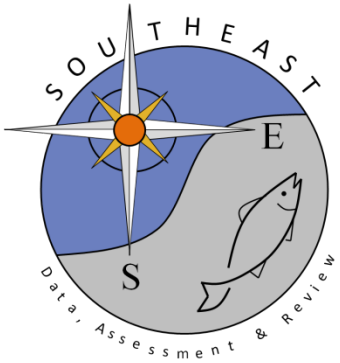
Examples of Stock Synthesis diagnostic methods and results implemented for previously completed North Atlantic shortfin mako Stock Synthesis model runs.

D. Courtney, F. Carvalho, H. Winker, and L. Kell

SEDAR65-RD13

Received: 6/24/2020



EXAMPLES OF DIAGNOSTIC METHODS IMPLEMENTED FOR PREVIOUSLY COMPLETED NORTH ATLANTIC SHORTFIN MAKO STOCK SYNTHESIS MODEL RUNS

D. Courtney[1], F. Carvalho[2], H. Winker[3], and L. Kell[4]

SUMMARY

A range of model diagnostics were implemented for three previously completed Stock Synthesis assessment models conducted for North Atlantic shortfin mako shark. The objectives were to evaluate stock assessment model fit to data, identify possible model misspecifications, and evaluate model prediction skill by implementing nine diagnostic approaches: 1) Simultaneous visualization of residuals from multiple catch per unit effort (CPUE) indices using Just Another Bayesian Biomass Assessment (JABBA) residual plots; 2) runs tests applied to individual CPUE indices; 3) runs tests applied to size composition data; 4) a runs test applied to estimated recruitment deviations; 5) retrospective analyses; 6) likelihood component profiles; 7) a deterministic age-structured production model (ASPM); 8)Markov Chain Monte Carlo (MCMC); and 9) hind-cast cross-validation. Overall, the diagnostic results were consistent and indicated that changes in abundance after the 1990's could not be explained by catches alone, that the abundance information, both absolute and relative, contained in the CPUE indices could not be interpreted without modelling fluctuations in recruitment, and that the recruitment deviation estimation had a large influence on the model fit.

RÉSUMÉ

Une gamme de diagnostics de modèles a été mise en œuvre pour trois modèles d'évaluation Stock Synthèse déjà achevés, qui ont été appliqués au requin-taupe bleu de l'Atlantique Nord. Les objectifs étaient d'évaluer l'ajustement du modèle d'évaluation des stocks aux données, d'identifier de possibles erreurs de spécification du modèle et d'évaluer les compétences de prédiction du modèle en mettant en œuvre neuf approches de diagnostic : 1) Visualisation simultanée des valeurs résiduelles de multiples indices de capture par unité d'effort (CPUE) à l'aide de diagrammes de valeurs résiduelles de JABBA (Just Another Bayesian Biomass Assessment) ; 2) « runs tests » appliqués à des indices de CPUE individuels ; 3) « runs tests » appliqués aux données de composition par taille ; 4) « runs test » appliqué aux écarts de recrutement estimées ; 5) analyses rétrospectives ; 6) profils de composante de la vraisemblance ; 7) modèle de production déterministe structuré par âge (ASPM) ; 8) méthodes de Markov Chain Monte Carlo (MCMC) ; et 9) validation croisée de la simulation rétrospective. Dans l'ensemble, les résultats du diagnostic étaient cohérents et indiquaient que les changements d'abondance après les années 1990 ne pouvaient pas être expliqués par les seules captures, que les informations sur l'abondance, tant absolues que relatives, contenues dans les indices de CPUE ne pouvaient pas être interprétées sans la modélisation des fluctuations du recrutement, et l'estimation de l'écart du recrutement avait eu une grande influence sur l'ajustement du modèle.

RESUMEN

Se implementó una gama de diagnósticos del modelo para los tres modelos de la evaluación de Stock Synthesis previamente completados para el marrajo dientuso del Atlántico. Los objetivos eran evaluar el ajuste del modelo de evaluación de stock a los datos, identificar posibles errores

[1] National Oceanographic and Atmospheric Administration, National Marine Fisheries Service, Southeast Fisheries Science Center, Panama City Laboratory, 3500 Delwood Beach Road, Panama City, Florida 32408, U.S.A. E-mail: Dean.Courtney@noaa.gov
[2] National Oceanographic and Atmospheric Administration, National Marine Fisheries Service, Pacific Islands Fisheries Science Center, Honolulu, HI 96816, USA
[3] DAFF, Department of Agriculture, Forestry and Fisheries, Private Bag X2, Rogge Bay 8012, South Africa.
[4] Centre for Environmental Policy, Imperial College London, London SW7 1NE

en las especificaciones y evaluar la capacidad de predicción del modelo implementando nueve enfoques de diagnóstico: 1) visualización simultánea de los valores residuales de múltiples índices de captura por unidad de esfuerzo (CPUE) usando diagramas residuales de Solo otra evaluación bayesiana de biomasa (JABBA), 2) "runs test" aplicados a índices individuales de CPUE, 3) "runs test" aplicados a datos de composición por tallas, 4) "runs test"aplicadas a desviaciones del reclutamiento estimadas, 5) análisis retrospectivos, 6) perfiles del componente de verosimilitud, 7) un modelo de producción estructurado por edad determinista (ASPM), 8) MCMC y 9) verificación cruzada de la simulación retrospectiva. En general, los resultados de los diagnósticos eran coherentes e indicaban que los cambios en la abundancia después de los 90 no podían explicarse solo por las capturas, que la información sobre abundancia, tanto absoluta como relativa, contenida en los índices de CPUE no podía interpretarse sin modelar las fluctuaciones en el reclutamiento y que la estimación de la desviación del reclutamiento tenía una gran influencia en el ajuste del modelo.

KEYWORDS

Stock assessment, Shark fisheries, Pelagic environment, Shortfin mako shark

1. Introduction

Several diagnostics have been evaluated for their utility to identify data conflicts and model misspecification within integrated stock assessment models (Carvalho *et al.* 2017). However, Carvalho *et al.* (2017) determined that there was no single diagnostic that worked well in all of the cases they evaluated. Instead, they recommend the use of a carefully selected range of diagnostics that proved to increase the ability to detect model misspecification, while acknowledging that the use of multiple diagnostics may increase the probability that a diagnostic test results in a false positive, i.e. fails the diagnostic when the model adequately fits the data.

This paper implements the key stock assessment diagnostics identified by Carvalho *et al.* (2017) to evaluate alternative Stock Synthesis model runs for North Atlantic shortfin mako. For improved interpretation, particular emphasis is placed on graphical visualization of these diagnostics, including implementation of the Just Another Bayesian Biomass Assessment (JABBA) residual plot (Winker *et al.* 2018) and a novel plot to illustrate runs test results for multiple data time series simultaneously. In addition, the set of diagnostic tests is extended by the hind-cast cross-validation approach (Kell *et al.* 2016) as a powerful tool to evaluate the predictive power of a stock assessment model, which is perhaps the most critical performance metric for adequate quota advice. A brief description of each diagnostic, its intended use, and a brief interpretation of its results are provided below as an aid to the Shark Working Group (Group) to determine if Stock Synthesis models fit the data adequately and that the models are well specified.

The diagnostics were implemented here for three previously completed North Atlantic shortfin mako shark Stock Synthesis models (model runs 1, 2, and 3) presented to the Group at its 2017 meeting (Anon. 2017, see their Section 4.3, and their Tables 6 – 8). Stock Synthesis model run 1 was the original Stock Synthesis model presented to the Group for the 2017 assessment of North Atlantic shortfin mako (Courtney *et al.* 2017a). Stock Synthesis model run 1 was updated by the Group to set natural mortality for males equal to that for females (Stock Synthesis model run 2). Stock Synthesis model run 2 was then updated by the Group to replace the Beverton-Holt (BH) stock recruitment relationship with the low fecundity stock-recruit (LFSR) relationship using Beta = 3 and sfrac = 0.171 (Stock Synthesis model run 3). The Group considered Stock Synthesis model run 3 to be the base Stock Synthesis model for the assessment. Stock status indicator trajectories for the three model runs were provided separately (Anon. 2017, their Figure 17). Kobe plots for the three model runs were provided separately (Anon. 2017, their Figures 18 – 19).

2. Methods

The methods section was structured according to nine diagnostic approaches (**Table 1**), which were either implemented in R (**Table A.1**), or through manual manipulation of the Stock Synthesis model (**Table A.2)**. Fishing fleets and surveys were labeled as defined in the original assessment (**Table B.1**). Some of the diagnostics below utilized asymptotic parameter estimation uncertainty and the resulting measures of precision obtained for derived quantities, which were obtained from Stock Synthesis output in the usual way by minimizing the negative of the log likelihood in AD Model Builder (ADMB; Fournier *et al.* 2011).

2.1 Diagnostic-1

The JABBA-residual plot (Winker *et al.* 2018) was applied to all CPUE indices fit in each Stock Synthesis model to quantitatively evaluate the randomness of all CPUE residuals combined. The diagnostic was adapted here for Stock Synthesis (HW) from JABBA (Winker *et al.* 2018) and implemented in R.

The diagnostic includes several features: 1) Color 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, with the area of each box indicating the strength of the discrepancy between CPUE series (larger boxes indicate a higher degree of conflicting information); 3) a loess smoother through all residuals, which highlights systematically auto-correlated residual patterns; and 4) the JABBA-Root-Mean-Squared-Error (JABBA-RMSE) fit to the loess smoother of all CPUE indices combined. The JABBA-RMSE can be interpreted analogously to a standard error. A relatively small JABBA-RMSE (≤ 0.3) is an indication of a good model fit to data (Winker *et al.* 2018).

2.2 Diagnostic-2

A runs test (Carvalho *et al.* 2017) was applied to the residuals of each CPUE index fit in the Stock Synthesis model in order to quantitatively evaluate the randomness of the time-series of CPUE residuals by fleet. The residuals runs test (Carvalho *et al.* 2017) was implemented using the function 'runs.test' in the R package 'tseries' (Trapletti, 2011). The runs test was described by Carvalho *et al.* (2017) as a nonparametric statistical randomness hypothesis test for a data sequence that calculates the 2-sided p-value of the Wald-Wolfowitz runs test. The diagnostic was adapted here (FC) from (Carvalho *et al.* 2017) and implemented in R.

New R plots were developed here (HW and FC) as an aid to visualize results obtained from residuals runs tests. The plots visually identified p-values obtained from the runs test for each series. The plots also identified individual time-series data points further than three standard deviations away from the mean (the three-sigma rule), which is another test used to detect non-random time series (e.g., see Anhøj and Olesen 2014).

2.3 Diagnostic-3

The runs test was also applied to the standardized residuals of the fit to length composition by fleet and year in order to quantitatively evaluate the randomness of the time-series of length composition residuals by fleet (Carvalho *et al.* 2017). Standardized residuals were obtained for each fleet using the Francis method (Carvalho *et al.* 2017, citing Punt 2017 their Table 2 equation 1.C; e.g., see Francis 2011, 2014, 2017).

2.4 Diagnostic-4

The runs test was also applied to recruitment deviations estimated in the Stock Synthesis model in order to quantitatively evaluate the randomness of the time-series of estimated recruitment deviations (Carvalho *et al.* 2017).

2.5 Diagnostic-5

Retrospective analysis is a way to detect bias and model misspecification (Hurtado-Ferro *et al.* 2014). A retrospective analysis as described in Carvalho *et al.* (2017) was applied to the Stock Synthesis model results. The diagnostic was adapted here (FC) from (Carvalho *et al.* 2017) and implemented in R.

The diagnostic was implemented here by sequentially eliminating the five most recent years of data from the full stock assessment model (a 5 year "peel") and then re-estimating all stock assessment model parameters from each peel and from the full model. The Mohn's rho statistic (Hurtado-Ferro *et al.* 2014; Carvalho *et al.* 2017) was calculated for ending year spawning stock size obtained from each peel relative to that obtained from the full model. Determining whether a given value of Mohn's rho indicates that an assessment exhibits a retrospective pattern is subjective. We followed the rule of thumb proposed by Hurtado-Ferro *et al.* (2014), i.e., values of Mohn's rho that fall outside the range (-0.15 to 0.20) can be interpreted as an indication of a retrospective pattern for long-lived species. In addition, the asymptotic 95% confidence intervals obtained for relative spawning stock size from each peel were compared to that of the full model.

2.6 Diagnostic-6

An R_0 likelihood component profile (Carvalho *et al.* 2017) was applied to the Stock Synthesis model results. The diagnostic was adapted here (FC) from (Carvalho *et al.* 2017) and implemented in R.

The diagnostic was implemented here by sequentially fixing the equilibrium recruitment parameter, R_0, on the natural log scale, $\log(R_0)$, to a range of values (4.8 to 6.8, step size of 0.1). This range included the parameter estimates obtained for $\log(R_0)$ from all three of the original Stock Synthesis model runs (5.6, 5.4, and 5.4, for model runs 1, 2, and 3, respectively). The maximum likelihood estimates of all other model parameters were obtained from Stock Synthesis output in the usual way, as described above.

The relative change in negative log-likelihood units over the range of fixed values for $\log(R_0)$ (the R_0 profile) was compared among the Stock Synthesis model likelihood components for CPUE, length-composition, and recruitment deviations using two diagnostic tests. First, a relatively large change in negative log-likelihood units along the R_0 profile was diagnostic of a relatively informative data source for that particular model. Second, a difference in the location of the minimum negative log-likelihood along the R_0 profile among data sources was diagnostic of either conflict in the data or model misspecification (or both).

2.7 Diagnostic-7

An age-structured production model (ASPM; Maunder and Piner 2015; Carvalho *et al.* 2017) was applied to the Stock Synthesis model results. The diagnostic was adapted here (DC) from (Carvalho *et al.* 2017) and implemented manually (**Table A.2**).

There is more than one way to implement this diagnostic, as discussed below. The diagnostic was implemented here by fixing selectivity to its estimated values in the full integrated stock assessment model, fixing recruitment equal to the stock recruitment curve obtained from the full integrated stock assessment model, and then estimating the remaining parameters of the stock assessment model. Trends in relative spawning stock size were compared from the full integrated stock assessment model and the ASPM. The asymptotic 95% confidence intervals obtained for relative spawning stock size were compared for the full stock assessment model and ASPM.

On the one hand, Carvalho *et al.* (2017) suggest that if the ASPM is able to fit well to the indices of abundance that have good contrast (i.e. those that have declining and/or increasing trends), then this is evidence of the existence of a production function, and the indices will likely provide information about absolute abundance. On the other hand, Carvalho *et al.* (2017) suggest that if there is not a good fit to the indices, then the catch data alone cannot explain the trajectories depicted in the indices of relative abundance. This can have several causes: (i) the stock is recruitment-driven; (ii) the stock has not yet declined to the point at which catch is a major factor influencing abundance; (iii) the base-case model is incorrect; or (iv) the indices of relative abundance are not proportional to abundance.

2.8 Diagnostic-8

Markov Chain Monte Carlo (MCMC) is used in ICCAT to develop Kobe plots and Kobe II risk matrices from individual and multiple model runs. MCMC diagnostics relative to parameter estimation uncertainty and the resulting measures of precision were adapted here (HW) from those previously implemented in JABBA (Winker *et al.* 2018).

Convergence of the MCMC samples to the posterior distribution was evaluated here with a visual inspection of the trace along with 'Heidelberger and Welch' and 'Geweke' tests implemented in the coda package (Plummer *et al.* 2006).

MCMC convergence was evaluated for a subset of estimated parameters and derived quantities obtained from Stock Synthesis output: 1) $\text{Log}(R_0)$, as described above, which determines the absolute scale of the population; 2) unfished spawning stock fecundity (SSF_0); 3) spawning stock fecundity at MSY (SSF_{MSY}); 4) fishing mortality at MSY (F_{MSY}); 5) ending year spawning stock fecundity relative to its value at MSY (SSF/SSF_{MSY}); and 6) ending year fishing mortality relative to its value at MSY (F/F_{MSY}).

2.9 Diagnostic-9

In addition to determining if the model fits the historical data, it is important to evaluate whether the model can replicate the future dynamics of the system, which is required to provide management advice. One diagnostic for this is model prediction skill. Model prediction skill was diagnosed here with hindcasting precision (Kell *et al.* 2016). Prediction skill of the stock assessment model was evaluated using a hindcast (Kell *et al.* 2016), where each assessment model was retrospectively re-run by tail cutting, i.e. removing recent years' data and the biomass trajectories projected up to the most recent year.

Model-free validation was adapted here for Stock Synthesis (LK) and implemented in FLR (Kell et al 2007) to compare the observed CPUE indices in the recent years (the input data) to their out of sample predicted values (the hindcast) calculated by multiplication of catchability and vulnerable biomass obtained from the stock assessment model one-step ahead predicted values from each hindcast for up to 15 years, as described in **Appendix C**.

Hindcast results were summarized using the Mean Absolute Scaled Error (MASE), as defined in **Appendix C**, which compared forecast accuracy to a naïve forecast equal to the last observation. A "naïve" forecast is where the predicted value is equal to the last observation, i.e. a random walk. Values greater than one indicate that the average one-step ahead forecasts from the naïve method perform better than the average forecast values under consideration. MASE also penalizes positive and negative errors and errors in large forecasts and small forecasts equally.

A benefit of the hindcast is that it can be used to compare different models and models run with different datasets. It can also be used for validation, i.e. to examine if a model family should be modified or extended, and is complementary to model selection and hypothesis testing. Model selection searches for the most suitable model within a family, whilst hypothesis testing examines if the model structure can be reduced. The inability of the model to predict observations, e.g., of CPUE indices, is a cause for concern since if data are regarded as being representative of the dynamics of the stock then they can be used as a validation measure. This is especially important where it is not possible to validate a model by comparing unobservable quantities such as stock biomass and fishing mortality.

Hindcasting can also be used for the comparison of different types of models using different data sets, i.e., to compare integrated models such as Stock Synthesis and biomass based models such as JABBA. Hindcasting can also be used to compare different model scenarios with different weighting, which is difficult to do with methods such as AIC.

3. Results

3.1 Diagnostic-1

JABBA-residual plot results are provided in **Figure 1**. All three models passed this diagnostic. The overall JABBA-RMSE for each run was 29%, but was slightly reduced from model run 1 (29.3 %) to model run 2 (29.2%) and model run 3 (29.1%). The JABBA-RMSE was less than 0.3, which indicated a generally good model fit to CPUE data (Winker *et al.* 2018). However, there appeared to be increasing variability in the residuals of model fit to CPUE over time.

3.2 Diagnostic-2

Runs test results applied to residuals from each CPUE index fit in the models are provided in **Figure 2**. The results for this diagnostic were mixed. There was no evidence (p-value > 0.05) to reject the hypothesis of randomly distributed residuals for most of the individual CPUE time series fit in the models. However, there was evidence (p-value < 0.05) to reject the hypothesis of randomly distributed residuals for one CPUE time series (S1_USA_LL_Log) in model run 2 and model run 3. There were also some years with residuals outside the three-sigma rule for two CPUE time series (S1_USA_LL_Log and S4_EU_POR_LL). Index S2 (USA_LL_Obs) was not fit in the model likelihood (lambda = 0) because of high variability in the index and because S2 described the same fishery as S1 (USA LL Log) (see **Table B.1)**.

3.3 Diagnostic-3

Runs test results applied to standardized residuals of the fit to length composition data by fleet are provided in **Figure 3**. All three models passed this diagnostic. There was no evidence (p-value > 0.05) to reject the hypothesis of randomly distributed residuals. Mean of the input and estimated length by fleet each year are provided in **Figure 4**.

3.4 Diagnostic-4

Runs test results applied to estimated recruitment deviations are provided in **Figure 5**. All three models failed this diagnostic. There was evidence (p-value < 0.05) to reject the hypothesis of randomly distributed recruitment deviations in model runs 1, 2, and 3. There was an apparent trend in the recruitment deviations with several recruitment deviations below the three-sigma rule followed by several above the three-sigma rule.

3.5 Diagnostic-5

Results from the retrospective analysis are provided in **Figure 6**. All three models passed this diagnostic. There was a consistent pattern of change in the 5-year peel relative to the full stock assessment model. However, the Mohn's rho statistic was 0.064, 0.063, and 0.055 for model runs 1, 2, and 3, respectively, which was between -0.15 and 0.20 and, consequently, indicated that the retrospective pattern was relatively small (Hurtado-Ferro *et al.* 2014; Carvalho *et al.* 2017). The ending values of all peels also fell within the asymptotic 95% confidence interval of the full stock assessment model, which also indicated that the retrospective pattern was within the asymptotic 95% margin of error obtained from the full stock assessment model.

3.6 Diagnostic-6

R_0 likelihood component profile results are provided in **Figure 7**. The results for this diagnostic were mixed.

First, there was a relatively larger change in the R_0 profile for estimated recruitment deviations (Recruitment) relative to the data likelihood components for length composition (Length_comp) and CPUE (Survey), respectively. This result indicated that the estimation of the recruitment deviations was relatively informative within the likelihood. Additionally, there was a relatively large change in the R_0 profile for two of the CPUE time series (S1_USA_LL_Log, S5_EU_ESP_LL) and two of the length compositions (F1_EU_LL, F4_USA_LL) (surveys and fleets as defined in **Table B.1**). This result indicated that these data sources were relatively more informative than the other data components included in the R_0 profile.

Second, differences in the location of the minimum value along the R_0 profile were observed among likelihood components for estimated recruitment deviations and the data likelihood components for CPUE and length composition. A minimum value was not identified for length composition. These results indicate that there was conflict among the different likelihood components in the best estimate of R_0. Differences in the location of the minimum value along the R_0 profile were also observed among individual CPUE indices and length composition data sources. A difference in the location of the minimum was identified for S1_USA_LL_Log, S5_EU_ESP_LL, and F1_EU_LL (surveys and fleets as defined in **Table B.1**). A minimum value was not identified for the other surveys or fleets. These differences in the location of the minimum value indicate that there was also conflict among individual CPUE indices and length composition data sources regarding the best estimate of R_0.

These results indicate that data weighting applied to conflicting CPUE indices and length composition data in the full integrated stock assessment model may have a large effect on model results, and should be evaluated carefully, as discussed below. The shape and location of the minimum also changed slightly among model runs, which may provide a useful diagnostic for future model development.

3.7 Diagnostic-7

ASPM results are provided in **Figure 8**. The results of this diagnostic were mixed. The models showed similar overall trend, however after the 1990's the ASPM showed a less steep decline in spawning stock size than the full integrated stock assessment model. The asymptotic 95% confidence intervals of relative spawning stock size did not overlap for many of the most recent years.

The differences observed here between the full integrated stock assessment model compared to the ASPM are explained by the estimated recruitment deviations in the full model (**Figure 8)**, which allowed for variability in age-0 recruitment (which can be interpreted as process error necessary to fit the observed trends CPUE). In contrast, the ASPM fixed recruitment to the assumed stock recruitment curve which limited inter-annual variability in age-0 recruitment and resulted in a relatively poor fit to the observed trends CPUE (**Figure 8**).

3.8 Diagnostic-8

MCMC diagnostics for each model run were evaluated with both a relatively short and a relatively long chain of 500,000 and one million random draws, respectively. Both chains removed the first 10,000 draws (a burn-in of 10,000) and saved every 1,000th draw after that (a thinning interval of 1,000). The results for this diagnostic were mixed.

A visual inspection of the MCMC trace (**Figures 9 and 10**) did not provide evidence to reject convergence of the MCMC samples (trace generally centered on the median, although there were some periods of deviation). Results of the Heidelberger-and-Welch and Geweke tests (**Tables 2 and 3**) were mixed. There was no evidence (Geweke p-value > 0.05; Heidelberger-and-Welch p-value > 0.05) to reject the hypothesis of convergence for most of the derived quantities examined with the long MCMC chain except for the derived quantities F_{MSY} and F-ratio for Run 2 (Geweke p-value < 0.05). In contrast, there was evidence (Geweke p-value < 0.05) to reject the hypothesis of convergence for most of the derived quantities examined with the short MCMC chain for Run 3 and for one of the derived quantities for Run 2.

Histograms of the MCMC posterior distributions (**Figures 11 and 12**) resulted in relatively narrow distributions for all derived quantities, and some MCMC distributions were slightly skewed in some model runs.

Kobe plots (**Figure 13**) indicated that Runs 1 and 2 overlapped and differed from Run 3. All model runs resulted in correlation between derived quantities F/F_MSY and SSF/SSF_MSY (banana shaped distribution).

3.9 Diagnostic-9

Figure C.1 shows the hindcast for Stock Synthesis Runs 1, 2 and 3. The results for this diagnostic were mixed.

MASE scores for the CPUE indices EU_ESP_LL, and JPN_LL were greater than one for model runs 1, 2, and 3 (fleets as defined in **Table B.1**). This diagnostic result indicated that the average one-step ahead naïve forecast was a better predictor than the stock assessment model for those indices, i.e. knowledge of resource dynamics in these cases did not help in prediction of those indices.

In contrast, MASE scores for the remaining CPUE indices were less than one for model runs 1, 2, and 3 (fleets as defined in **Table B.1**). This diagnostic result indicated that the stock assessment model was a better predictor than the average one-step ahead naïve forecast for those indices. As noted above, USA_LL_Obs was not fit in the model likelihood (See **Table B.1**). The CPUE indices for CTP_LL was also relatively short (6 years) compared to the other CPUE indices.

4. Discussion

The nine stock assessment diagnostics provide an objective aid to evaluate stock assessment model fit to data and to identify possible model misspecification (**Table 1**). An expectation is that if the data used in the model are informative, and model fit to data is sufficient, then a carefully selected range of model diagnostics should be met. In contrast, if a diagnostic fails, then continued model development or a re-evaluation of the input data may be necessary. If time or resources do not allow for continued model development, then the implications of failing a diagnostic should be discussed in order to identify potential model weaknesses and priorities for future model development. A disadvantage of using multiple diagnostics is the possibility of obtaining a false positive diagnostic result, failing to pass a diagnostic when there is no actual underlying failure in the model fit to data. Consequently, failing a diagnostic is not necessarily the same as model failure, but rather an aid in model interpretation and future development.

The diagnostics identified above assume that the model has been evaluated for convergence (e.g., **Table A.3**), and that a thorough visual inspection of model residuals has already been conducted to identify goodness of fit (e.g. a lack of pattern in the residuals) or model misspecification (e.g. patterns in the residuals). Other, perhaps better, convergence and model specification diagnostics may be available from the published scientific literature but were not evaluated here due to time constraints. This document focused on implementing published diagnostics methods for identifying model misspecification available in Carvalho *et al.* (2017) and Winker *et al.* (2018) and on hindcasting (Kell *et al.* 2016).

4.1 Diagnostics 1–4

The JABBA-residual and the residuals runs test diagnostics are designed to accompany a visual inspection of time series residuals and provide and objective method for determining the statistical significance of any non-random patterns observed in the residuals. Their utility is as an objective aid for determining when model development is sufficiently complete to fit the time series data used in the model. A rule of thumb used here is that if the time series data used in the model are valid, then model development should continue until the JABBA-RMSE < 0.3 and there are no significant runs in the residuals identified with a runs test (p-value > 0.05). If these conditions are not met, then additional model development may be needed, informed by observed patterns in the residuals, until the diagnostics are met.

For each of the three models runs examined here, the JABBA-RMSE was less than 0.3, which indicated a reasonably good fit to CPUE. Individual CPUE and length composition data used in the model appeared to be fit by the model reasonably well and did not have significant runs in the residuals identified with a runs test, except for CPUE (S1_USA_LL_Log) in model runs 2 and 3.

In contrast, there was a significant non-random pattern in the time-series of estimated recruitment deviations for each of the model runs. An accepted model development strategy is to add additional model structure informed by observed patterns in the residuals until all of the time series diagnostics are met. Consequently, these results suggest that additional model development or a re-evaluation of the input data may improve diagnostic results for the time-series of estimated recruitment deviations and residuals of the fit to CPUE S1_USA_LL_Log.

4.2 Diagnostic-5

The retrospective analysis diagnostic is designed to evaluate the potential for bias in relative spawning stock size provided from integrated stock assessment model results. Mohn's rho statistic will be large, either positive or negative, when there is a consistent pattern of change in the peel relative to the full assessment. For the three model runs examined here, there was a consistent pattern in the peel. However, the pattern was relatively small (Mohn's rho statistic was between -0.15 and 0.20).

4.3 Diagnostic-6

The R_0 profile is designed to identify conflict among likelihood components fit within an integrated stock assessment model relative to all of the assumptions made within that particular model. The equilibrium recruitment parameter, R_0, determines the absolute size (scale) of the population. Consequently, conflicts among likelihood components in the estimation of R_0 can have a large effect on model results depending upon the data weighting applied among the conflicting likelihood components in the model. If the data are assumed to be valid, then the expectation is that model development should continue until there are no data conflicts identified by the R_0 profile diagnostic. However, in practice this can be difficult to accomplish, in which case the diagnostic can be used to identify conflict among likelihood components and the possible implications of alternative data weighting applied among the conflicting likelihood components in the model.

For the three models runs examined here, there was a relatively large change in the total likelihood obtained for estimated recruitment deviation compared to those for fits to data (CPUE and length composition). This result suggests that the estimation of recruitment deviations has a large influence in the model likelihood, and is consistent with the results of diagnostics 1-4.

There were also differences in the location of the minimum negative log-likelihood along the R_0 profile observed among data likelihood components for all model runs. This result identifies data conflict in the model and indicates that data weighting applied among the conflicting likelihood components may have a large effect on model results and should be evaluated carefully (e.g., see Punt 2017). For the Stock Synthesis model runs evaluated here (Courtney *et al.* 2017a; Anon 2017), the influence of conflicting data within the model likelihood was reduced

using a two-stage Francis (2011) data "right-weighting" approach. The approach was implemented for each of the three models to iteratively tune (right-weight) variance adjustment factors for fleet-specific relative abundance indices (CPUE) externally to the model (Stage 1) and fleet-specific size data distributions (length composition) within the Stock Synthesis model (Stage 2) as described in Courtney *et al.* (2017a, 2017b).

The R_0 profile showed evidence that there was a conflict between the CPUE and the size data. The two-stage Francis approach seemed to have reduced the conflict, but not completely eliminated it. It is important to note that correctly specified stock assessment models generally require more information than is available for most by-catch species, such as shortfin mako sharks. For example, many of the size data available for this stock assessment were incomplete across fisheries, and the time series suffered from low sample sizes and inconsistencies across years. Diagnosing which of many confounded model processes lead to the data conflicts is difficult even for stock assessments of targeted species. However, despite the identified conflict between CPUE and size data, the diagnostics presented here showed that the model produced a good fit for the CPUE series, which is consistent with the recommendations of Francis (2011, 2014, 2017), namely "do not let other data stop the model from fitting abundance data well".

4.4 Diagnostic-7

The rational for the ASPM diagnostic is to determine if the indices of abundance used in an integrated stock assessment model provide information about the resulting time series of relative stock size estimated within the model. Following Carvalho *et al.* (2017), if the ASPM cannot mimic the CPUE indices fit in the stock assessment model, then either the stock is recruitment-driven (unlikely for a long lived species), catch levels have not been high enough to have a detectable impact on the population (unlikely for a fully exploited long lived species), the production function assumed within the integrated model (which includes the combined effect of the stock recruitment relationship, natural mortality, and life history) is incorrect, or the index of relative abundance is uncertain or not proportional to abundance.

The expectation is that model development should continue (for example by adding more model structure – or re-evaluating input data) until indices of abundance used in an integrated stock assessment model provide all of the information about the relative trend in abundance estimated within the model. However, in practice this is difficult to accomplish and may never be possible.

Overall, the ASPM for the three models runs examined here follows the trend from the full-integrated stock assessment. However, after the 1990's it seems that the changes in the abundance indices cannot be explained by the catches alone. These results indicate that the abundance information, both absolute and relative, contained in the CPUE indices cannot be interpreted without accounting for the fluctuations in recruitment. This result is consistent with the results of diagnostics 1 – 4, and 6, in suggesting that the recruitment deviation estimation has a large influence on the model fit to CPUE.

A separate ASPM is required to test if the information about scale in the integrated model is coming from the composition data. In order to check whether the stock is recruitment-driven involves fitting the ASPM with the recruitment deviates fixed at the values estimated in the base case integrated model. If the ASPM with recruitment fixed at the values identified from the full integrated stock assessment model values is still not able to capture the population trajectory estimated in the integrated model, then it can be concluded that the information about scale in the integrated model is coming from the composition data. This diagnostic was not implemented due to time constraints.

4.5 Diagnostic-8

The MCMC diagnostic results indicate that chain length may be an important consideration, especially for model run 3 which implemented the LFSR stock recruitment relationship. In practice, it may be impractical to implement a long MCMC chain during an ICCAT assessment meeting, for example if analyses are conducted on a laptop during the assessment meeting.

The skewed distribution in the derived quantity for SSF/SSF_{MSY} may also be an important consideration, for example if MCMC is used to produce KOBE plots and KOBE II tables of SSF/SSF_{MSY} probabilities. In particular, the median from a skewed MCMC distribution for SSF/SSF_{MSY} may differ from the parameter estimate for SSF/SSF_{MSY} obtained by ADMB.

4.6 Diagnostic-9

A hind-cast cross-validation diagnostic identified that all three models had poor prediction skill for two of the five indices. An explanation may be that either the indices are not proportional to relative abundance or that there are processes that are not being accounted for in the model structure. In the latter case this could be due to recruitment dynamics, or changes in spatial and temporal distribution or catchability. This could be investigated by considering a range of scenarios based on alternative datasets and model structures. Hindcasting could then be used to identify the best performing scenarios (e.g., choice of models and data) by comparing predictions with observations (**Appendix C**).

References

Anhøj, J. and 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(11):e113825. Available: https://doi.org/10.1371/journal.pone.0113825 (May 2019).

Anon. 2017. Report of the 2017 Shortfin Mako Assessment Meeting (Madrid, Spain 12–16 June 2017). Collect. Vol. Sci. Pap. ICCAT, 74(4):1465–1561.

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

Courtney, D. 2016. Preliminary Stock Synthesis model runs conducted for North Atlantic blue shark. SCRS/2015/151. Collect. Vol. Sci. Pap. ICCAT, 72(5):1186–1232.

Courtney, D., Cortés, E., and Zhang, X. 2017a. Stock synthesis model runs conducted for North Atlantic shortfin mako shark. SCRS/2017/125. Collect Vol. Sci. Pap. ICCAT 74(4):1759–1821.

Courtney, D., Cortés, E., Zhang, X. and Carvalho, F. 2017b. Stock Synthesis model sensitivity to data weighting: An example from preliminary model runs previously completed for North Atlantic blue shark. SCRS/2016/066. Collect Vol. Sci. Pap. ICCAT 73(8):2860–2890.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J., 2011. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim. Methods Softw. 27:233–252.

Francis, R. I. C. C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68:1124–1138.

Francis, R. I. C. C. 2014. Replacing the multinomial in stock assessment models: a first step. Fish. Res. 151:70–84.

Francis, R. I. C. C. 2017. Revisiting data weighting in fisheries stock assessment models. Fish. Res. 192:5–15.

Hurtado-Ferro, F., Szuwalski, C. S., Valero, J. L., Anderson, S. C., Cunningham, C. J., Johnson, K. F., Licandeo, R., McGilliard, C. R., Monnahan, C. C., Muradian, M. L., Ono, K., Vert-Pre, K. A., Whitten, A. R., and Punt, A. E. 2014. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES J. Mar. Sci. 72:99–110.

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

Kell, L. T., Mosqueira, I., Grosjean, P., Fromentin, J.-M., Garcia, D., Hillary, R., Jardim, E., Mardle, S., Pastoors, M. A., Poos, J. J., Scott, F., and Scott, R. D. 2007. FLR: An open-source framework for the evaluation and development of management strategies. ICES J. Mar. Sci. 64:640–646.

Maunder, M. N., Piner, K. R. 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72:7–18.

Methot Jr., R. D. 2015. User manual for Stock Synthesis model version 3.24s, Updated February 11, 2015. NOAA Fisheries, Seattle, WA.

Methot Jr., R. D., and Wetzel, C. R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142:86–99.

Plummer, M., Best, N., Cowles, K., and Vines, K. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News 6:7–11

Punt, A. E., 2017. Some insights into data weighting in integrated stock assessments. Fish. Res. 192:52–65.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available: https://www.R-project.org/ (May 2019).

Taylor, I. G., *et al.* 2018. r4ss: R Code for Stock Synthesis. R package version 1.30.2. Available: https://github.com/r4ss (May 2019).

Trapletti, A. 2011. tseries: Time series analysis and computational finance. Rpackage version 0. 10-25. Available: http://CRAN.R-project.org/package=tseries (May 2019).

Winker, H., Carvalho, F., and Kapur, M. 2018. JABBA: Just Another Bayesian Biomass Assessment. Fish. Res. 204:275–288.

Table 1. Summary of diagnostics results: 1) green, all three models passed; 2) yellow, some model diagnostics were failed, while others were passed, so the results for this diagnostic were mixed; and 3) red, all three models failed this diagnostic.

Diagnostic-1 (JABBA-residual plot and RMSE of CPUE residuals)
All three models passed this diagnostic.

Diagnostic-2 (Runs test of CPUE residuals)
The results for this diagnostic were mixed.

Diagnostic-3 (Runs test of size composition residuals)
All three models passed this diagnostic.

Diagnostic-4 (Runs test of recruitment deviations)
All three models failed this diagnostic.

Diagnostic-5 (Retrospective patterns and Mohn's Rho test)
All three models passed this diagnostic.

Diagnostic-6 (R_0 likelihood component profile)
The results for this diagnostic were mixed.

Diagnostic-7 (ASPM)
The results of this diagnostic were mixed.

Diagnostic-8 (MCMCs)
The results for this diagnostic were mixed.

Diagnostic-9 (Hind-cast cross-validation)
The results for this diagnostic were mixed.

Table 2. Geweke and Heidelberger-and-Welch test results for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C) MCMC short chain (as described in the methods).

A. Run 1

	Median	LCI	UCI	Geweke.p	Heidelberger.p
ln_R_0	5.59	5.50	5.72	0.631	0.471
SSF_0	1368	1244	1563	0.620	0.445
SSF_{MSY}	575	524	656	0.643	0.472
F_{MSY}	0.06	0.06	0.07	0.245	0.120
B-ratio	1.23	1.12	1.39	0.628	0.606
F-ratio	3.37	2.61	4.19	0.413	0.336

B. Run 2

	Median	LCI	UCI	Geweke.p	Heidelberger.p
ln_R_0	5.47	5.35	5.64	0.267	0.798
SSF_0	1209	1079	1440	0.230	0.827
SSF_{MSY}	508	454	606	0.206	0.805
F_{MSY}	0.06	0.06	0.07	0.02 *	0.109
B-ratio	1.26	1.13	1.48	0.102	0.393
F-ratio	3.19	2.26	4.07	0.190	0.636

C. Run 3

	Median	LCI	UCI	Geweke.p	Heidelberger.p
ln_R_0	5.43	5.34	5.56	0.004 **	0.299
SSF_0	1165	1061	1324	0.004 **	0.314
SSF_{MSY}	607	553	689	0.003 **	0.292
F_{MSY}	0.05	0.05	0.06	0.105	0.238
B-ratio	1.00	0.90	1.12	0.006 **	0.289
F-ratio	3.99	2.90	4.98	0.035 *	0.205

* p-value < 0.05
** p-value < 0.01
*** p-value < 0.001

Table 3. Geweke and Heidelberger-and-Welch test results for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C) MCMC long chain (as described in the methods).

A. Run 1

	Median	LCI	UCI	Geweke.p	Heidelberger.p
ln_R_0	5.60	5.50	5.73	0.686	0.117
SSF_0	1375	1249	1580	0.667	0.113
SSF_{MSY}	578	526	663	0.688	0.129
F_{MSY}	0.06	0.06	0.07	0.952	0.078
B-ratio	1.24	1.12	1.40	0.879	0.241
F-ratio	3.38	2.49	4.21	0.700	0.362

B. Run 2

	Median	LCI	UCI	Geweke.p	Heidelberger.p
ln_R_0	5.47	5.36	5.64	0.250	0.654
SSF_0	1213	1083	1442	0.283	0.693
SSF_{MSY}	509	455	607	0.266	0.655
F_{MSY}	0.06	0.06	0.07	0.046 *	0.267
B-ratio	1.27	1.13	1.47	0.084	0.144
F-ratio	3.17	2.24	4.04	0.041 *	0.291

C. Run 3

	Median	LCI	UCI	Geweke.p	Heidelberger.p
ln_R_0	5.43	5.33	5.55	0.767	0.277
SSF_0	1162	1056	1315	0.803	0.290
SSF_{MSY}	605	550	685	0.773	0.280
F_{MSY}	0.05	0.05	0.06	0.937	0.573
B-ratio	0.99	0.90	1.11	0.468	0.102
F-ratio	4.01	3.00	5.00	0.354	0.601

* p-value < 0.05
** p-value < 0.01
*** p-value < 0.001

A. Run 1

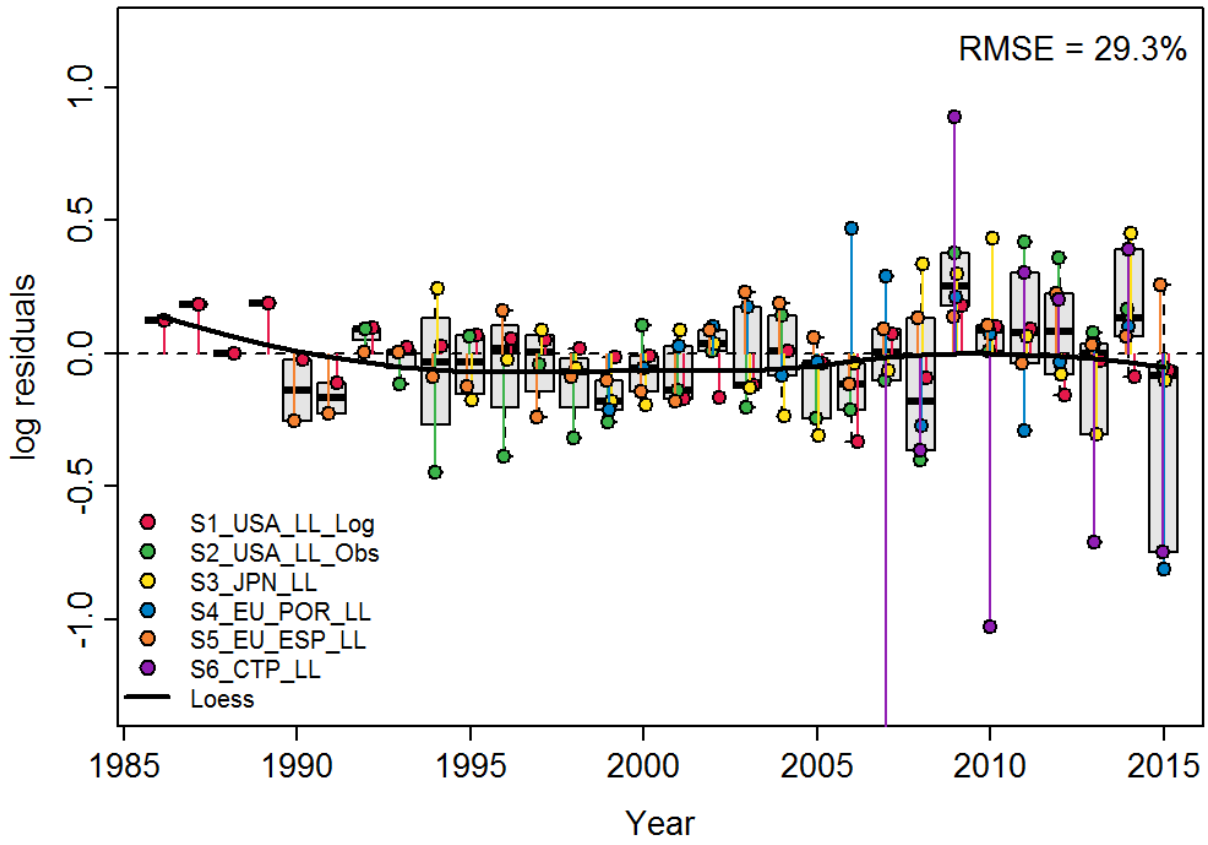


Figure 1. JABBA residual diagnostic plots for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C). Surveys as defined in **Table B.1**. Boxplots as described in the methods section above.

B. Run 2

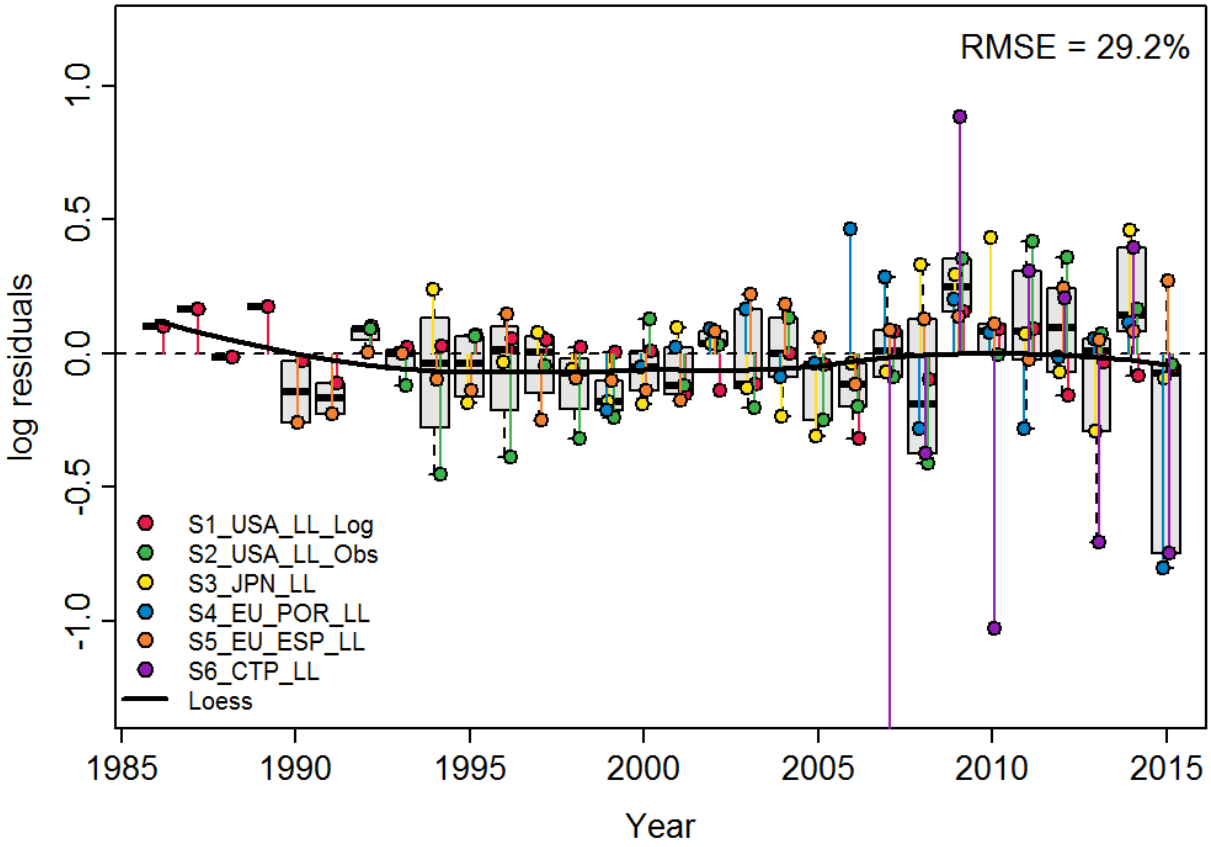


Figure 1. Continued.

C. Run 3

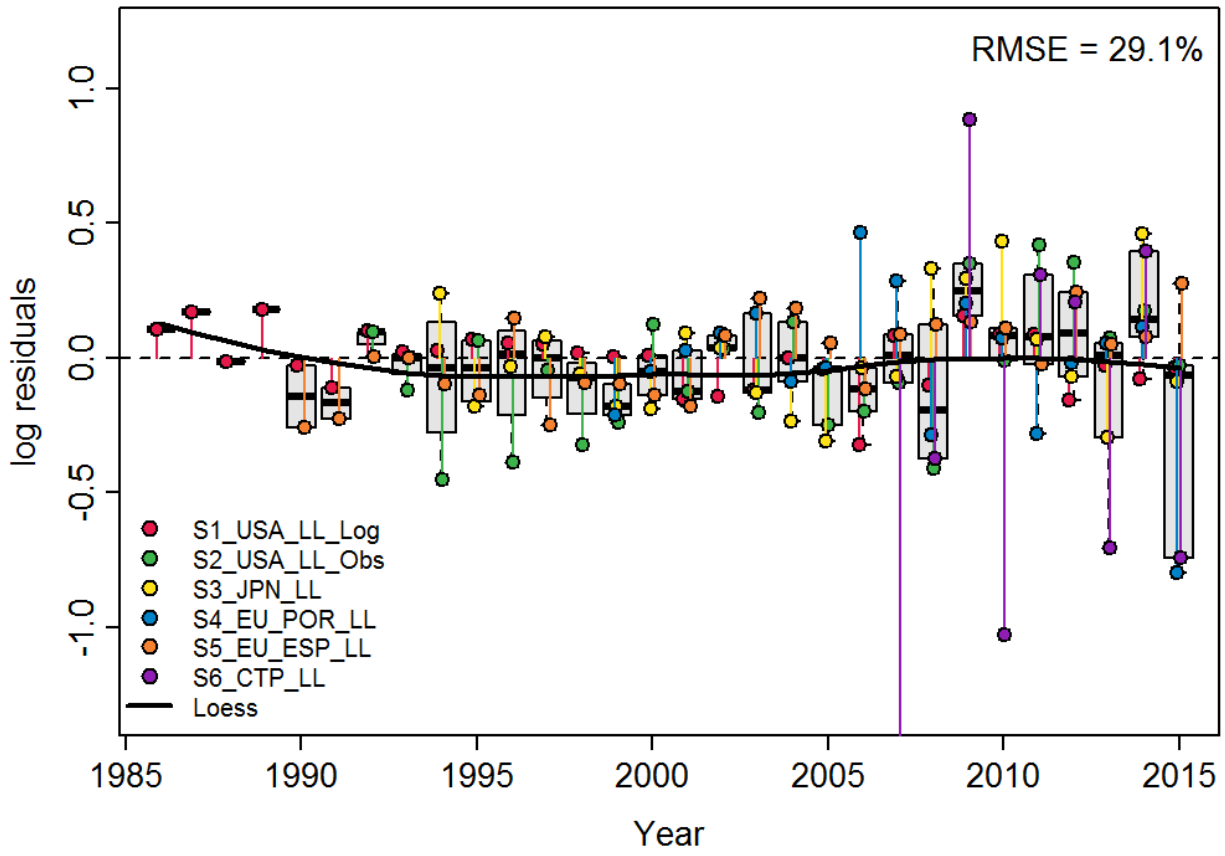


Figure 1. Continued.

A. Run 1

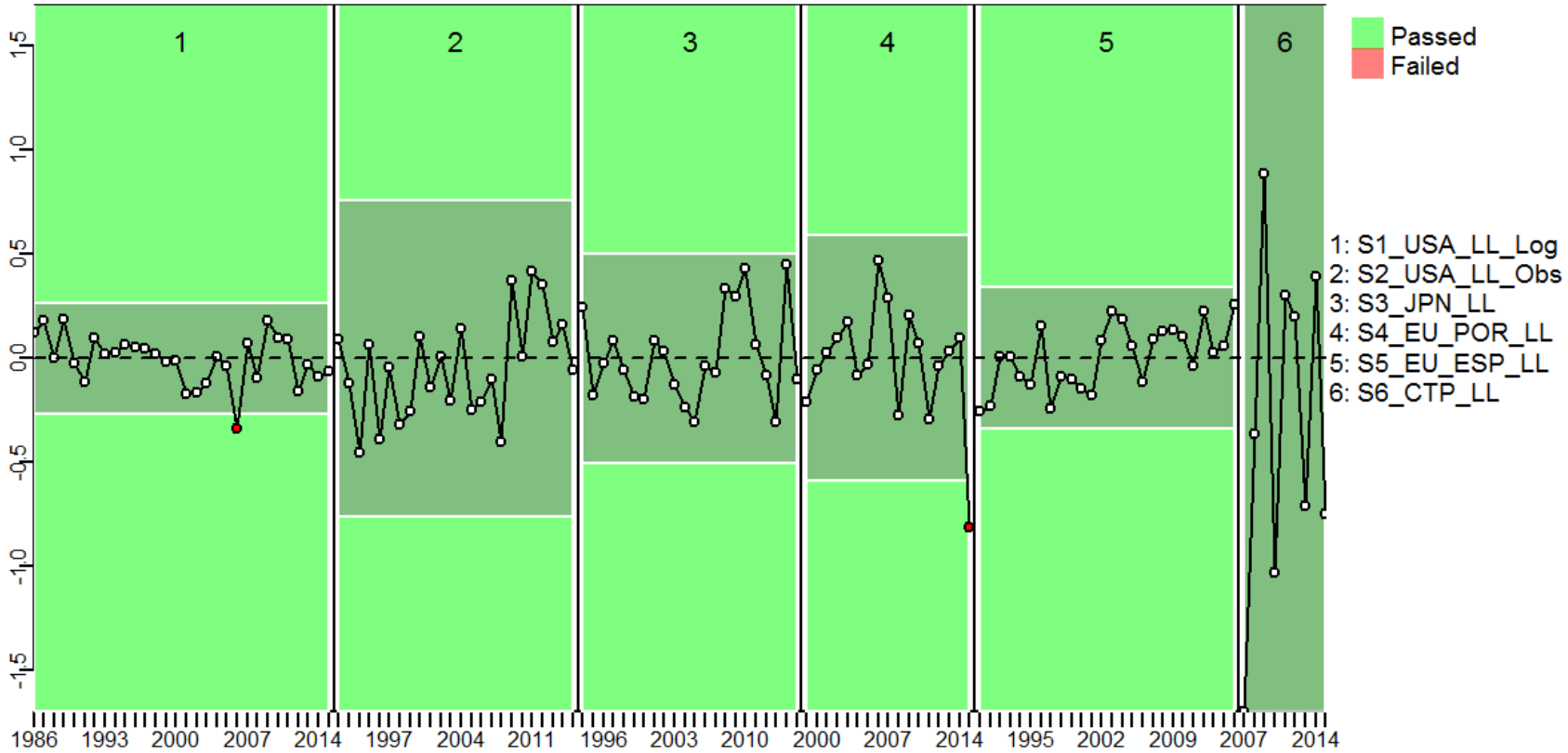


Figure 2. Runs tests on fits to CPUE for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C). Surveys as defined in **Table B.1**. Green background indicates no evidence ($p > 0.05$) to reject the hypothesis of a randomly distributed time-series of residuals. Red background indicates evidence ($p < 0.05$) to reject the hypothesis of randomly distributed residuals. Inner panels represent three standard errors from the mean. Red circles identify a residual greater than three standard errors from the mean.

B. Run 2

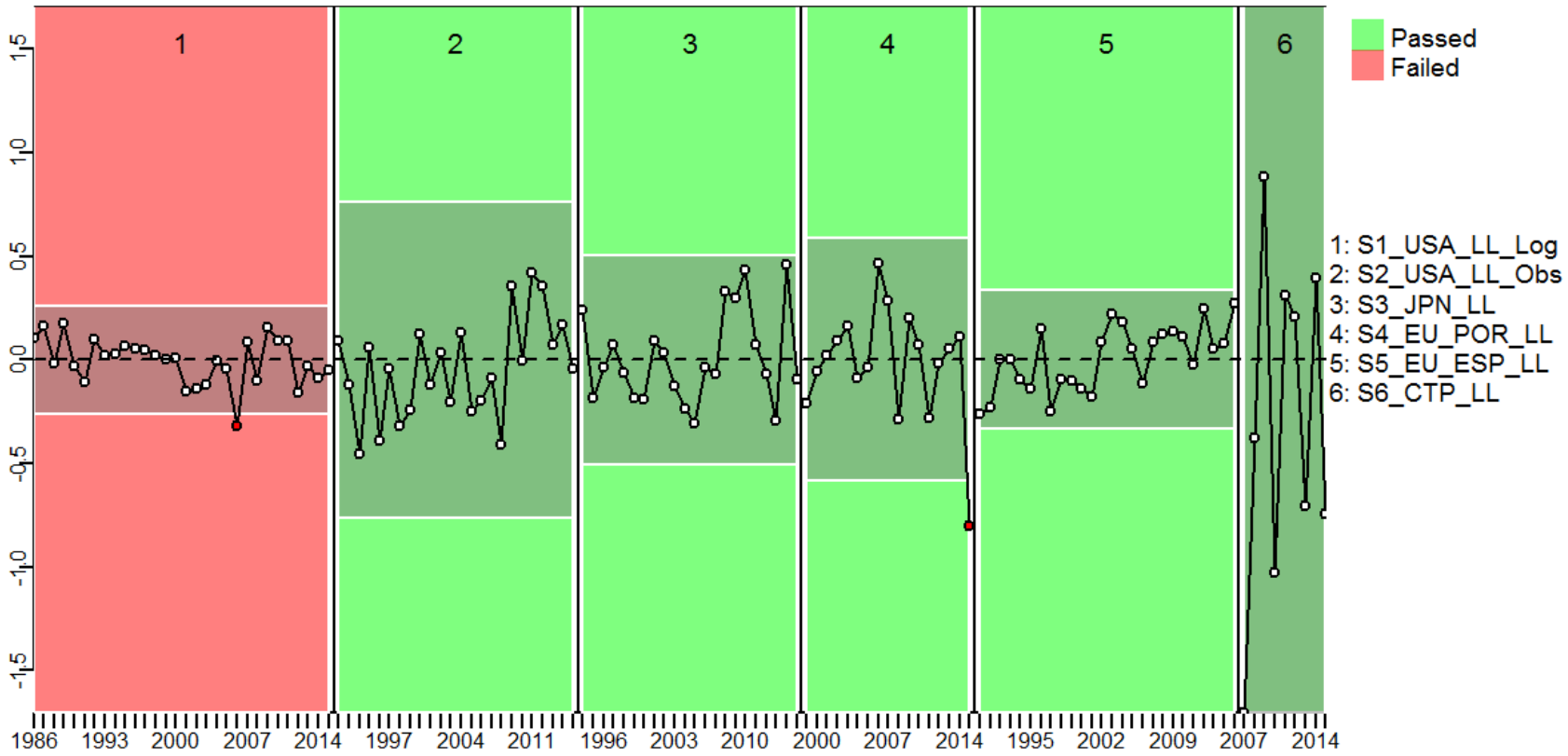


Figure 2. Continued.

C. Run 3

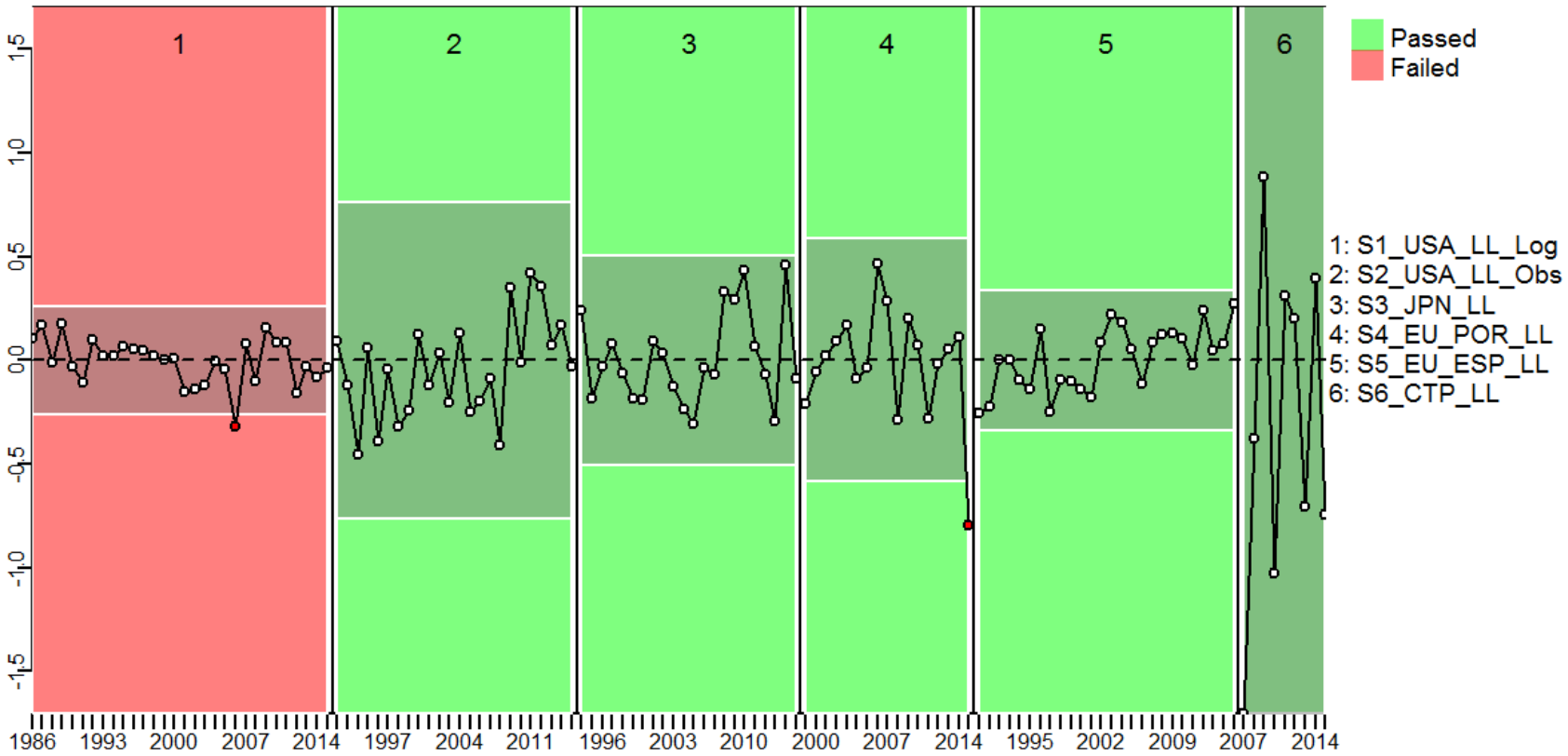


Figure 2. Continued.

A. Run 1

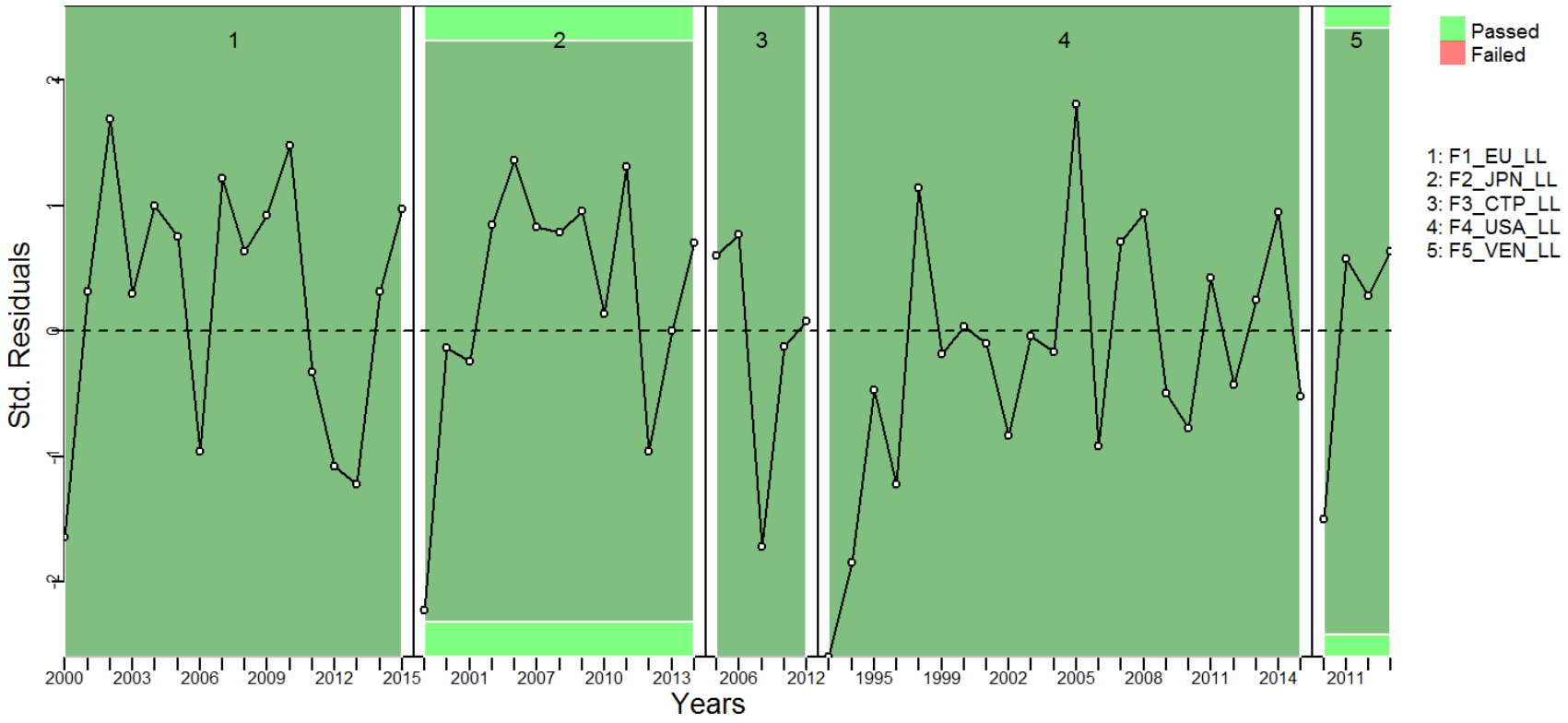


Figure 3. Runs tests on fits to mean length for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C). Fleets as defined in **Table B.1**. Green background indicates no evidence ($p > 0.05$) to reject the hypothesis of a randomly distributed time-series of residuals. Red background indicates evidence ($p < 0.05$) to reject the hypothesis of randomly distributed residuals. Inner panels represent three standard errors from the mean. Red circles identify a residual greater than three standard errors from the mean.

B. Run 2

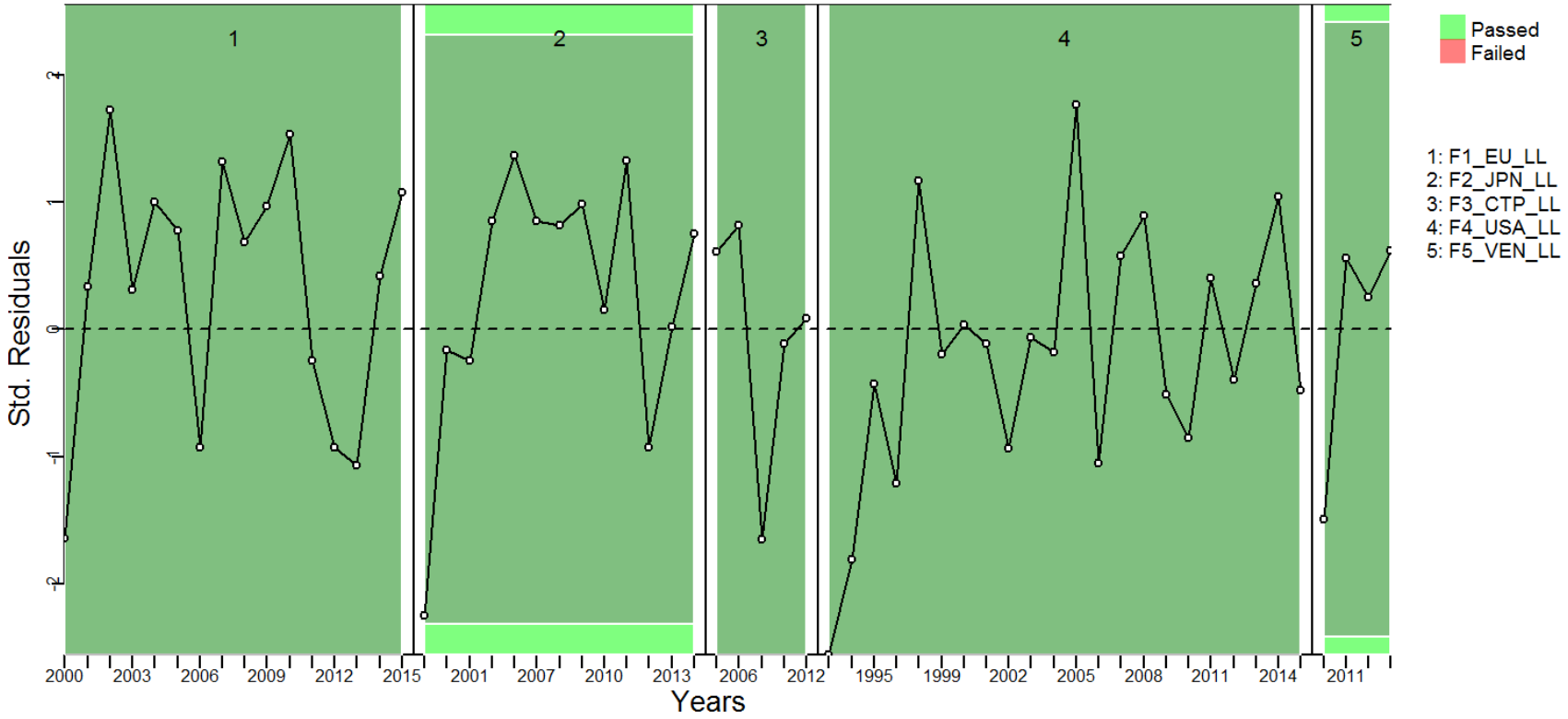


Figure 3. Continued.

C. Run 3

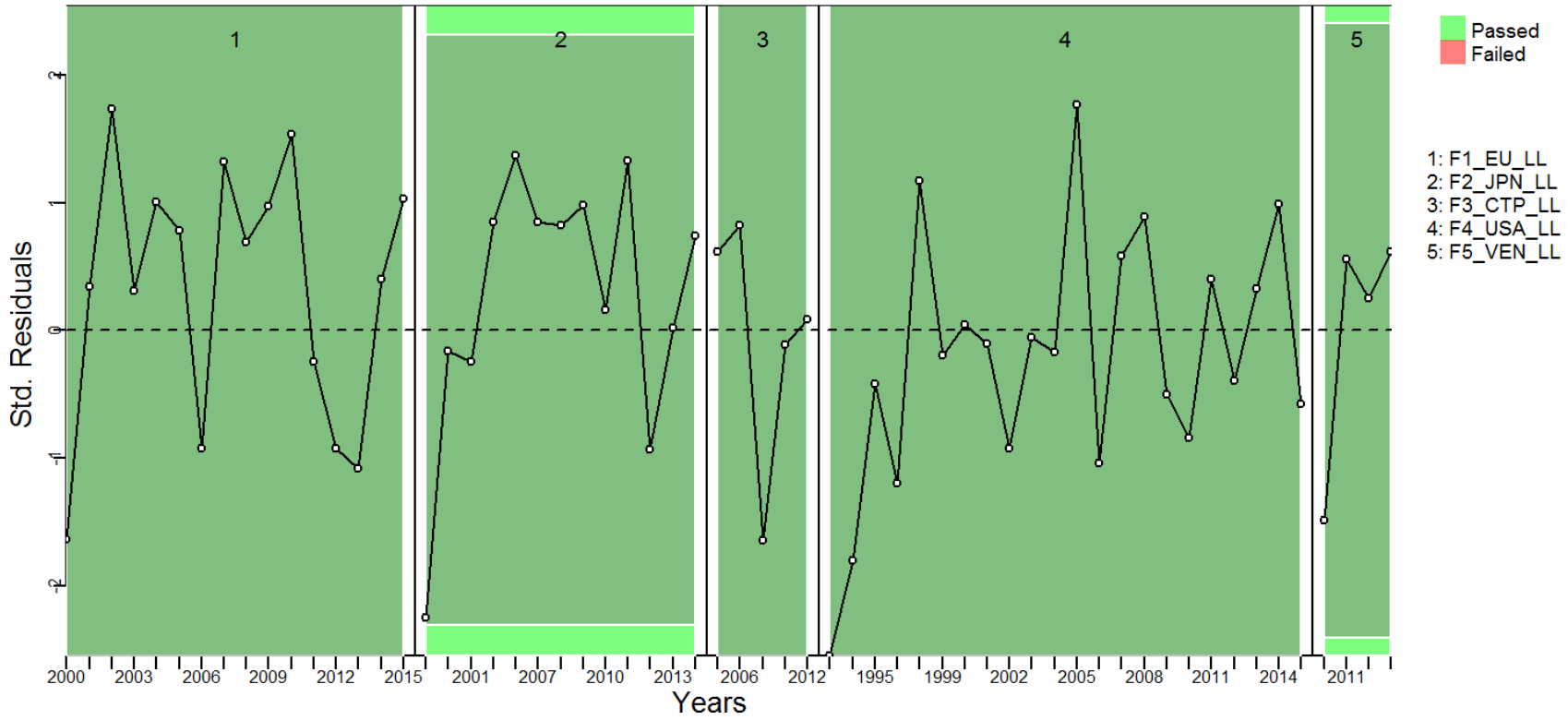


Figure 3. Continued.

A. Run 1

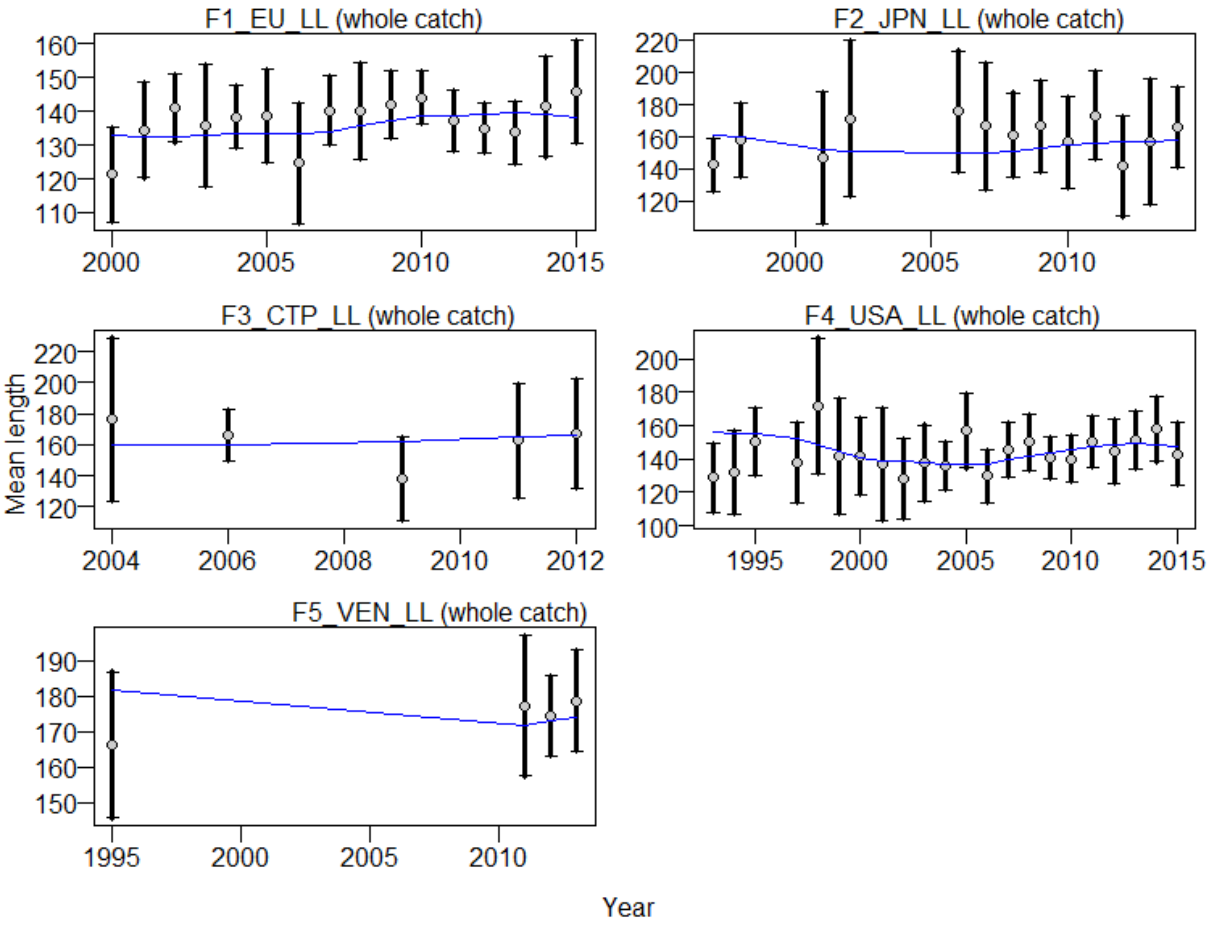


Figure 4. Mean length for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C). Fleets as defined in **Table B.1**. Each subplot represents the input mean length by fleet (grey circle) and asymptotic 95% CI bars for years with input data along with the estimated mean length by fleet (blue line) from the integrated stock assessment model.

B. Run 2

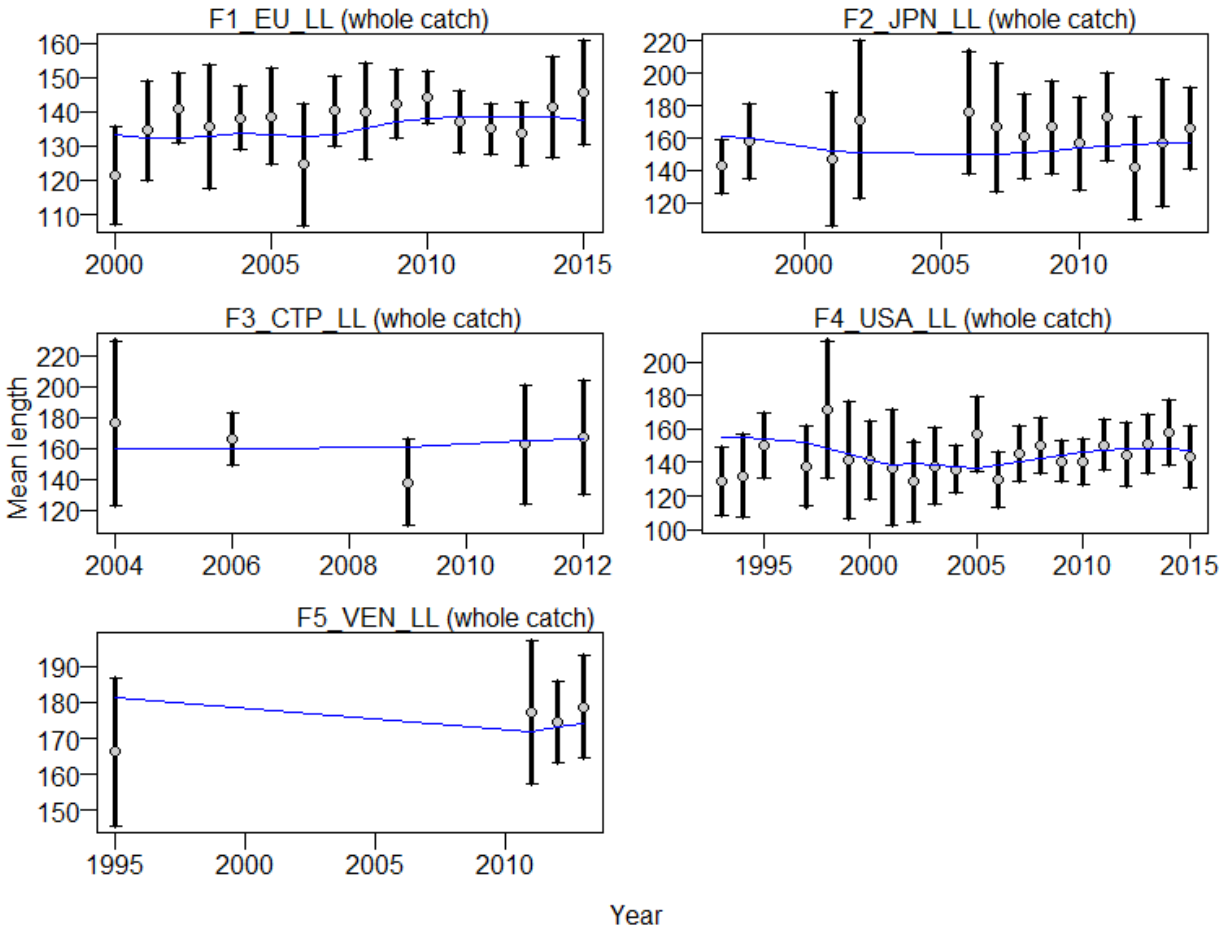


Figure 4. Continued.

C. Run 3

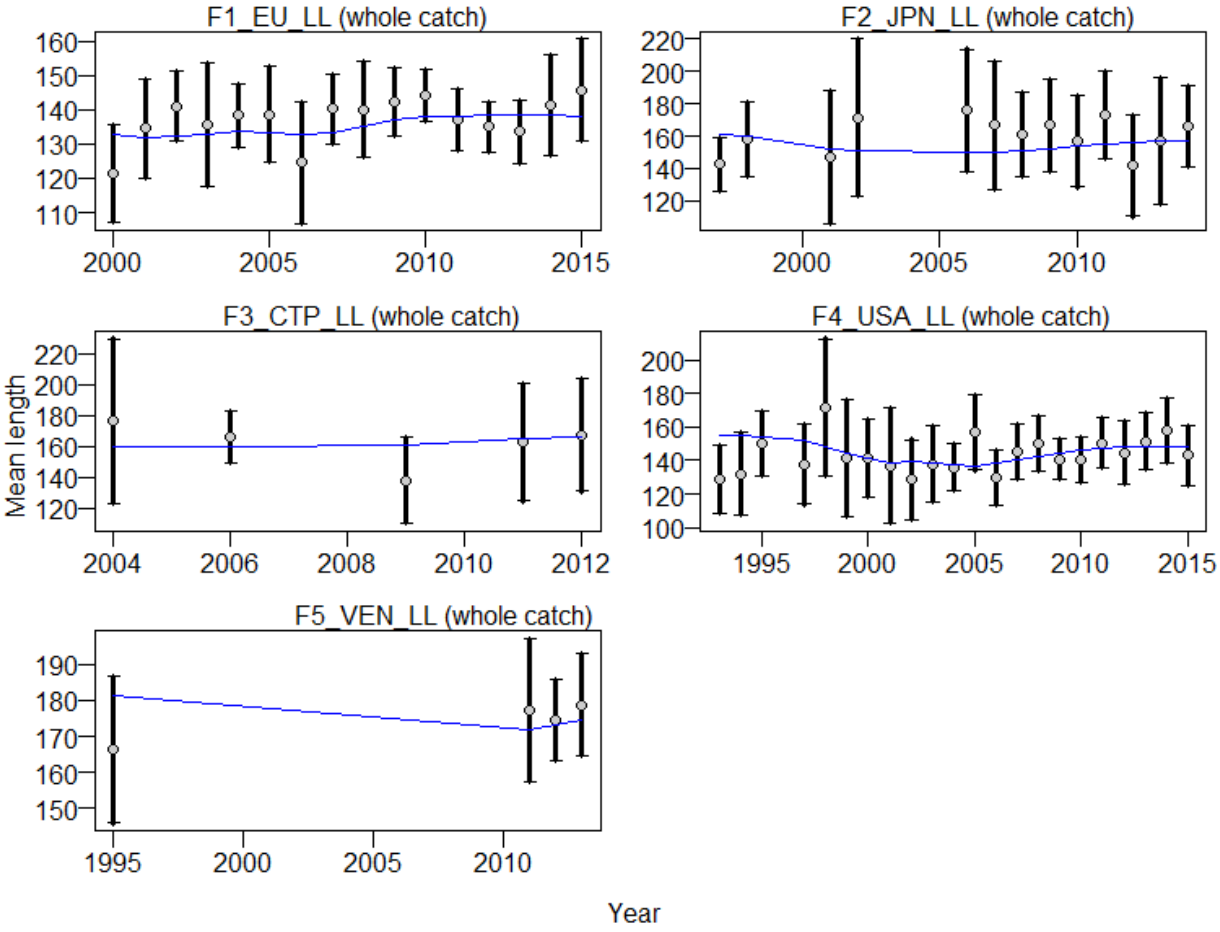


Figure 4. Continued.

A. Run 1

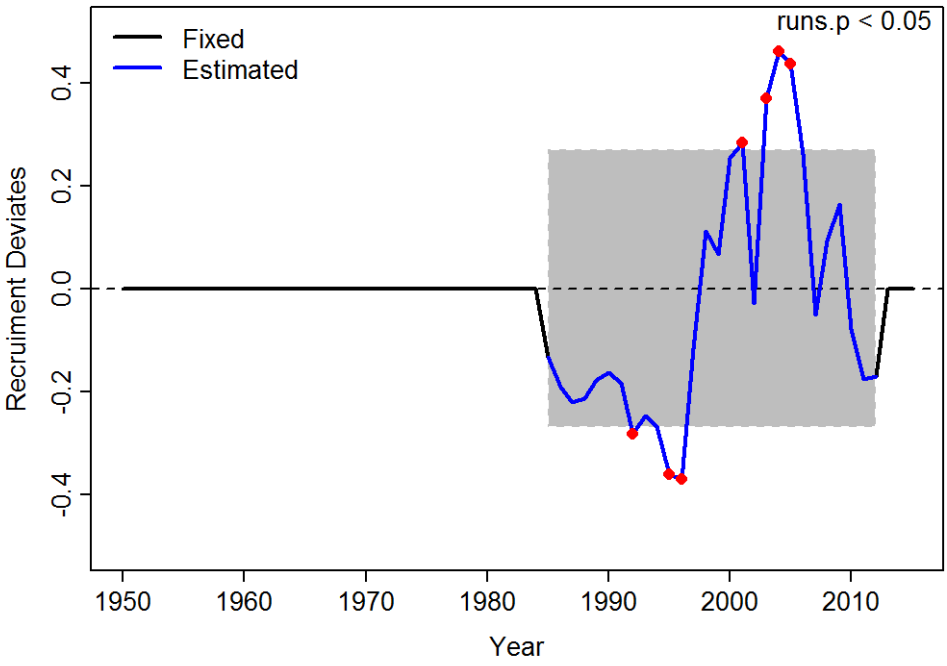


Figure 5. Runs tests on estimated recruitment deviations for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C) provide evidence (runs.p < 0.05) to reject the hypothesis of randomly distributed estimated recruitment deviations. Inner panels represent three standard errors from the mean. Red circles identify a residual greater than three standard errors from the mean.

B. Run 2

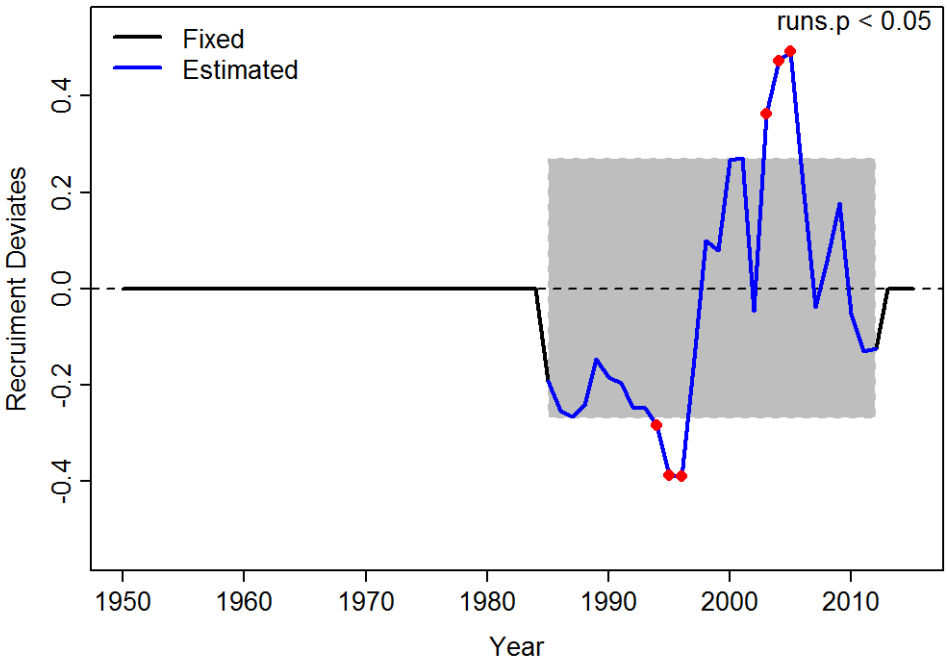


Figure 5. Continued.

C. Run 3

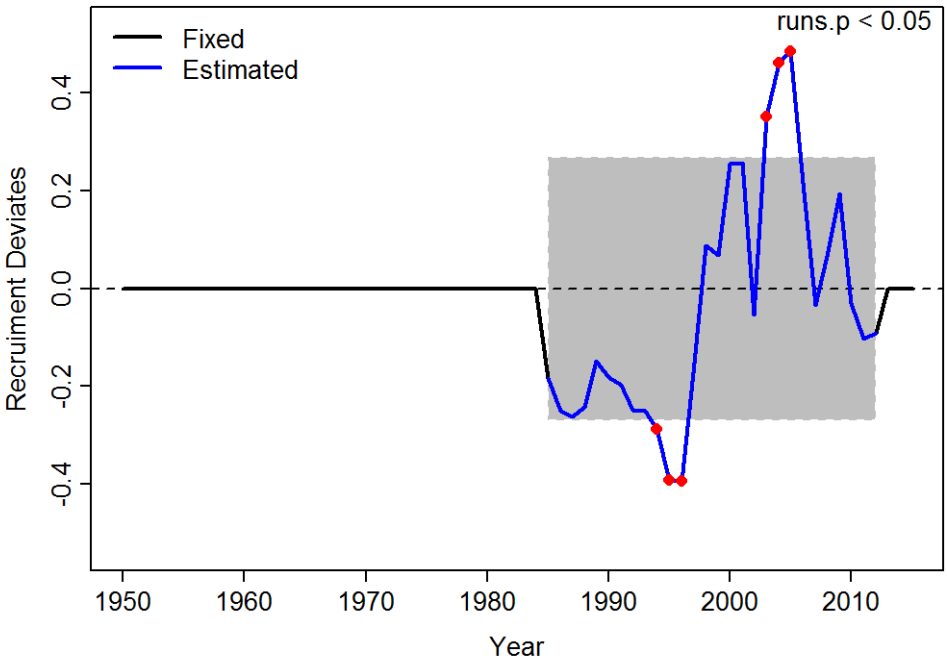


Figure 5. Continued.

A. Run 1

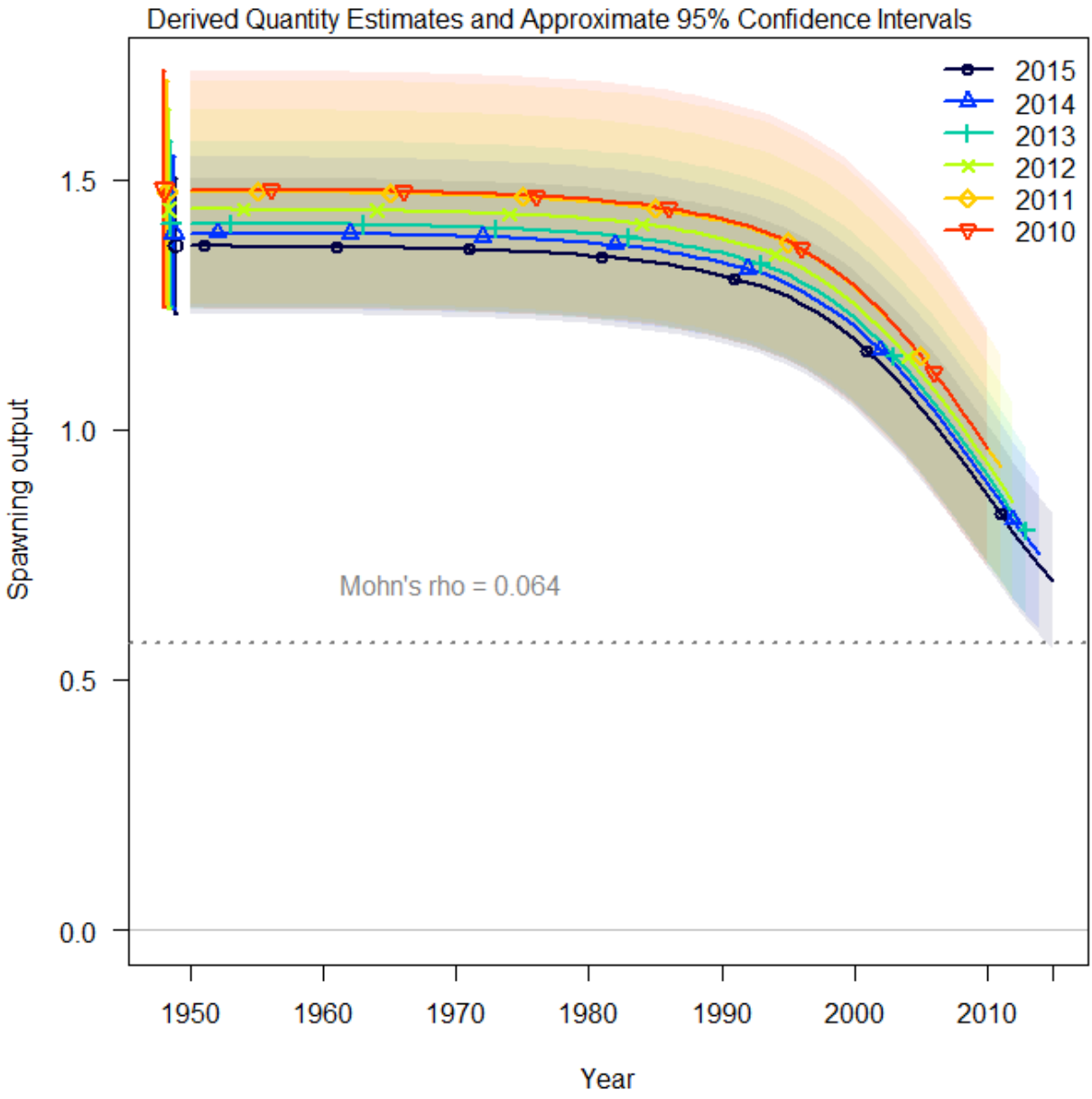


Figure 6. Retrospective analysis of spawning output for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C). Spawning output is spawning stock fecundity (SSF) in units of millions of pups produced. The stippled line is SSF at MSY. The Mohn's rho statistic is provided for the 5year peel.

B. Run 2

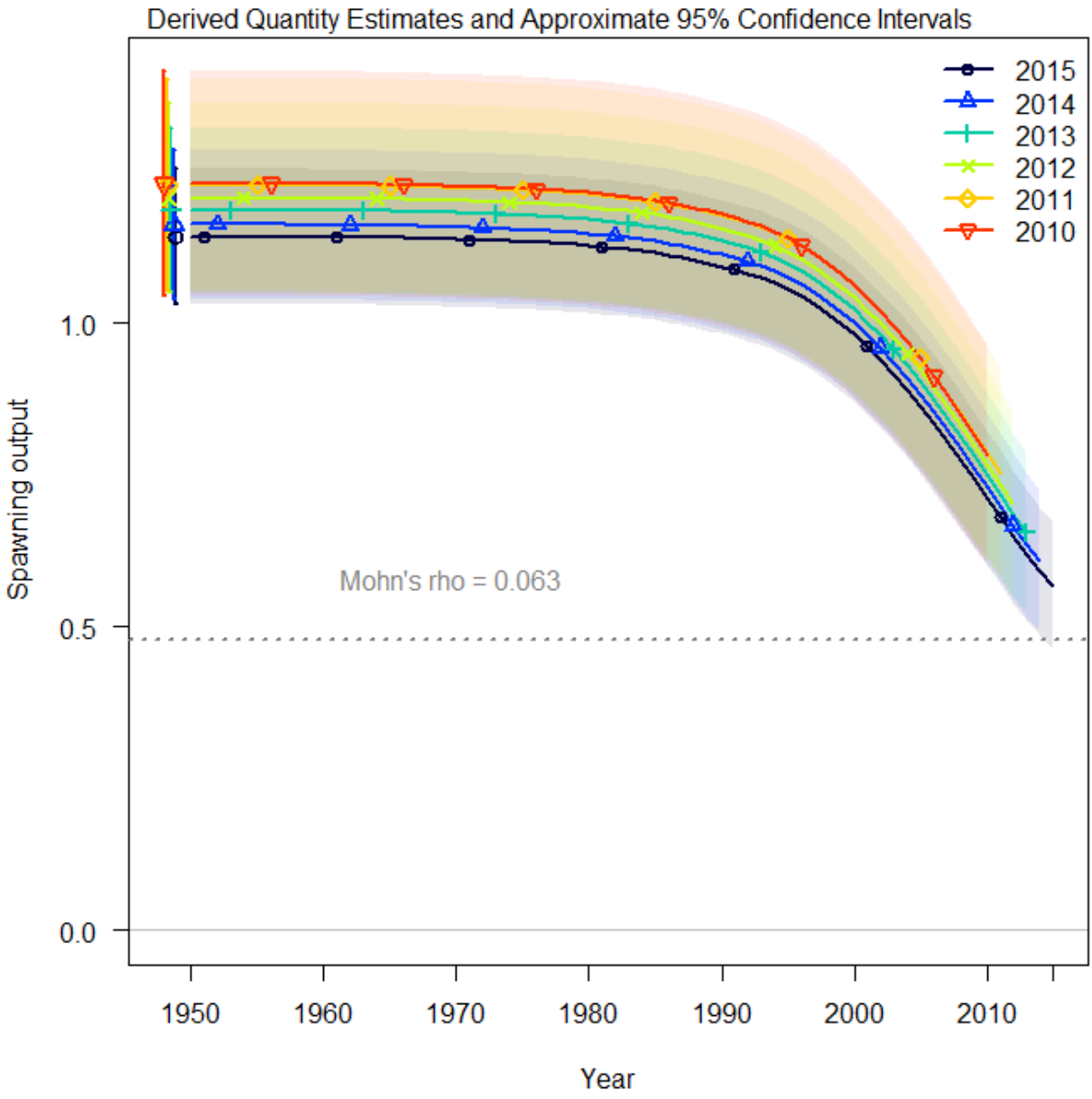


Figure 6. Continued.

C. Run 3

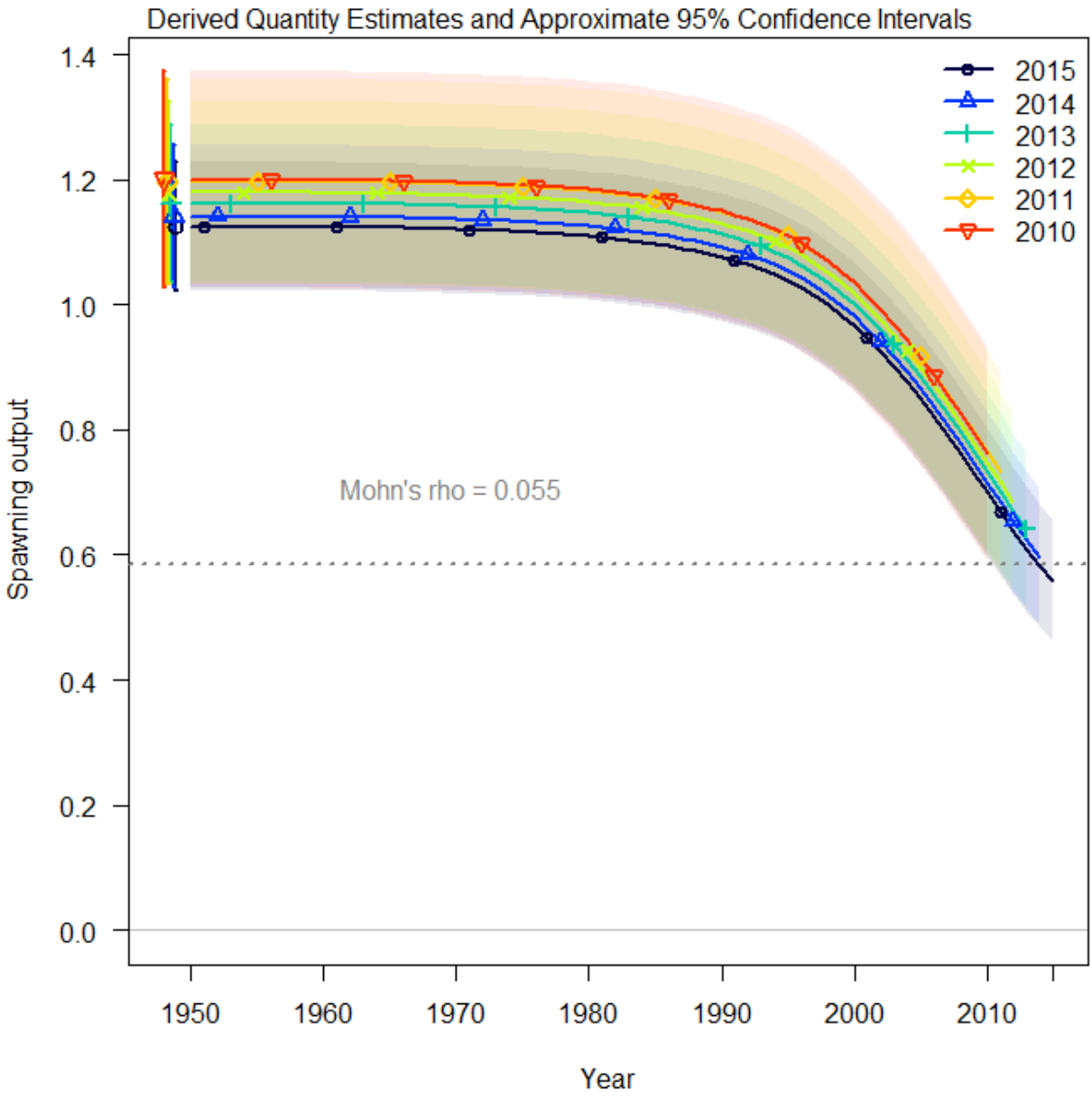


Figure 6. Continued.

A. Run 1

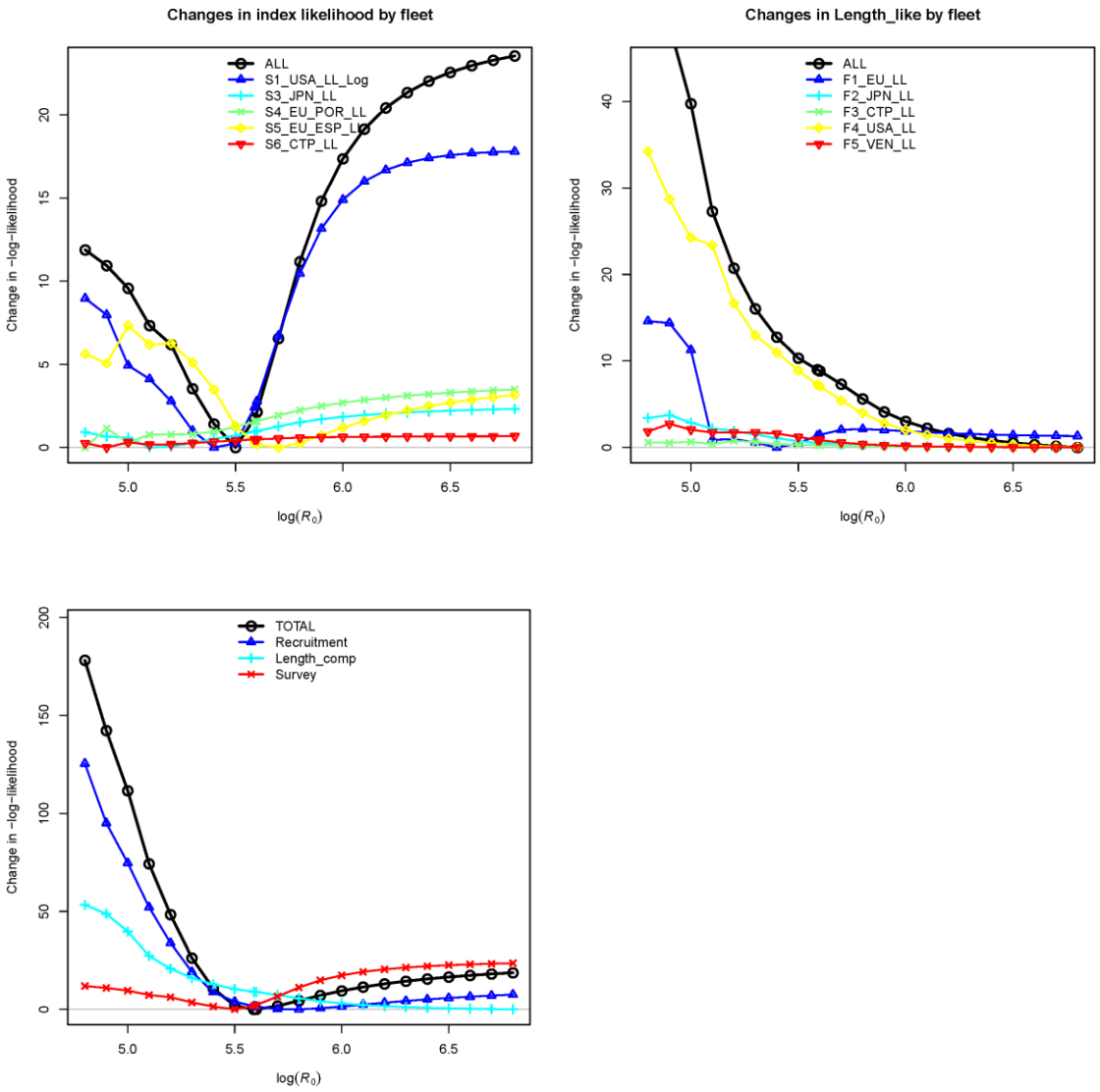


Figure 7. R_0 profiles for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C). Fleets as defined in **Table B.1.**

B. Run 2

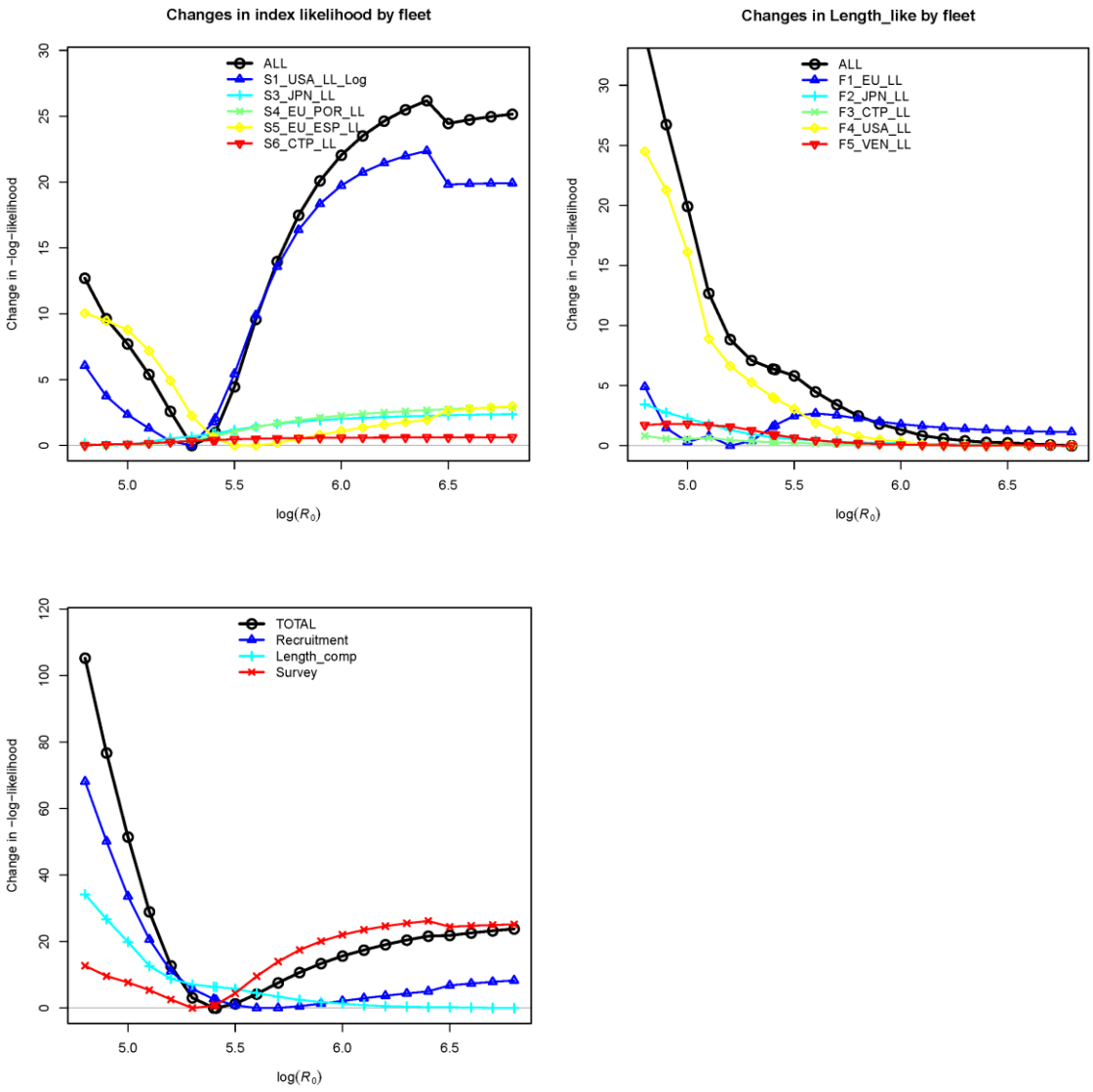


Figure 7. Continued.

C. Run 3

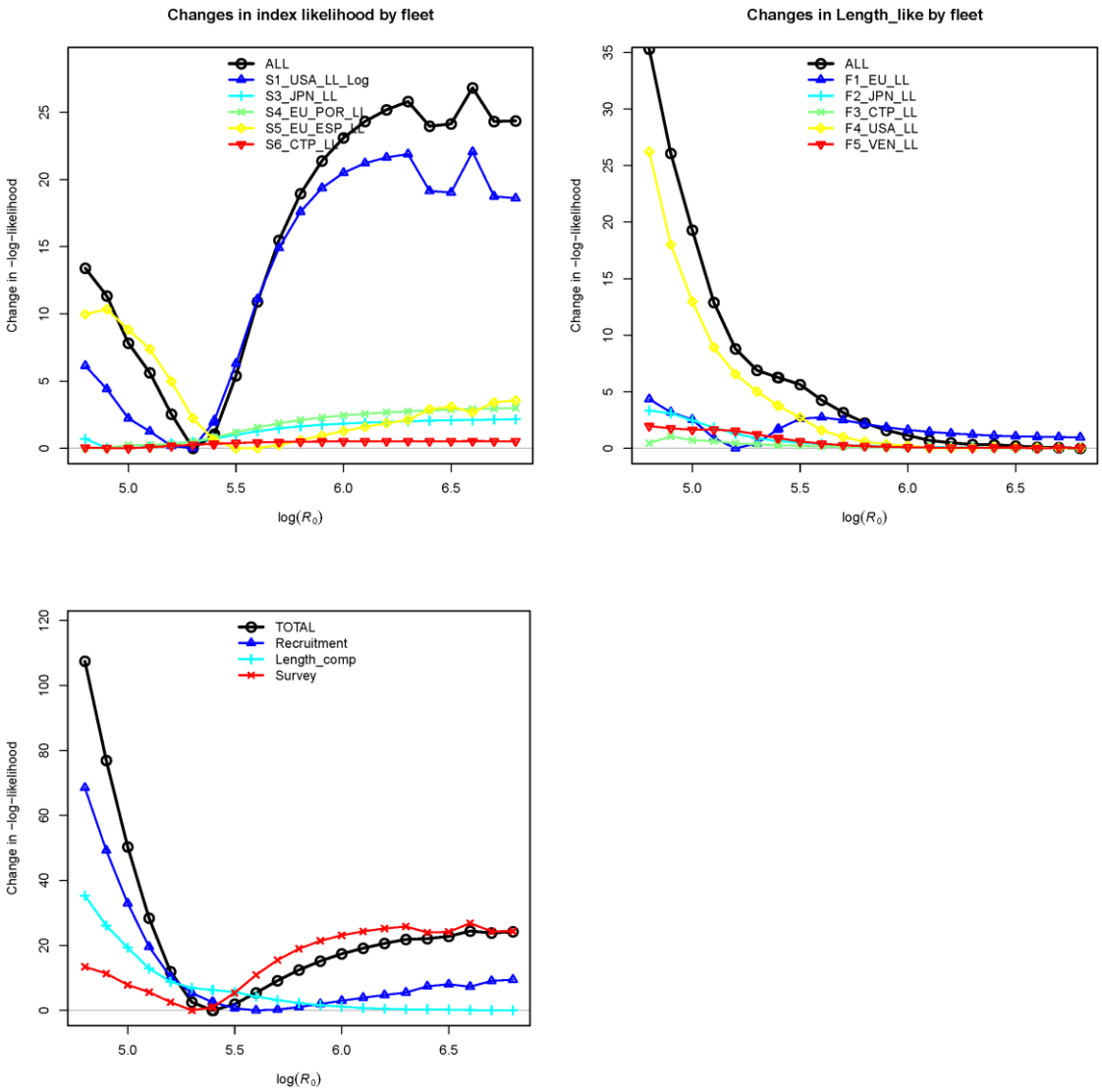


Figure 7. Continued.

A. Run 1

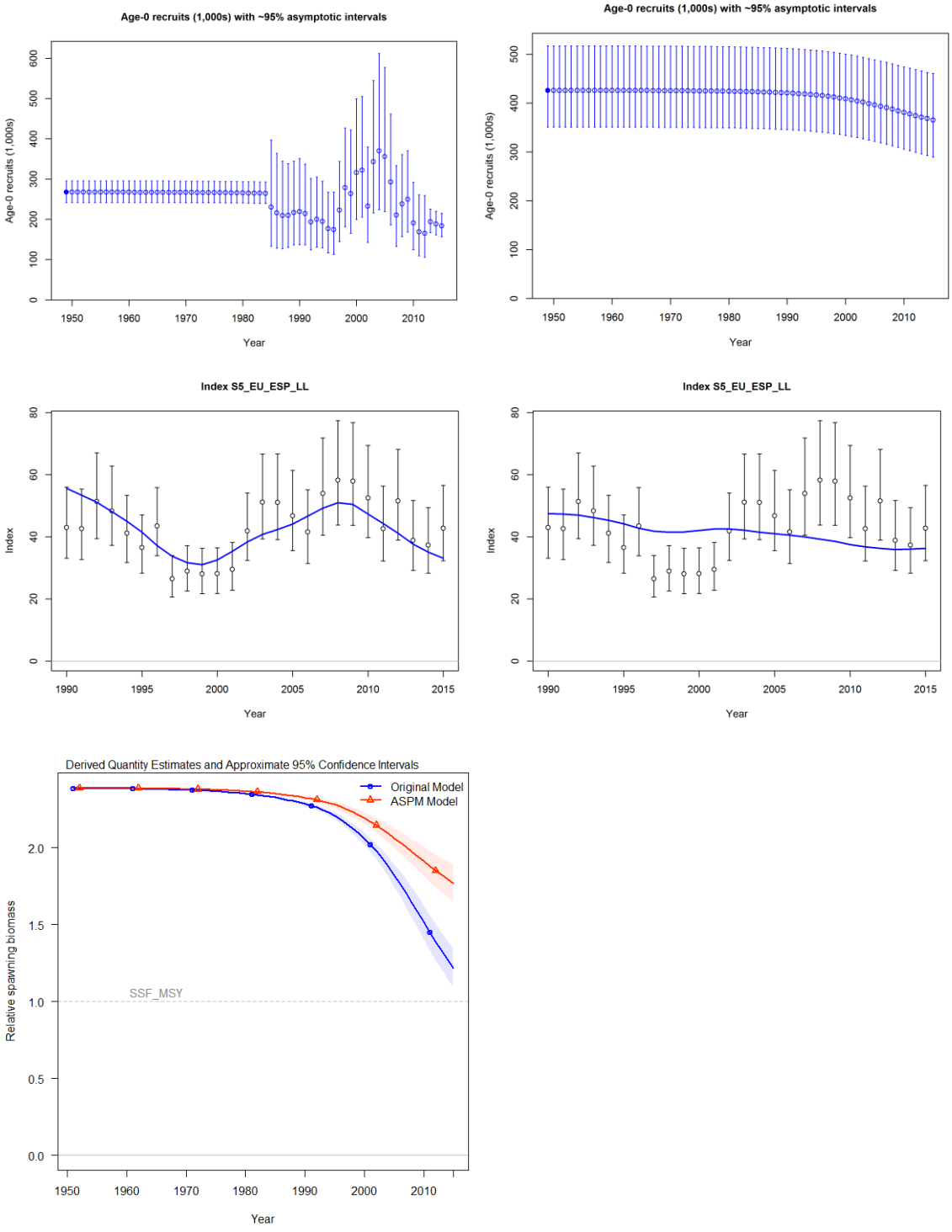


Figure 8. Age structured production model (ASPM) diagnostic for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C). Integrated stock assessment model age-0 recruits (upper left panel), ASPM model age-0 recruits (upper right panel), integrated stock assessment model fit to CPUE (EU ESP LL, middle left panel), ASPM model fit to CPUE (EU ESP LL, middle right panel), and relative spawning output (SSF/SSF_MSY, lower left panel) defined here as spawning stock fecundity (SSF), in units of pups produced per female, relative to SSF at MSY (SSF/SSF_MSY) from both the integrated stock assessment model and ASPM. Fleets as defined in **Table B.1.**

B. Run 2

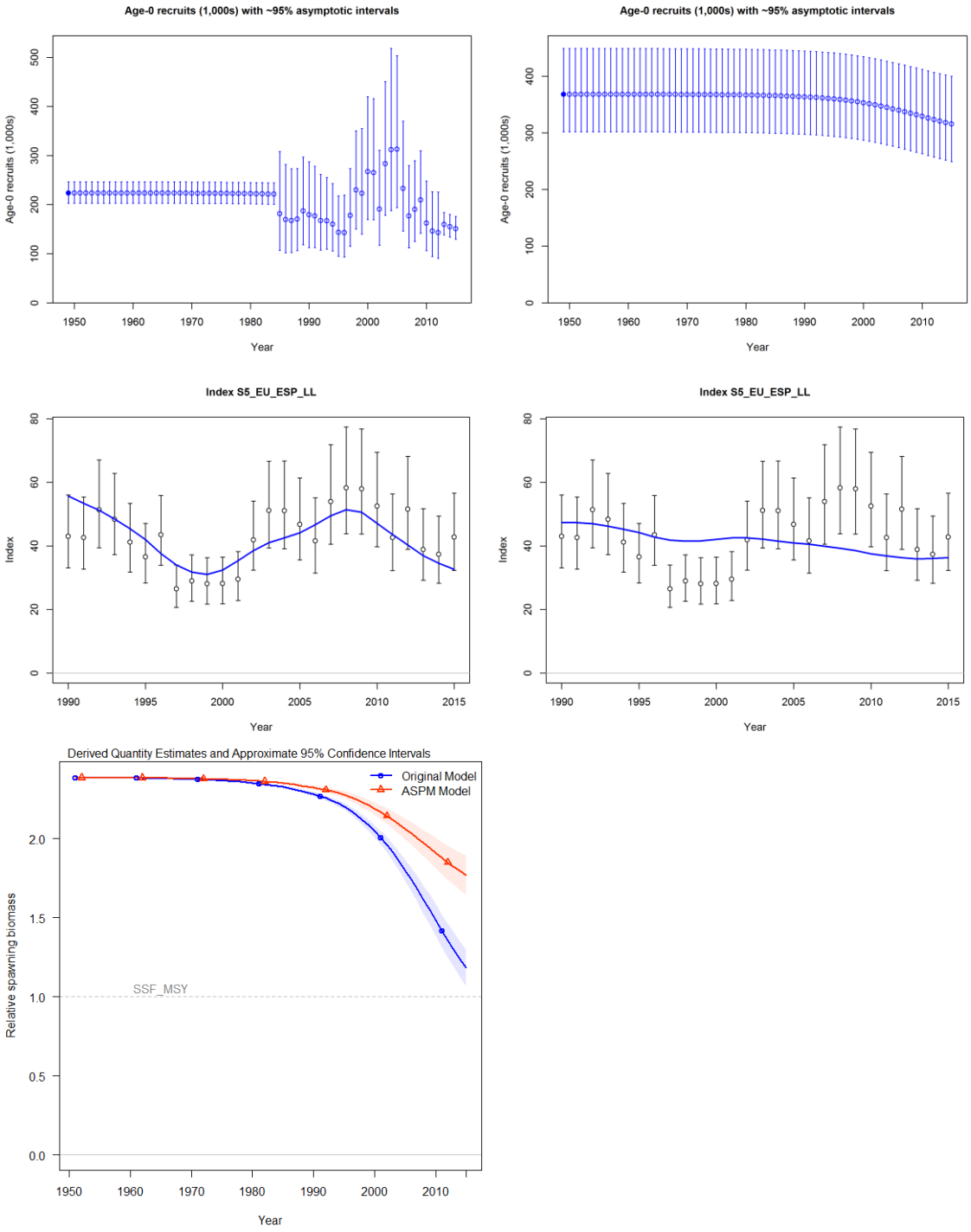


Figure 8. Continued.

C. Run 3

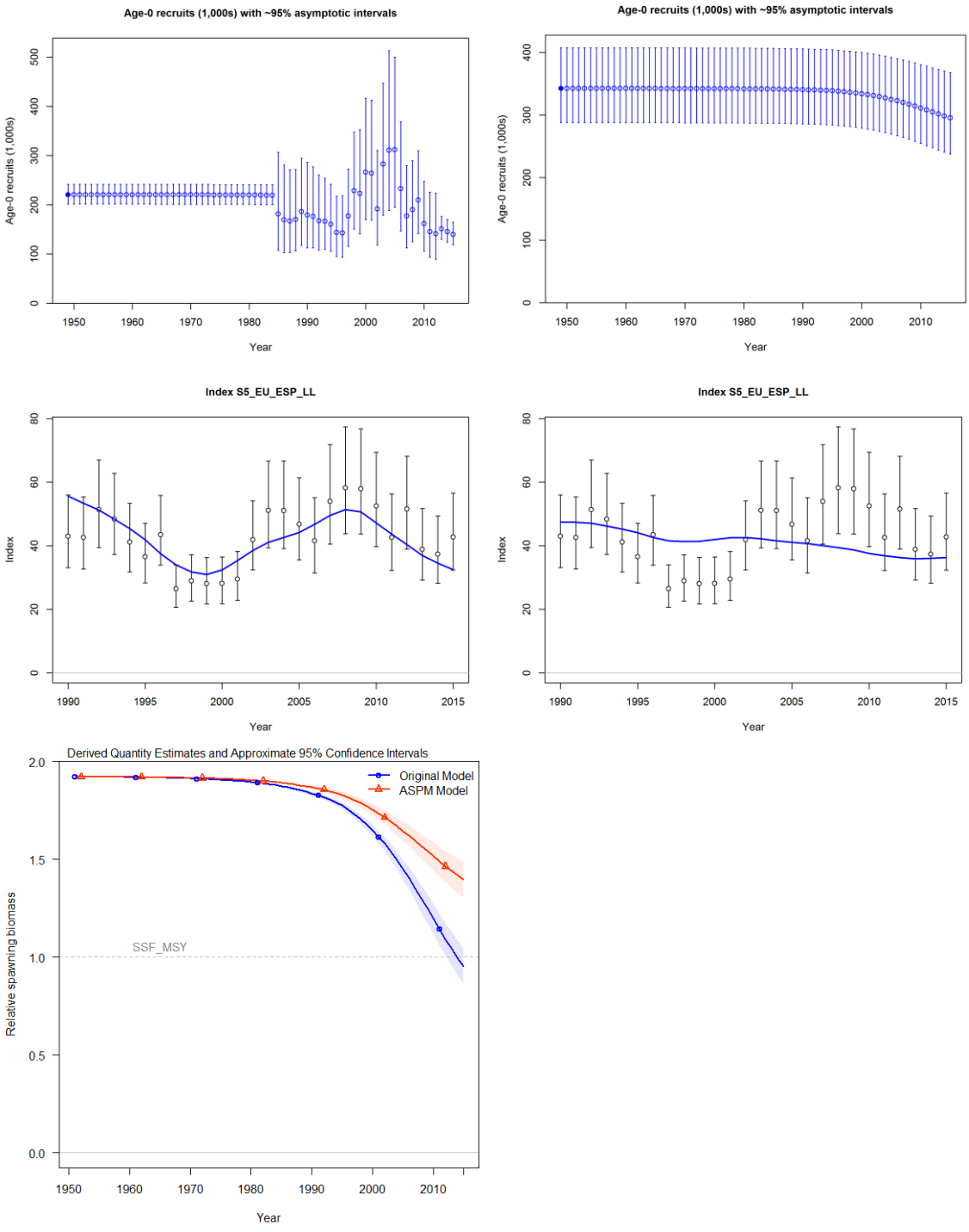


Figure 8. Continued.

A. Run 1

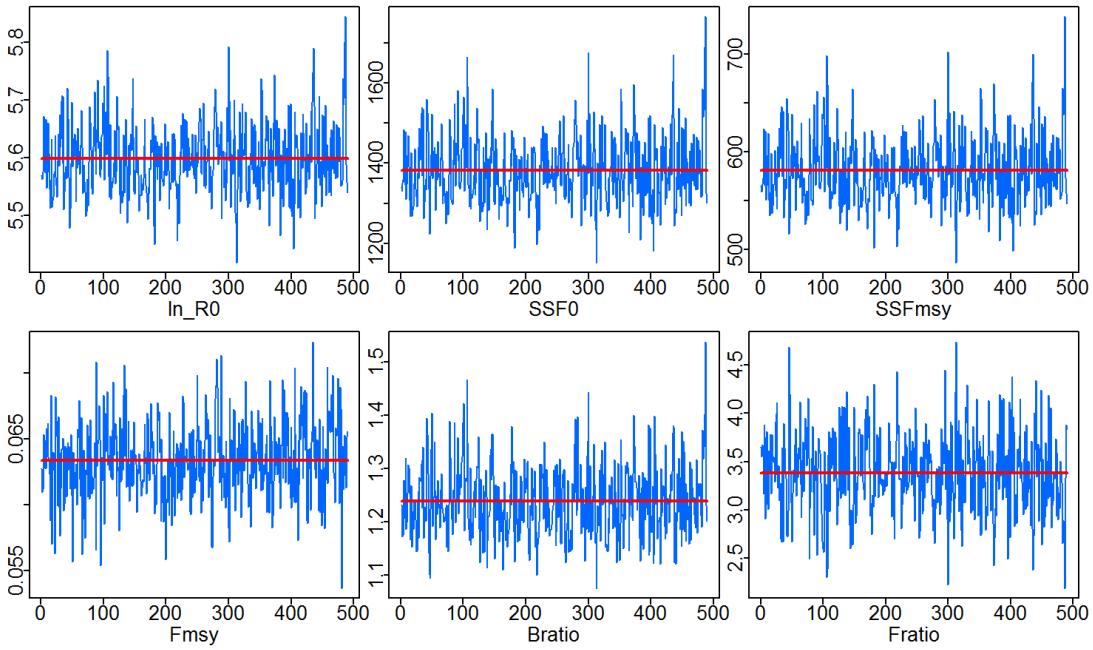


Figure 9. Short MCMC chains (as described in the methods) for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C) obtained for estimated parameters and derived quantities from Stock Synthesis output: 1) $\log(R_0)$; 2) unfished spawning stock fecundity (SSF_0); 3) spawning stock fecundity at MSY (SSF_{MSY}); 4) fishing mortality at MSY (F_{MSY}); 5) ending year spawning stock fecundity relative to its value at MSY (SSF/SSF_{MSY}); and 6) ending year fishing mortality relative to its value at MSY (F/F_{MSY}).

B. Run 2

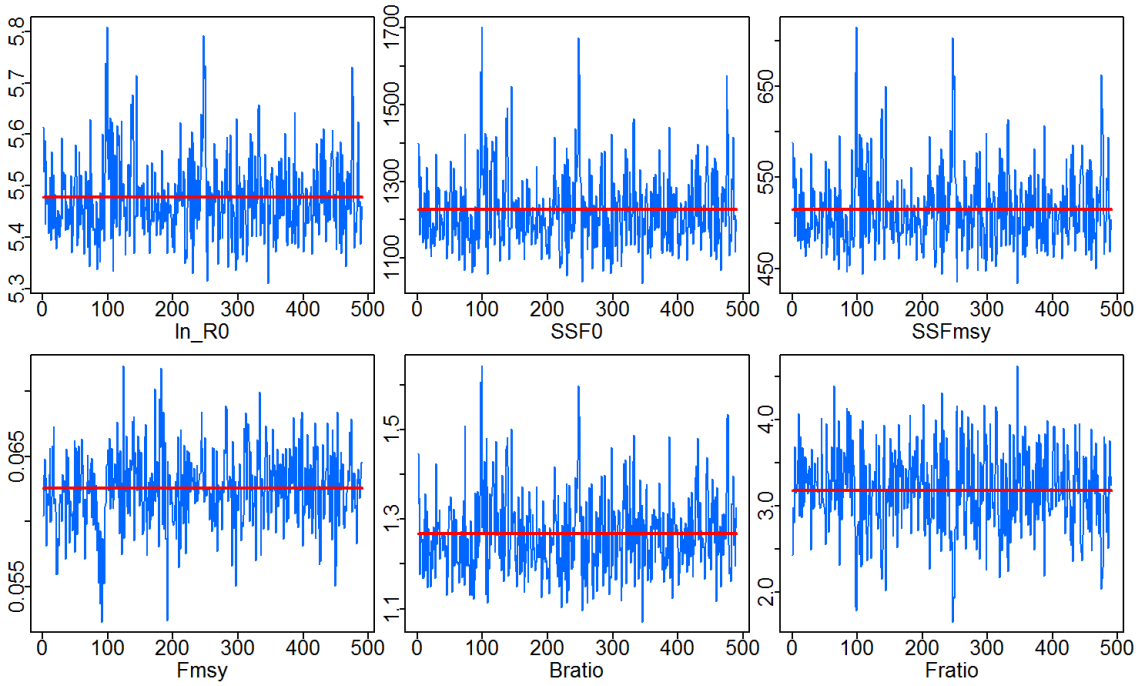


Figure 9. Continued.

C. Run 3

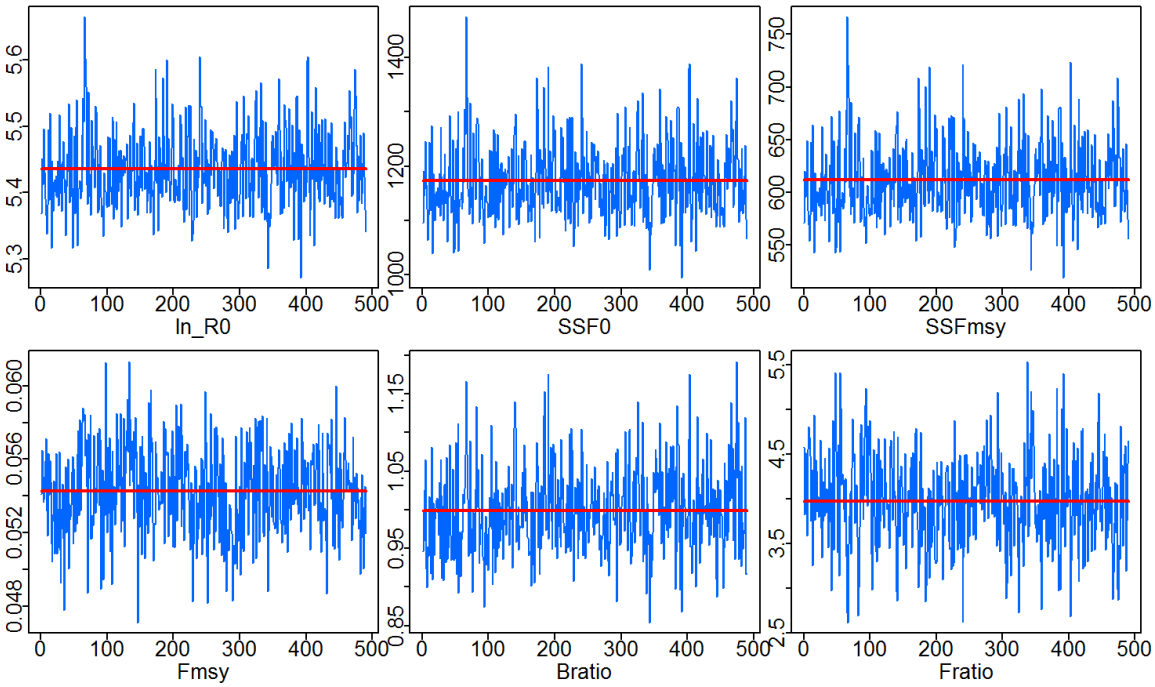


Figure 9. Continued.

A. Run 1

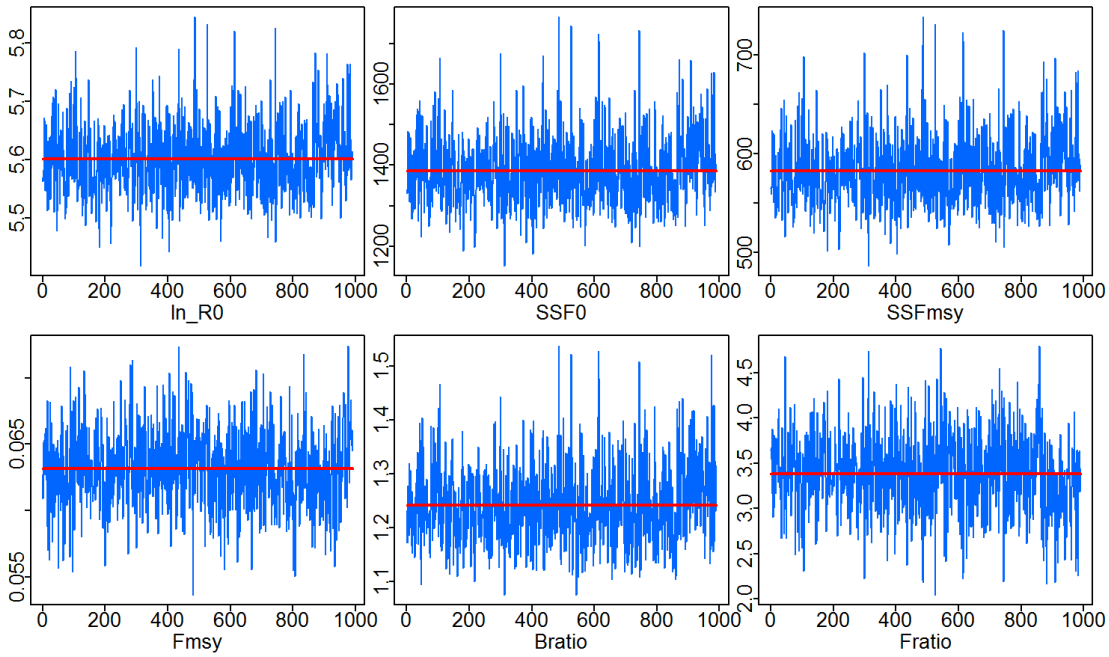


Figure 10. Long MCMC chains (as described in the methods) for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C) obtained for estimated parameters and derived quantities from Stock Synthesis output: 1) $\log(R_0)$; 2) unfished spawning stock fecundity (SSF_0); 3) spawning stock fecundity at MSY (SSF_{MSY}); 4) fishing mortality at MSY (F_{MSY}); 5) ending year spawning stock fecundity relative to its value at MSY (SSF/SSF_{MSY}); and 6) ending year fishing mortality relative to its value at MSY (F/F_{MSY}).

B. Run 2

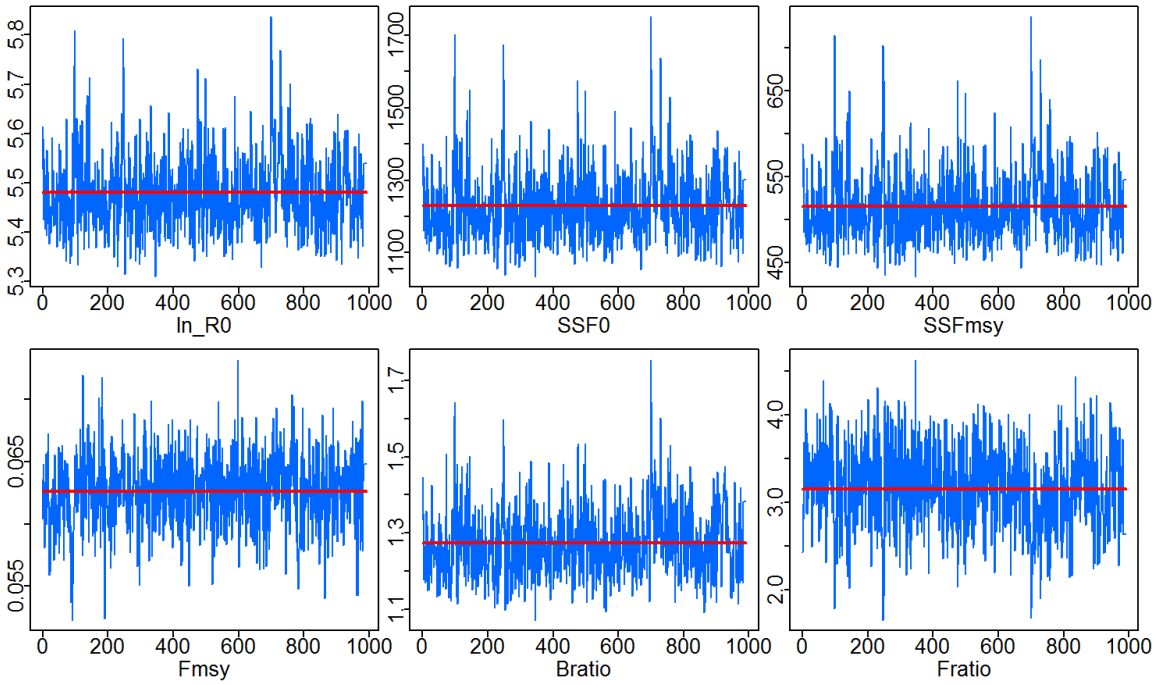


Figure 10. Continued.

C. Run 3

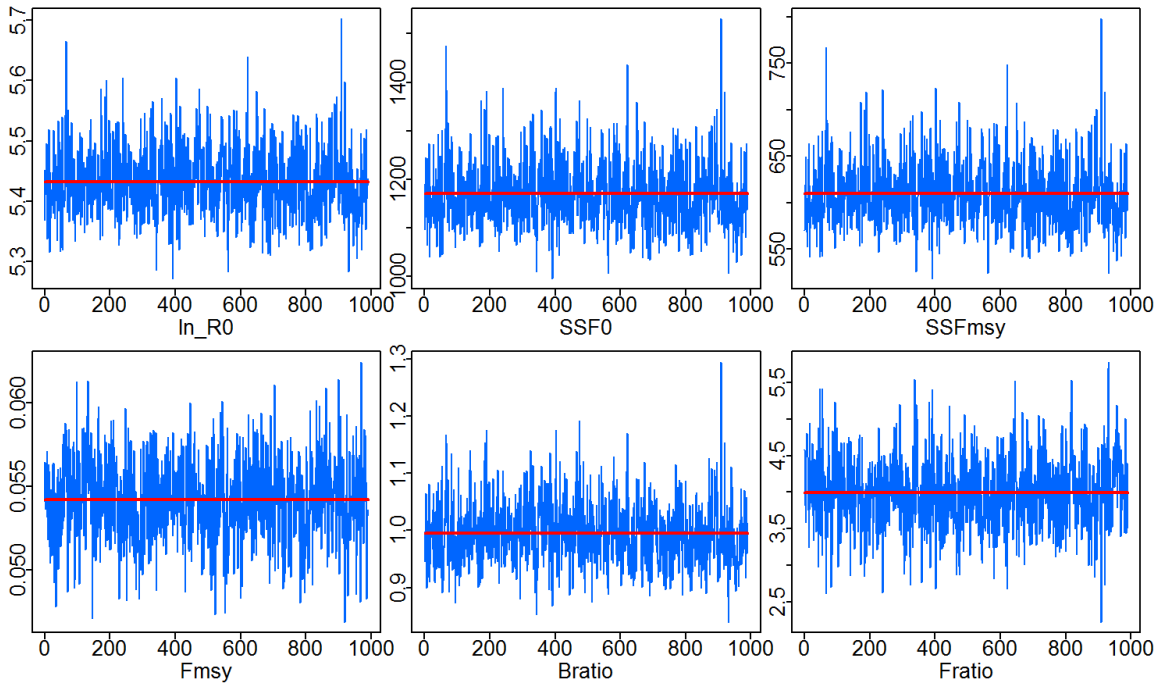


Figure 10. Continued.

A. Run 1

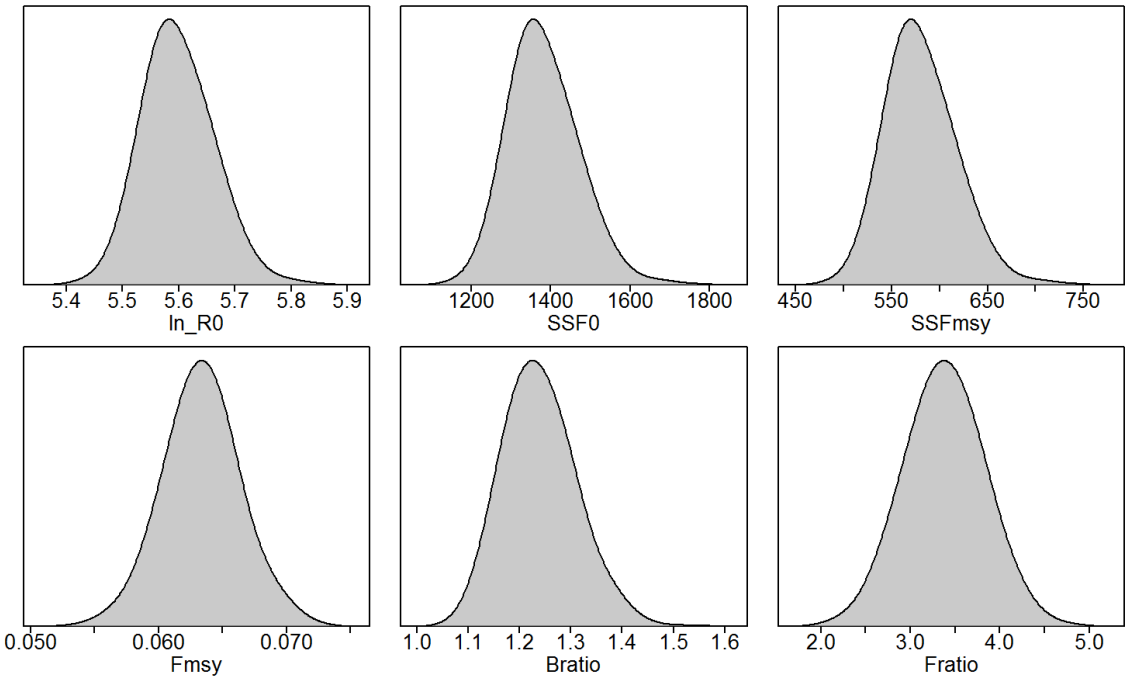


Figure 11. Short MCMC chain (as described in the methods) posterior distributions for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C) obtained for estimated parameters and derived quantities from Stock Synthesis output: 1) $\log(R_0)$; 2) unfished spawning stock fecundity (SSF_0); 3) spawning stock fecundity at MSY (SSF_{MSY}); 4) fishing mortality at MSY (F_{MSY}); 5) ending year spawning stock fecundity relative to its value at MSY (SSF/SSF_{MSY}); and 6) ending year fishing mortality relative to its value at MSY (F/F_{MSY}).

B. Run 2

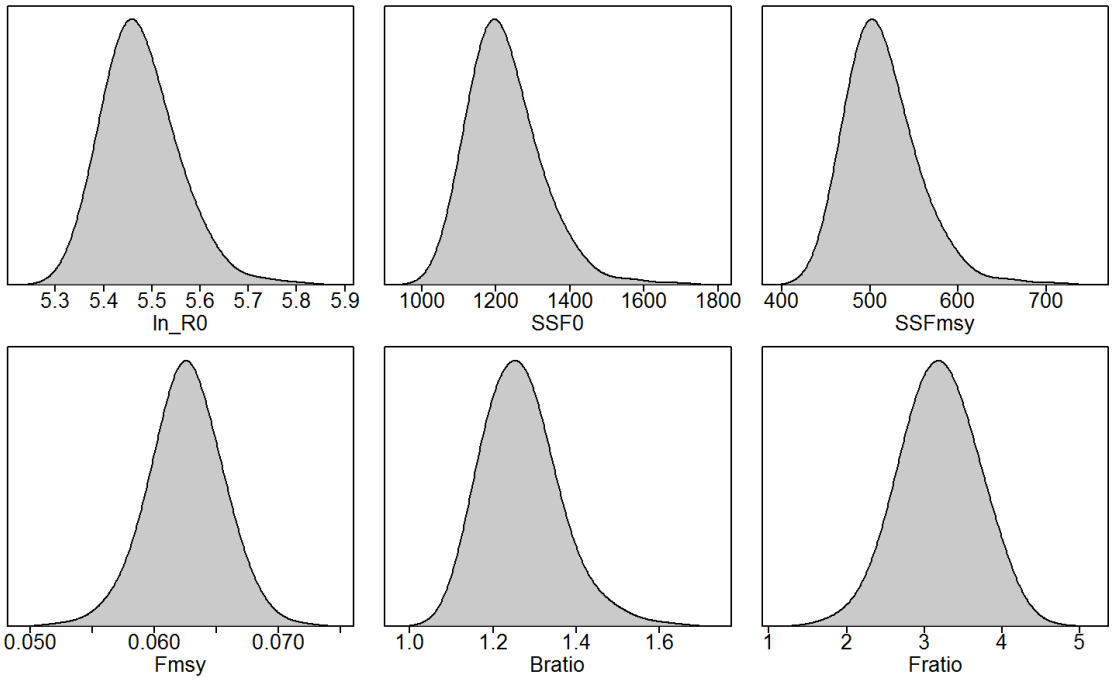


Figure 11. Continued.

C. Run 3

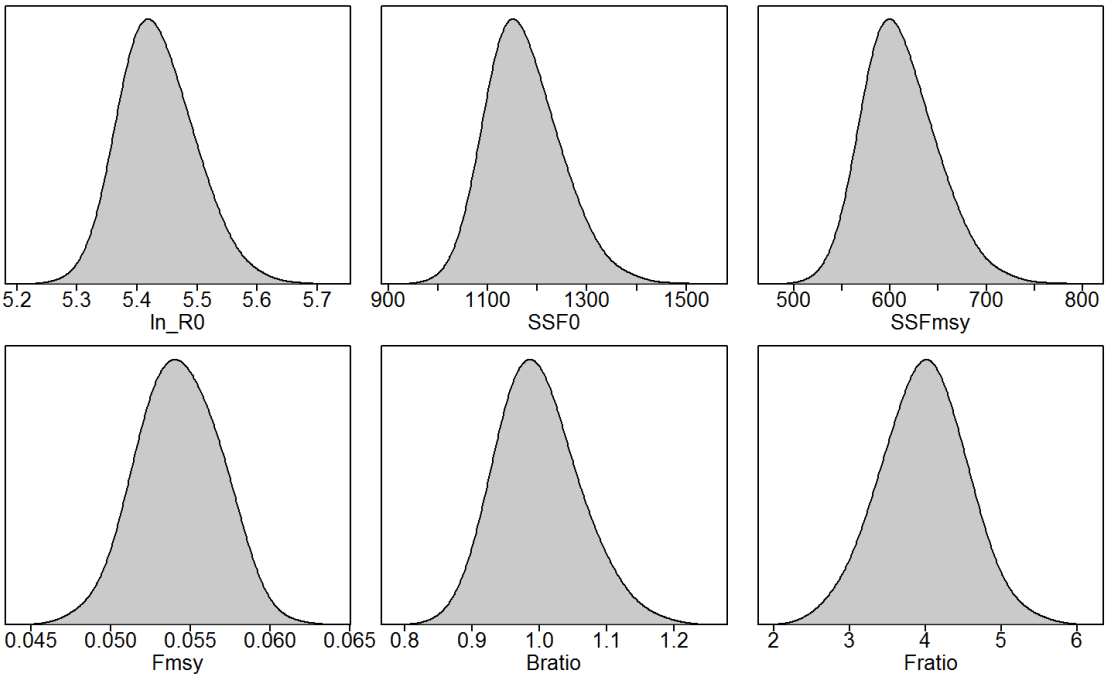


Figure 11. Continued.

A. Run 1

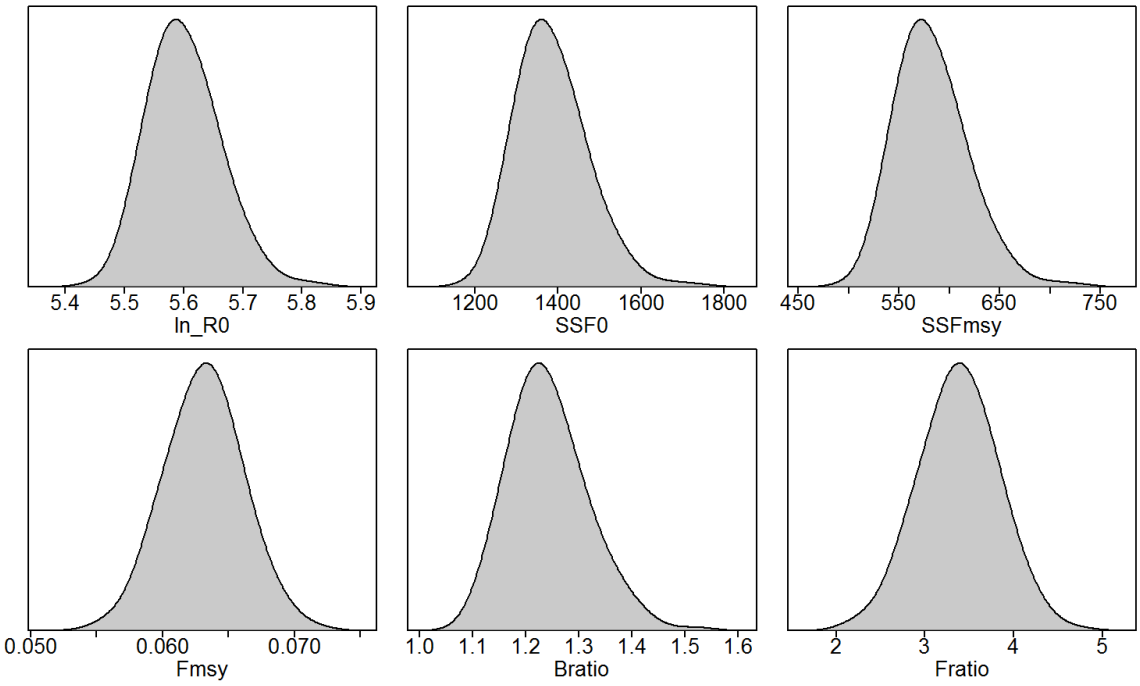


Figure 12. Long MCMC chain (as described in the methods) posterior distributions for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C) obtained for estimated parameters and derived quantities from Stock Synthesis output: 1) $\log(R_0)$; 2) unfished spawning stock fecundity (SSF_0); 3) spawning stock fecundity at MSY (SSF_{MSY}); 4) fishing mortality at MSY (F_{MSY}); 5) ending year spawning stock fecundity relative to its value at MSY (SSF/SSF_{MSY}); and 6) ending year fishing mortality relative to its value at MSY (F/F_{MSY}).

B. Run 2

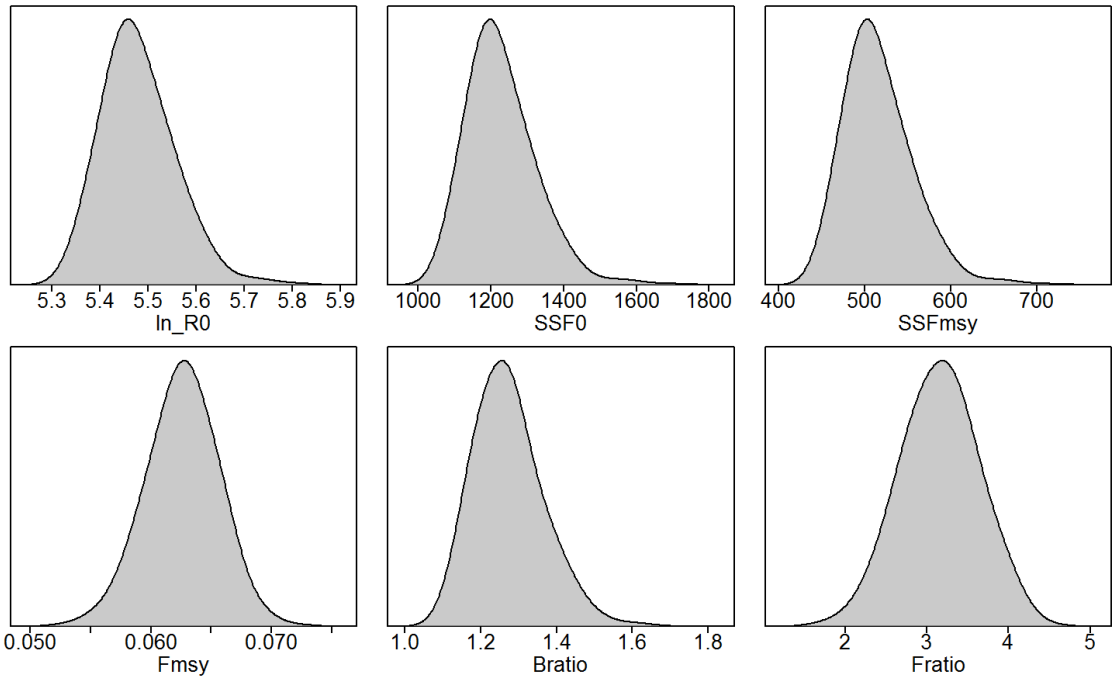


Figure 12. Continued.

C. Run 3

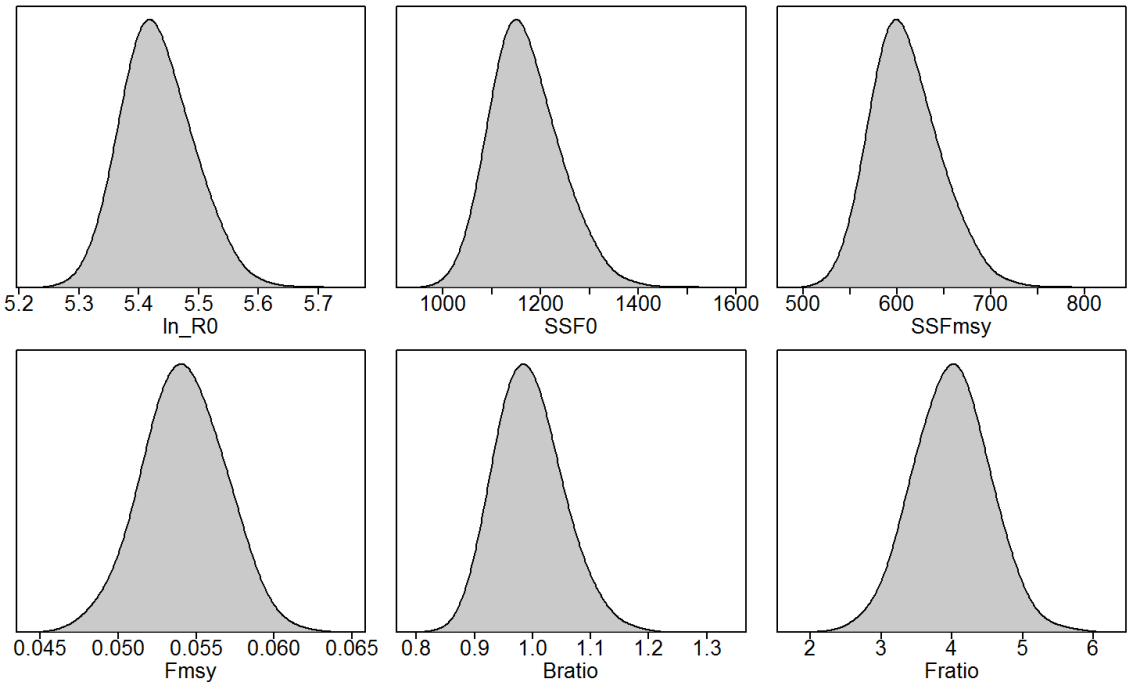


Figure 12. Continued.

A. MCMC short chain

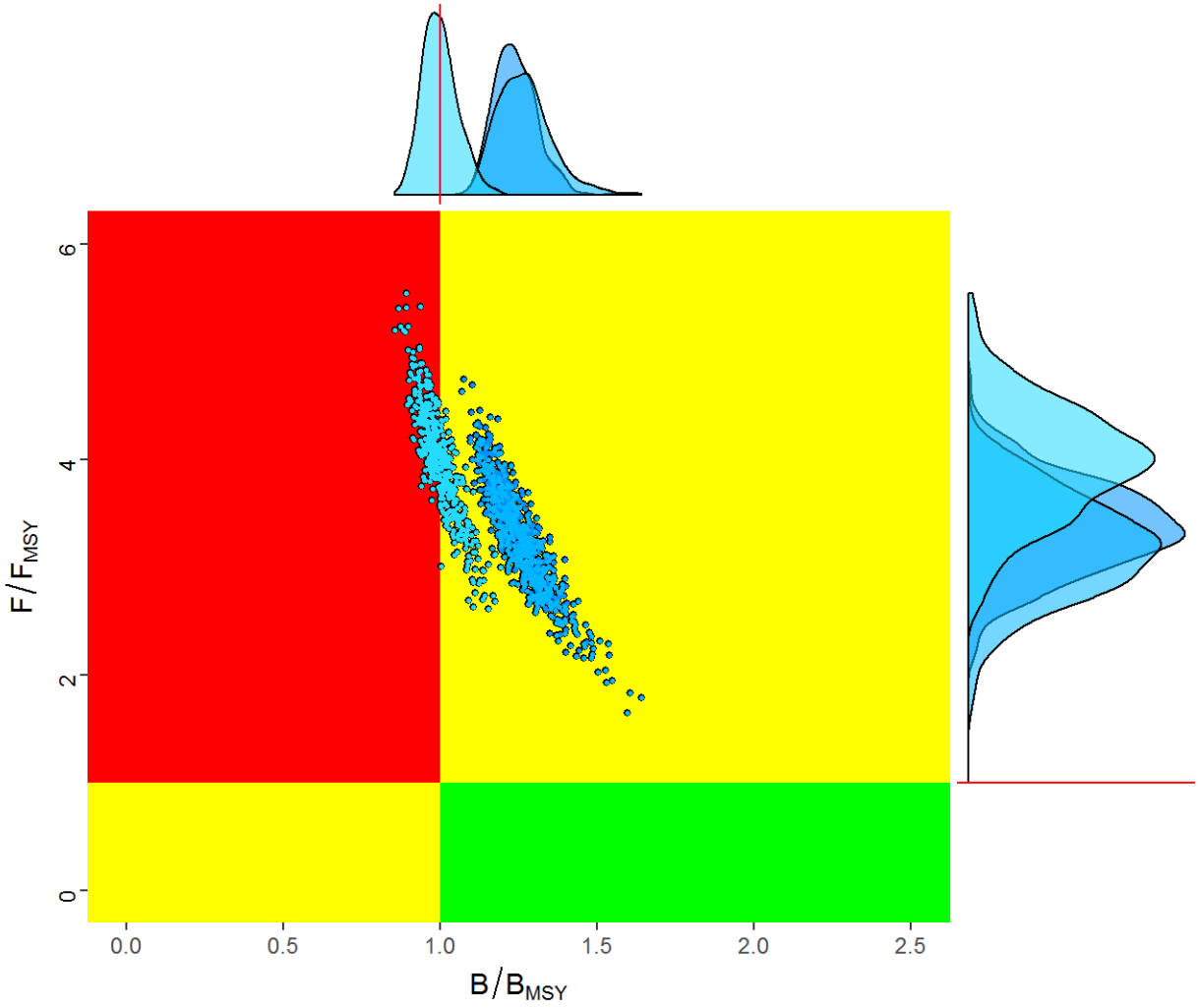


Figure 13. Kobe plot of MCMC short chain (Panel A) and long chain (Panel B) for Stock Synthesis model runs 1, 2, and 3. Runs 1 and 2 overlapped and differed from Run 3 (MCMC chain length as described in the methods).

B. MCMC long chain

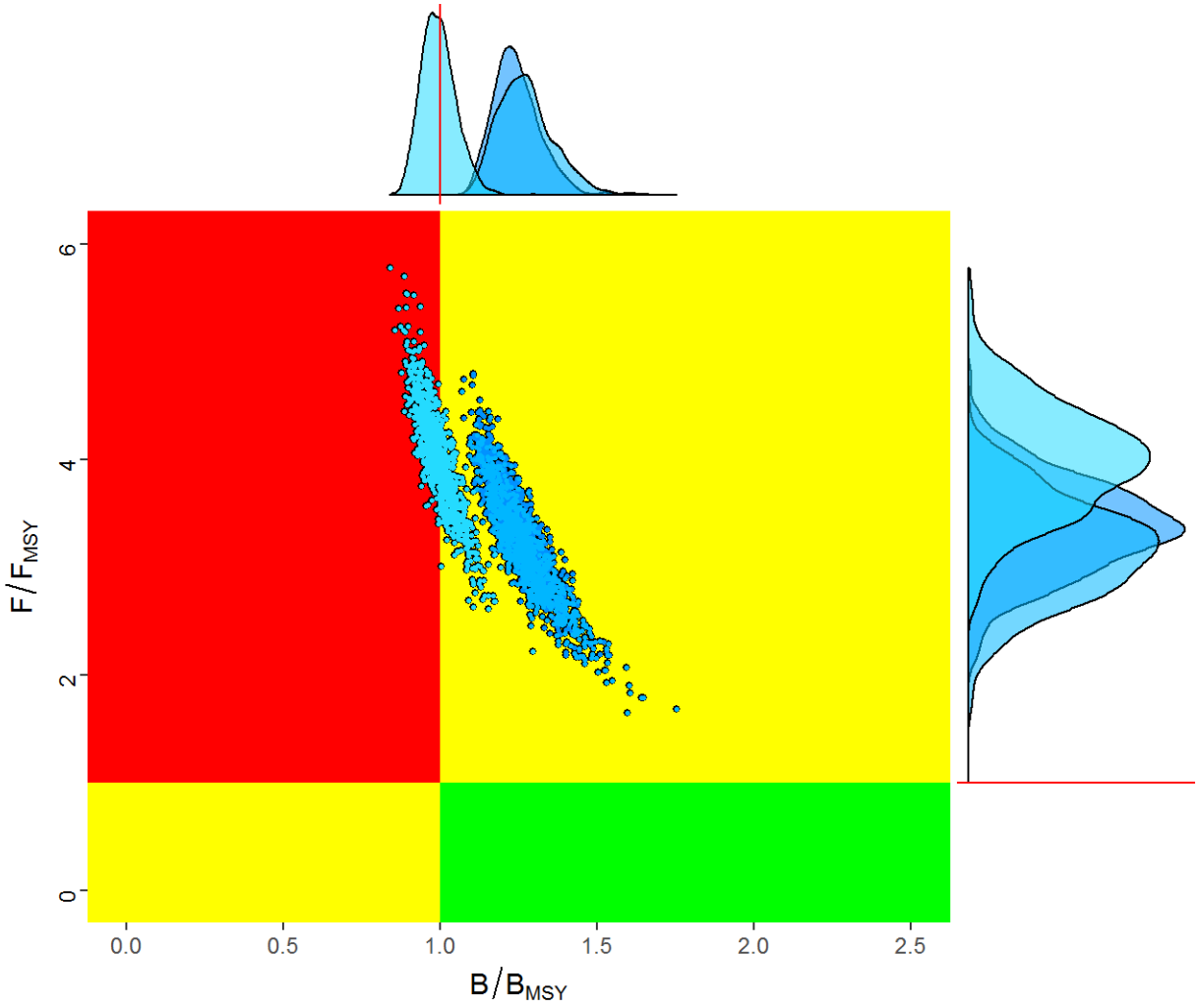


Figure 13. Continued.

Appendix A.

Table A.1. Software (and versions) used for diagnostics.

Diagnostic	R[1] and FLR[2]	Pseudo-code	r4ss[3]	Stock Synthesis[4]
Diagnostic-1	R (version 3.3.3)		version 1.24.0	(version 3.24U)
Diagnostic-2	R (version 3.3.3)		version 1.24.0	(version 3.24U)
Diagnostic-3	R (version 3.3.3)		version 1.24.0	(version 3.24U)
Diagnostic-4	R (version 3.3.3)		version 1.24.0	(version 3.24U)
Diagnostic-5	R (version 3.4.4)		version 1.30.2	(version 3.24U)
Diagnostic-6	R (version 3.4.4)		version 1.30.2	(version 3.24U)
Diagnostic-7		Manually implemented (see **Table A.2)**. Summarized in R (version 3.4.4)	version 1.30.2	(version 3.24U)
Diagnostic-8	FLR			
Diagnostic-9	FLR			

[1] R (R Core Team 2018). Available: https://www.R-project.org (May 2019).
[2] Fisheries Library in R (FLR; Kell *et al.* 2007). Available: http://www.flr-project.org (May 2019).
[3] R code for Stock Synthesis (r4ss; Taylor *et al.* 2018). Available: https://github.com/r4ss/r4ss (May 2019).
[4] Stock Synthesis (Methot and Wetzel 2013). Available: https://vlab.ncep.noaa.gov/web/stock-synthesis/home (May 2019).

Table A.2. *Diagnostic-7* Pseudo-code for ASPM as implemented here (Stock Synthesis v324U).

- Copy the necessary files from the full model run at the converged solution and rename
    - control.ss_new, data.ss_new, forecast.ss_new, starter.ss_new, ss.exe, ss.par
    - Rename files from *.ss_new to *.ss
- Set all estimated recruitment deviations equal to zero
    - Edit ss3.par; set rec-devs equal to 0.
- Read initial estimated parameter values directly from the edited ss3.par file
    - Edit starter.ss; read from par file.
- Fix all selectivity parameters to their estimated values
    - Formatted selectivity values at the converged solution are in control.ss_new.
    - Rename control.ss_new file to the control.ss file,
    - Edit control.ss; set selectivity parameter estimation phase to a negative value.
- Turn off length comp data likelihood components
    - Edit control.ss; set length comp likelihood lambda(s) to 0.
- Set recruitment equal to that obtained from the stock recruitment curve
    - Edit control.ss; set recruitment parameter estimation phase to negative value.
    - Edit control.ss; set recruitment likelihood lambda to 0.
- Run the full model
- Run the ASPM
- Compare model results in R

An alternative form of the APSM, as described in the discussion, can be obtained by setting all estimated recruitment deviations equal to their estimated values in the ASPM.

Table A.3. Example of Stock Synthesis convergence diagnostics presented to the Shark Working Group during a previous North Atlantic blue shark stock assessment (Courtney 2016, their section 2.3.3).

- Model convergence was based on whether or not the Hessian inverted (i.e., the matrix of second derivatives of the likelihood with respect to the parameters, from which the asymptotic standard error of the parameter estimates is derived).
- Other convergence diagnostics were also evaluated.
    - Excessive CVs on estimated quantities (>> 50%) or a large final gradient (>1.00E-05) were indicative of uncertainty in parameter estimates or assumed model structure.
    - The correlation matrix was also examined for highly correlated (> 0.95) and non-informative (< 0.01) parameters.
    - Parameters estimated at a bound were a diagnostic for possible problems with data or the assumed model structure.

Appendix B.

Table B.1. Time series of catch, relative abundance, and length composition data considered for use in the North Atlantic shortfin mako SS3 model runs (adapted from Courtney *et al.* 2017a, their Table 1)

Time series	Symbol	Catch (t) and abundance (numbers or biomass)	Name	Definition	Length composition (10 cm FL bins)
1	F1	Catch (t)	EU LL	EU España + Portugal Longline (1950-2015)	EU España + Portugal LL (1997-2015)
2	F2	Catch (t)	JPN LL	Japan Longline(1971-2015)	Japan LL (1997-2015)
3	F3	Catch (t)	CTP LL	Chinese Taipei Longline (1981-2015) [1]	Chinese Taipei LL (2004-2015)
4	F4	Catch (t)	USA LL	USA Longline (1982-2015)	USA LL (1992-2015)
5	F5	Catch (t)	VEN LL	Venezuela Longline (1986-2015)	Venezuela LL (1994-2013)
6	F6	Catch (t)	CAN LL	Canada Longline (1995-2015)	Mirror USA LL (F4)
7	F7	Catch (t)	MOR LL	Morocco Longline (1961-2015) [1]	Mirror EU LL (F1)
8	F8	Catch (t)	USA RR	USA Recreational (1981-2015)	Mirror USA LL (F4)
9	F9	Catch (t)	BEL LL	Belize Longline (2009-2015)	Mirror VEN LL (F5)
10	F10	Catch (t)	MOR PS	Morocco Purse Seine (2011-2015)	Mirror EU LL (F1)
11	F11	Catch (t)	CPR LL	China PR Longline (2000-2015)	Mirror CTP LL (F3)
12	F12	Catch (t)	OTH	Other (1982-2015)	Mirror CTP LL (F3)
13	S1	Relative abundance (numbers)	USA LL Log	USA Longline-Logbook (1986-2015)	Mirror USA (F4)
14	S2	Relative abundance (numbers)	USA LL Obs	USA Longline-Observer (1992-2015) [2]	Mirror USA (F4)
15	S3	Relative abundance (numbers)	JPN LL	Japan Longline (1994-2015)	Mirror JPN (F2)
16	S4	Relative abundance (biomass)	EU POR LL	EU Portugal Longline (1999-2015)	Mirror EU (F1)
17	S5	Relative abundance (biomass)	EU ESP LL	EU España Longline (1990-2015) [3]	Mirror EU (F1)
18	S6	Relative abundance (numbers)	CTP LL	Chinese Taipei Longline (2007-2015)	Mirror CTP (F3)

[1] Not ICCAT Task I - Finalized catch data for this assessment was obtained from the 2017 Shortfin Mako Data Preparatory meeting.
[2] Index S2 (USA LL Obs) was not fit in the model likelihood (lambda = 0) because of high variability in the index and because S2 describes the same fishery as S1 (USA LL Log).
[3] Index S5 was obtained in weight.

Appendix C. Prediction Skill

The provision of fisheries management advice requires the assessment of stock status relative to reference points, the prediction of the response of a stock to management, and checking that predictions are consistent with reality. To evaluate uncertainty, often a number of scenarios are considered corresponding to alternative model structures and dataset choices (Hilborn 2016). It is difficult, however, to empirically validate the various stock assessment models as it is seldom possible to observe fish populations directly. Stock assessments, however, are sometimes proven to be wrong in retrospect, due to poor model assumptions or to data that do not reflect the key processes (Schnute and Hilborn 1993). Therefore techniques such as retrospective analysis, where a model is fitted to increasing periods of data, provide a diagnostic to identify systematic inconsistencies (Mohn 1999).

A key concept to understand therefore in the evaluation of prediction is the concept of *skill*. A prediction is said to have skill if it improves upon a naive baseline. A naive baseline is the predictive performance that you could achieve without really having any expertise in the subject. For instance, in weather forecasting a naive forecast could be the weather tomorrow will be the same as today. Prediction skill is a statistical measure of the accuracy of a forecast compared to an observation or estimate of the actual value of what is being predicted (Huschke 1959), and can be used to compare alternative models or observations to a reference set of estimates or data (e.g., Jin *et al.* 2008; Weigel *et al.* 2008; Balmaseda *et al.* 1995).

If data are regarded as being representative of the dynamics of the stock then they can be used as a model-free validation measure (Hjorth 1993), and the best performing scenarios (e.g., choice of models and data) can be identified by comparing predictions with observations. In contrast quantities such as stock biomass and fishing mortality are model estimates not data and so cannot actually be observed, so if estimates from a stock assessment model are compared this is model-based validation.

Various criteria are available for estimating prediction skill (see Hyndman and Koehler 2006). One commonly used measure is root-mean-square error (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

RMSE, however, is an inappropriate and misinterpreted measure of average error (Willmott and Matsuura 2005) because it is a function of three characteristics of a set of errors, as it varies with the variability within the distribution of error magnitudes, the square root of the number of errors ($n^{1/2}$), as well as with the average-error magnitude (MAE).

MAE is a more natural measure of average error and unlike RMSE is unambiguous.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

Scaling the average errors using the Mean Absolute Scaled Error (MASE) allows forecast accuracy to be compared across series on different scales.

$$MASE = \frac{\sum_{i=1}^{n} |e_i|}{\frac{n}{n-1}\sum_{i=2}^{n} |Y_i - Y_{i-1}|}$$

A scaled error is less than one if it arises from a better forecast than the average one-step naïve forecast. Where a "naïve" forecast is where the forecast is equal to the last observation, i.e. a random walk. Values greater than one indicate that one-step forecasts from the naïve method perform better than the forecast values under consideration. MASE also penalizes positive and negative errors and errors in large forecasts and small forecasts equally.

Prediction skill of the assessment was evaluated using a hindcast (Kell *et al.* 2016), where an assessment model is retrospectively re-run by tail cutting, i.e. removing recent years' data and the biomass trajectories projected up to the most recent year.

To conduct model-free validation the abundance indices in the recent years were removed and their predicted values calculated by multiplication of catchability and vulnerable biomass. **Figure C.1** shows the hindcast for Stock Synthesis model runs 1, 2 and 3; the red points are the observed CPUE values, and the black points the one-step ahead predicted values from each hindcast, ran for each of up to 15 years.

For all runs with fleets as defined in **Table B.1**, the CPUE index EU_ESP_LL, and JPN_LL MASE scores were greater than 1. This indicates that a random walk is a better predictor than these indices. The inference is that these indices are simply adding noise rather than information on stock trends in these runs.

A. Run 1

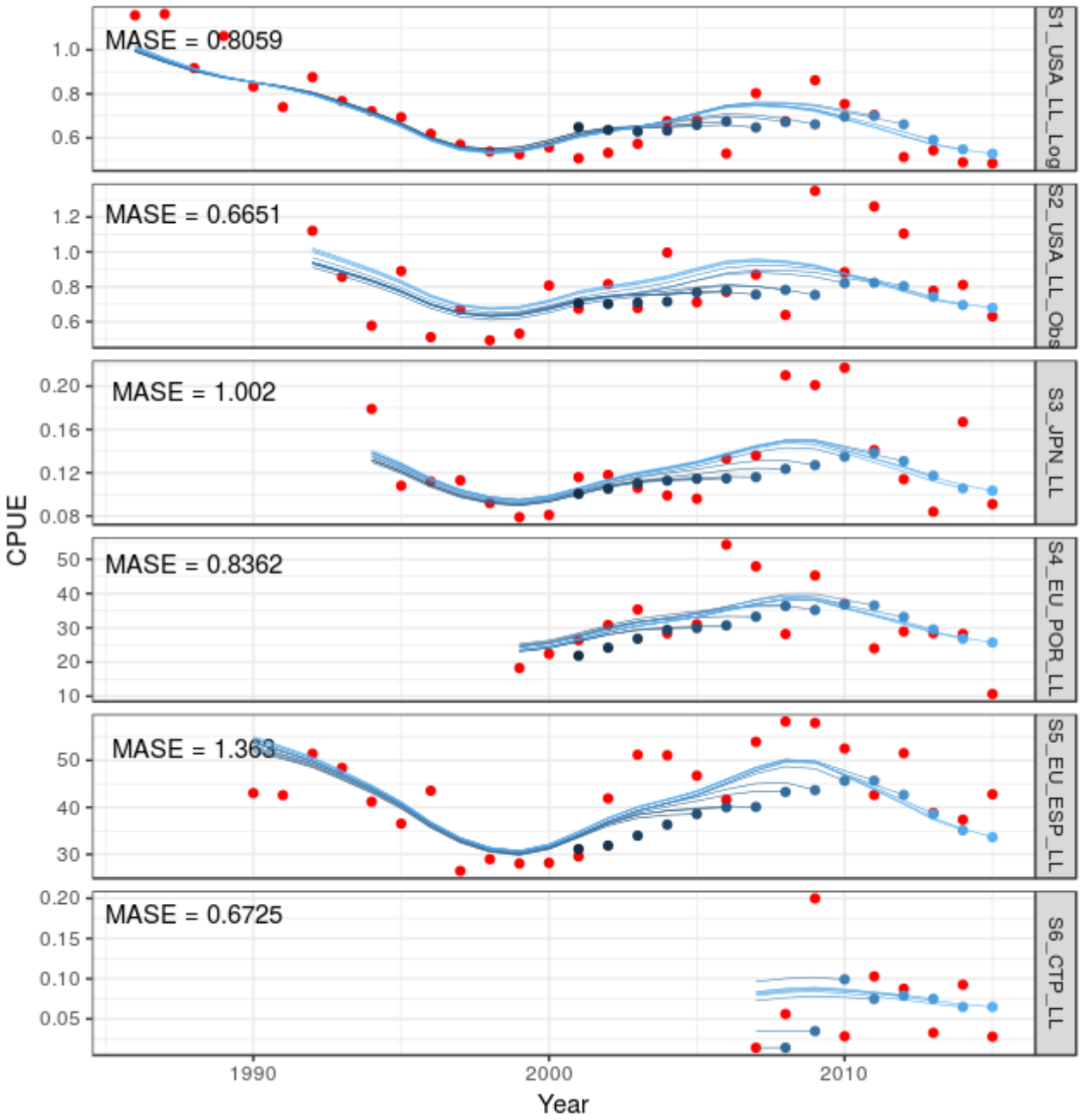


Figure C.1. Hindcast for Stock Synthesis model runs 1 (Panel A), 2 (Panel B), and 3 (Panel C); the red points are the observed CPUE values, and the blue points are the one-step ahead predicted values from each hindcast, ran for up to 15 years.

B. Run 2

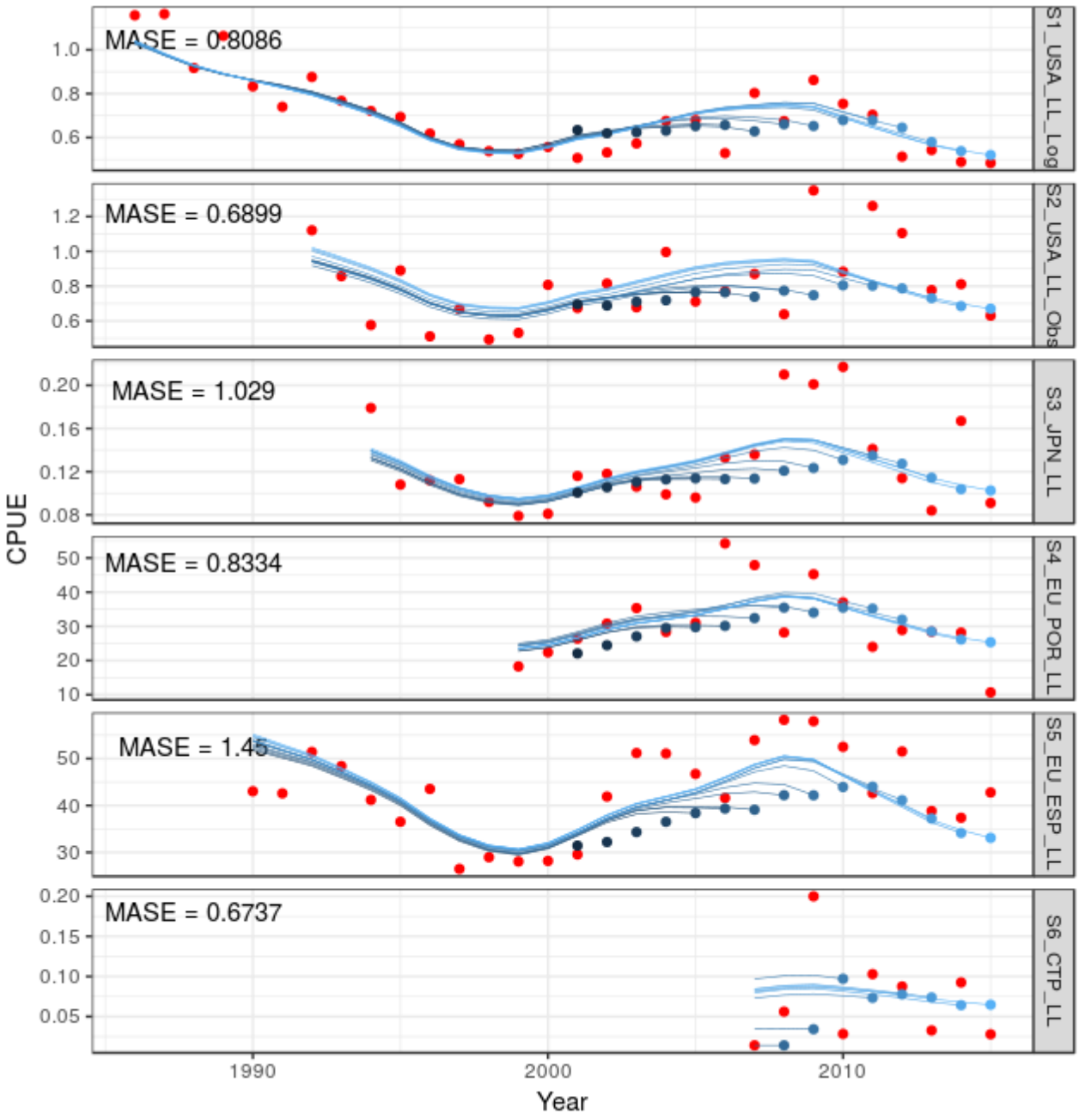


Figure C.1. Continued.

C. Run 3

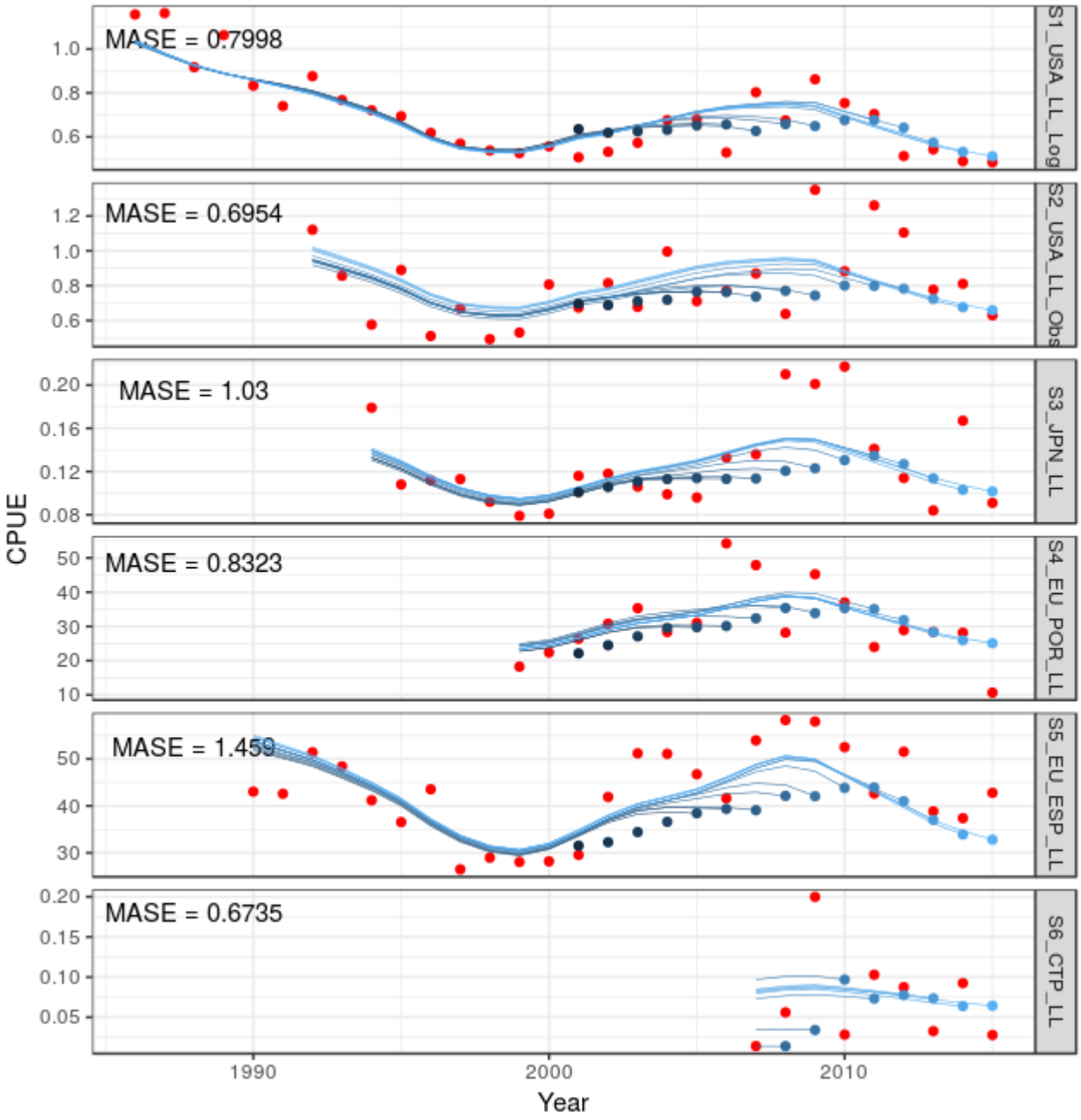


Figure C.1. Continued.

References

Balmaseda, M. A., Davey, M. K., and Anderson, D. L. 1995. Decadal and seasonal dependence of ENSO prediction skill. J. Clim. 8:2705–2715.

Hilborn, R., 2016. Correlation and causation in fisheries and watershed management. Fisheries 41:18–25.

Hjorth, J. U. 1993. Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap. CRC Press.

Huschke, R. E. 1959. Glossary of meteorology. American Meteorological Society.

Hyndman, R. J., and Koehler, A. B. 2006. Another look at measures of forecast accuracy. Int. J. Forecasting 22:679–688.

Jin, E. K., *et al.* 2008. Current status of ENSO prediction skill in coupled ocean-atmosphere models. Clim. Dyn. 31:647–664.

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

Mohn, R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56:473–488.

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

Weigel, A., Liniger, M., and Appenzeller, C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134:241–260.

Willmott, C. J., and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30:79–82.