Residual Diagnostics for Indian Ocean yellowfin tuna Stock Synthesis models

Henning Winker

Summary

This paper presents additional residual diagnostics for the 2018 Stock Synthesis 4-Area reference case model and the newly proposed 2019 2-Area reference model for Indian Ocean yellowfin tuna. The set of residual diagnostics comprise runs tests to evaluate the randomness of residuals, three-sigma limits to identify outliers, and "JABBA residual plots" together with associated Residual-Mean-Squared (RMSE) values. In addition, a summary of runs test results is provided for the 24 Stock Synthesis models of the 2018 reference grid. The residual diagnostic results indicate a slightly higher runs test passing rate for the 2018 4-Area reference case model, but showed overall no clear evidence in favor of either reference case model. On closer inspection of the individual grid model runs, the scenarios that estimate a second catchability coefficient (q2) were generally associated lower runs test passing rates for CPUE residual than alternative formulations. The generally poor runs test passing rates of mean lengths residuals may be interpreted in support of the relative low weight that currently assigned to the length composition for both reference case models.

Introduction

Integrated assessment models, such as Stock Synthesis (Methot and Wetzel, 2013), have enabled stock assessment scientists to draw simultaneous inference from multiple data sources. These can include abundance indices, length or age composition and tagging data. However, fitting multiple data series can results conflicts among the data sources as well as conflicts between data and assumptions about population dynamics. Whereas some of these conflicts may arise from sampling errors and systemically biased input data, others may be a symptom of model misspecification. Analysis of residuals is a critical element of model diagnostics, especially when evaluated in conjunction with other recommended diagnostic approaches, such as log-likelihood profiling, retrospective analysis (Carvalho et al., 2017) and recent developments of hind-casting cross-validation techniques to evaluate the predictive skill (Kell et al., 2016).

This paper provides additional residual diagnostics to complement the routine residual analysis as implemented in the R package 'r4ss' (Taylor et al., 2013). These comprise runs tests to evaluate the randomness of residuals (Carvalho et al. 2017), three-sigma limits to identify outliers (Anhøj and Olesen, 2014) and "JABBA residual plots" together with Residual-Mean-Squared (RMSE) values (Winker et al., 2018). This set of residual diagnostics are applied to the 2018 Stock Synthesis reference case (Fu et al., 2018), the newly proposed 2019 2-Area reference model that removes the influence of tagging data (IOTC-2019-WPTT21-50). In addition, a summary of runs test results for the 24 Stock Synthesis models included in the 2018 reference grid for Indian Ocean yellowfin tuna (Fu et al. 2018).

Material and Methods

Stock assessment outputs

The 2018 Stock Synthesis reference was fitted to four abundance indices (Regions 1-4)with a steepness h = 0.8, single catchability (q1) for longline, tag-release mortality of 28.5% and tag data not down-weighed (lamda = 1). The proposed 2019 reference model RC2 assumes two Regions and

is fitted to two joint indices I1 (joint for Regions 1-2) and I2 (joint for Regions 3-4), based on assumption that there is no movement between the two areas and removing the influence of tagging data. The 2018 reference grid were build based on factorial design and comprised of 24 Stock Synthesis model of the reference case.

Each Stock Synthesis model output was provided in a folder that was labelled according to the model name. The data under assessment were the (i) residuals of observed and expected CPUE indices by area (I1-I4) and quarter (Q1-Q4), (ii) residual the observed and predicted mean length of the size composition by fleet (F1-F25) and quarter (Q1-Q4) and (iii) annual recruitment residuals. The data series were imported from the Report.sso file into the software environment R with the SS_output() function from the R package 'r4ss' (Taylor et al., 2013). The mean length residuals were constructed by customizing source code from 'r4ss'.

Residual diagnostics
Runs tests were applied as a residual diagnostic test to quantitatively evaluate the randomness of the time series of CPUE and mean length residuals on log scale by fleet and quarter (Carvalho et al. 2017). In addition, runs tests were applied to the estimated recruitment deviations. The runs tests were implemented using the function *runs.test* in the R package *tseries* (Trapletti, 2011), considering the 2-sided p-value of the Wald-Wolfowitz runs test. This test is non-parametric and only considers whether the sequence of residuals is positive or negative, but not their absolute values. The runs test results were visualized using a specifically designed plot that illustrates which time series passed or failed the runs test and highlights individual time-series data points fall outside the three-sigma limits (e.g. Anhøj and Olesen 2014).

JABBA-residual plots used to graphically evaluate potential conflict among CPUE residuals from the different indices and mean length residuals from the different fleets. The JABBA-residual plot shows: (1) colour coded lognormal residuals of observed versus predicted CPUE indices by fleet, (2) boxplots indicating the median and quantiles of all residuals available for any given year; the area of each box indicates the strength of the discrepancy between CPUE series (larger box means higher degree of conflicting information), (3) a loess smoother through all residuals which highlights systematically auto-correlated residual patterns, and (4) depicts the Residual-Mean-Squared-Errors (RMSE) for all residuals shown in the plot (Winker et al., 2018).

Results and Discussions
The 2018 4-Area reference model passed 11 out of 12 runs tests on quarterly CPUE indices with exception of Index 2 in Quarter 1 (Figure 1; Table 1), while the 2019 2-Area reference model passed 6 out of 8. The two instances where the runs test failed to reject the null hypothesis of a non-random residual pattern were Index 1 in quarter 2 and Index 2 in quarter 4 (Figure 2; Table 2), with the latter clearly indicating extended "runs" with negative recruitment residuals (Figure 2). The RMSE values for the CPUE residuals ranged from about 28-36% for the 2018 reference case (Figure 3) and from about 22%-26% for the 2019 reference case (Figure 4), where ranges reflect the variations in RMSE values among the four quarters. However, despite the lower RMSE values, the CPUE fits from 2019 Reference model appear to be associated with an overall stronger correlated residual pattern (Figure 4). The summary of CPUE runs test results for the 2018 grid model showed that quarter 1 was associated with substantially poorer passing rates than the quarters (Table 1). A second emerging pattern was relatively poorer passing of models that estimated an additional catchability (q2).

Runs tests conducted on meal length residuals only passed for 11 of 25 (44%) fleets in the case of the 2018 reference case (Figure 5) and for 10 of 25 fleets in (40%) in the case of the 2019 reference case (Figure 6). The highest p-values and thus best evidence for random residual pattern was generally associated with Fleet 1 (Table 2). By contrast, several mean length residual indicated strong systematic patterns, which resulted in consistent runs test failures for all 2018 grid models that were associated with p-values ($p < 0.001$). Examples of particularly poor mean lengths residual diagnostics include fleets F6-11, F13, F19 and F22 (Figures 5-6; Table 2).

The runs test for the recruitment deviations passed for 2018 reference model (Figure 7), but failed for the 2019 reference model (Figure 8). The recruitment residuals from the 2019 reference model showed a notable systematic pattern in during initiation phase and in most years. The latter is possibly a result limited information in the data, which should be carefully evaluated due it likely influence on projections. Across the 2018 grid, the reference case is only one of four model that passed the runs tests (Table 1), indicating that the runs test may be sensitive to small variations in the recruitment deviations, particularly at the start and towards the terminal year.

Table 1: Summary of runs test results of recruitment residuals and CPUE residuals for indices I1-I5 and quarters Q1-Q4. Green fields and red fields denote if the null hypothesis of a random residual distribution cannot be rejected (pass) and was rejected (fail; $p < 0.05$), respectively.

Model	Rec	I1.Q1	I2.Q1	I3.Q1	I4.Q1	I5.Q1	I1.Q2	I2.Q2	I3.Q2	I4.Q2	I5.Q2	I1.Q3	I2.Q3	I3.Q3	I4.Q3	I5.Q3	I1.Q4	I2.Q4	I3.Q4	I4.Q4	I5.Q4
RC2_ref	**0.000**	**0.028**	**0.538**				**0.845**	**0.28**				**0.337**	**0.978**				**0.302**	**0.036**			
io_h80_q1_tm30_dw1	**0.089**	**0.136**	**0.017**	**0.221**	**0.052**		**0.124**	**0.513**	**0.601**	**0.373**		**0.478**	**0.886**	**0.429**	**0.804**		**0.804**	**0.311**	**0.624**	**0.854**	
io_h70_q1_tm10_dw1	0.011	0.038	0.202	0.221	0.052		0.309	0.056	0.941	0.919		0.478	0.374	0.396	0.804		0.271	0.904	0.624	0.809	
io_h70_q1_tm10_dw2	0.005	0.139	0.124	0.221	0.009		0.361	0.673	0.941	0.76		0.478	0.967	0.396	0.853		0.925	0.197	0.624	0.774	
io_h70_q1_tm30_dw1	0.132	0.037	0.017	0.221	0.052		0.309	0.143	0.601	0.602		0.478	0.948	0.804	0.673		0.804	0.311	0.591	0.618	
io_h70_q1_tm30_dw2	0.016	0.038	0.124	0.221	0.009		0.309	0.056	0.601	0.76		0.478	0.575	0.804	0.853		0.925	0.467	0.385	0.774	
io_h70_q2_tm10_dw1	0.023	0.231	0.039	0.221	0.052	0.48	0.499	0.056	0.941	0.919	0.724	0.513	0.886	0.396	0.804	0.001	0.573	0.946	0.624	0.809	0.157
io_h70_q2_tm10_dw2	0.013	0.244	0.025	0.221	0.009	0.48	0.499	0.056	0.601	0.76	0.724	0.56	0.907	0.804	0.853	0.001	0.96	0.453	0.385	0.774	0.157
io_h70_q2_tm30_dw1	0.031	0.244	0.08	0.221	0.009	0.48	0.9	0.056	0.601	0.602	0.724	0.756	0.615	0.804	0.853	0.001	0.898	0.989	0.591	0.601	0.157
io_h70_q2_tm30_dw2	0.054	0.063	0.03	0.221	0.009	0.48	0.9	0.303	0.941	0.76	0.724	0.56	0.967	0.396	0.853	0.001	0.96	0.197	0.624	0.774	0.157
io_h80_q1_tm10_dw1	0.005	0.037	0.202	0.221	0.011		0.309	0.056	0.601	0.919		0.24	0.374	0.396	0.775		0.271	0.904	0.624	0.898	
io_h80_q1_tm10_dw2	0.002	0.038	0.025	0.221	0.009		0.309	0.08	0.941	0.76		0.478	0.907	0.396	0.804		0.925	0.453	0.624	0.774	
io_h80_q1_tm30_dw2	0.011	0.037	0.039	0.221	0.009		0.309	0.056	0.941	0.76		0.478	0.575	0.396	0.804		0.925	0.467	0.624	0.774	
io_h80_q2_tm10_dw1	0.012	0.06	0.202	0.221	0.011	0.48	0.499	0.035	0.601	0.919	0.724	0.756	0.168	0.396	0.775	0.001	0.519	0.904	0.624	0.898	0.157
io_h80_q2_tm10_dw2	0.026	0.244	0.025	0.258	0.009	0.48	0.499	0.056	0.941	0.76	0.724	0.56	0.907	0.396	0.804	0.001	0.96	0.453	0.624	0.774	0.157
io_h80_q2_tm30_dw1	0.045	0.063	0.08	0.467	0.009	0.48	0.15	0.411	0.601	0.919	0.724	0.099	0.718	0.162	0.804	0.001	0.96	0.13	0.624	0.601	0.001
io_h80_q2_tm30_dw2	0.026	0.244	0.039	0.221	0.009	0.48	0.499	0.035	0.941	0.76	0.724	0.56	0.374	0.396	0.853	0.001	0.96	0.453	0.624	0.774	0.157
io_h90_q1_tm10_dw1	0.013	0.037	0.202	0.221	0.052		0.309	0.056	0.601	0.76		0.24	0.374	0.804	0.853		0.271	0.904	0.385	0.854	
io_h90_q1_tm10_dw2	0.006	0.037	0.089	0.221	0.009		0.309	0.724	0.601	0.76		0.234	0.855	0.804	0.853		0.673	0.197	0.385	0.774	
io_h90_q1_tm30_dw1	0.041	0.136	0.017	0.221	0.052		0.124	0.143	0.601	0.373		0.478	0.886	0.429	0.804		0.853	0.311	0.624	0.854	
io_h90_q1_tm30_dw2	0.012	0.037	0.124	0.221	0.009		0.309	0.08	0.601	0.76		0.478	0.575	0.429	0.853		0.925	0.467	0.385	0.774	
io_h90_q2_tm10_dw1	0.006	0.06	0.202	0.221	0.052	0.48	0.388	0.056	0.601	0.919	0.724	0.56	0.25	0.804	0.853	0.001	0.573	0.467	0.385	0.854	0.157
io_h90_q2_tm10_dw2	0.012	0.244	0.03	0.221	0.009	0.48	0.499	0.056	0.601	0.76	0.724	0.56	0.575	0.804	0.853	0.001	0.96	0.453	0.385	0.774	0.157
io_h90_q2_tm30_dw1	0.055	0.244	0.08	0.221	0.052	0.48	0.9	0.056	0.601	0.919	0.724	0.756	0.506	0.478	0.853	0.001	0.96	0.946	0.591	0.45	0.157
io_h90_q2_tm30_dw2	0.031	0.244	0.03	0.221	0.009	0.48	0.499	0.035	0.601	0.76	1	0.56	0.675	0.429	0.853	0.001	0.96	0.467	0.385	0.774	0.001

Table 2: Summary of runs test results of mean lengths residuals for fleets F1-F255. Green fields and red fields denote if the null hypothesis of a random residual distribution cannot be rejected (pass) and was rejected (fail; $p < 0.05$), respectively.

Model	F1	F2	F3	F4	F5	F6	F7	F10	F11	F12	F13	F14	F15	F8	F16	F17	F19	F20	F21	F22	F23	F24	F25
RC2_ref	**0.076**	**0.039**	**0**	**0.001**	**0.007**	**0.008**	**0**	**0.001**	**0**	**0.051**	**0**	**0.001**	**0.001**	**0.705**	**0.851**	**0.751**	**0.001**	**0.208**	**0.348**	**0**	**0.916**	**0.665**	**0.48**
io_h80_q1_tm30_dw1	**0.374**	**0.018**	**0**	**0.317**	**0.049**	**0.008**	**0**	**0**	**0**	**0.046**	**0**	**1**	**0.02**	**0.705**	**0.786**	**0.571**	**0.001**	**0.572**	**0.073**	**0.007**	**0.532**	**0.796**	**0.902**
io_h70_q1_tm10_dw1	0.208	0.018	0	0.317	0.049	0.008	0	0	0	0.046	0	1	0.02	0.705	0.704	0	0.001	0.193	0.016	0.001	0.57	0.39	0.414
io_h70_q1_tm10_dw2	0.12	0.009	0	0.317	0.023	0.008	0	0.005	0	0.046	0	1	0.014	0.705	0.704	0.001	0.001	0.966	0.066	0	0.57	0.01	0.296
io_h70_q1_tm30_dw1	0.374	0.018	0	0.317	0.049	0.008	0	0	0	0.898	0	0.748	0.014	0.705	0.718	0.344	0.001	0.208	0.066	0.007	0.532	0.001	0.902
io_h70_q1_tm30_dw2	0.12	0.039	0	0.317	0.017	0.008	0	0.001	0	0.975	0	0.748	0	0.705	0.704	0.098	0.001	0.328	0.155	0.001	0.57	0.103	0.296
io_h70_q2_tm10_dw1	0.095	0.018	0	0.317	0.045	0.008	0	0	0	0.046	0	1	0.02	0.705	0.704	0	0.001	0.193	0.057	0	0.57	0.39	0.414
io_h70_q2_tm10_dw2	0.075	0.039	0	0.317	0.017	0.008	0	0.004	0	0.943	0	0.748	0	0.705	0.704	0.001	0.001	0.328	0.008	0	0.634	0.103	0.296
io_h70_q2_tm30_dw1	0.374	0.018	0	0.317	0.045	0.008	0	0	0	0.898	0	0.748	0.02	0.705	0.634	0.344	0.001	0.328	0.073	0.011	0.916	0.001	0.902
io_h70_q2_tm30_dw2	0.301	0.137	0	0.317	0.02	0.008	0	0.003	0	0.046	0	0.748	0.014	0.705	0.536	0.084	0.001	0.966	0.055	0.001	0.344	0.01	0.296
io_h80_q1_tm10_dw1	0.208	0.003	0	0.317	0.049	0.008	0	0	0	0.046	0	1	0.02	0.705	0.704	0	0.001	0.193	0.016	0.001	0.57	0.39	0.296
io_h80_q1_tm10_dw2	0.12	0.039	0	0.317	0.02	0.008	0	0.004	0	0.046	0	1	0.014	0.513	0.704	0	0.001	0.966	0.06	0.001	0.57	0.01	0.014
io_h80_q1_tm30_dw2	0.12	0.039	0	0.317	0.017	0.008	0	0.001	0	0.046	0	1	0.014	0.705	0.704	0.061	0.001	0.966	0.155	0.001	0.57	0.01	0.014
io_h80_q2_tm10_dw1	0.095	0.003	0	0.317	0.053	0.008	0	0	0	0.046	0	1	0.02	0.705	0.704	0	0.001	0.193	0.057	0.001	1	0.39	0.296
io_h80_q2_tm10_dw2	0.12	0.039	0	0.317	0.02	0.008	0	0.001	0	0.046	0	1	0.014	0.513	0.704	0	0.001	0.966	0.008	0	0.634	0.01	0.014
io_h80_q2_tm30_dw1	0.078	0.003	0	0.317	0.053	0.008	0	0	0	0.132	0	1	0.014	0.705	0.718	0.347	0.001	0.529	0.073	0.005	0.56	0.796	0.414
io_h80_q2_tm30_dw2	0.12	0.039	0	0.317	0.008	0.008	0	0.001	0	0.046	0	1	0.014	0.705	0.704	0.061	0.001	0.572	0.06	0.001	0.916	0.103	0.296
io_h90_q1_tm10_dw1	0.208	0.018	0	0.317	0.049	0.008	0	0	0	0.975	0	1	0.397	0.705	0.704	0	0.001	0.328	0.016	0.001	0.57	0.39	0.414
io_h90_q1_tm10_dw2	0.076	0.022	0	0.001	0.002	0.008	0	0.004	0	0.943	0	0.748	0	0.705	0.704	0	0.001	0.328	0.06	0.001	0.175	0.086	0.296
io_h90_q1_tm30_dw1	0.078	0.018	0	0.317	0.053	0.008	0	0	0	0.044	0	1	0.02	0.705	0.786	0.571	0.001	0.572	0.073	0.011	0.532	0.796	0.902
io_h90_q1_tm30_dw2	0.121	0.039	0	0.317	0.017	0.008	0	0.001	0	0.975	0	0.077	0.014	0.705	0.704	0.098	0.001	0.328	0.06	0.001	0.332	0.01	0.296
io_h90_q2_tm10_dw1	0.095	0.003	0	0.317	0.053	0.008	0	0	0	0.975	0	1	0.397	0.705	0.704	0	0.001	0.328	0.057	0.001	1	0.39	0.414
io_h90_q2_tm10_dw2	0.076	0.074	0	0.317	0.02	0.008	0	0.002	0	0.943	0	1	0.014	0.705	0.704	0	0.001	0.328	0.008	0	0.916	0.103	0.013
io_h90_q2_tm30_dw1	0.374	0.018	0	0.317	0.053	0.008	0	0	0	0.898	0	1	0.011	0.705	0.536	0.571	0.001	0.328	0.073	0.007	0.916	0.001	0.902
io_h90_q2_tm30_dw2	0.076	0.039	0	0.001	0.019	0.008	0	0	0	0.943	0	1	0	0.705	0.704	0.071	0.001	0.328	0.06	0.001	0.17	0.01	0.296

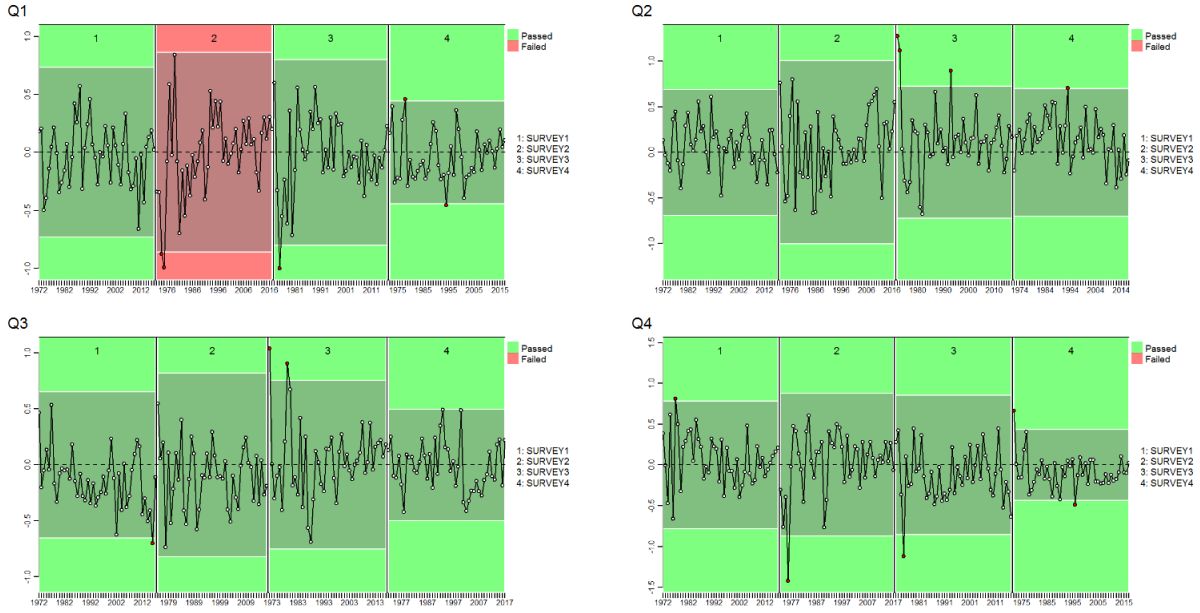

Figure 1*: Runs test plot of CPUE residuals by quarter (Q1-Q4) for the 2018 reference case model io_h80_q1_tm30_dw1. Green panels and red panels denotes if the null hypothesis of a random residual distribution cannot be rejected (pass) or is rejected (fail; p < 0.05), respectively. The inner shaded area denote the 3sigma-limits and red observations identify a specific year with residuals falling outside than these threshold limits.*

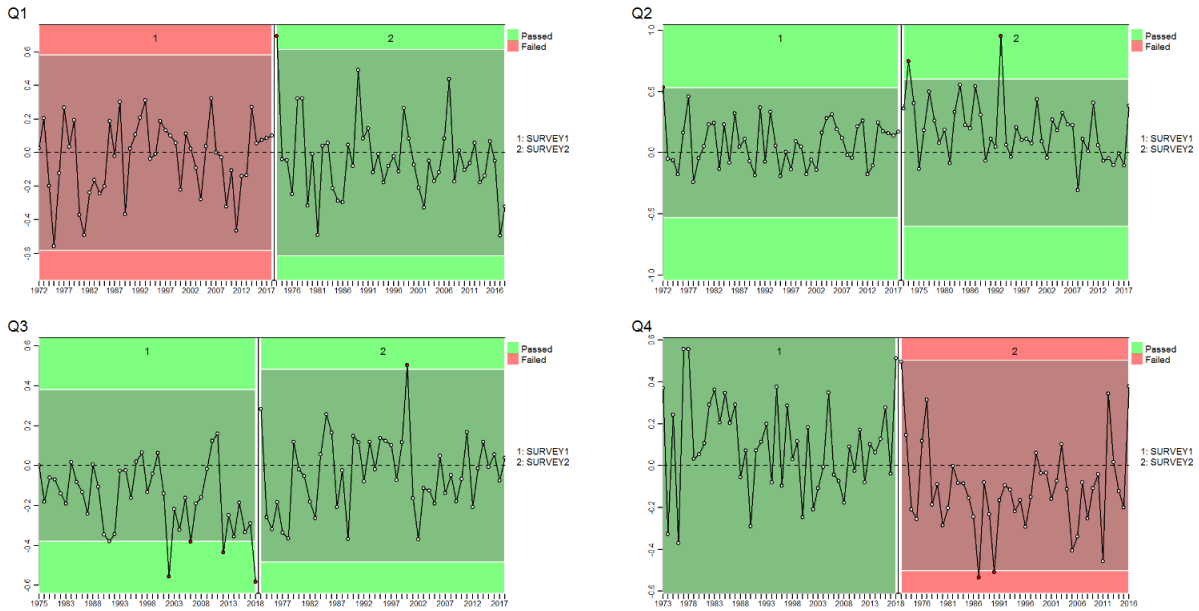

***Figure 2**: Runs test plot of CPUE residuals by quarter (Q1-Q4) for the 2019 reference case model RC2_ref. Green panels and red panels denotes if the null hypothesis of a random residual distribution cannot be rejected (pass) or is rejected (fail; $p < 0.05$), respectively. The inner shaded area denote the 3sigma-limits and red observations identify a specific year with residuals falling outside than these threshold limits.*

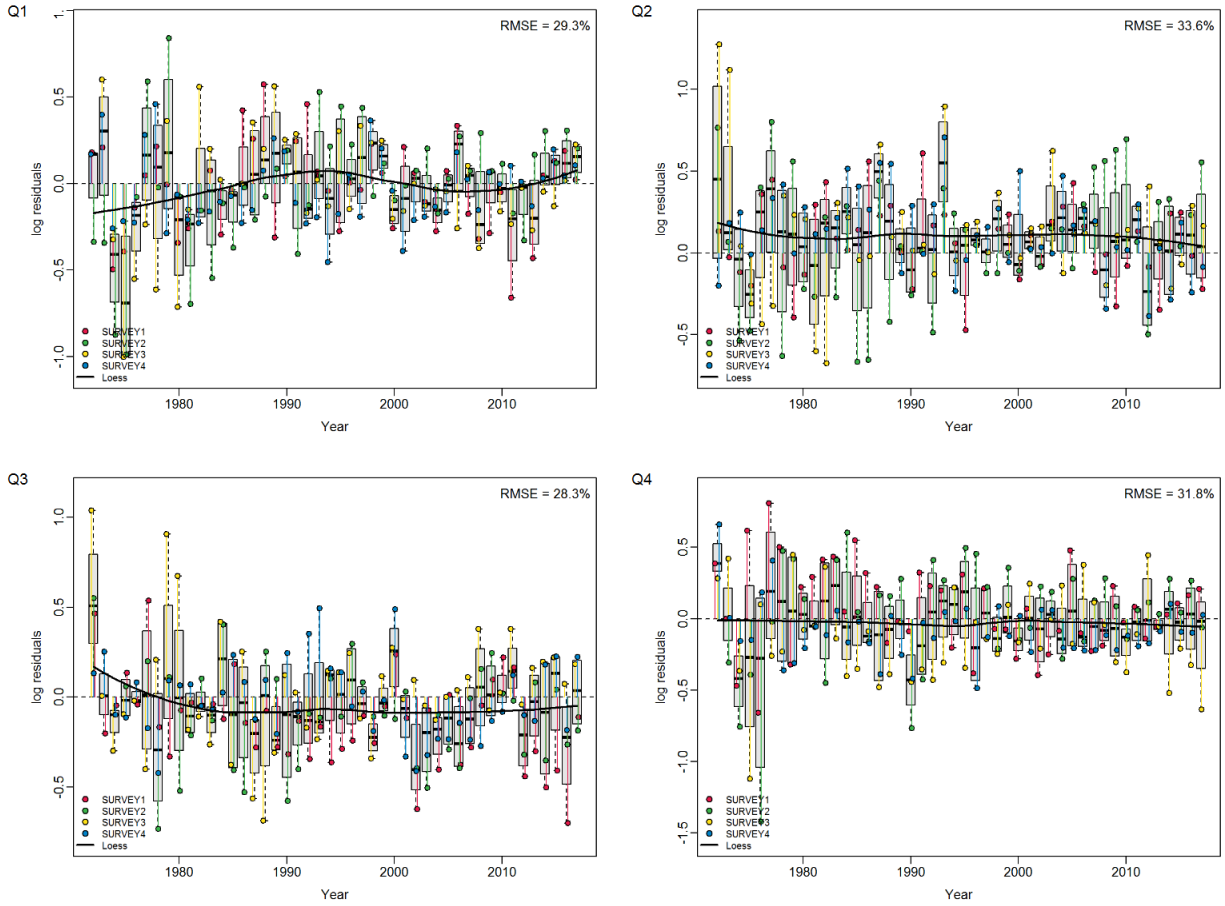

***Figure 3**: JABBA-type residual plot, showing CPUE residuals by quarter (Q1-Q4) for the 2018 reference model io_h80_q1_tm30_dw1 and Residual-Mean-Squared-Error (RMSE%) for all indices combined. Boxplots indicate the median and quantiles of all residuals available for any given quarter, and solid black lines indicate a loess smoother through all residuals.*

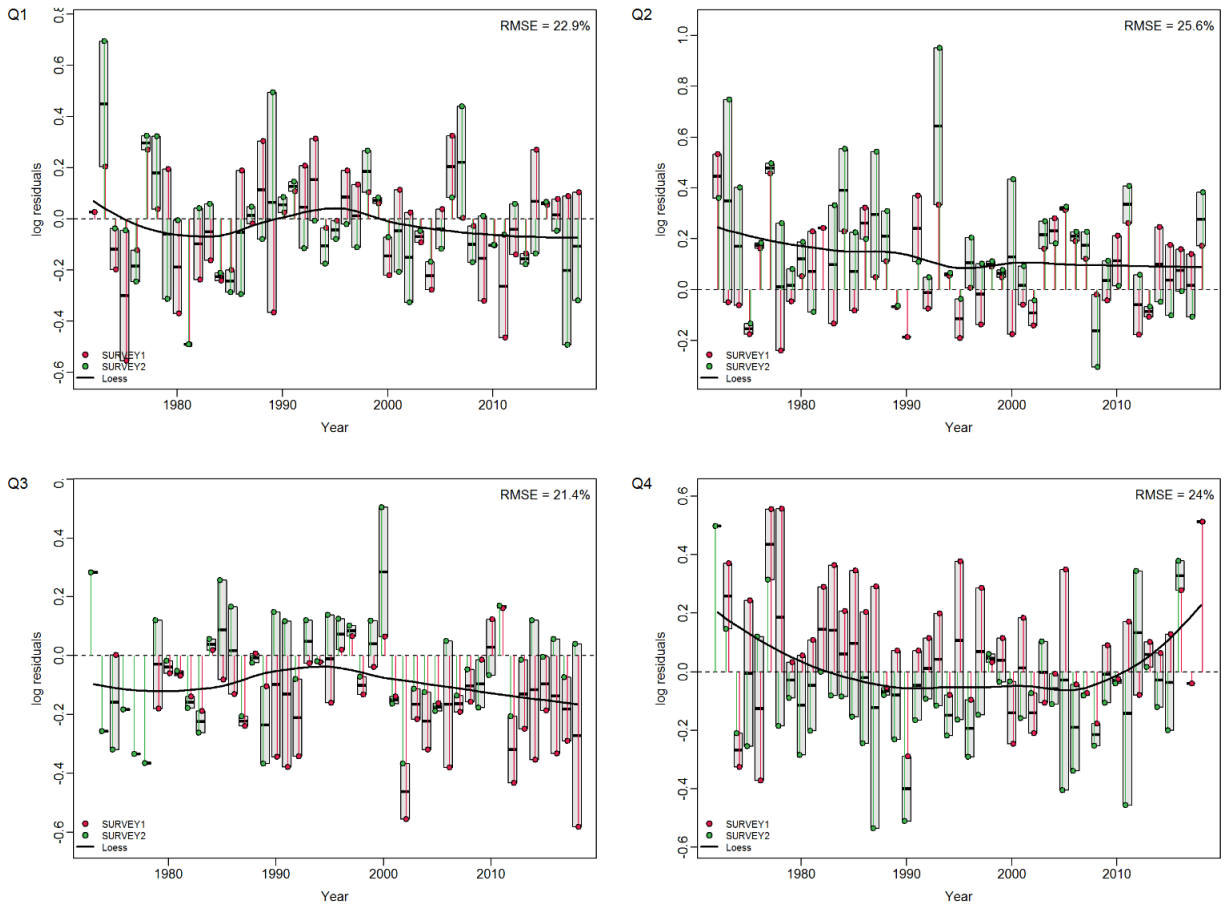

***Figure 4**: JABBA-type residual plot, showing CPUE residuals by quarter (Q1-Q4) for the 2019 reference case model RC2_ref and Residual-Mean-Squared-Error (RMSE%) for all indices combined. Boxplots indicate the median and quantiles of all residuals available for any given quarter, and solid black lines indicate a loess smoother through all residuals.*

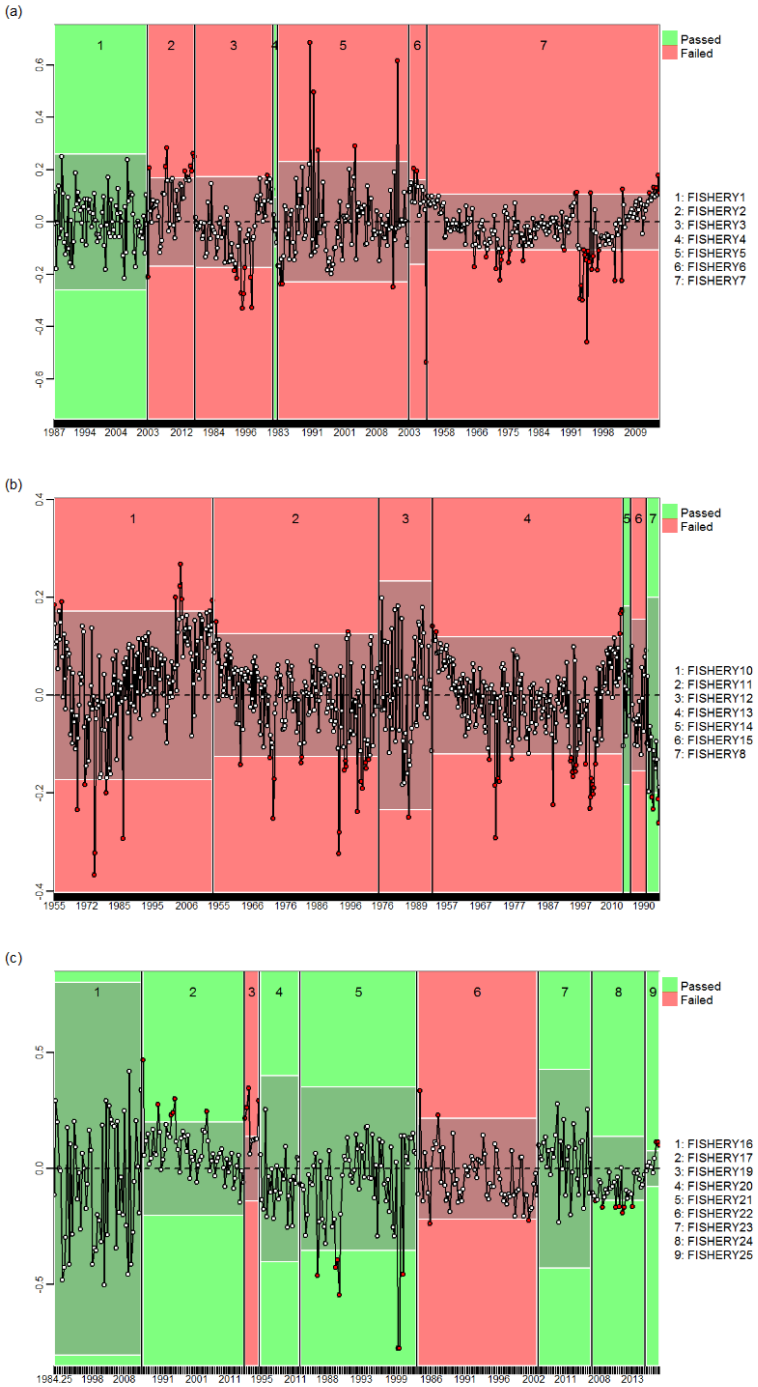

***Figure 5**: Runs test plot of mean length residuals for the 2018 reference case model io_h80_q1_tm30_dw1. Green panels and red panels denote if the null hypothesis of a random residual distribution cannot be rejected (pass) or is rejected (fail; p < 0.05), respectively. The inner shaded area denote the 3sigma-limits and red observations identify a specific year with residuals falling outside than these threshold limits.*

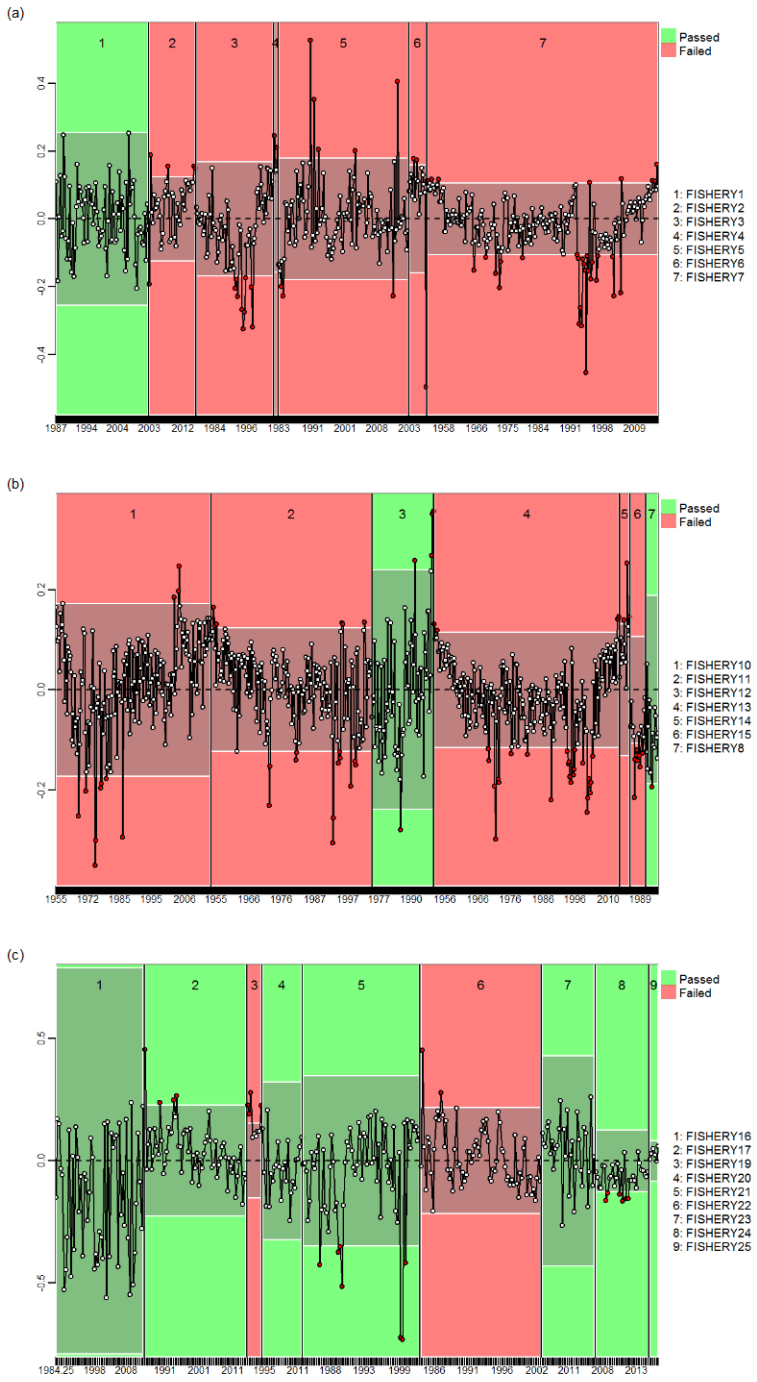

Figure 6*: Runs test plot of mean length residuals for the 2019 reference case model RC2_ref. Green panels and red panels denote if the null hypothesis of a random residual distribution cannot be rejected (pass) or is rejected (fail; p < 0.05), respectively. The inner shaded area denote the 3sigma-limits and red observations identify a specific year with residuals falling outside than these threshold limits.*

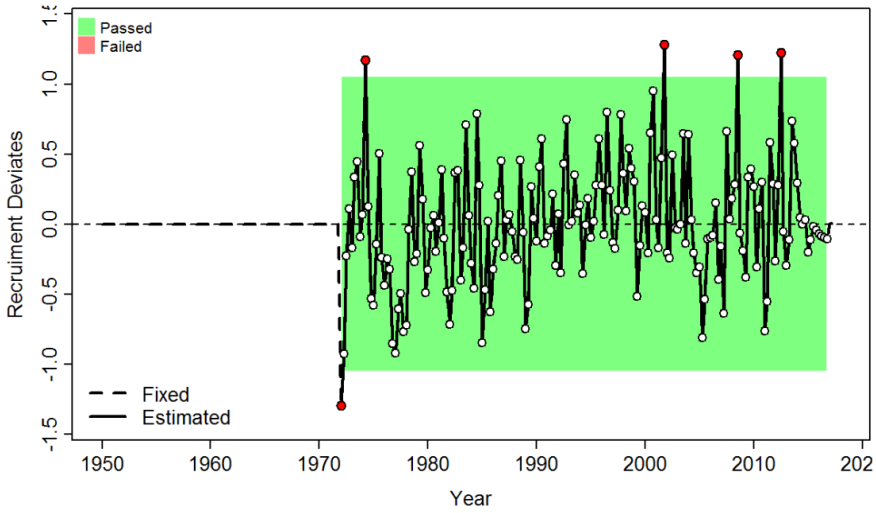

***Figure 7**: Runs test plot of recruitment residuals for 2018 reference case model io_h80_q1_tm30_dw1. Green panels and red panels denote if the null hypothesis of a random residual distribution cannot be rejected (pass) or is rejected (fail; p < 0.05), respectively. The borders of the colored area denote the 3sigma-limits and red observations identify a specific year with residuals falling outside than these threshold limits.*

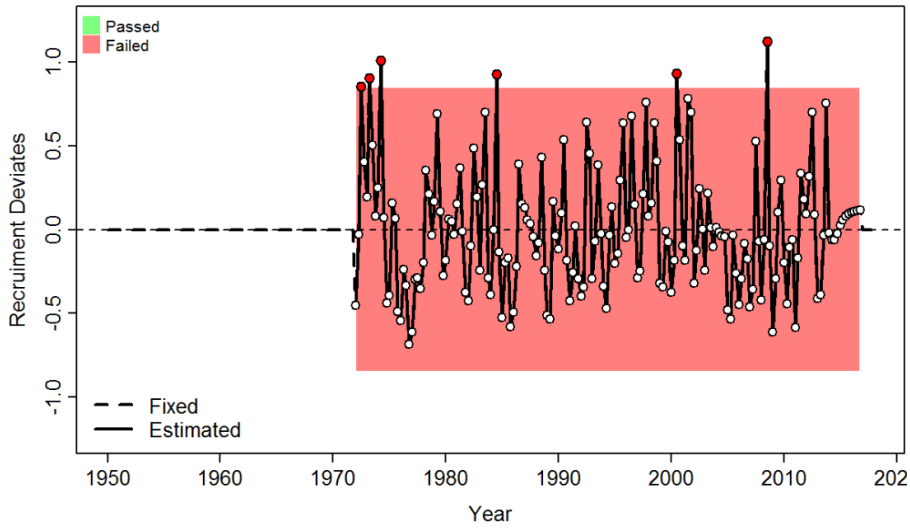

***Figure 8**: Runs test plot of recruitment residuals for the 2019 reference case model RC2_ref. Green panels and red panels denote if the null hypothesis of a random residual distribution cannot be rejected (pass) or is rejected (fail; p < 0.05), respectively. The borders of the colored area denote the 3sigma-limits and red observations identify a specific year with residuals falling outside than these threshold limits.*

References

Anhøj, J., Olesen, A.V., 2014. Run charts revisited: A simulation study of run chart rules for detection of non-random variation in health care processes. PLoS One 9, 1–13. doi:10.1371/journal.pone.0113825

Carvalho, F., Punt, A.E., Chang, Y.J., Maunder, M.N., Piner, K.R., 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. 192, 28–40. doi:10.1016/j.fishres.2016.09.018

Fu, D., Merino, G., Langley, A.D., Ijurco, A.U., 2018. Preliminary Indian Ocean yellowfin tuna stock assessment 1950-2017 (Stock Synthesis). IOTC-WPTT20 33.

Kell, L.T., Kimoto, A., Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fish. Res. 183, 119–127.

Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99. doi:http://dx.doi.org/10.1016/j.fishres.2012.10.012

Taylor, I.G., Stewart, I.J., Hicks, A.C., Garrison, T.M., Punt, A.E., Wallace, J.R., Wetzel, C.R., Thorson, J.T., Takeuchi, Y., Ono, K., Monnahan, C.C., Stawitz, C.C., Teresa, Z.A., Whitten, A.R., Johnson, K.F., Emmet, R.L., Anderson, S.C., Iantaylornoaagov, M.I.T., 2013. Package 'r4ss': R Code for Stock Synthesis.

Winker, H., Carvalho, F., Kapur, M., 2018. JABBA: Just Another Bayesian Biomass Assessment. Fish. Res. 204, 275–288. doi:http://doi.org/10.1016/j.fishres.2018.03.01