year	Number of Releases	Number of recaps after *x* quarters			
			x=0	*x*=4	*x*=6	*x*=8
1-9	2006.75	1467.4	49	4	2	1
10-19	2007.00	772.1	23	1	1	1
20-23	2005.00	14.4	5	4	4	2
24-32	2005.25	996.1	100	78	58	34
33-43	2005.50	2513.5	440	249	151	75
44-54	2005.75	2541.7	542	317	191	108
55-65	2006.00	9662.1	2723	1056	655	412
66-75	2006.25	8778.9	2596	1085	650	317
76-84	2006.50	5322.2	1648	440	324	175
85-90	2006.75	66.7	23	8	3	2
91-103	2007.00	1249.1	398	119	67	41
104-115	2007.25	5644.7	1494	427	271	161
116-127	2005.00	400.1	181	45	26	18
128-129	2005.25	1.4	0	0	0	0
130-131	2005.75	2.9	2	1	0	0

Table 8. Summary of tag recapture timing by fleet of recapture.

		Number of recaps after *x* quarters			
Fleet #	Fleet Name	*x*=0	*x*=4	*x*=6	*x*=8
1	GI1a	156	57	14	7
2	HD1a	59	28	14	11
3	LL1a	7	2	1	1
4	OT1a	3	1	0	0
5	BB1b	249	1	0	0
6	PSFS1b	2156	1782.5	1518.4	936.9
7	LL1b	51	40	33	17
8	PSLS1b	6939.2	1839.1	770.1	392.1
9	TR1b	32	5	2	1
10	LL2	36	20	15	9
13	LL4	1	1	0	0
16	PSFS2	168.6	113.5	99.8	43.7
17	PSLS2	607	113.7	65.3	23.9
18	TR2	56	28	18	3
19	PSFS4	5	5	4	4
20	PSLS4	26	1	1	1

Figures

Figure 1. Four region spatial stratification of the Indian Ocean for the basic IOYT assessment model (from Fu et al., 2021).

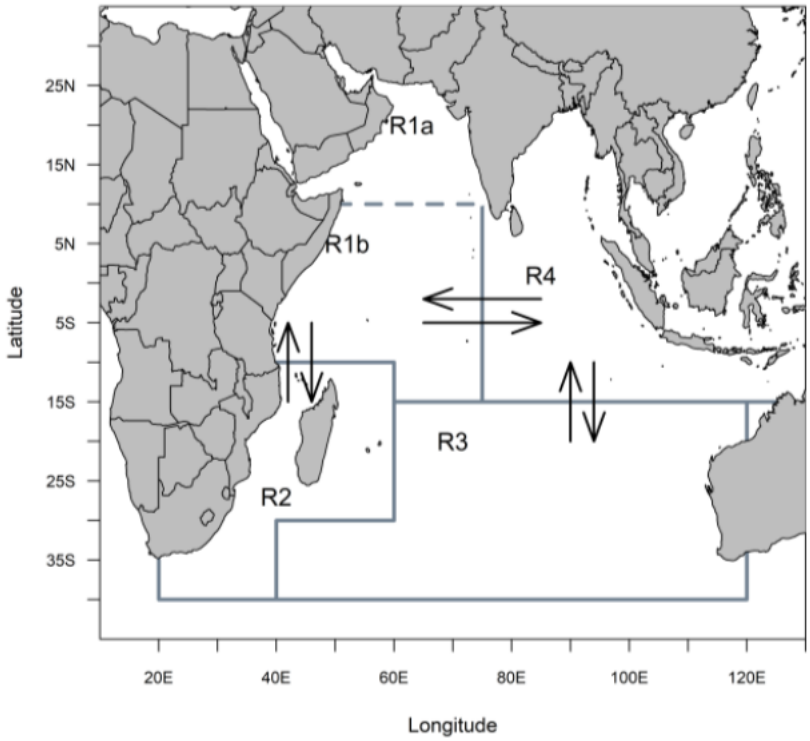

Figure 2. Length composition sample sizes used in B models. Points represent fleet-/quarter-specific sample sizes whereas the dashed red lines represent time-averaged, fleet-specific sample sizes.

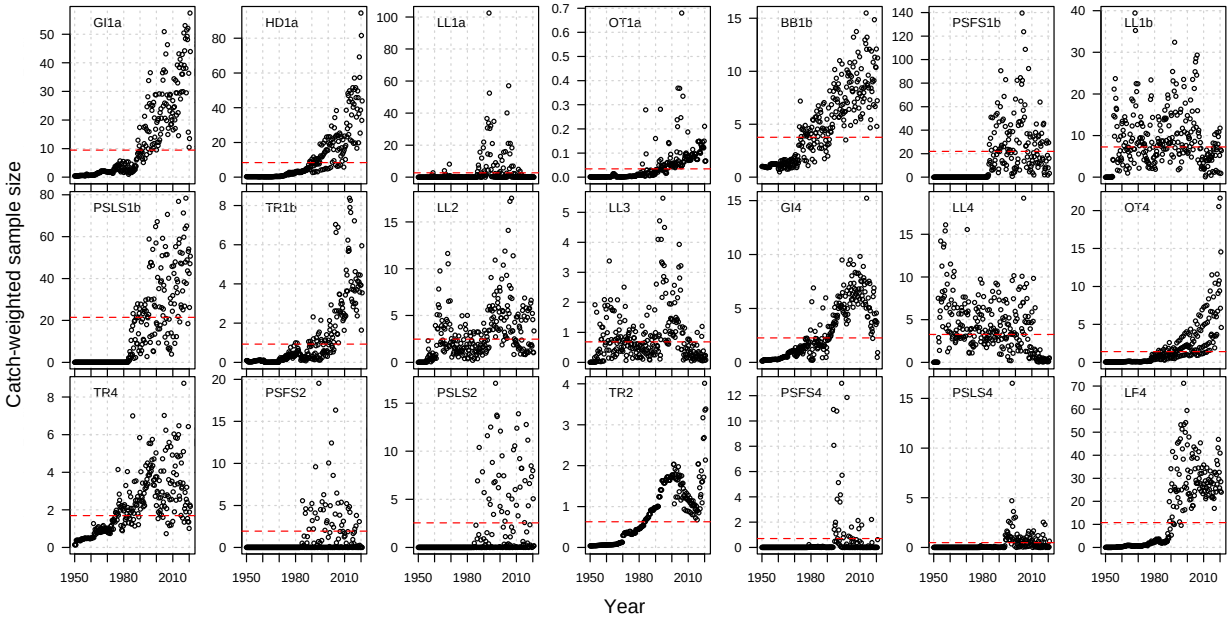

Figure 3. Survey indices used in (a) the assessment model, (b) the one-area models, and (c) the two-area models.

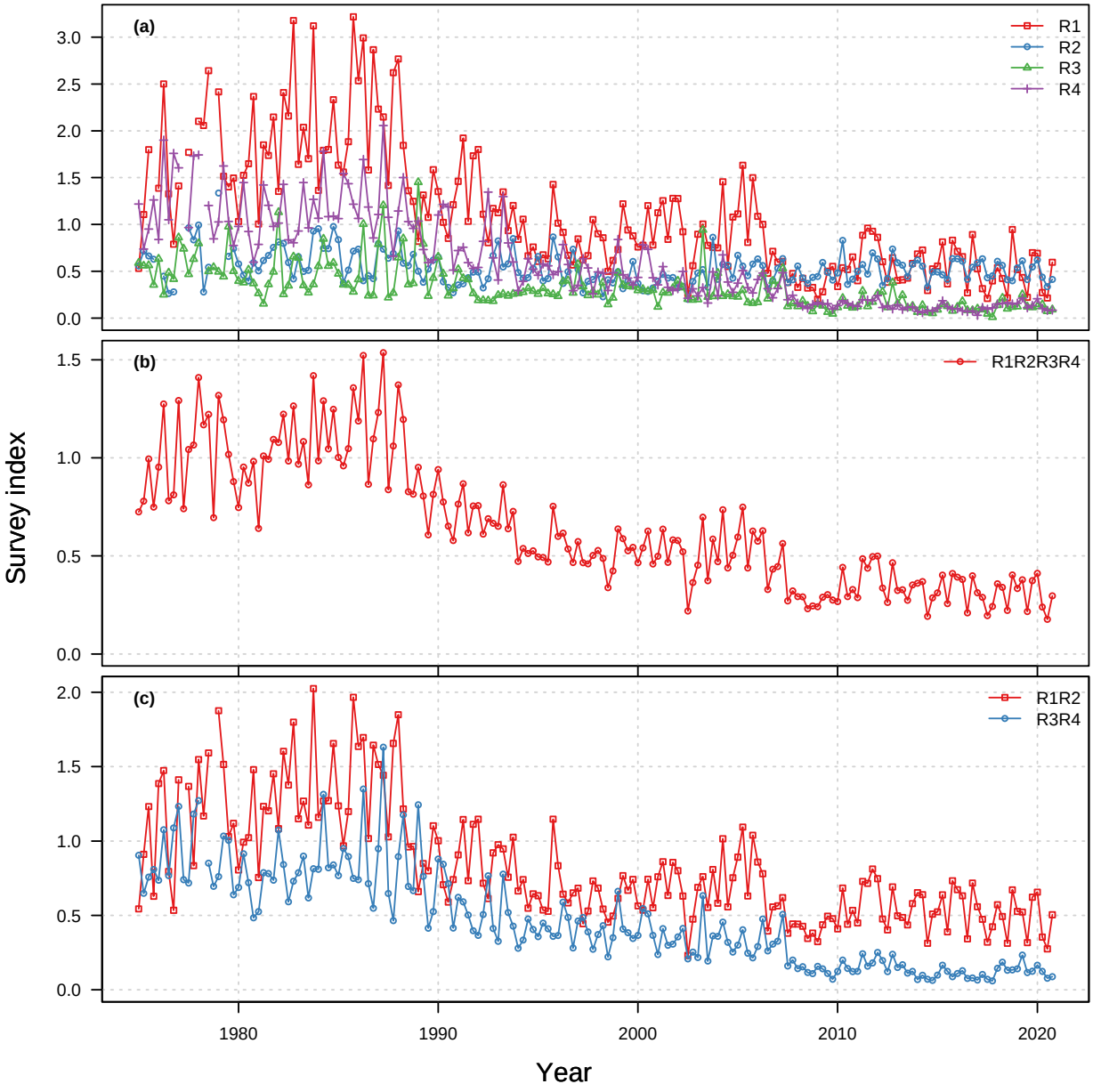

Figure 4. IOYT catches from combined R1/R2 and R3/R4 areas, 1950-2020.

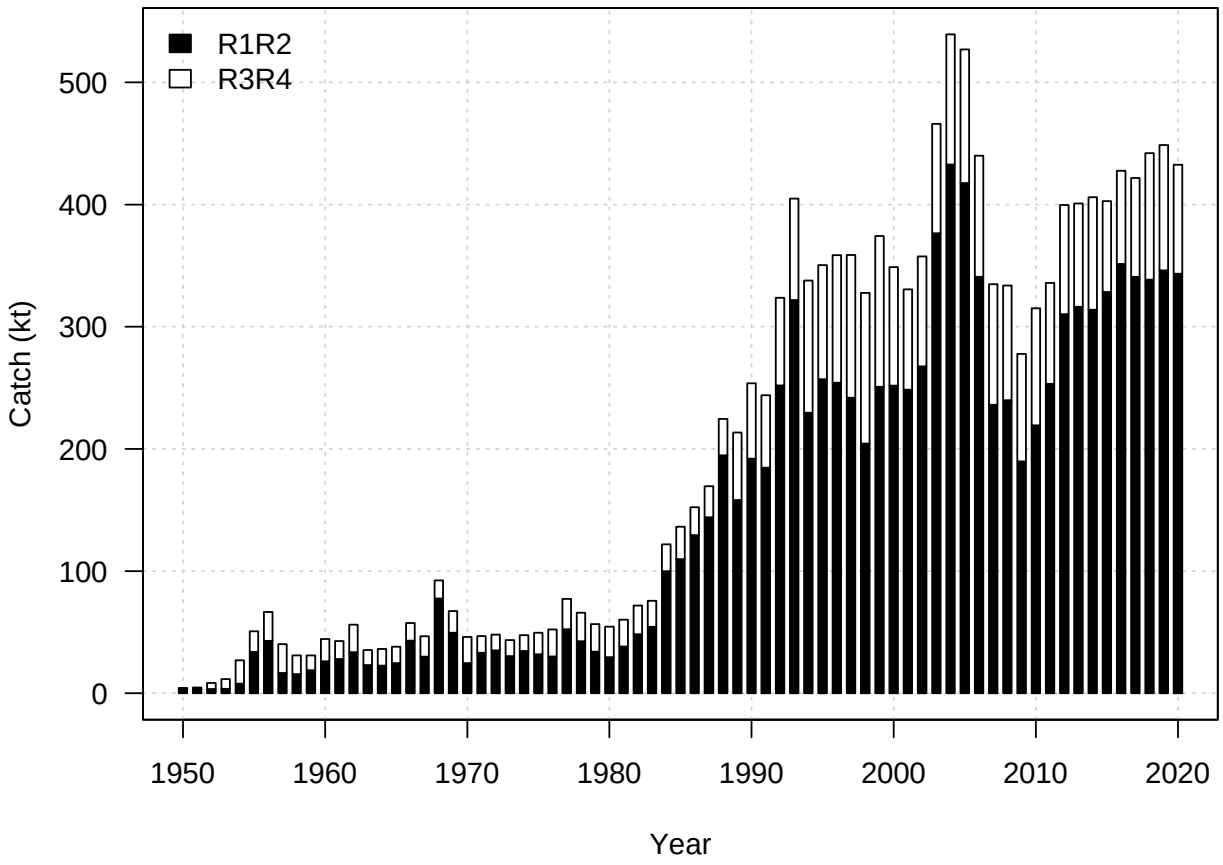

Figure 5. Fits of 1-area models to combined survey indices.

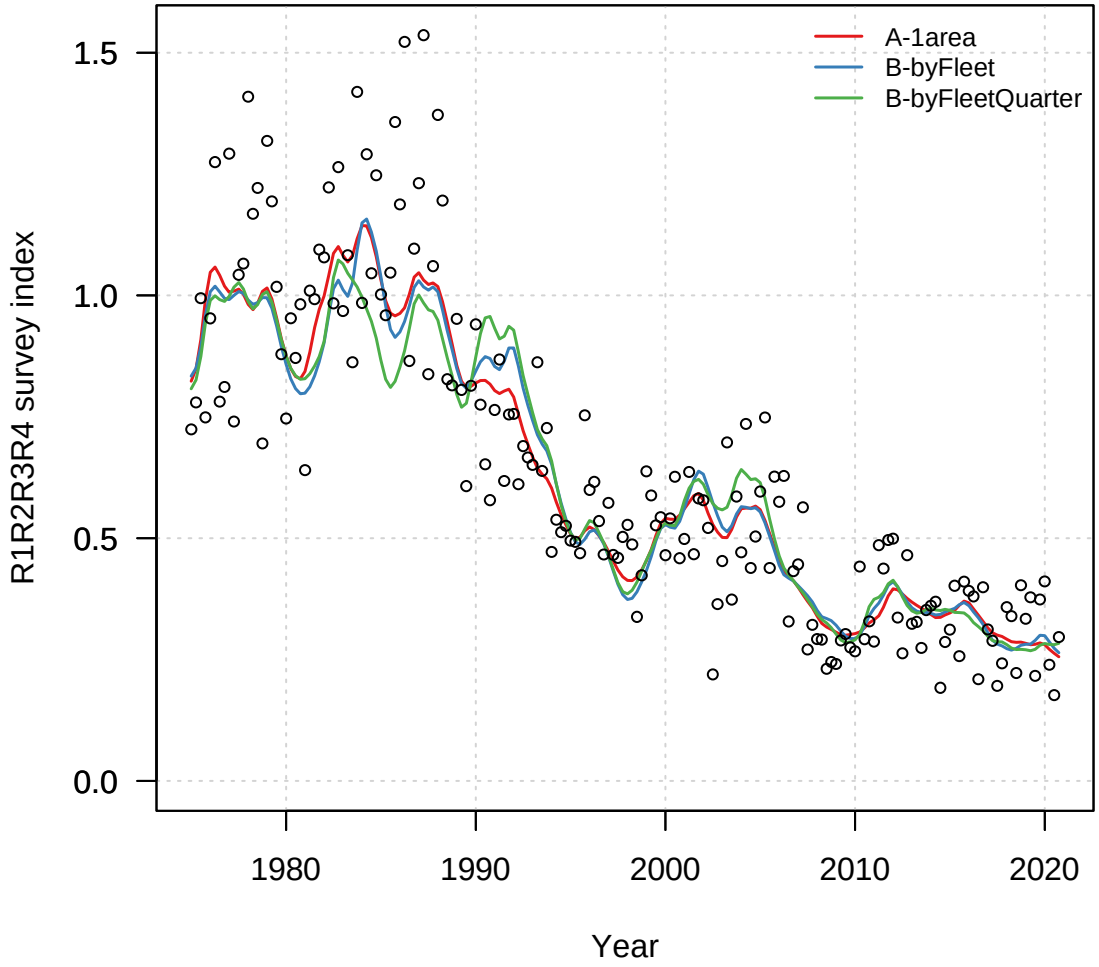

Figure 6. Survey index residuals for one-area models. Residuals are calculated as (ln(Exp)-ln(Obs))/SE. Residuals from common calendar years are alternatively coloured black and white.

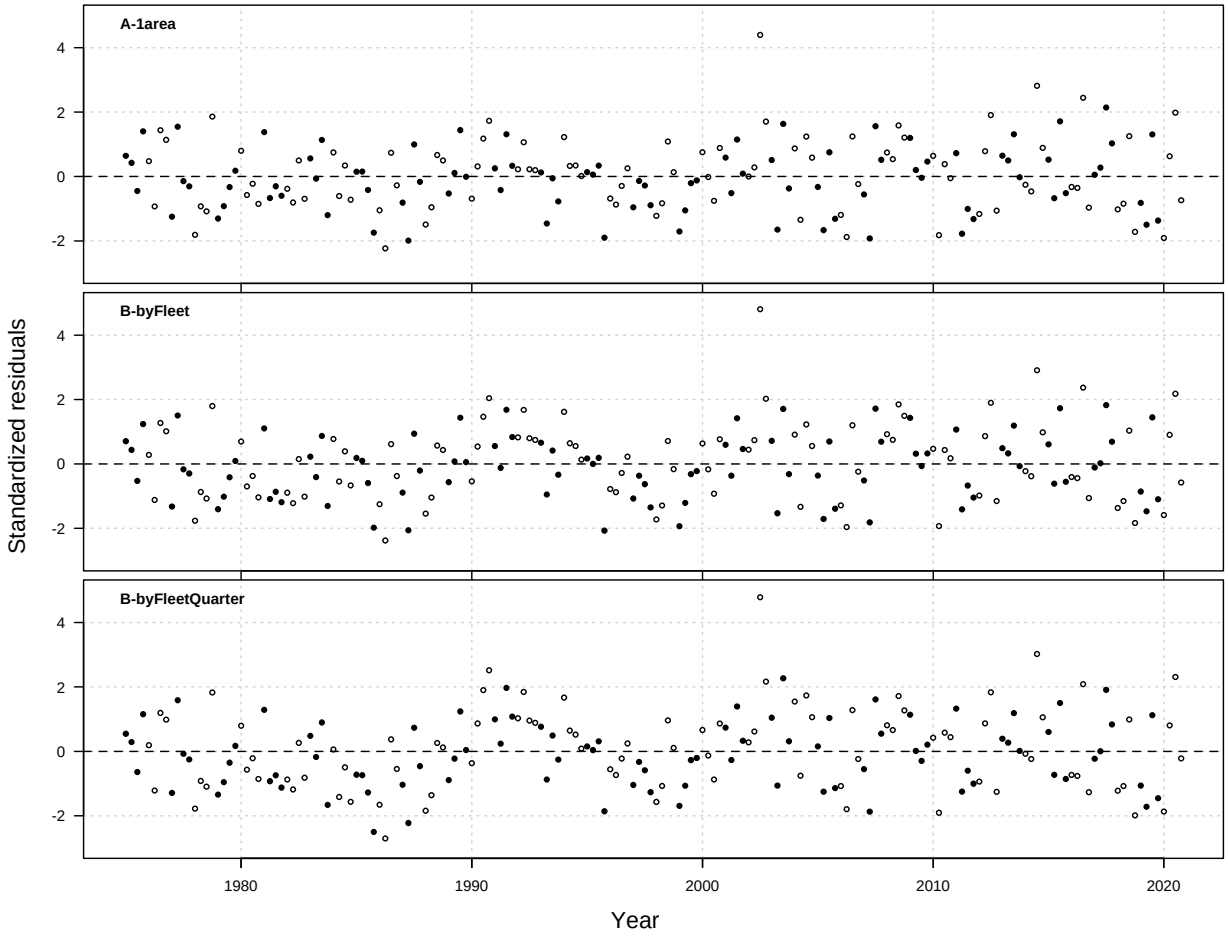

Figure 7. Autocorrelation function plots of residuals from 1-area model fits to survey index.

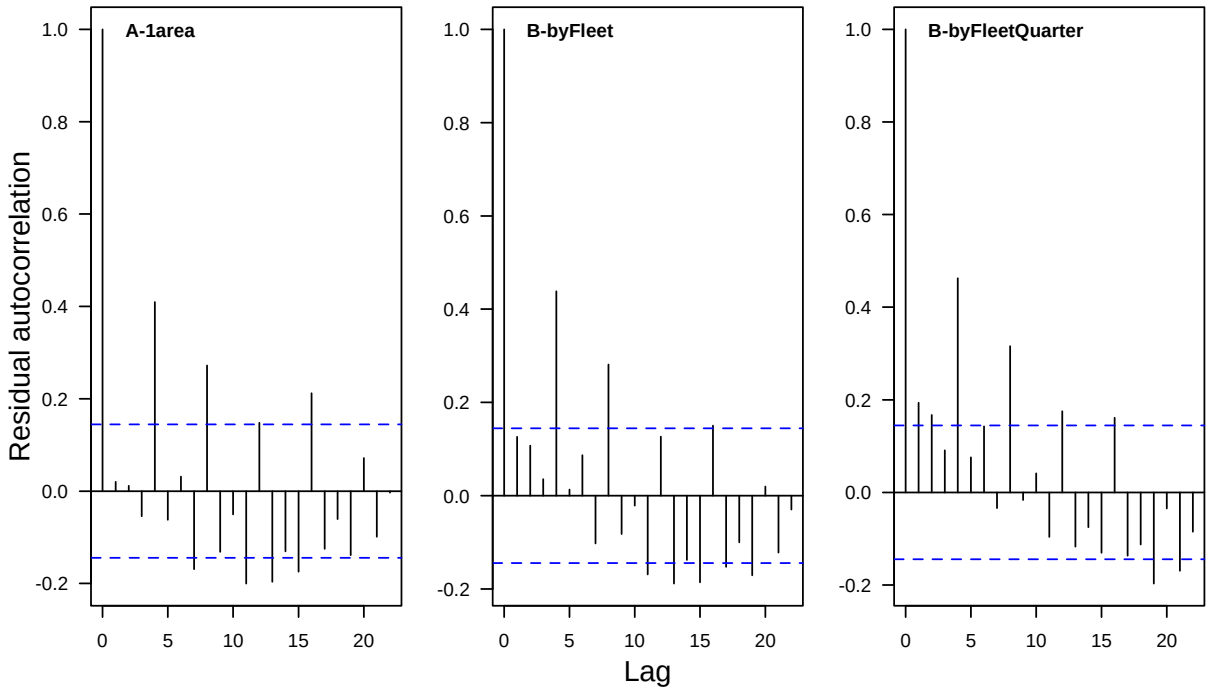

Figure 8. Observed (grey bars) and predicted (green line) temporally aggregated length compositions (in 4 cm intervals) by fishery for the A-1area model.

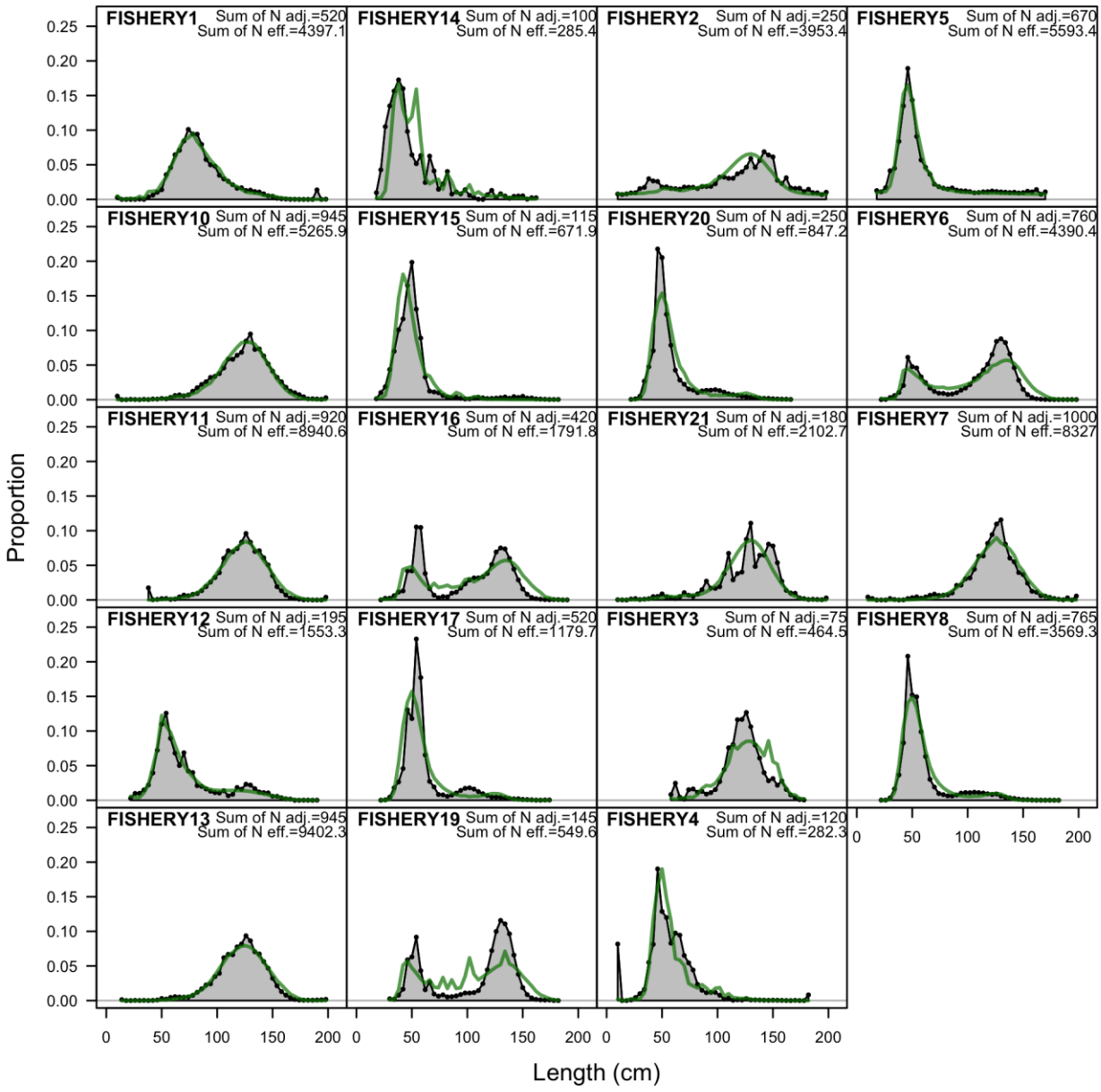

Figure 9. Relative residuals from fits of A-1area model to the length compositions for longline fisheries.

Figure 10. Relative residuals from fits of A-1area model to the length compositions for four purse seine fisheries.

Figure 11. Observed (grey bars) and predicted (green line) temporally aggregated length compositions (in 4 cm intervals) by fishery for the B-byFleet model.

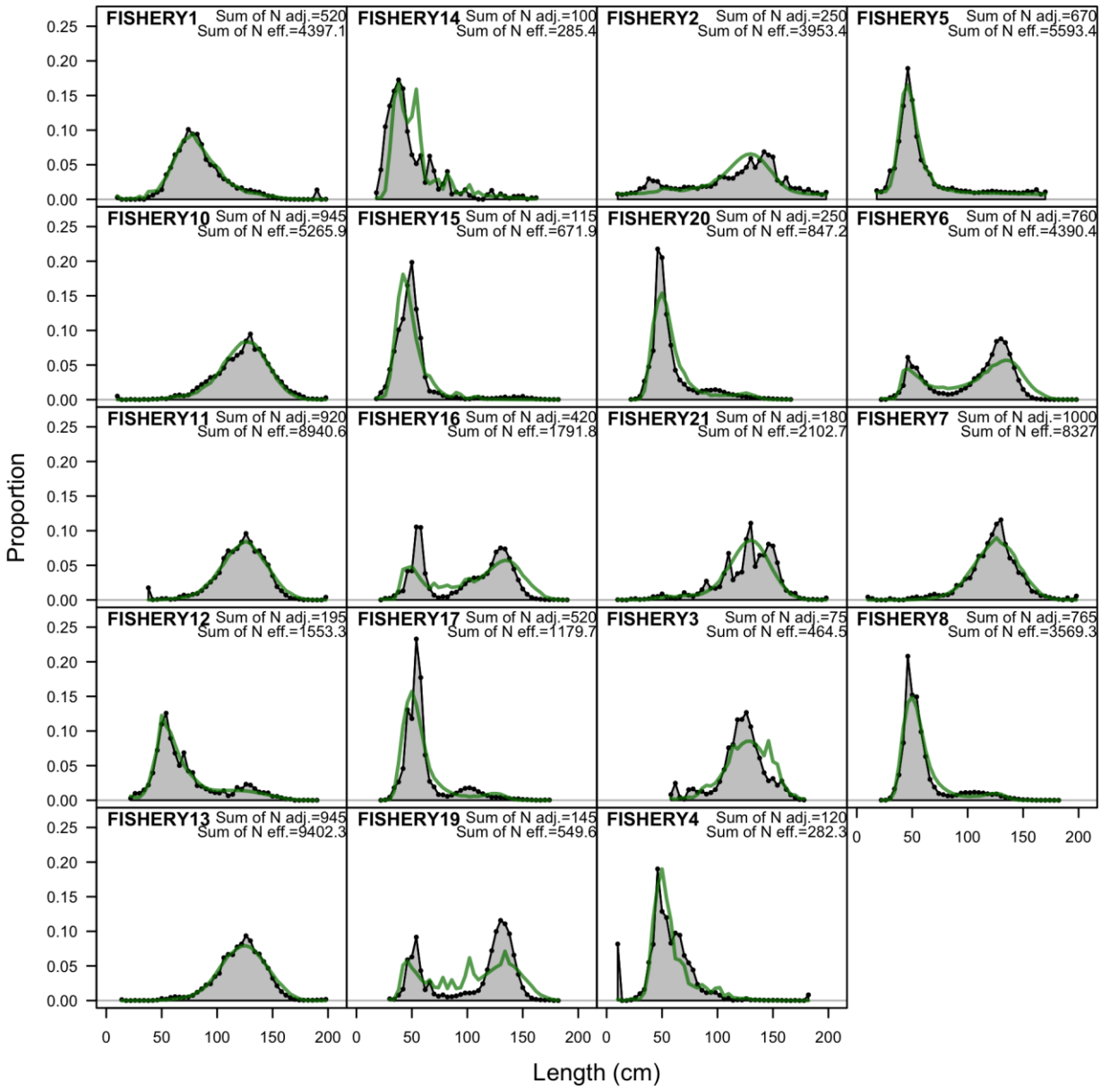

Figure 12. Observed (grey bars) and predicted (green line) temporally aggregated length compositions (in 4 cm intervals) by fishery for the B-byFleetQuarter model.

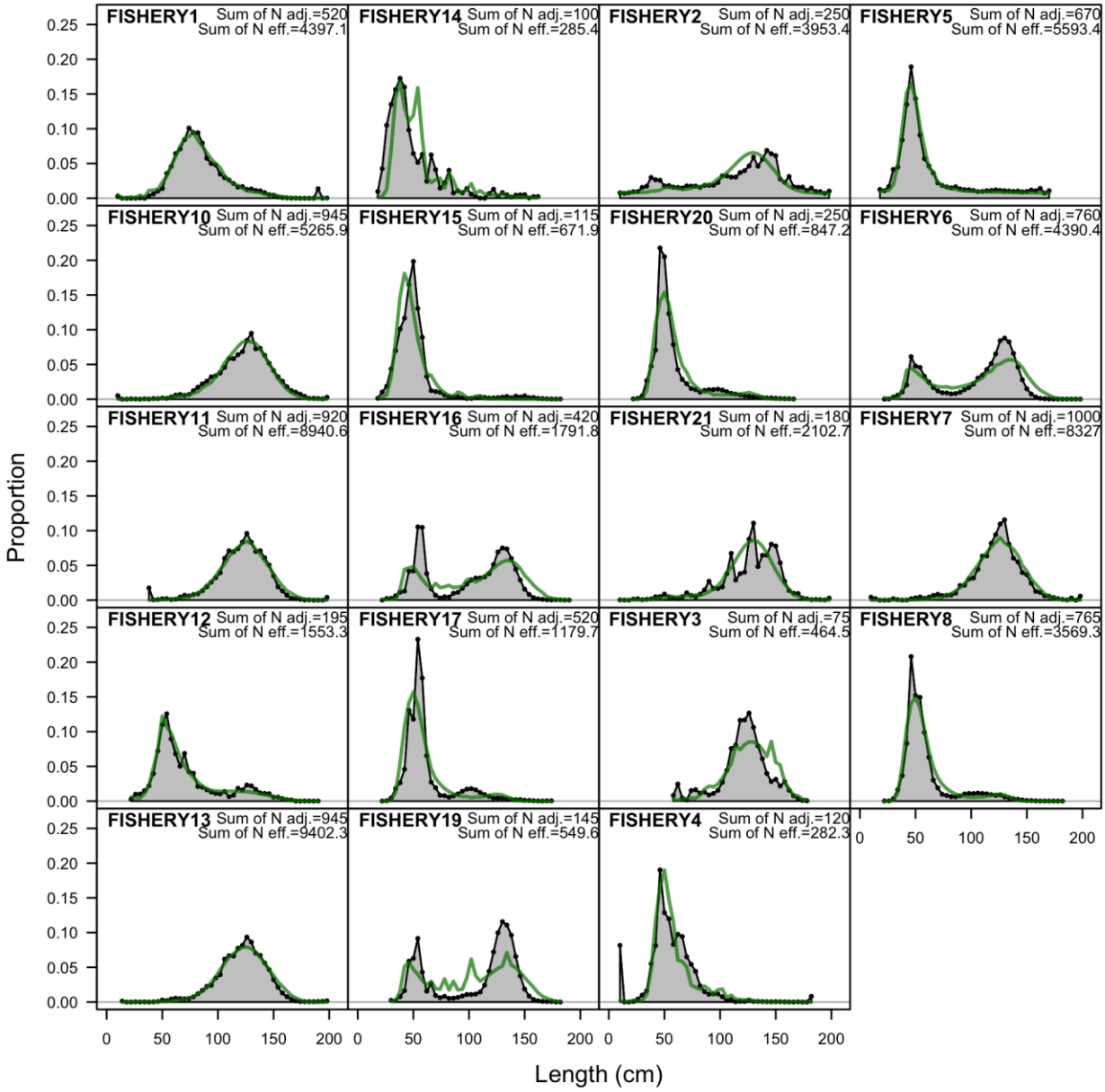

Figure 13. Relative residuals from fits of the B-byFleet model to the length compositions for longline fisheries.

Figure 14. Relative residuals from fits of the B-byFleet model to the length compositions for four purse seine fisheries.

Figure 15. Relative residuals from fits of the B-byFleetQuarter model to the length compositions for longline fisheries.

Figure 16. Relative residuals from fits of the B-byFleetQuarter model to the length compositions for four purse seine fisheries.

Figure 17. Estimated spawning biomass from the assessment model and seven alternative models.

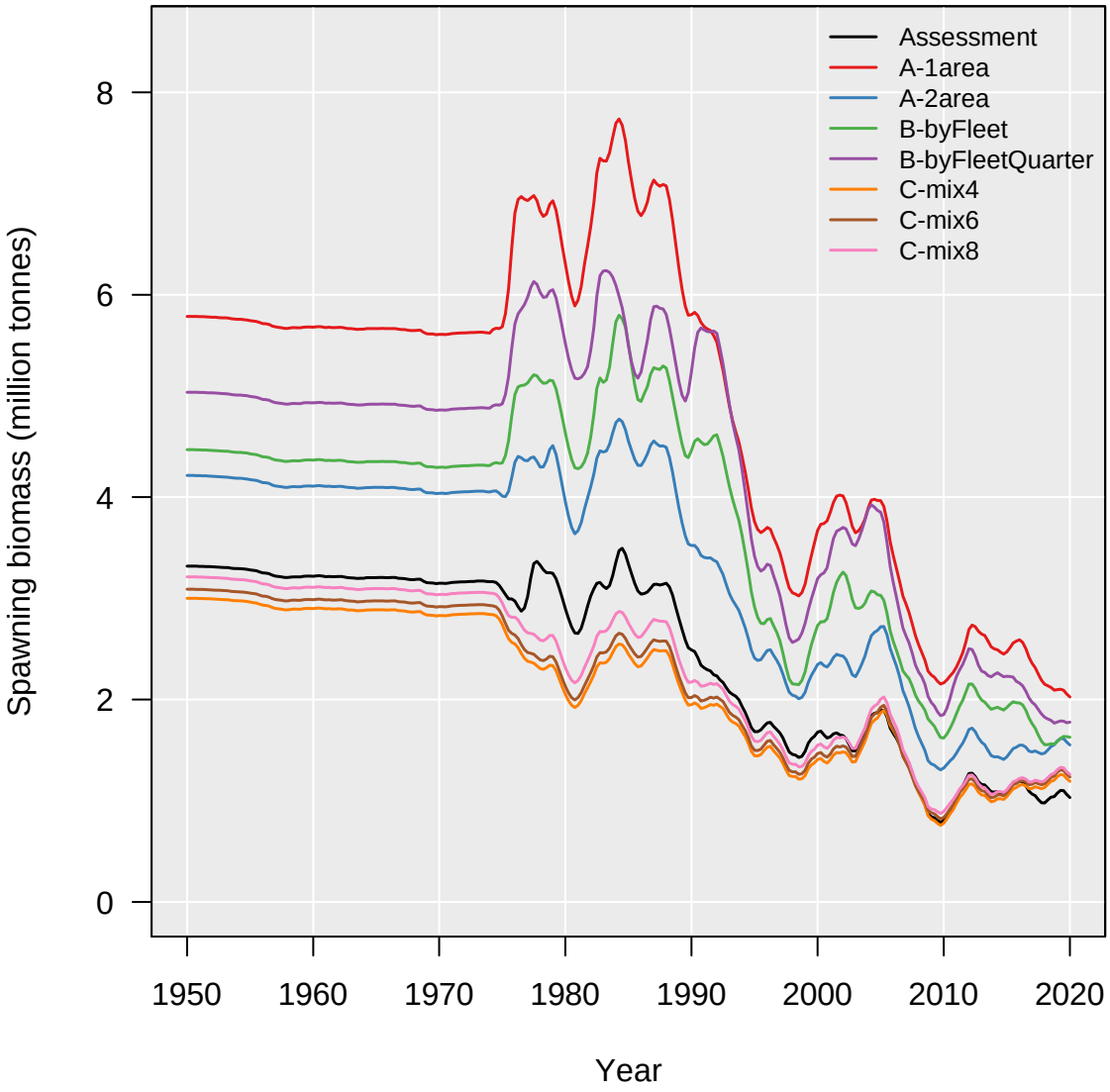

Figure 18. Model A-2area fits to survey indices.

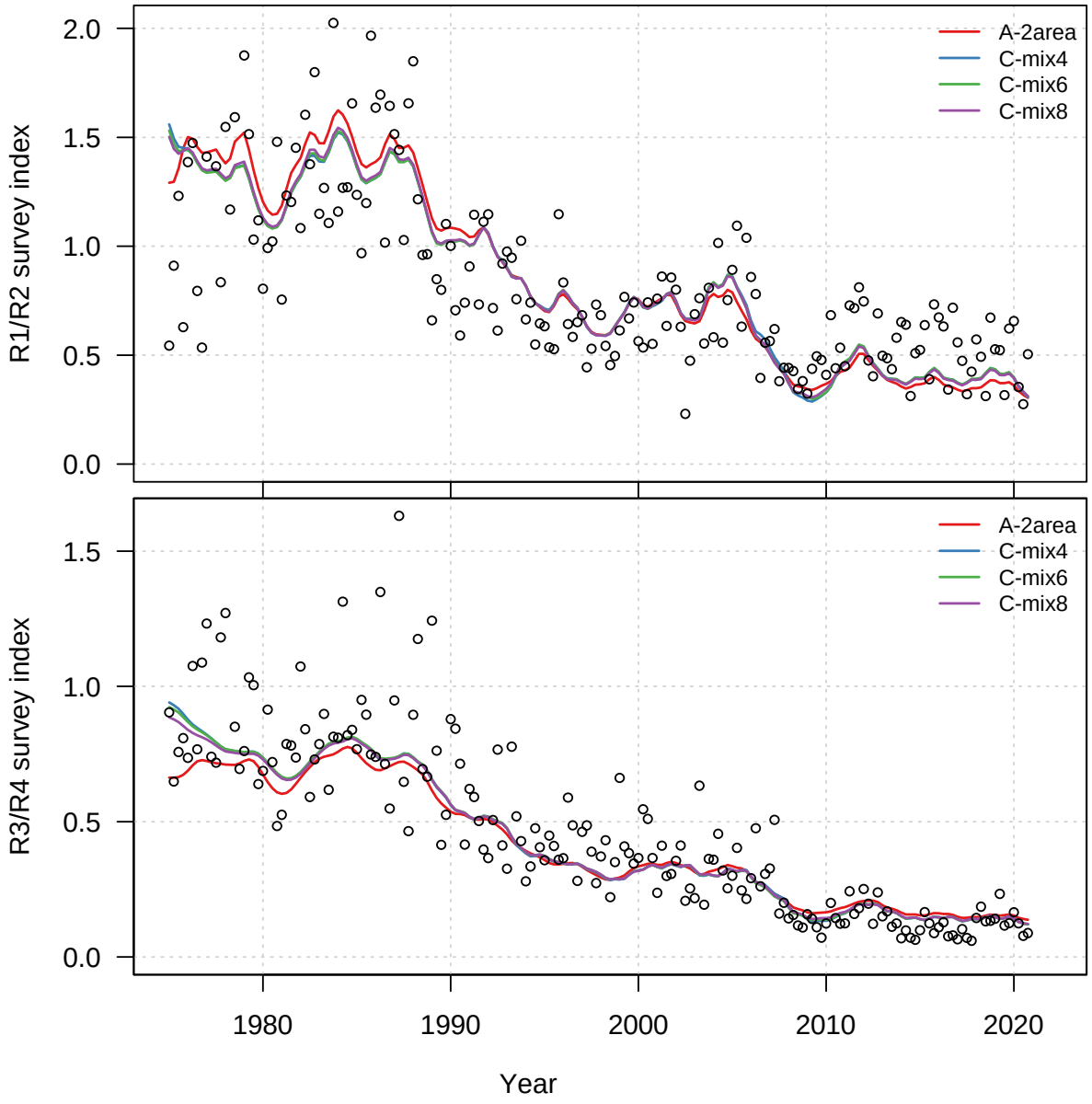

Figure 19. Survey index residuals by area (columns) and model (rows) for each two-area model. Residuals are calculated as (ln(Exp)-ln(Obs))/SE. Residuals from common calendar years are alternatively coloured black and white.

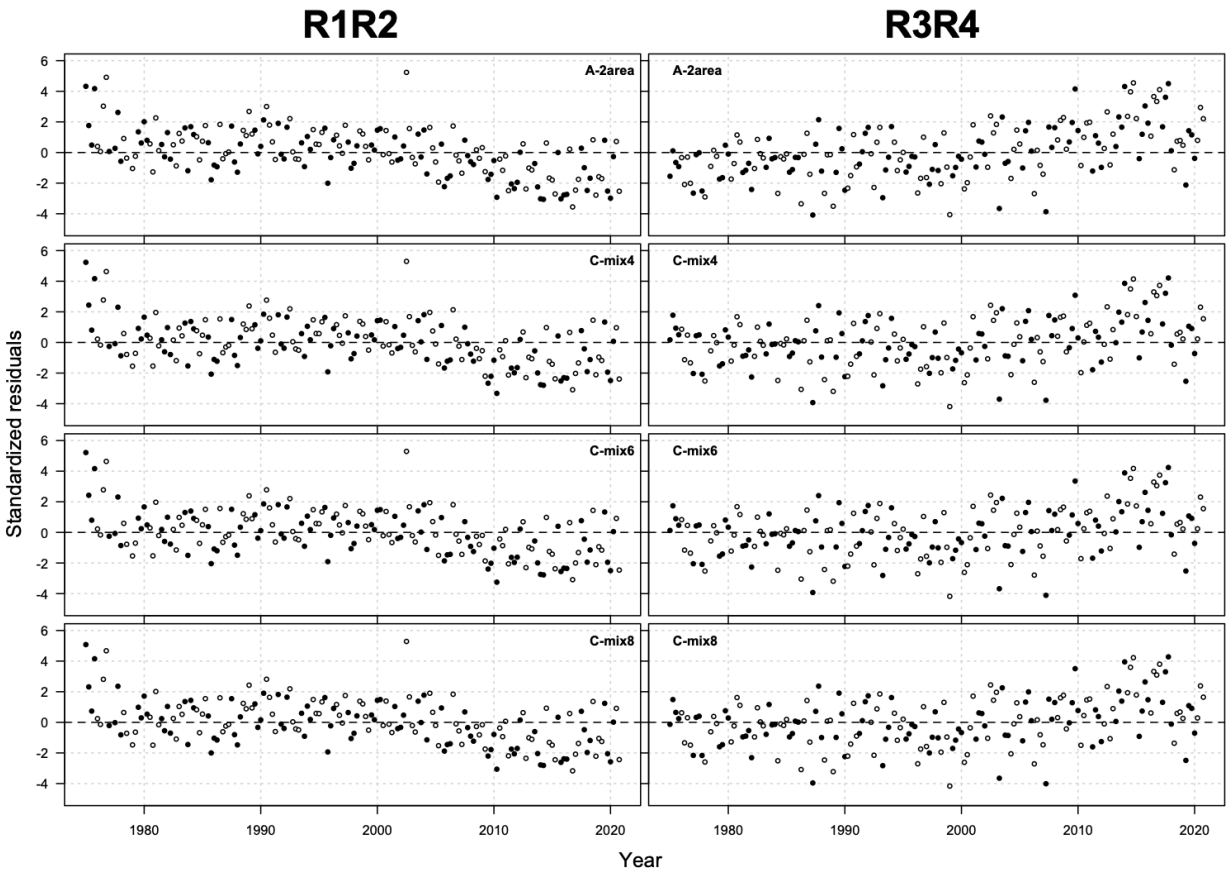

Figure 20. Observed (grey bars) and predicted (red line) temporally aggregated length compositions (in 4 cm intervals) by fishery for the A-2area model.

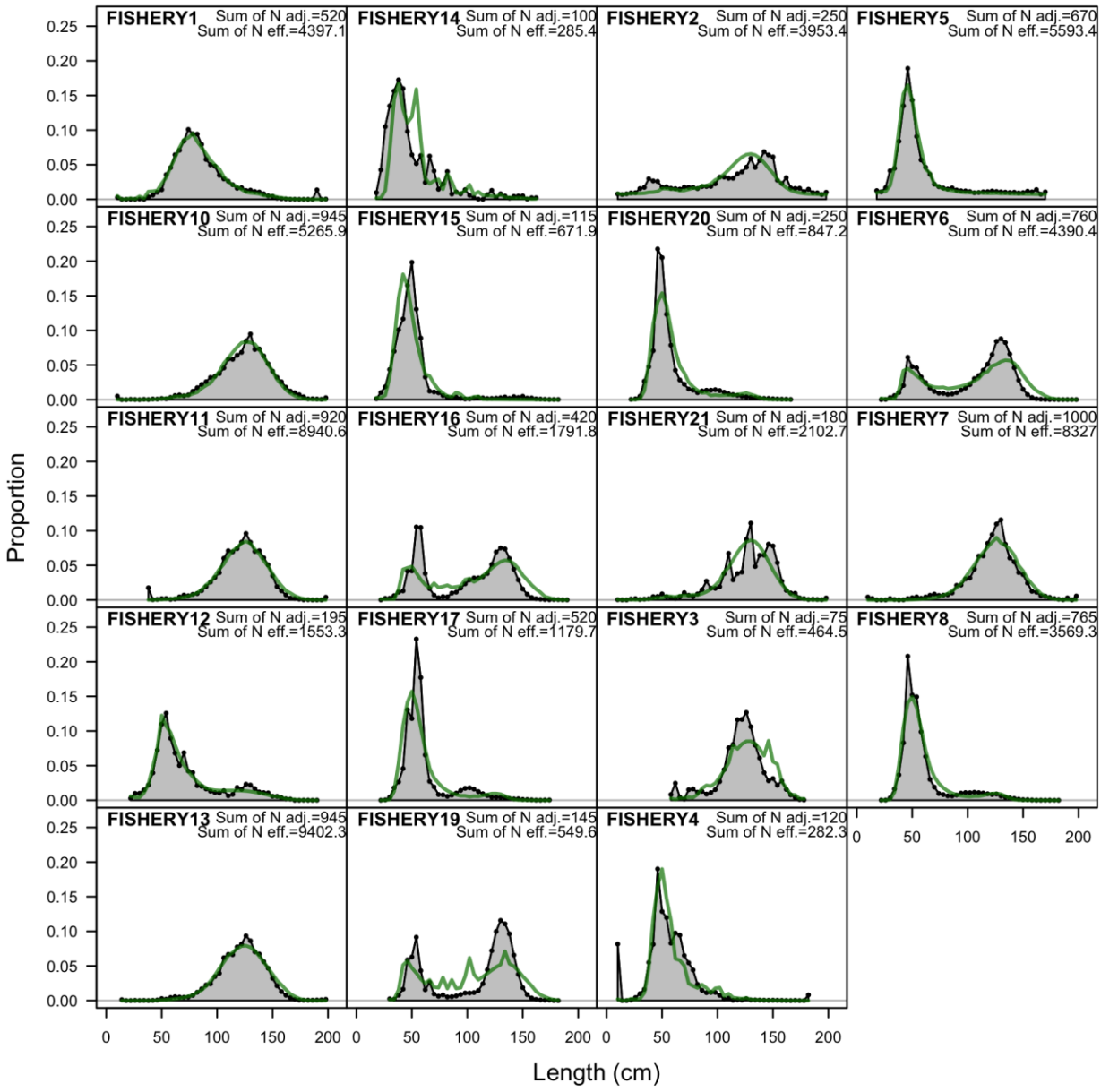

Figure 21. Relative residuals from fits of A-2area model to the length compositions for longline fisheries.

Figure 22. Relative residuals from fits of A-2area model to the length compositions for four purse seine fisheries.

Figure 23. Post-latency tag recaptures aggregated across tag groups for the assessment model and three alternative two-area models with mixing latency periods of 4, 6, and 8 seasons. Bars and lines represent observed and expected tag recaptures, respectively.

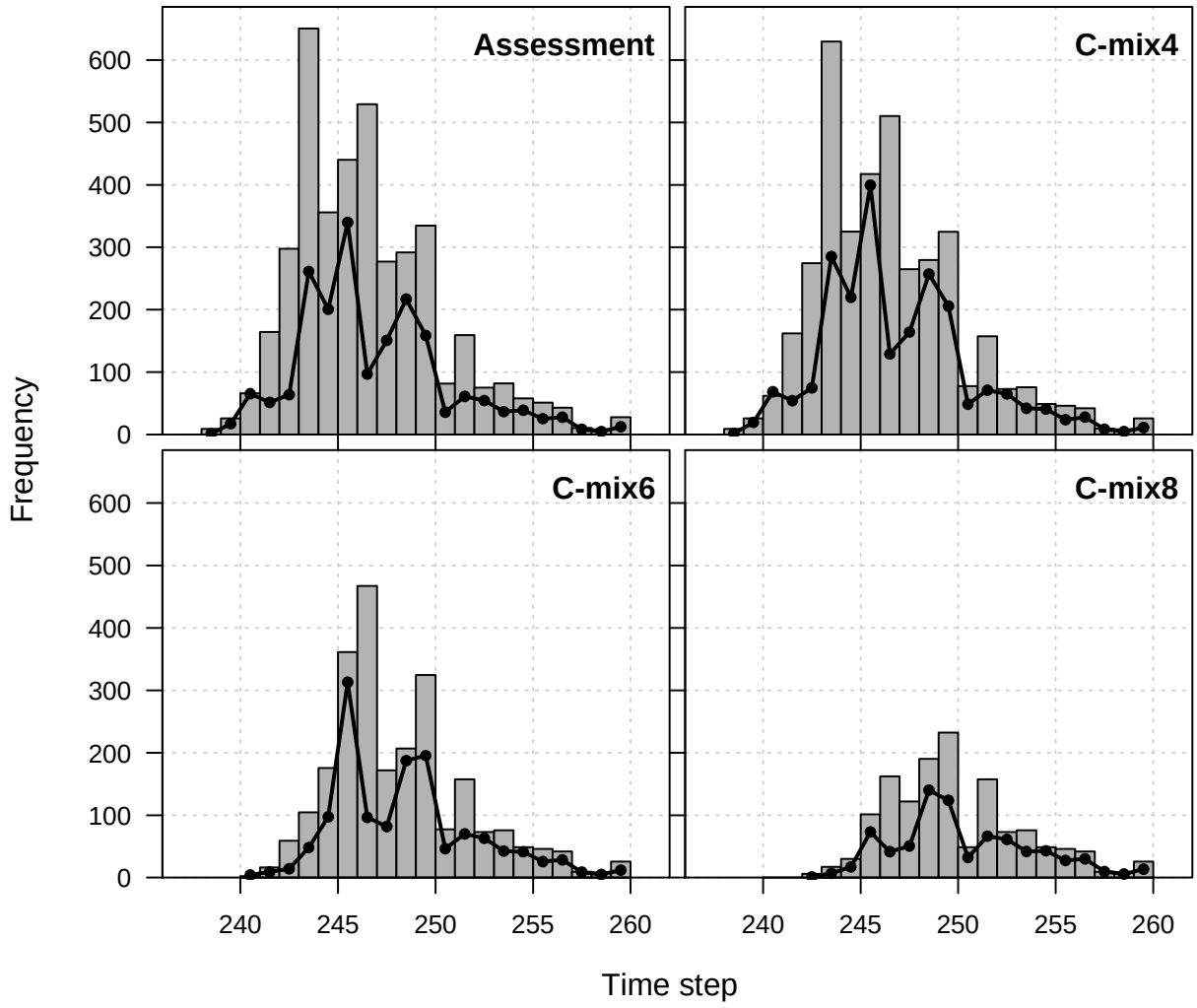

Figure 24. Residuals for post-latency tag recaptures from the C-mix4 model.

Figure 25. Residuals for post-latency tag recaptures from the C-mix6 model.

Figure 26. Residuals for post-latency tag recaptures from the C-mix8 model.

Figure 27. Estimated spawning biomass by area from the assessment model and four alternative two-area models.