Collect. Vol. Sci. Pap. ICCAT, 70(5): 2069-2081 (2014)

MODEL DIAGNOSTICS FOR STOCK SYNTHESIS 3: EXAMPLES FROM THE 2012 ASSESSMENT OF COBIA IN THE U.S. GULF OF MEXICO

Shannon L. Cass-Calay[1], Jakob C. Tetzlaff[1], Nancie J. Cummings[1] and J. Jeffery Isely[1]

SUMMARY

This document summarizes common model diagnostics available for Stock Synthesis 3 and describes their interpretation. Examples of model misspecification are described and the resulting diagnostics are illustrated. Solutions to improve model performance are also discussed.

RÉSUMÉ

Le présent document récapitule les diagnostics du modèle communs disponibles pour Stock Synthèse 3 et décrit leur interprétation. Des exemples d'erreurs de spécification du modèle sont décrits et les diagnostics résultants sont illustrés. Les solutions visant à améliorer les performances du modèle sont également discutées.

RESUMEN

Este documento resume los diagnósticos del modelo comunes disponibles para Stock Synthesis 3 y se describe su interpretación. Se describen ejemplos de especificaciones erróneas del modelo y se ilustran los diagnósticos resultantes. Se discuten también soluciones para mejorar el rendimiento del modelo.

KEYWORDS

Stock Synthesis, Stock Assessment, Model Diagnostics

1. Introduction

Stock Synthesis (SS) is an integrated stock assessment model which is widely used in western and southeastern United States, and increasingly used throughout the world (i.e. ICES 2012). SS has a statistical framework that allows the user to calibrate a population dynamics model using a variety of fisheries and survey data (Methot 2011). SS is a flexible model which can be specified using multiple genders, growth morphs, and/or stocks within one or more areas, and can be parameterized using an age-based structure or size-based structure. SS allows the user to include ageing error, and estimate growth, the spawner-recruitment relationship, and movement between areas. Due to its flexibility, SS can be parameterized to mimic many commonly used assessment models such as ASPIC, age-structured production models, statistical catch-at-age models and virtual population analysis. SS also includes an integrated projection routine which allows the uncertainty in estimated parameters to be propagated to the projections and the calculation of management references, thus facilitating risk analyses, including projections of possible annual catch limits.

A comprehensive collection of R functions (r4SS; http://code.google.com/p/r4ss/) is available to summarize and plot model results, manipulate files, and visualize model parameterizations. R4SS produces a variety of model diagnostics. Unfortunately, many preliminary SS model configurations, particularly complex ones, exhibit poor diagnostic behaviors which are difficult to interpret, and correct. This contributes to the extensive "learning curve" for SS.

To facilitate the use of SS for ICCAT assessments, we illustrate example diagnostics that suggest improper model parameterization or structure, and describe the cause of the poor model behavior. When possible, we also include an improved diagnostic for contrast. We also include a checklist that describes the steps required to adequately, and efficiently evaluate the performance of SS model runs.

[1] U.S. Department of Commerce, NOAA Fisheries, Southeast Fisheries Science Center, Miami Laboratory, 75 Virginia Beach Drive, Miami, Florida 33149 U.S.A. Email: Shannon.Calay@noaa.gov

2. Materials and methods

During 2012-13, the Southeast Fisheries Science Center completed an assessment of cobia (*Rachycentron canadum*) in the U.S. Gulf of Mexico. The assessment was conducted using Stock Synthesis 3.24h (beta), a version of stock synthesis that allows discard-only fleets. The data inputs are briefly summarized below. A detailed description can be found in SEDAR (2013). Model results and diagnostics were developed using r4SS, a comprehensive set of functions in R for summarizing and plotting SS results (http://code.google.com/p/r4ss/).

2.1 Life history

Life history data used in the assessment included natural mortality, growth, maturity, and fecundity.

- A single von Bertalanffy equation was used to model growth of cobia for both sexes. The von Bertalanffy parameters were estimated within the SS model.
- A fixed weight-length relationship ($W=aFL^b$) was used for both sexes combined.
- The age at 50% maturity was fixed at 2, and fish age-3+ fish were assumed to be fully mature.
- Fecundity was assumed to be directly proportion to female weight in the SS model.
- The sex ratio was fixed at 60% females for all ages.
- A point estimate of natural mortality (M=0.38 y^{-1}) was used to scale the age-based estimates of natural mortality (Lorenzen 1996).

2.2 Landings and Discards

Catch and discards series used in the assessment included:

- Commercial landings (1927-2011) in metric tons.
- Recreational landings (1950-2011) in 1000s of fish.
- Commercial discards (1993-2011) in metric tons. Discard mortality rate = 5%.
- Recreational discards (1981-2011) in 1000s of fish. Discard mortality rate = 5%.
- Shrimp bycatch (1972-2011) was modeled as a discard only fleet (Discard Mortality = 100%). Shrimp Bycatch was assumed to be a function of the shrimp fishing fleet effort and was modeled using a super-year approach (Methot 2011).

2.3 Length Composition

Due to a lack of direct observations of age composition, the cobia stock assessment model used a length-based structure. Sources of length composition included:

- Commercial length composition (1983-2011).
- Reef fish observer program length composition (2006-2011).
- Recreational length composition (1979-2011).
- SEAMAP trawl survey length composition (1989-2011) was assumed to be representative of the shrimp fishery.

2.4 Age conditioned on length (ALKs)

Recreational age composition data were made conditional on length. Using these conditional age compositions has the advantage of linking age data directly to length data (essentially creating an age-length key). As a result, the data contain more detailed information about the relationship between size and age and so provides a stronger ability to estimate growth parameters, especially the variance of size-at-age.

2.5 Indices

Two relative indices of abundance were used in the stock assessment. Both indices are fishery-dependent and both provide indices of abundance for the recreational fishery for cobia in the Gulf of Mexico.

- Marine Recreational Fishery Statistics Survey (MRFSS), an index of total catch.
- Headboat Survey, an index of landings.

3. Results and Discussion

The R script r4ss automatically plots model results and diagnostics for SS. A thorough examination of these figures may reveal improper model behavior resulting from model misspecification. For example, abnormal recruitment patterns may occur, stock productivity estimates may be unrealistically high or low, selectivity patterns may be nonsensical and drastic increases/decreases in stock biomass may occur in a single year. Examples of these problematic behaviors follow.

A small number of cobia are captured each year as bycatch to the shrimp trawl fishery in the Gulf of Mexico. Observations of length composition in this fishery suggest that most cobia in shrimp trawls are small, but some large fish are also captured, apparently while "riding" then net. When the selectivity of the shrimp bycatch "fleet" is misspecified, allowing unrealistically high numbers of large fish to be intercepted, unusual model results occur. Upon the initiation of the shrimp bycatch fleet, an improbable depletion of the spawning stock biomass occurs (**Figure 1**) and the equilibrium SPR drops precipitously (**Figure 2**).

These improbable results are also accompanied by unusual pattern in recruitment, specifically a "boom and bust" behavior that is unlikely given the biology of cobia (**Figure 3**).

Examinations of the fit to the length (or age) composition can also be informative. A sudden lack of fit to length composition can be caused by the improper specification of a management regulation (**Figure 4**). In this example, the model was specified to re-estimate the retention function following the implementation of a minimum size limit in 1983. However, the length composition data suggests that the management regulation was ignored for a number of years, causing large residuals in the length composition from 1983-1986 (**Figure 4**). The model was improved by retaining the pre-regulatory retention function for several years after the imposition of the minimum size regulation, allowing the change in retention to coincide with the observed shift in length composition.

The covariance matrix may also contain evidence of improper model formulation. In particular, these tables should be scrutinized for evidence of high coefficients of variation and strong correlations between selectivity patterns and growth parameters. In this example, there was evidence of a high degree of confounding in estimates of shrimp selectivity (Shrimp_5) and the growth parameter (Lmin; **Table 1**). Also, there was high correlation between the shrimp selectivity parameters themselves (Shrimp_2, Shrimp_3, Shrimp_4; **Table 2**). As a result, this model could not reliably estimate the double-normal selectivity pattern of the shrimp bycatch fishery.

Trace plots are a useful tool which show the parameter estimates relative to the phase of estimation. In a well behaved model, the parameters should not change a great deal after the final phase of estimation. In this example (**Figure 5**), it is clear that model parameters varied substantially following the final phase of estimation. This undesirable result was alleviated by reconfiguring the phases of estimation.

A thorough evaluation of model sensitivity to key assumptions, data weighting choices and alternative data inputs may reveal an undesirable degree of model sensitivity. In this example (**Figure 6**) the model converged on alternative solutions when the weighting of the shrimp fishery length composition data was changed. This unstable model result warrants a thorough examination of the model components to more properly specify component weighting. A reiterate reweighting procedure, such as (Francis 2011) could be instructive.

Stock synthesis includes several automated routines to evaluate model stability. One, the "Jitter" analysis allows the user to examine the effect of varying model input parameters on model results. A well behaved model should converge on a global solution across a reasonable range of input parameters (**Table 3**).

Likelihood profiling is also an automated routine in Stock Synthesis. This tool allows the user to evaluate model performance across a range of values of an input parameter (e.g. steepness, R0, sigma-R). Ideally the profile should be a smooth functional shape. Abnormal model behavior is indicated by numerous spikes and saw-toothed profiles. One should take care to evaluate the likelihood profile with sufficient precision. In this example, a likelihood profile on steepness was run using the range 0.2-1.0. At a precision of 0.05, the profile appears smooth (**Figure 7**). However, a finer precision, 0.01 reveals model instability (i.e. numerous peaks in the profile).

In this case, the poor performance was caused by growth and selectivity parameter estimates which were highly correlated with each other.

This improper model behavior is also evident in plots of the maximum likelihood estimates of various model parameters across a range of steepness (**Figure 8**).

The performance of stock synthesis can also be evaluated by plotting the distribution of parameter estimates and derived quantities across bootstrapped replicates. In stock synthesis, the variances of all the estimated parameters are propagated to the bootstraps. In a well behaved model, the maximum likelihood estimate should be similar to the mean of the bootstraps. In this example (**Figure 9 and Figure 10**), 1000 bootstraps are summarized. Note the deviations between the maximum likelihood estimate (MLE) and the mean for several model parameters. This indicates that the model performance in not ideal.

Stock synthesis models that estimate steepness merit additional caution because steepness is often inestimable. In this example (**Figure 10**), the mean steepness estimate of the bootstraps differs substantially from the maximum likelihood estimate. Furthermore, a large number of bootstrap runs estimate the steepness at the theoretical maximum, 1.0. The behavior of model runs at steepness = 1.0 often exhibit unrealistic behaviors (e.g. FSPR20% unsustainable) that may influence other parameter estimates.

Our experiences with the stock assessment of Gulf of Mexico cobia encouraged us to create a checklist to facilitate an efficient and complete evaluation of model diagnostics. These are summarized in **Table 4**.

Acknowledgments

The authors would like to acknowledge the assistance of Dr. Rick Methot and Ian Taylor who graciously reviewed numerous preliminary SS runs of Gulf of Mexico cobia. We would also like to thank the SEDAR 28 participants including data providers, analysts and reviewers.

References

ICES. 2012. Report of the Working Group on the Assessment of Southern Shelf Stocks of Hake, Monk and Megrim. ICES Document CM 2012/ACOM: 11.

FRANCIS. R.I.C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. of Fish. Aquat. Sci. 68: 1124-1138.

LORENZEN, K. 1996. The relationship between body weight and natural mortality in juvenile and adult fish: a comparison of natural ecosystems and aquaculture. Journal of Fish Biology, 49: 627–642.

METHOT, R.D. 2011. User manual for Stock Synthesis: model version 3.23b. Nov 7, 2011. NOAA Fisheries Service, Seattle, WA.

Table 1. Mean, standard deviation, CV and correlation of model parameters.

Parameter	value	std.dev	CV	Lmin	Lmax	VonBK	Cvyoung	Cvold	R0	Steepness
Lmin	35.73	2.04	0.06	1						
Lmax	120.11	5.79	0.05	0.4571	1					
VonBK	0.26	0.03	0.13	-0.6474	-0.9027	1				
Cvyoung	0.24	0.01	0.06	-0.4066	0.186	-0.1793	1			
Cvold	0.14	0.02	0.13	-0.1195	-0.7891	0.5285	-0.3645	1		
R0	6.80	0.11	0.02	-0.2606	-0.3373	0.2691	0.0926	0.2921	1	
Steepness	0.90	0.11	0.13	0.1931	0.2734	-0.2294	-0.0438	-0.2361	-0.9254	1
Q_parm[1]	1.86	0.12	0.06	0.342	0.0144	-0.1622	-0.2258	0.1744	-0.115	-0.0376
Com_Selex1	88.41	2.57	0.03	0.133	-0.0014	-0.0941	-0.0505	0.0943	-0.027	0.018
Com_Selex2	6.31	0.23	0.04	-0.2481	-0.1393	0.2307	0.0286	0.0139	0.0896	-0.0779
Rec_Selex1	74.73	1.72	0.02	0.4295	0.1224	-0.3613	-0.108	0.1649	-0.1317	0.1259
Rec_Selex2	25.71	1.98	0.08	-0.1278	-0.2289	0.2456	-0.0443	0.0884	0.1185	-0.1224
Shrimp_1	32.72	28.77	0.88	-0.0018	-0.0012	0.0017	-0.0002	0.0006	0.0005	-0.0004
Shrimp_2	-2.53	4.21	-1.67	0.0008	0.0005	-0.0007	0.0001	-0.0003	-0.0002	0.0002
Shrimp_3	-2.45	277.06	-113.00	0	0	0	0	0	0	0
Shrimp_4	-2.25	161.87	-72.03	-0.0001	-0.0001	0.0001	0	0	0	0
Shrimp_5	0.57	1.04	1.82	0.868	0.3188	-0.4758	-0.5274	-0.0064	-0.2204	0.1551
Shrimp_6	-1.83	0.22	-0.12	0.0379	0.0454	-0.0861	-0.031	0.0573	0.1027	-0.0244

Table 2. Mean, standard deviation, CV and correlation of model selectivity parameters.

name	value	std.dev	CV	Q	C_Selex1	C_Selex2	R_Selex1	R_Selex2	Shrimp_1	Shrimp_2	Shrimp_3	Shrimp_4	Shrimp_5	Shrimp_6
Q_parm[1]	1.86	0.12	0.06	1										
Com_Selex1	88.41	2.57	0.03	0.0623	1									
Com_Selex2	6.31	0.23	0.04	-0.0733	0.5354	1								
Rec_Selex1	74.73	1.72	0.02	0.1831	0.2162	-0.2227	1							
Rec_Selex2	25.71	1.98	0.08	-0.0172	-0.0007	0.0473	0.3073	1						
Shrimp_1	32.72	28.77	0.88	-0.0017	-0.0003	0.0005	-0.0009	0.0003	1					
Shrimp_2	-2.53	4.21	-1.67	0.0013	0.0001	-0.0002	0.0004	-0.0001	-0.7169	1				
Shrimp_3	-2.45	277.06	-113.00	0	0	0	0	0	1	-0.7169	1			
Shrimp_4	-2.25	161.87	-72.03	0	0	0	-0.0001	0	-0.0078	-0.6916	-0.0078	1		
Shrimp_5	0.57	1.04	1.82	0.4848	0.1182	-0.2023	0.3666	-0.0925	-0.0013	0.0006	0	-0.0001	1	
Shrimp_6	-1.83	0.22	-0.12	0.4164	0.0418	-0.0354	0.1086	-0.0258	-0.0008	0.0006	0	0	0.1412	1

Table 3. The results of a “Jitter” analysis evaluating the effect of varying input parameters on model results.

Run	Likelihood	SSB unfished	SSB 2011	Depletion
1	1127.85	7277	2180	0.30
2	1127.85	7277	2180	0.30
3	1126.94	7253	2212	0.30
4	1127.85	7277	2180	0.30
5	1127.85	7277	2180	0.30
6	1126.68	7260	2240	0.31
7	1126.68	7260	2240	0.31
8	1126.94	7253	2212	0.30
9	1127.85	7277	2180	0.30
10	1126.94	7253	2212	0.30
11	1126.68	7260	2240	0.31
12	1127.22	7235	2213	0.31
13	1127.85	7277	2180	0.30
14	1126.94	7253	2212	0.30
15	1126.94	7253	2212	0.30
16	1127.85	7277	2180	0.30
17	1126.94	7253	2212	0.30
18	1127.85	7277	2180	0.30
19	1126.68	7260	2240	0.31
20	1127.22	7235	2213	0.31
21	1127.22	7235	2213	0.31
22	1127.85	7277	2180	0.30
23	1126.94	7253	2212	0.30
24	1127.22	7235	2213	0.31
25	1126.68	7260	2240	0.31
26	1126.94	7253	2212	0.30
27	1126.94	7253	2212	0.30
28	1127.85	7277	2180	0.30
29	1127.22	7235	2213	0.31
30	1126.68	7260	2240	0.31
31	1127.22	7235	2213	0.31
32	1127.22	7235	2213	0.31
33	1126.94	7253	2212	0.30
34	1127.22	7235	2213	0.31
35	1127.22	7235	2213	0.31
36	1126.68	7260	2240	0.31
37	1127.85	7277	2180	0.30
38	1127.85	7277	2180	0.30
39	1126.68	7260	2240	0.31
40	1131.13	7257	2195	0.30
41	1126.94	7253	2212	0.30
42	1127.85	7277	2180	0.30
43	1126.68	7260	2240	0.31
44	1127.22	7235	2213	0.31
45	1126.68	7260	2240	0.31
46	1126.94	7253	2212	0.30
47	1127.85	7277	2180	0.30
48	1127.85	7277	2180	0.30
49	1132.46	7295	2168	0.30
50	1127.22	7235	2213	0.31

Table 4. A flowchart to facilitate an efficient and complete evaluation of Stock Synthesis model diagnostics

SS Model Diagnostics

1. Does the model run?
    a. No → use echo input to debug
    b. Yes → continue
2. Does the hessian converge?
    a. No → check warning file, check estimated parameters in report file
    b. Yes → continue
3. Are there any parameters on bounds?
    a. No → continue
    b. Yes → change starting values/change bounds/add priors/simplify parameterization → rerun
4. Plot model output. Anything obviously wrong? Examples: productivity way too low, selectivity patterns that don't make sense, drastic decrease/increase in biomass in a single year, abnormal recruitment patterns (boom/bust).
    a. No → continue
    b. Yes → go through report file to diagnose (depends on problem)
5. Examine parameter estimates. Plot parameter distributions along with starting values, bounds, and priors. Do parameters appear well estimated?
    a. No → check bounds, check priors, check phase of estimation
    b. Yes → continue
6. Look at trace plots of parameter estimates relative to phase of estimation? Do model parameters change considerably in the final phase?
    a. No → continue
    b. Yes → try alternative phases: for example, important scaling parameters like unfished recruitment and catchability might be estimated in the first phase, recruitment deviates estimated added in the second phase, and selectivity added in the final phase.
7. Look at mean and standard deviation of estimated parameters. Is CV of estimated parameters less than 1?
    a. No → is there data to inform parameter?
        i. No → change bounds/add informative prior/fix parameter
        ii. Yes → check correlation matrix
    b. Yes → continue
8. Are any of the parameters highly correlated?
    a. No → continue
    b. Yes → why? Does one of the parameters require an informative prior?
9. Plot model fits to data and diagnostics. Is model fitting data reasonably?
    a. No → diagnose the problem.
    b. Yes → continue
10. Check for model stability to initial starting parameters using Jitter analysis. Does model converge to a "global" solution?
    a. No → identify why.
        i. look at which likelihood components are changing
        ii. Evaluate the phases of estimation
        iii. Plot distribution of estimated parameters over all model runs
    b. Yes → continue (try again with larger deviation from starting values)
11. Profile leading model parameters such as stock-recruitment parameters (steepness/R0) or natural mortality. Was the profile smooth?
    a. No → Plot estimated parameters as a function of profiled leading parameter
        i. Do any of the parameters hit bounds across the runs? Do any of the parameters bounce between alternative solutions? Do some parameters show similar patterns?
            1. Yes → may not have enough data to inform all estimated parameters: add informative priors/reduce the number of estimated parameters.
    b. Yes → Does profile show leading parameter is well estimated? Do the different data components show similar signals?
        i. No → parameter may require informative prior or need to be fixed
        ii. Yes → profile at finer scale
            1. Does profile remain smooth?
                a. Yes → continue

12. Evaluate model sensitivity to key model assumptions, data weighting choices, and alternative data inputs. Was model highly sensitive to any key model assumptions or certain data sources?
    a. No → continue
    b. Yes → Is model specified correctly? Are assumptions appropriate? Is model over-parameterized? Should data be re-weighted?
13. Evaluate model sensitivity to the most recent years of data using a retrospective analysis. Did the retrospective analysis reveal any inconsistencies in the data?
    a. No → continue
    b. Yes → identify source of the retrospective pattern
14. Evaluate model uncertainty using bootstrap approach. Plot distribution of parameter estimates and derived quantities from bootstrapped runs. Compare MLE of parameter estimates to mean of bootstrap results. Are parameters or derived quantities well estimated when data is resampled?
    a. No → do distributions show multi-modality or high proportion of bounding?
        i. Yes → may not have enough data to inform all estimated parameters: add informative priors/reduce the number of estimated parameters.
    b. Yes → continue
15. Evaluate model convergence using MCMC approach. Use standard approaches to evaluating MCMC results: look at trace plots/plot posterior distributions/compare MLE to mean of posterior distribution. Does MCMC converge on a single solution? Are MLEs of parameters/derived quantities similar to mean of posterior distributions?
    a. No →
    b. Yes → continue

Model Misspecified Base Model

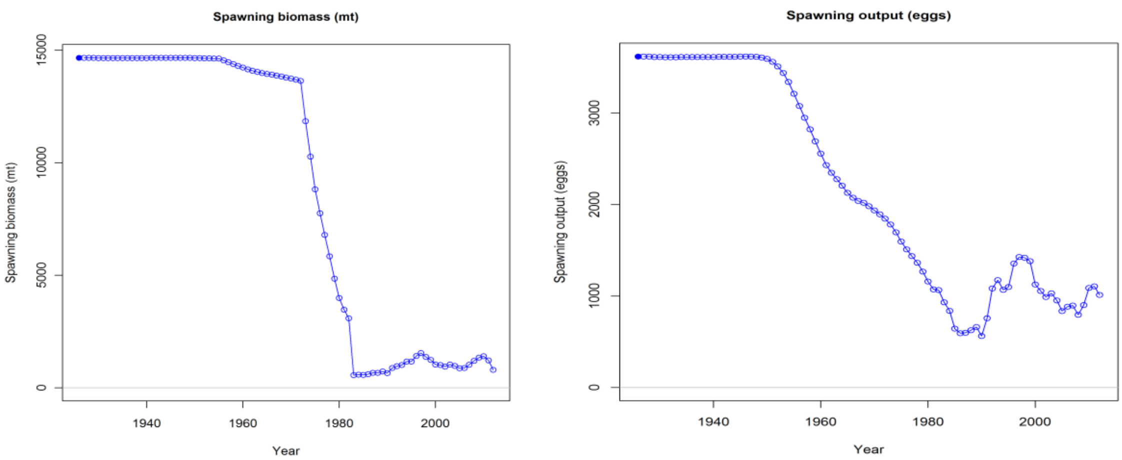

Figure 1. An improbable drop in spawning stock biomass resulting from model misspecification of the shrimp bycatch selectivity (left) and the improved behavior of the corrected model (right).

Figure 2. An improbable drop in equilibrium SPR resulting from model misspecification (left) and the improved behavior of the corrected model (right).

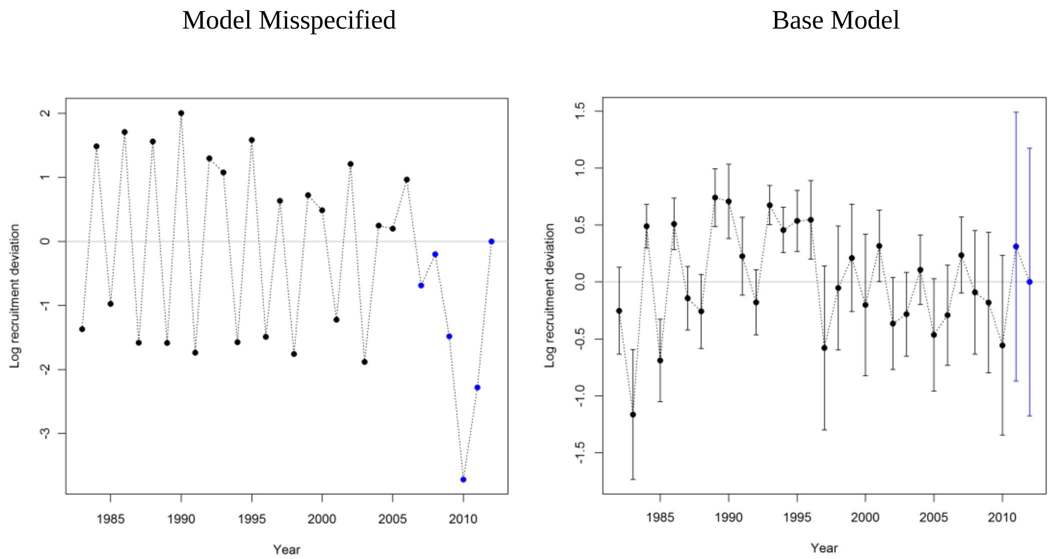

Figure 3. Unusual "boom and bust" recruitment deviations caused by model misspecification (left) and the improved behavior of the corrected model (right).

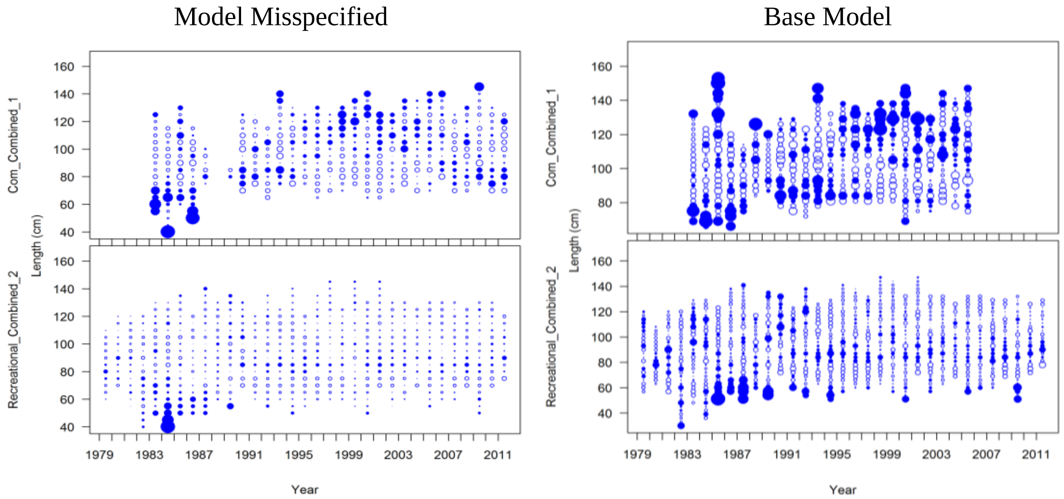

Figure 4. Pearson residuals for the fit to the length composition resulting from model misspecification (left) and the improved behavior of the base model (right).

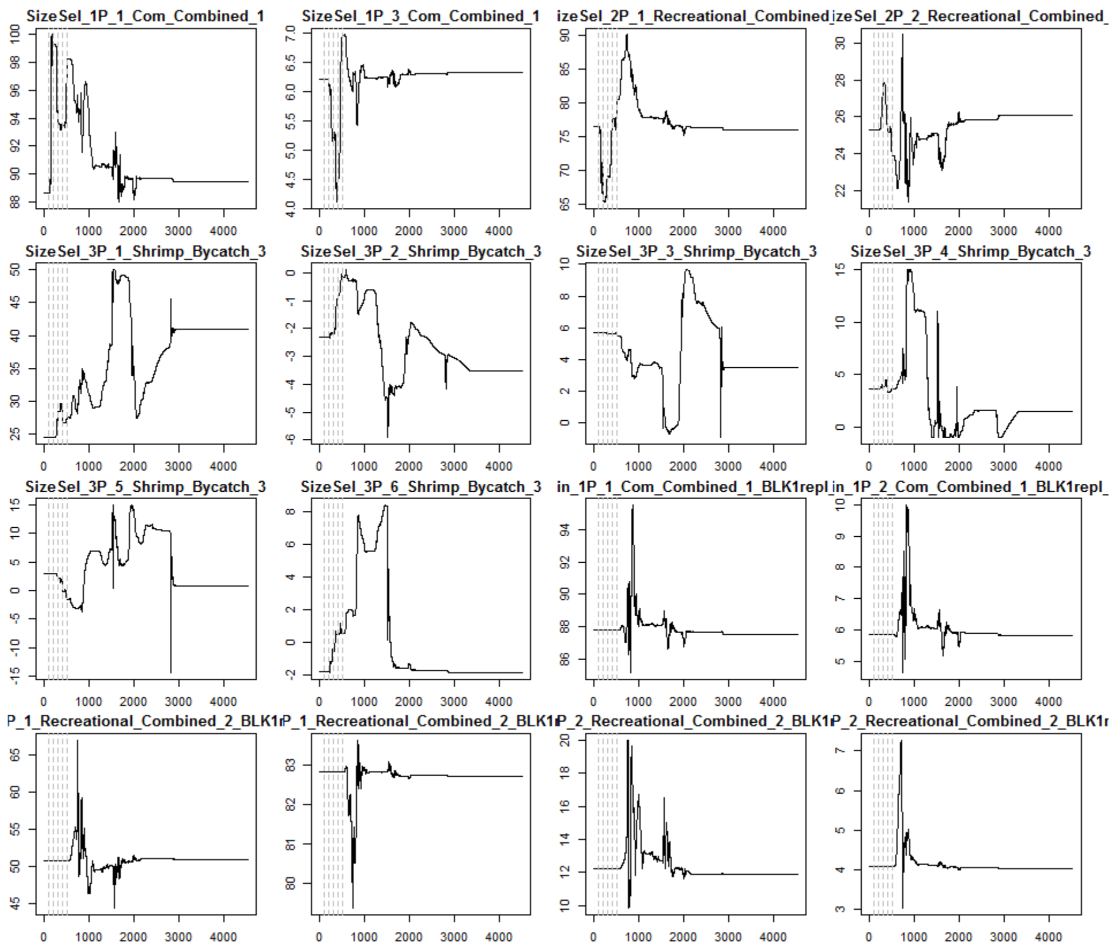

Figure 5. Trace plots of model parameters relative to the phase of estimation (dashed vertical lines). Parameter estimates should not vary substantially following the final phase of estimation.

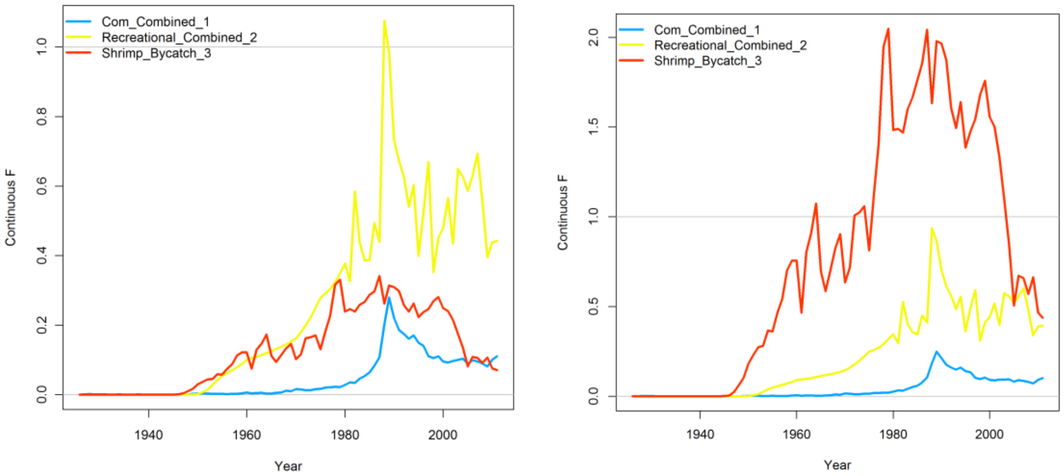

Figure 6. This model converged on an alternative solution when length composition data was reweighted.

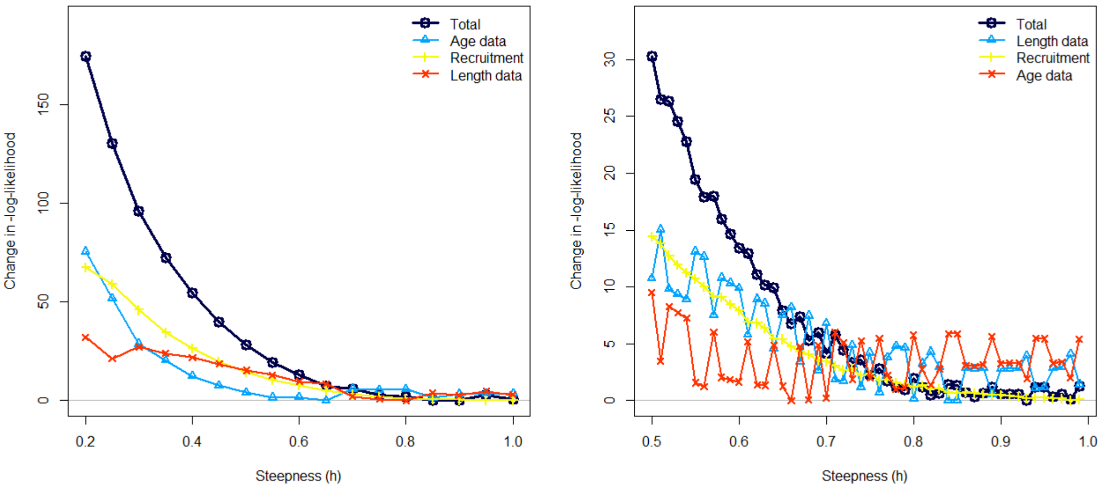

Figure 7. Likelihood profile for steepness at intervals of 0.05 (left) and 0.01 (right).

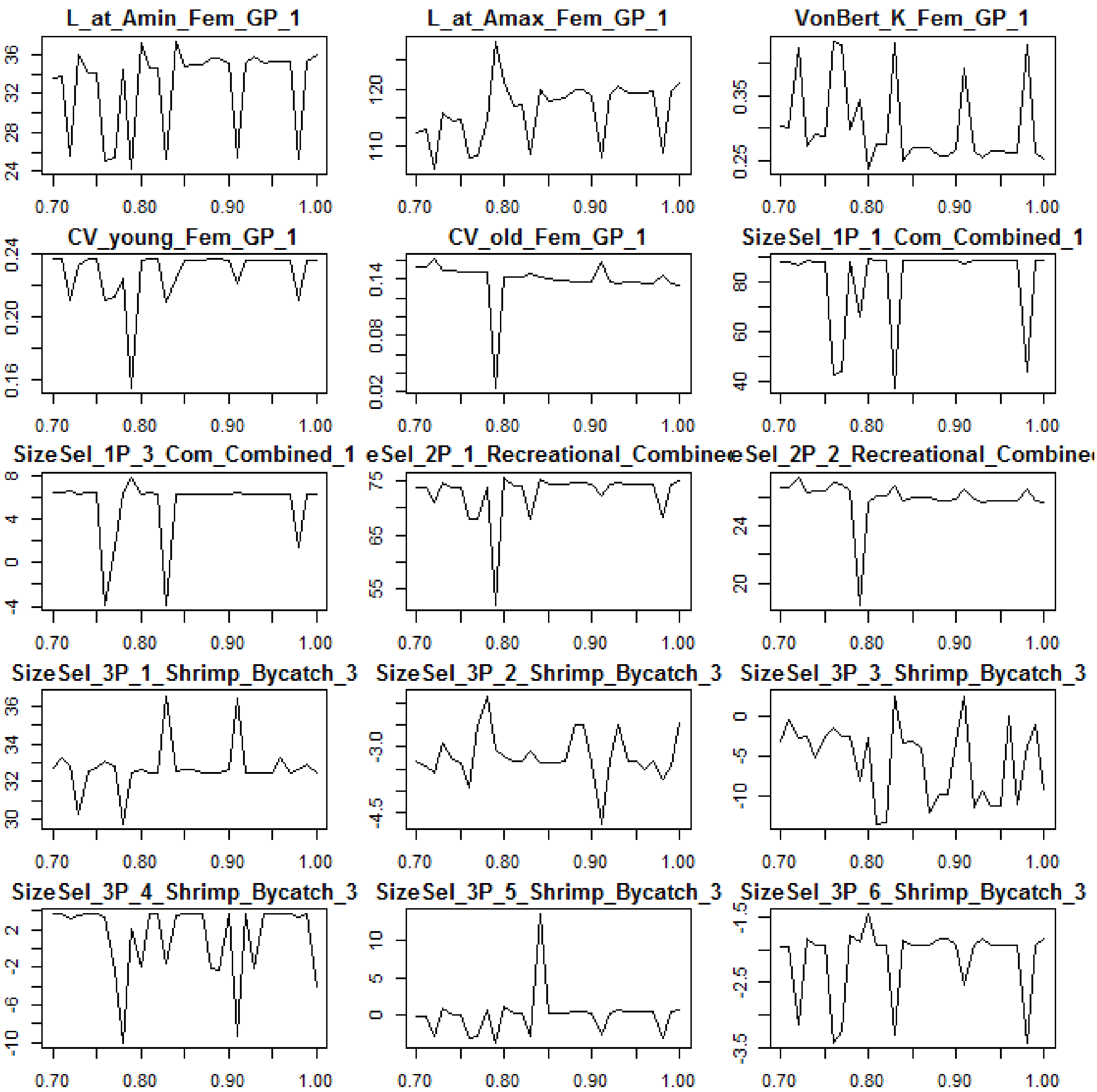

Figure 8. Plot of MLE of parameters (y-axis) across a range of values for steepness (x-axis).

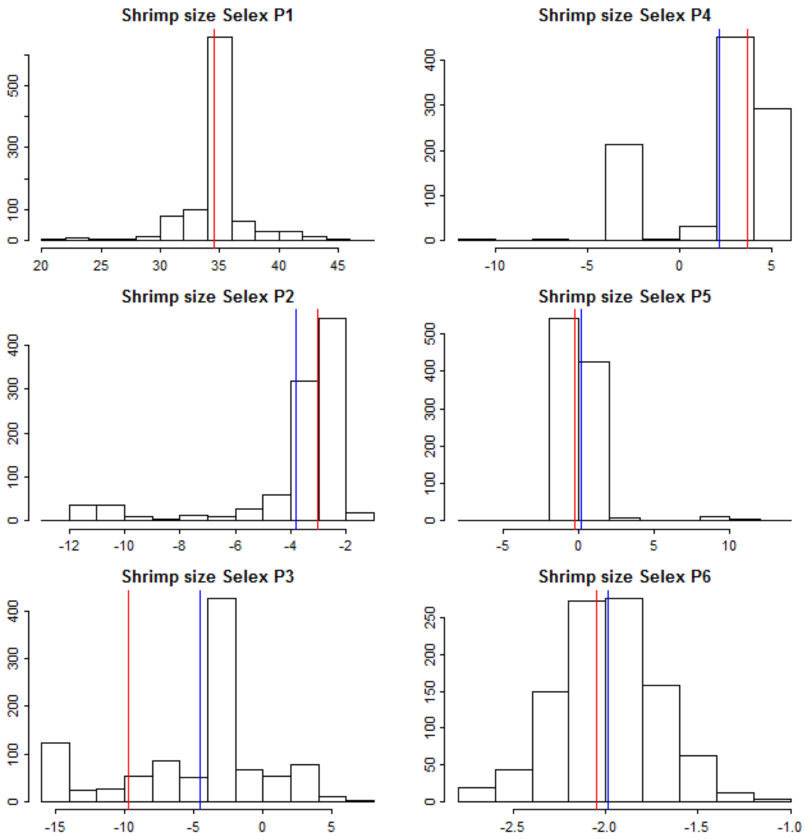

Figure 9. Distribution of estimated shrimp selectivity parameters from 1000 bootstrap replicates. Blue lines represent mean estimates from the bootstrap samples, red lines represent the point estimate of the parameters from the base model.

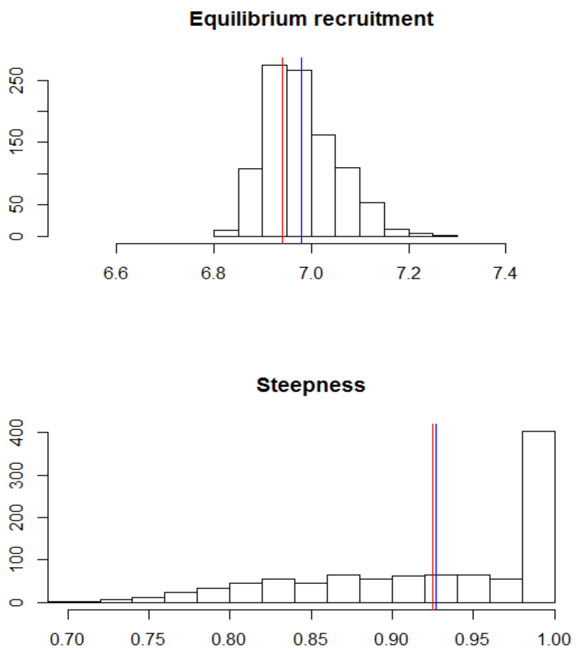

Figure 10. Distribution of estimated equilibrium recruitment and steepness from 1000 bootstrap samples. Blue lines represent mean estimates from the bootstrap samples, red lines represent the point estimate of the parameters from the base model.