SCRS/2021/166 Collect. Vol. Sci. Pap. ICCAT, 78(9): 16-62 (2021)

VALIDATION OF ALTERNATIVE STOCK ASSESSMENT HYPOTHESES: NORTH ATLANTIC SHORTFIN MAKO SHARK

Laurence T. Kell[1]

SUMMARY

A multi-model approach for evaluating uncertainty is presented for North Atlantic Shortfin Mako. Several stock assessment methods were used to explore assumptions and uncertainty about biological parameters, reported catch, effort and length data. Methods include trend analysis, length-based indicators, and catch-only, Bayesian state space biomass dynamic, and integrated statistical age-based assessment models. A variety of diagnostics are available to examine goodness of fit, however, it is difficult to compare models with different data sets or structures. Particularly, as residual patterns can be removed by adding more parameters than justified by the data, and retrospective patterns removed by ignoring the data. Therefore, hindcasting was used to estimate prediction skill, a measure of an estimate's accuracy compared to its observed value unknown by the model. Consideration of prediction skill allows data conflicts and model misspecification to be explored. The next steps are to use the methods developed in this study as an objective and transparent way to evaluate and weight stock assessment scenarios and to evaluate the impacts of uncertainty, and the benefits of reducing risk by improving data and knowledge.

RÉSUMÉ

Une approche pluri-modèles visant à évaluer l'incertitude est présentée pour le requin-taupe bleu de l'Atlantique Nord. Plusieurs méthodes d'évaluation du stock ont été utilisées pour étudier les postulats et l'incertitude concernant les paramètres biologiques, la capture déclarée, l'effort et les données de tailles. Ces méthodes incluent l'analyse des tendances, des indicateurs basés sur les longueurs et divers modèles d'évaluation : fondé uniquement sur les captures, de dynamique de la biomasse état-espace de type bayésien et statistique intégré basé sur les âges. Plusieurs diagnostics sont disponibles pour examiner la qualité de l'ajustement. Toutefois, il est difficile de comparer des modèles avec des jeux de données ou structures différents. Et notamment du fait que les schémas résiduels peuvent être supprimés en ajoutant plus de paramètres que ceux justifiés par les données et que les schémas rétrospectifs peuvent être supprimés en ignorant les données. Par conséquent, la simulation rétrospective a été utilisée pour estimer la capacité de prédiction, une mesure de la précision de l'estimation, par rapport à sa valeur observée inconnue par le modèle. L'étude de la capacité de prédiction permet d'analyser les conflits entre les données et les erreurs de spécification du modèle. Les prochaines étapes viseront à utiliser les méthodes développées dans cette étude comme un moyen objectif et transparent pour évaluer et pondérer les scénarios de l'évaluation du stock et évaluer les impacts de l'incertitude et les avantages de réduire le risque en améliorant les données et les connaissances.

RESUMEN

Se presenta un enfoque de varios modelos para evaluar la incertidumbre para el marrajo dientuso del Atlántico norte. Se utilizaron varios métodos de evaluación de stock para explorar los supuestos y la incertidumbre acerca de los parámetros biológicos, la captura declarada, el esfuerzo y los datos de talla. Los métodos incluyen análisis de tendencias, indicadores basados en la talla, y modelos de evaluación solo de captura, de dinámica de biomasa bayesiano de estado espacio, y basado en la edad estadístico integrado. Están disponibles diversos diagnósticos para examinar la bondad del ajuste, sin embargo, es difícil comparar modelos con diferentes conjuntos de datos o estructuras. En particular, los patrones residuales pueden

[1] Centre for Environmental Policy, Imperial College London, London, United Kingdom.

eliminarse añadiendo más parámetros que los justificados por los datos, y los patrones retrospectivos pueden eliminarse ignorando los datos. Por tanto, se utilizó la simulación retrospectiva para estimar la capacidad de predicción, una medida de la precisión de una estimación en comparación con valor observado desconocido por el modelo. La consideración de la capacidad de predicción permite explorar los conflictos entre los datos y la especificación errónea del modelo. Los próximos pasos serán utilizar los métodos desarrollados en este estudio como una forma objetiva y transparente para evaluar y ponderar los escenarios de la evaluación del stock y evaluar el impacto de la incertidumbre y los beneficios de reducir el riesgo mejorando los datos y los conocimientos.

KEYWORDS

Biomass Dynamic, Diagnostics, Catch Only, Integrated Assessment,
Length-Based Indicators, Prediction Skill, Stock Assessment, Trend Analysis

1. Introduction

There are uncertainties associated with every phase of the stock assessment process, ranging from the collection of data, choice of assessment model, model assumptions, interpretation of risk, and the implementation of management advice (Jardim, 2021). Models should therefore be validated to increase confidence in their outputs and trust among the public, stake and asset-holders and policymakers (Saltelli, et al., 2020). Therefore, the aim of this work is to extend the diagnostics of Carvalho et al. (2020) which attempt to identify "a best assessment" for integrated assessments to validate alternative model structure and scenarios.

As the stock assessment process becomes more complex, there are concerns about a lack of transparency, because of the many internal, implicit and often poorly documented assumptions, and a lack of access as only a few highly skilled experts can run the models (Hilborn, 2003). We, therefore, used a range of simple models and a common set of diagnostics based on Carvalho et al., (20201) to evaluate the impact of assumptions about and quality of catch, catch per unit effort (CPUE), and biological parameters on model estimates. Methods include trend analysis (Sherley, et al., 2020), length-based indicators (Cope and Punt 2009, Carvalho et al, 2018, Shephard et al., 2018, Miethe, et al., 2019), catch-only methods (Froese et. al. 2019, Sharma et al., 2021), biomass dynamic assessment models (e.g. JABBA, Winker et al., 2018), and integrated statistical analysis that combines multiple sources of data into a single model (Methot and Wetzel, 2013).

The diagnostics presented allow alternative assessments to be compared and the impact of different datasets and assumption to be evaluated. The approach will, therefore, assist the ICCAT Scientific Committee to identify the impact of uncertainty on the risk of failing to achieve management objectives.

As a worked example, we use the 2017 North Atlantic shortfin mako shark (SMA-N) assessment conducted by the Scientific Committee of the International Commission for the Conservation of Atlantic Tuna (ICCAT, 2018). We consider additional scenarios to those ran in 2017, and use the hindcast to evaluate prediction skill (Kell et al., 2021).

2. Material and Methods

The main assessment method used to develop advice for SMA-N assessment was Stock Synthesis; an integrated statistical analysis that combines multiple sources of data into a single model. Datasets include records of catches and landings, indices of abundance based on CPUE and samples of length composition. Stock Synthesis can be configured in multiple ways, allowing for a range of scenarios to be developed to reflect uncertainty. However, problems remain when using integrated assessments due to lack of information to estimate key parameters, missing data, inadequate theory, latent state variables, and unpredictable future elements (Gass, 1983).

The main decisions made in the 2017 assessment were related to the choice of i) catch scenarios, ii) indices of abundance based on standardised catch per unit effort (CPUE), and iii) biological parameters. These were reviewed by Mejuto et al., (2021) who noted that the historical series of total catch (i.e. Task 1) considered in the base case assessment scenario (1950-2015) may have greatly underestimated the level of historical catches during several initial decades taking into consideration the history of the fisheries, fleet capacity and historical fishing effort, an alternate catch scenario (1971-2015) estimated on ratios among species could overestimate in an important amount the catches during the final period, that there were limitations regarding the CPUE series used as indices of abundance; and high uncertainty about key biological parameters selected.

A reference or **Base Case** was chosen based on the 2017 ICCAT assessment, then scenarios were developed for factors based on alternative catch histories and biological parameters. Rather than considering alternative CPUE scenarios based on down weighting these series were used in a model-free hindcasting to estimate prediction skill a measure of the accuracy of an estimate compared to its observed value unknown by the model.

The main diagnostics used in stock assessment are residuals to check the fit and retrospective analysis to check for stability. However, the best way to get rid of a residual pattern is to overfit and the best way to get rid of a retrospective pattern is to ignore the data. Neither approach can therefore be used to validate a model, since this requires assessing whether it is plausible that a system equivalent to the model generated the data. It is also difficult to compare models based on metrics such as AIC when they have different structures. We therefore use the hindcast to estimate prediction skill, a measure of the accuracy of an estimate compared to its observed value that is not known by the model, to explore data conflicts and potential model misspecification.

Sharks are long-lived and so the impact of the current age structure, particularly if dome-shaped selectivity is assumed, may not be seen until many years in the future. This makes it difficult to make projections, therefore consequences for management advice was based on yield-per-recruit and biomass-based production functions instead of attempting to provide catch advice by developing a Kobe strategy matrix.

2.1 Assessment Inputs

The data available for use in the SMA-N assessment are fisheries dependent and include time series of total catch, standardised CPUE, and length composition samples.

2.1.1 Catch

Six alternative time series of total catch were considered (**figure 1**). As well as the total catch series assumed in the Stock Synthesis base case (ICCAT, 2017), two additional catch scenarios were considered by the WG. These were C1) where it was assumed that the aggregated Task 1 catch by year between 1950-1970 and probably the 1980's, was underestimated; and C2) based on a hypothesized relationship between the catch of main species reported and SMA catches between 1971-2015, but based on where it was likely assumed that Task 1 had been overestimated, probably by up to around 1000 t/yr, in the most recent years of that series for some main fleets. Three additional catch scenarios based on Mejuto et al. (2021) were also considered (C3_11, C3_12 and C3_13). These are based on assuming a range of observed average nominal catch rates of 1.0, 1.5.and 2.0 fish per thousand hooks, respectively, (mean weight of 50.0 kg RW per fish) in those long line fleets-years with known fishing activity and fishing effort available but without SMA historical catches reported. Catches by fleet as assumed by the working group are shown in **figure 2.**

The length composition by sex are summarised in **figure 3**, the vertical lines show length at which 50% of individuals reach maturity. Most females caught are immature, i.e. the selection pattern is dome shaped.

Indices of abundance are summarised in **figure 4**. The pairwise scatter plots (**figure 5**) look at correlations between the indices, and **Figure 6** shows the correlation matrix for the indices, blue indicate a positive correlation and red negative. Cross correlations between indices, to identify potential lags due to year-class effects or spatial distribution, are shown in **figure 7**. These plots all show that the indices appear to have similar trends and apart from the Portuguese long line, where lags are seen, are catching similar components of the population, i.e. there is no apparent structure due to age or spatial effects.

2.1.2 Biology

In many by catch fisheries there is a high degree of uncertainty about the biological parameters and their effect on productivity. The range of models for natural mortality-at-age (**figure 8)**, and length-at-age (**figure 9)** were considered by the working group. Natural mortality-at-length is shown in **figure 10** for the Gislason et al., (2008) and Lorenzen (1996) natural mortality models.

The main differences between the Stock Synthesis scenarios configured by the WG was the choice of M and the assumed stock recruitment relationship (ICCAT, 2017). The corresponding vectors-at-age are summarised for Stock Synthesis runs 1, 2 and 3 are summarised in **figure 11**.

2.3 Assessment Methods

Five different methods were used to screen the datasets and assess the stocks. A Bayesian state-space trend analysis tool (JARA), designed to objectively incorporate uncertainty into the IUCN Red Listing evaluation process, was used to evaluate trends in the indices. Length-based indicators (LBIs) were computed to look at information in the length data. While three different assessment methods were used to estimate stock status; i.e. catch-only, Bayesian state space biomass dynamic, and integrated assessment models.

Three main Stock Synthesis model runs were developed by the Group in 2017. The base case (run 1) was updated to set natural mortality for males equal to that for females in run 2. While for run 3 the BH stock recruitment relationship was replaced with the Low Fecundity Spawner Recruitment (LFSR).

The catch only method (COM) was implemented using the Bayesian state space biomass dynamic model JABBA using priors for initial and final depletion rather than fitting to indices of abundance. This allowed an exploration of the impact of different catch scenarios on stock trends and productivity and reference points. Priors were either based on a heuristic or were known, i.e. taken from the Stock Synthesis assessment runs. The biomass dynamic assessment model with CPUE indices was then used to explore the same catch scenarios and assumption about productivity (i.e. r) on stock trends and reference points.

2.3.1 Reference Points

Clarke and Hoyle (2014) recommended a tiered framework for establishing limit reference points (LRPs), depending on the quality of the assessment: namely

(1) For those elasmobranchs evaluated using a stock assessment model for which there is confidence that the stock-recruitment relationship is appropriately specified, use a fishing mortality-based LRP of F_{msy};

(2) In cases where a stock assessment model was used but the stock-recruitment relationship is highly uncertain, consider SPR-based LRP such as $F_{60\%SPR}$;

(3) When stock assessments are not available, or when the results are not considered robust, use risk-based fishing mortality LRP benchmarks (F_{msm}, F_{lim} and F_{crash}), as used in Australia (Zhou et al. 2011).

WCPFC for deriving risk-based reference points assumed that the population dynamics could be described by a Graham-Schaefer production model where F_{msm}=F_{msy}, F_{lim}=1.5 F_{msm}, and F_{crash}=2F_{msm}= =r_{max} (Zhou et al., 2011). “msm” stands for “maximum sustainable mortality” for non-retained bycatch, but it is equivalent to MSY for commercial species. Hence F_{msm} is identical to F_{msy}.

In this work we limit our explorations to MSY based reference points and a biomass limit based on 20% of virgin biomass.

2.3.2 Length-based indicators

Various length-based indicators are used by ICES for screening length compositions and to classify stocks according to conservation and sustainability status, yield optimisation and MSY (**Table 1**).

There are three elements in making an indicator operational, the indicator itself, a reference point, and a reference level. LBIs based on lower percentiles of the length frequency distribution are for the conservation of immature fish; these include $L_{25\%}$ (the 25th percentile of the length distribution) and L_c (the length at 50\% of modal abundance). Indicators based on the central tendencies are proxies for MSY and include: L_{mean} (mean length of individuals > L_c); L_{maxY} (the length class with maximum biomass in catch); and L_{bar} (the mean length). Those based on upper percentiles are for the conservation of larger individuals and are: L_{max5} (mean length of largest 5%); $L_{max95\%}$ (95th percentile); and P (the proportion of individuals above L_{opt}). Potential reference points based on life history parameters are $L_{infinity}$, L_{mat}, L_{opt}, and $L_{F=M}$.

2.3.3 Scenarios

Six scenarios were considered for catch; The SS base case catch, C1 and C2 as used by the WG, and C3_11, C3_12 and C3_13 of Mejuto et al., (2021). Three additional scenarios were considered in the Stock Synthesis assessment, and four in the biomass dynamic assessments to evaluate assumptions about productivity.

For Stock Synthesis productivity these were implemented by modifying the biological parameters (see **Figure 21** below). In the low productivity scenarios slow growth and high survival was modelled by scaling k and $L_{infinity}$ of the von Bertalanffy growth model and M-at-age by 0.9 and age at 50% maturity was increased by 2 ages. In the high productivity scenario growth and M were scaled by 1.25 and age at 50% maturity was reduced by two ages. A so call North Pacific scenario (Anon, 2018) was also modelled where maturity and growth was set to that of the North Pacific stock as described and compared in Mejuto et al., (2021).

For the biomass dynamic-based models the r prior varied by setting it to 0.03, i.e. based on Cortés (2019), or 0.75, 1.5 and 2 times the Cortés value.

2.4 Retrospective Analysis

A standard diagnostic tool in stock assessment is retrospective analysis (Hurtado-Ferro et al., 2015). This involves sequentially removing all data from the most recent period, refitting the model, and comparing terminal year estimates to the full model using the relative error (Mohn’s rho). The diagnostic is widely used and in Europe, it is often the "key diagnostic" for accepting or rejecting a model.

Retrospective analysis has been extended to include stock forecasts, where the terminal year estimates are projected using assumptions about future catches, recruitment, biological parameters and the vulnerability of the stock to fishing (Brooks and Legault, 2016). Stability and a reduction in variance, however, can be achieved at the expense of accuracy and bias, for example by shrinking terminal estimates towards recent historical values. Also, the best way to remove a retrospective pattern is to ignore the data. It is not possible, however, to validate a model if bias is unknown and the quantity used for validation is not observable (Hodges and Dewar, 1992).

2.5 Hindcast

A hindcast procedure rather than a traditional retrospective analysis was used for validation (see Kell, et al., 2021 for full details). In a hindcast like a retrospective analysis all observations for a year are sequentially removed from the terminal year backwards (i.e. peeled), the model is then refitted to the truncated series and observations compared to model estimates. Validation requires that the system be observable and measurable and so observations should be used for cross-validation unless model estimates are known to be very close to their true values. This is unlikely to be the case in stock assessment, meaning that bias is difficult to quantify. For example, a reduction in mean squared error (a measure of variance) can be achieved by shrinkage at the expense of prediction skill. The absence of retrospective patterns in model-based quantities, therefore, while reassuring is not sufficient for validation. Therefore, validation should, be conducted using prediction skill based on observations. We, therefore, used the models to generate pseudo data for the CPUE and compare these to observations.

2.6 Productivity

In age-structured models, density dependence is mainly accounted for by the stock-recruitment relationship. Cury et al. (2014), however, showed that in most cases the stock-recruitment relationship used to estimate productivity and determine reference points, has poor estimation/predictive power and the environment has a larger effect on productivity, a result confirmed by other studies (e.g. Szuwalski et al., 2015; Free et al., 2019), and observed 100 years ago by Hjort (1914). In ICCAT assessments, however, growth, maturation and natural mortality are assumed not to have varied despite the significant changes in the environment and stock biomass seen. Therefore as well as summarising the expected productivity in the form of the production function we also summarise the recruitment deviates for Stock Synthesis and compare these to the estimates of process error for the biomass dynamic assessment.

3. Results

In the results we concentrate on examples of how the various models and diagnostics can assist the assessment process, rather than argue about current stock status.

3.1 Length-based indicators

Figure 12 presents fits to the indices from the trend analysis, the ribbons show the common trend and the bars the observed indices with their 95% confidence intervals. The indices all, apart from Chinese Taipei long line, show a common trend, i.e. a decline in the 1990s followed by a recovery in the 2000s before the another decline. Based on the historical indices a long-term decline is forecast, however, this is beyond the range of the data and so is not intended as a prediction just a summary of current state. The residuals are summarised in **figure 13**. Red backgrounds indicate where there are strong residual patterns, assessed by fewer crossings and longer runs than expected (Carvalho et al., 2017). There are strong patterns and for example the USA long-line is suggesting recent lower stock levels, although this may be due to recent under-reporting (Cortes, 2017).

Figure 14 summarises the length-base indicators derived from the Stock Synthesis length composition data. The quality of data samples varies by fleet (column). Fleet 4 appears to have the more complete dataset. All indices show a general upward trend, i.e. an increase in mean size of the indicator. Problems are that length samples only cover a short period, their quality could be improved, and their impact on the assessment is likely to be less than the assumptions about growth.

3.2 Catch-only method

The catch only model (COM) was used to explore the impact of the different catch scenarios on stock trends. The assessment was conducted using a biomass dynamic model, with priors for initial and final depletion, rather than fitting to indices of abundance. The depletion priors were either based on a heuristic or were known, i.e. taken from the Stock Synthesis assessment runs. Four different priors were used for population growth rate (r), to evaluate the impact of the assumed biological characteristics. The figures summarise trends (**figure 15)**, and the production functions (**figure 16)** by catch scenarios (column) and type of depletion prior (row). The heuristics tend to suggest a more depleted stock, but trends are similar. C1 where the catch was inflated suggests a lower abundance than Stock Synthesis.

3.3 Biomass dynamic assessment

The biomass dynamic assessment model was fitted to the indices of abundance. The parameter estimates and correlations between them are summarised in **Figure 17** for the scenario where the catch was the same as that used in Stock Synthesis and the r prior was 0.3. The posteriors for virgin biomass (K), final depletion (psi) and r are shown in **figure 18**. The trends and production functions are shown in **figures 19** and **20** respectively. These also show the target reference points, based on MSY. It appears that the assumed catch has a bigger impact on productivity than the priors for r/. The trends in biomass relative to B_{MSY} is similar for all scenarios, i.e. a decline after 1980, a rebuilding after 1990, and a subsequent decline.

3.4 Stock Synthesis

Twelve Stock synthesis runs were conducted for three catch and four productivity scenarios, and the trends in SSB relative to B_{MSY} are summarised in **Figure 21**. These show smooth trends, i.e. the stock is stable until the 1990s after which there is a steady decline. These are quite different from the biomass dynamic assessment results. Production functions from the Stock Synthesis runs for the low and high productivity scenarios are shown in **figure 22,** and the trends in SSB in **figure 23.**

3.5 Model free hindcasts

The model-free hindcasts are shown in **figures 24** through **33** for the JABBA and Stock Synthesis model runs, each plot compares the model-free hindcast for an individual CPUE series. A MASE scores < 1 indicates that a model has better prediction skill than the naïve baseline forecast, these are shaded green for ease of comparison. **Figures 24 to 28** are for the JABBA assessments and the most accurate predictions were observed for Spanish and US long line, while the Japanese and Portuguese long lines appear to be mainly adding noise, since a random walk performs better than the index in all cases for the Japanese long line, and for the Portuguese fleet MASE~1. There was little difference between the catch scenarios or r priors, suggesting that the data do not have much information.

For Stock Synthesis **Figures 29 to 33** the US longline did not have prediction skill for the low productivity scenarios, Japanese long line again had poor prediction skill, although the Portuguese longline now had prediction skill.

3.6 Residuals

The residuals are plotted in **Figures 34** and **35** for all JABBA and Stock Synthesis scenarios, no real patterns are seen in the plots and there is little difference between the scenarios.

3.7 Mohn's rho

Mohn's rho is plotted in **Figures 36** and **37** for all JABBA and Stock Synthesis scenarios, values are between -0.15 and 0.2 so there is no strong retrospective pattern.

3.8 Surplus production

Figures 38 and **39** show the surplus production against biomass for all JABBA and Stock Synthesis scenarios. The patterns are quite different. As the stock declines, i.e. the trajectory passes from right to left. For Jabba in the early period there are no indices of abundance and so the trend shows no variability due to process error, once the indices are available for tuning surplus production varies around 0. This is in marked contrast to Stock Synthesis. Stock synthesis gave similar results to the catch-only method, namely stable biomass with a more recent decline. In contrast JABBA which fits mainly to the CPUE, predicts a "double dip". The main difference between JABBA and Stock Synthesis is the use of the length composition data, however LBIs suggest that there is little information in the length compositions and so the difference is likely to be due to assuming age dynamics.

3.9 Process Error

Figure 40 Summarises the recruitment deviates by scenario for Stock Synthesis and **Figure 41** process error by scenario for JABBA.

Discussion and Conclusions

In the 2019 shortfin mako shark stock assessment (ICCAT, 2019) it was noted that different assumptions and modelling frameworks led to different outcomes, and that the projections conducted by the SCRS did not include uncertainties about future elements related to growth, age-at-maturity, natural mortality, the stock-recruitment relationship, and selectivity and catch rates by fleets. This study, therefore employed a variety of models and scenarios to explore the impact of uncertainty on historical trends in stock status and estimates of productivity. Methods used include trend analysis, length-based indicators, as well as catch only, biomass dynamic, and integrated statistical age-based assessment models. A major difference between the biomass dynamic and the integrated assessment methods is that the biomass dynamic models only use series of catch and CPUE, while the integrated assessment model the age structure and use length data. In an integrated model the indices of abundance largely determine the trends in abundance, while the length compositions helps determine the absolute level.

Trends estimated by the integrated models were similar across scenarios. The main difference was the level of the stock relative to reference points. A reason for this is because the catch scenarios evaluated by the group were relatively similar and did not consider uncertainty about historical under-reporting Mejuto et al., (2021). Al thought, Cortés (2017) noted that changes in reporting practices as a result of the implementation of several logbook programs historically, and perhaps a tendency to under-report bycatch over time as fishers develop a growing perception that those reports result in increasingly restrictive management measures may have affected the logbook index to some extent.

The productivity of a stock and maximum sustainable yield (MSY) reference points are determined by parameters that are hard to estimate (e.g. natural mortality-at-age and the stock recruitment relationship) and assumptions about growth and fecundity. Furthermore, the ages which are vulnerable to fishing, can have a large effect on estimates of stock status and reference points, but is often difficult to estimate from length data (Carruthers et al. 2018). In biomass based models productivity is modelled by an explicit production function which requires the estimation, fixing or providing priors for fewer parameters. Variability in the expected productivity is modelled by variability in recruitment and hence year-class strength in the integrated model and by process error term in the biomass dynamic model.

The main difference between scenarios was due to the assumed biological parameters which had a large impact on estimates of B_{MSY}. The length composition data only cover the recent period and the length-based analysis showed that they appear to have little information on changes in exploitation level and sampling is poor for some fleets. In contrast to the integrated model the biomass dynamic model showed greater variability in the trends, both when the decline first occurred, and it's rate, there is also a recovery in the 1990s that was not seen in the integrated model.

The difference between the integrated and biomass dynamic trends is due to variability in productivity. In the former this is modelled by year-class strength (i.e. recruitment), while in the latter by a process error term. The variability in recruitment was substantially larger at 40% than process error at 10%. Therefore, the dynamics in stock synthesis are determined mainly by process error, rather than a production function. This means that it is unlikely that the observed catches and expected surplus production alone can explain trends in the abundance indices. This can have several causes, including that stock dynamics are recruitment-driven, the indices of relative abundance are not proportional to abundance, the model is incorrectly specified; or the data are biased (Minte-Vera et al., 2017). The assumptions about the biology had an impact on the assessment and whether the stock was overfished.

The main diagnostics used in stock assessments are residuals to check the fit and retrospective analysis to check for stability. However, the best way to get rid of a residual pattern is to overfit and the best way to get rid of a retrospective pattern is to ignore the data. Neither approach can therefore be used to validate a model, since this requires assessing whether it is plausible that a system identical to the model generated the data (Thygesen, et al., 2017). We therefore propose that hindcasting be used to estimate prediction skill, a measure of the accuracy of an estimate compared to its observed value that is not known by the model, to explore data conflicts and model misspecification. This is important since historical dynamics in the integrated model were driven by recruitment which was highly variable between years and also showed patterns over time. This means that it will be difficult to make predictions into the future.

References

Anonymous. 2018. Stock Assessment of shortfin mako shark in the North Pacific Ocean Through 2016. WCPFC-SC14-2018/ SA-WP-11: 121pp.

Brooks, E. N. and Legault. C. M. Retrospective forecasting—evaluating performance of445stock projections for new england groundfish stocks. Canadian Journal of Fisheries and446Aquatic Sciences, 73(6):935–950, 2016.

Carruthers, T., Kell, L. and Palma, C., 2018. Accounting for uncertainty due to data processing in virtual population analysis using Bayesian multiple imputation. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(6), pp.883-896.

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

Carvalho, F., Lee, H.H., Piner, K.R., Kapur, M. and Clarke, S.C., 2018. Can the status of pelagic shark populations be determined using simple fishery indicators?. Biological Conservation, 228, pp.195-204.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R. and Maunder, M.N., 2021. A cookbook for using model diagnostics in integrated stock assessments. *Fisheries Research*, *240*, p.105959.

Clarke, S., and Hoyle, S. 2014. Development of limit reference points for elasmobranchs. WCPFCSC10-2014/ MI-WP-07. Scientific Committee Tenth Regular Session, Majuro, Republic of the Marshall Islands, 6-14 August 2014. 43 pp.

Cope, J. and Punt, A.E. 2009. Length-Based Reference Points for Data-Limited Situations: Applications and Restrictions. Marine and Coastal Fisheries: Dynamics, Management, and Ecosystem Science 1:169–186

Cortés, E. 2017. Stock status indicators of mako sharks in the western North Atlantic Ocean based on the US pelagic longline logbook and observer programs. Collect. Vol. Sci. Pap. ICCAT, 74(4): 1639-1663.

Cortés E. 2019. Preliminary estimates of population dynamics parameters of porbeagle shark in the western North Atlantic Ocean. SCRS/2019/087.

Cury, P. M., Fromentin, J.-M., Figuet, S., and Bonhommeau, S. (2014). Resolving hjort's dilemma how is recruitment related to spawning stock biomass in marine fish? Oceanography, 27(4):42–47.

Free, C. M., Thorson, J. T., Pinsky, M. L., Oken, K. L., Wiedenmann, J., and Jensen, O. P. (2019). Impacts of historical warming on marine fisheries production. Science, 363(6430):979 LP – 983.

Froese, R., Zeller, D., Kleisner, K. & Pauly, D. 2012. What catch data can tell us about the status of global Fisheries. Marine Biology, 159(6): 1283–1292.

Gass, Saul I. 1983. Decision-aiding models: validation, assessment, and related issues for policy analysis. Operations Research, 31(4), 603–631.

Gislason, H., Daan, N., Rice, J. and Pope, J., 2008. Does natural mortality depend on individual size. ICES CM.

Hilborn. R. 2003. The state of the art in stock assessment: where we are and where we are going. 482 Scientia Marina, 67(S1):15–20, 2003.

Hilborn, R. (2001). Calculation of biomass trend, exploitation rate, and surplus production from survey and catch data. Canadian Journal of Fisheries and Aquatic Sciences, 58(3):579–584.

Hjort, J. (1914). Fluctuations in the great fisheries of northern europe viewed in the light of biological research. ICES.

Hodges, J. S. and Dewar, J. A. Is it you or your model talking?: A framework for model483validation. 1992

Hurtado-Ferro, F., Szuwalski, C. S., Valero, J. L., Anderson, S. C., Cunningham, C. J., Johnson, K., Licandeo, R., McGilliard, C. R., Monnahan, C. C., Muradian, M. L. et al. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock489assessment models.ICES Journal of Marine Science, 72(1):99–110, 2015.

ICCAT, 2017. Report Of The 2017 Iccat Shortfin Mako Assessment Meeting. SCRS/2017/007 Collect. Vol. Sci. Pap. ICCAT, 74(4): 1465-1561 (2017)

ICCAT, 2019. Report of the 2019 shortfin mako shark stock assessment update meeting. SCRS/2019/008 Collect. Vol. Sci. Pap. ICCAT, 76(10): 1-77 (2020)

Jardim, E., Azevedo, M., Brodziak, J., Brooks, E.N., Johnson, K.F., Klibansky, N., Millar, C.P., Minto, C., Mosqueira, I., Nash, R.D. and Vasilakopoulos, P., 2021. Operationalizing ensemble models for scientific advice to fisheries management. *ICES Journal of Marine Science*, *78*(4), pp.1209-1216.

Kell, L. T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., and Fu, D. Validation of stock assessment methods: is it me or my model talking?, *ICES Journal of Marine Science*, 2021;, fsab104, https://doi.org/10.1093/icesjms/fsab104

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(4), pp.627-642.

Miethe, T., Reecht, Y. and Dobby, H., 2019. Reference points for the length-based indicator L max5% for use in the assessment of data-limited stocks. ICES Journal of Marine Science, 76(7), pp.2125-2139.

Minte-Vera, C.V., Maunder, M.N., Aires-da-Silva, A.M., Satoh, K. and Uosaki, K., 2017. Get the biology right, or use size-composition data at your own risk. *Fisheries research*, *192*, pp.114-125.

Mejuto, J., Fernández-Costa, J., Ramos-Cartelle, A., García-Cortés, B. and Carroceda, A. 2021. Plausibility and uncertainty of basic data and parameter selection on stock assessments: a review of some input data used in the 2017 assessment of the shortfin mako (*Isurus oxyrinchus*) of the northern Atlantic stock. Collect. Vol. Sci. Pap. ICCAT, Vol. 78(5): 119-170.

Methot Jr, R.D. and Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fisheries Research, 142, pp.86-99.

Rice, J., 2021. Stock assessment blue shark (Prionace glauca) in the Indian Ocean using Stock Synthesis. *Working Party on Ecosystems and Bycatch. IOTC document IOTC–2021–WPEB17(AS)-15:*

Saltelli A., Bammer G., Bruno I., Charters E., Di Fiore M., Didier E., Espeland W. N., Kay J., Piano S. L., Mayo D., et al. Five ways to ensure that models serve society: a manifesto, 2020.

Sharma, R., Winker, H., Levontin, P., Kell, L., Ovando, D., Palomares, M.L., Pinto, C. and Ye, Y., 2021. Assessing the Potential of Catch-Only Models to Inform on the State of Global Fisheries and the UN's SDGs. *Sustainability*, *13*(11), p.6101.

Shephard, S., Davidson, I.C., Walker, A.M. and Gargan, P.G., 2018. Length-based indicators and reference points for assessing data-poor stocks of diadromous trout Salmo trutta. Fisheries research, 199, pp.36-43.

Sherley, R.B., Winker, H., Rigby, C.L., Kyne, P.M., Pollom, R., Pacoureau, N., Herman, K., Carlson, J.K., Yin,.S., Kindsvater, H.K. and Dulvy, N.K., 2020. Estimating IUCN Red List population reduction: JARA—A decision-support tool applied to pelagic sharks. Conservation Letters, 13(2), p.e12688.

Szuwalski, C. S., Vert-Pre, K. A., Punt, A. E., Branch, T. A., and Hilborn, R. (2015). Examining common assumptions about recruitment: A meta-analysis of recruitment dynamics for worldwide marine fisheries. Fish and Fisheries, 16(4):633–648.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K. and Nielsen, A., 2017. Validation of ecological state space models using the Laplace approximation. *Environmental and Ecological Statistics*, *24*(2), pp.317-339.

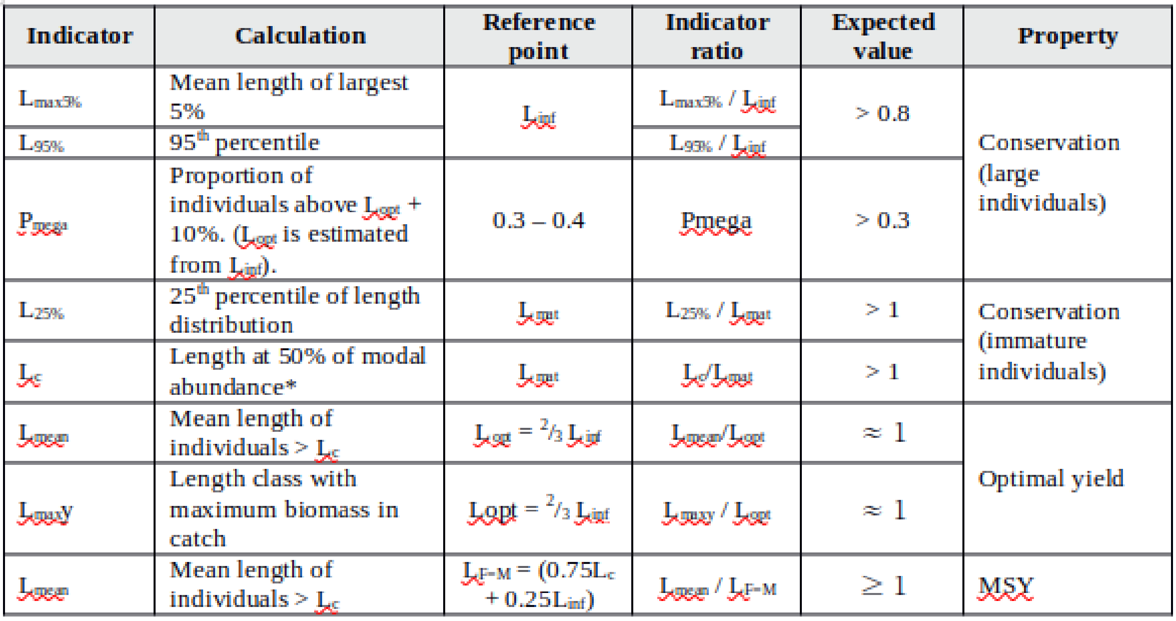
Zhou, S., Smith, A. D. M. M., and Fuller, M. 2011. Quantitative ecological risk assessment for fishing effects on diverse data-poor non-target species in a multi-sector and multi-gear fishery. Fisheries Research, 112: 168–178.

Zhou, S., Deng, R., Hoyle, S. and Dunn, M., 2019. Identifying appropriate reference points for elasmobranchs within the WCPFC. Report to Western and Central Pacific Commission, Pohnpei, Federated States of Micronesia.

Winker, H., Carvalho, F. and Kapur, M., 2018. JABBA: just another Bayesian biomass assessment. Fisheries Research, 204, pp.275-288.

Weigel, AP, Liniger, MA, & Appenzeller, C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Quarterly Journal of the Royal Meteorological Society, 134(630), 241–260.

Table 1. Length-based indicators



Indicator	Calculation	Reference point	Indicator ratio	Expected value	Property
$L_{max5\%}$	Mean length of largest 5%	L_{inf}	$L_{max5\%}$ / L_{inf}	> 0.8	Conservation (large individuals)
$L_{95\%}$	95th percentile		$L_{95\%}$ / L_{inf}		
P_{mega}	Proportion of individuals above L_{opt} + 10%. (L_{opt} is estimated from L_{inf}).	0.3 – 0.4	Pmega	> 0.3	
$L_{25\%}$	25th percentile of length distribution	L_{mat}	$L_{25\%}$ / L_{mat}	> 1	Conservation (immature individuals)
L_c	Length at 50% of modal abundance*	L_{mat}	L_c/L_{mat}	> 1	
L_{mean}	Mean length of individuals > L_c	$L_{opt} = {}^2/_3\ L_{inf}$	L_{mean}/L_{opt}	≈ 1	Optimal yield
L_{maxy}	Length class with maximum biomass in catch	$Lopt = {}^2/_3\ L_{inf}$	L_{maxy} / L_{opt}	≈ 1	
L_{mean}	Mean length of individuals > L_c	$L_{F=M} = (0.75L_c + 0.25L_{inf})$	L_{mean} / $L_{F=M}$	≥ 1	MSY

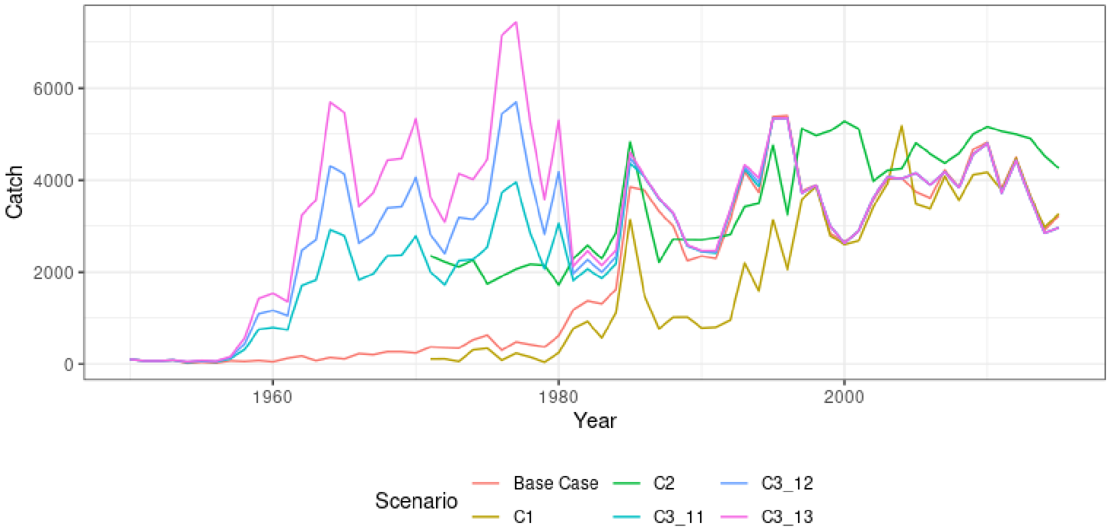


Figure 1. Alternative catch scenarios.

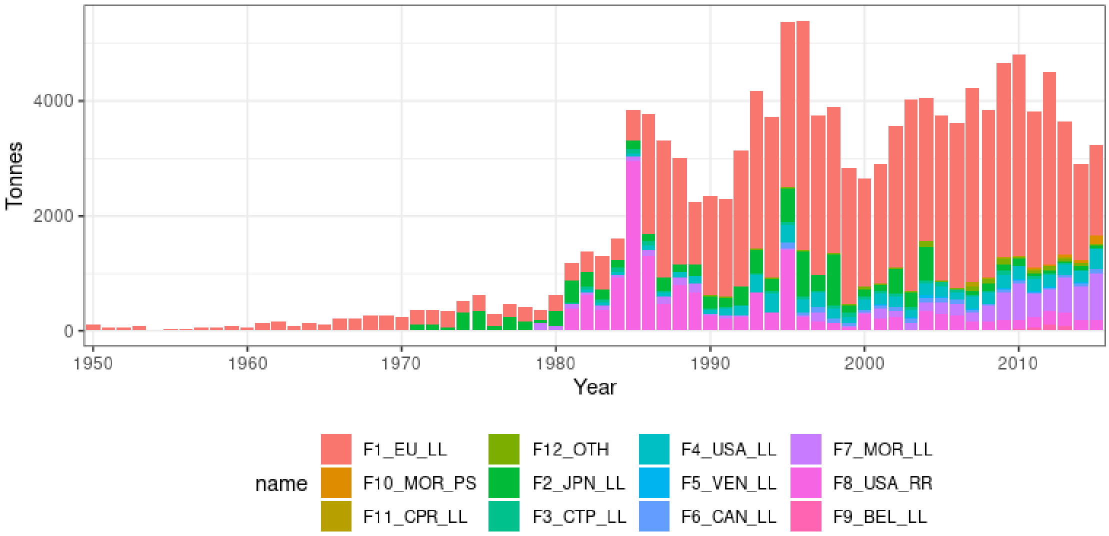


Figure 2. Catch by fleet for working group base case.

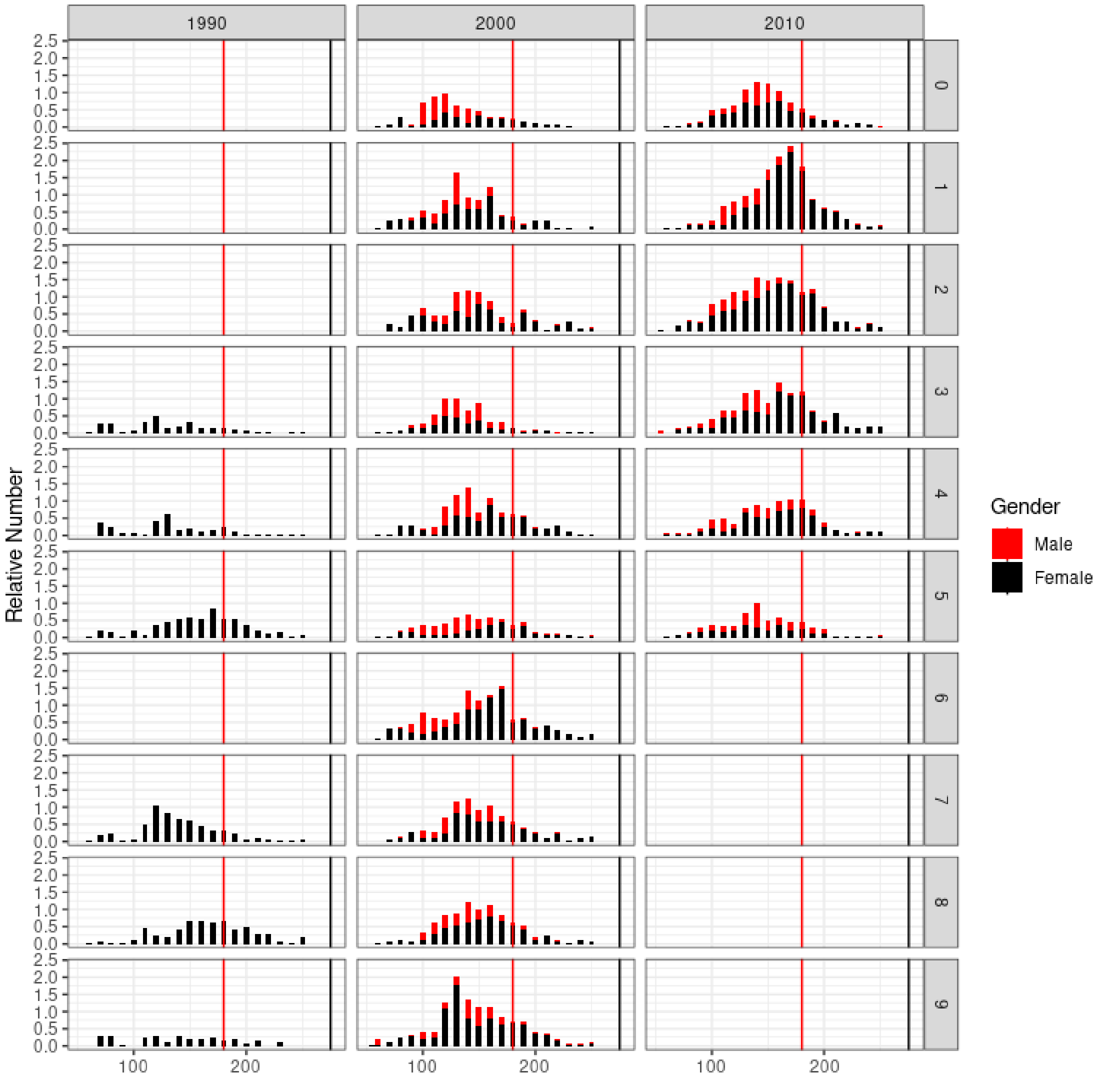


Figure 3. Length compositions, length at 50% maturity is indicated.

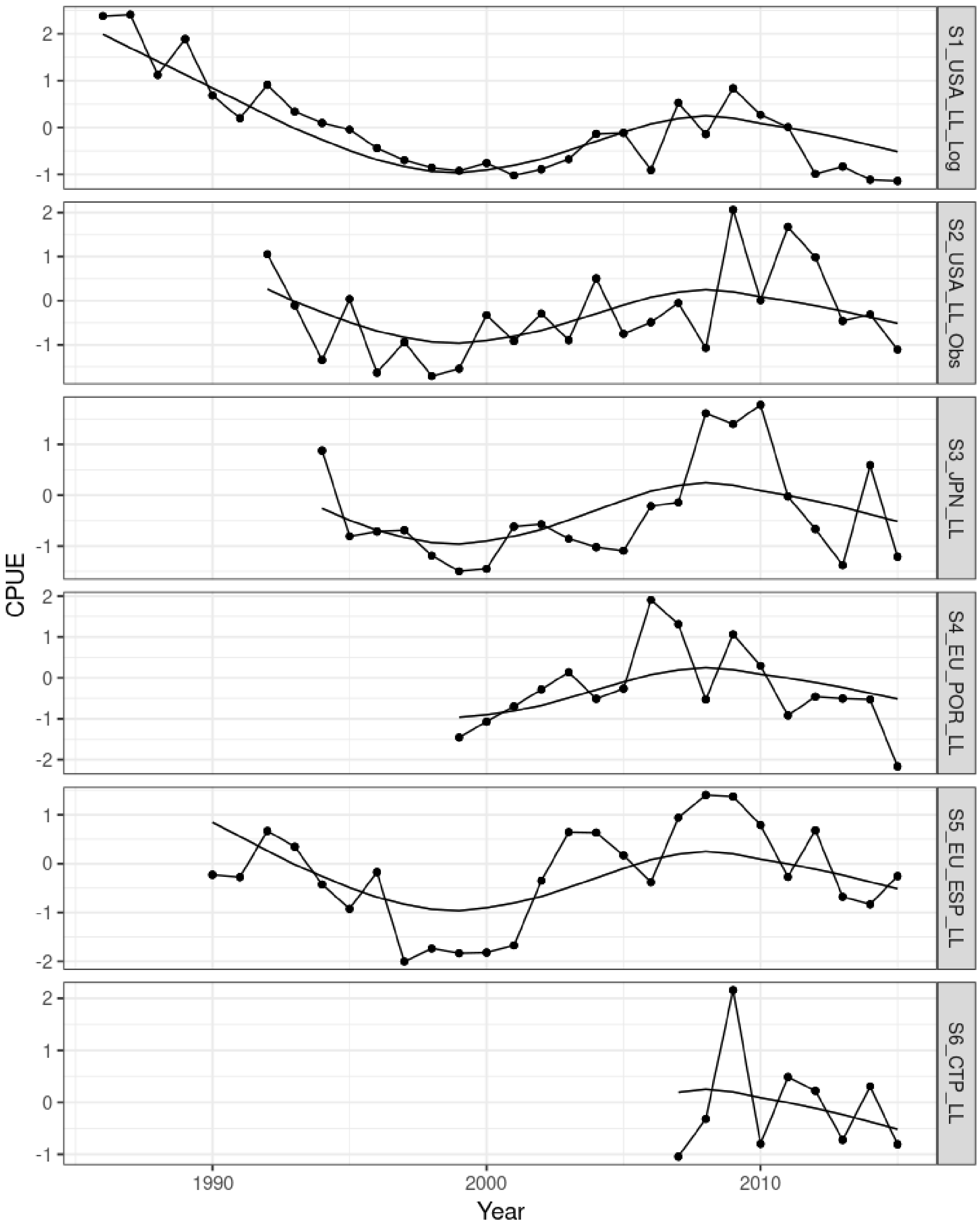


Figure 4. Time series of CPUE indices, continuous black line is a lowess smother showing the average trend by area (i.e. fitted to year with series as a factor).

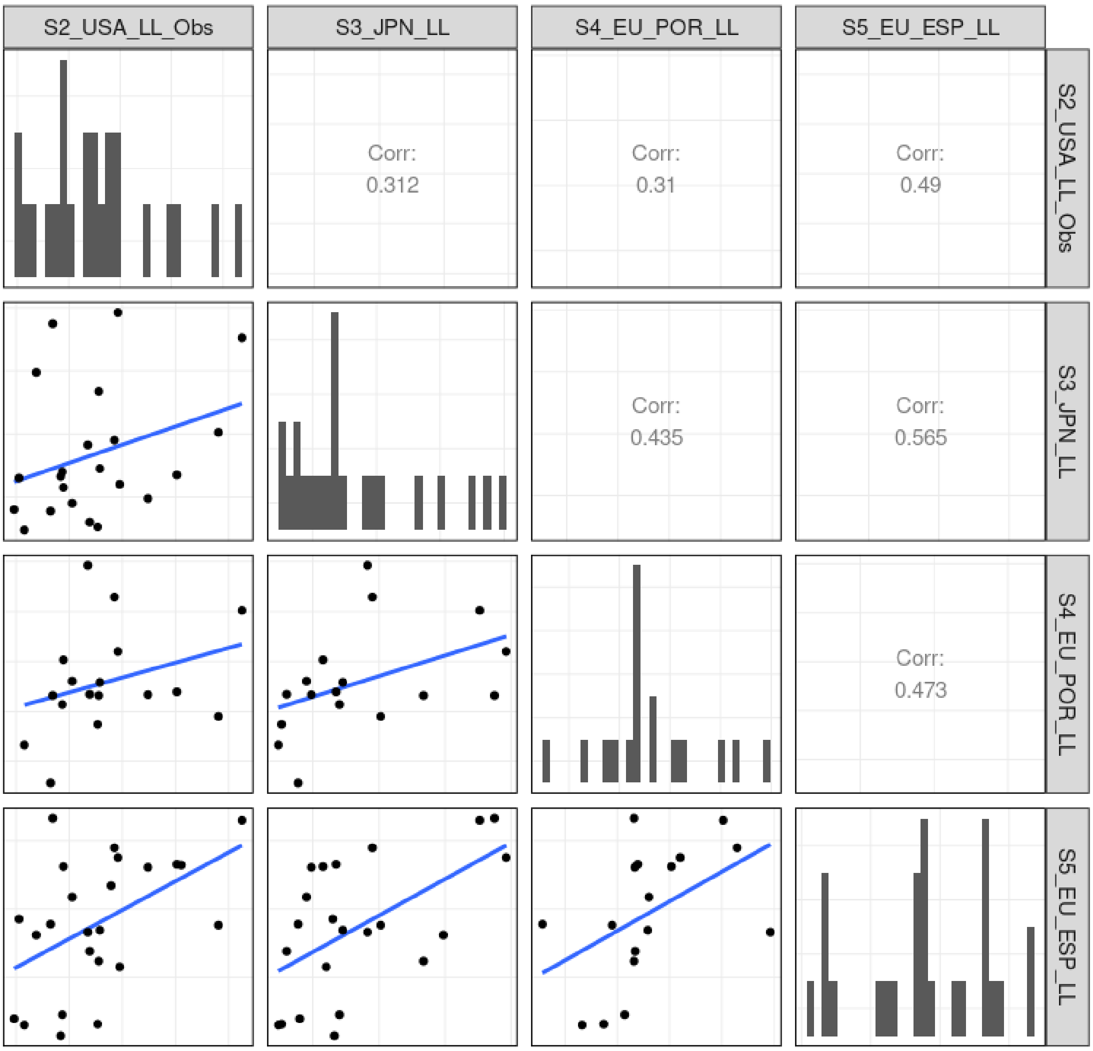
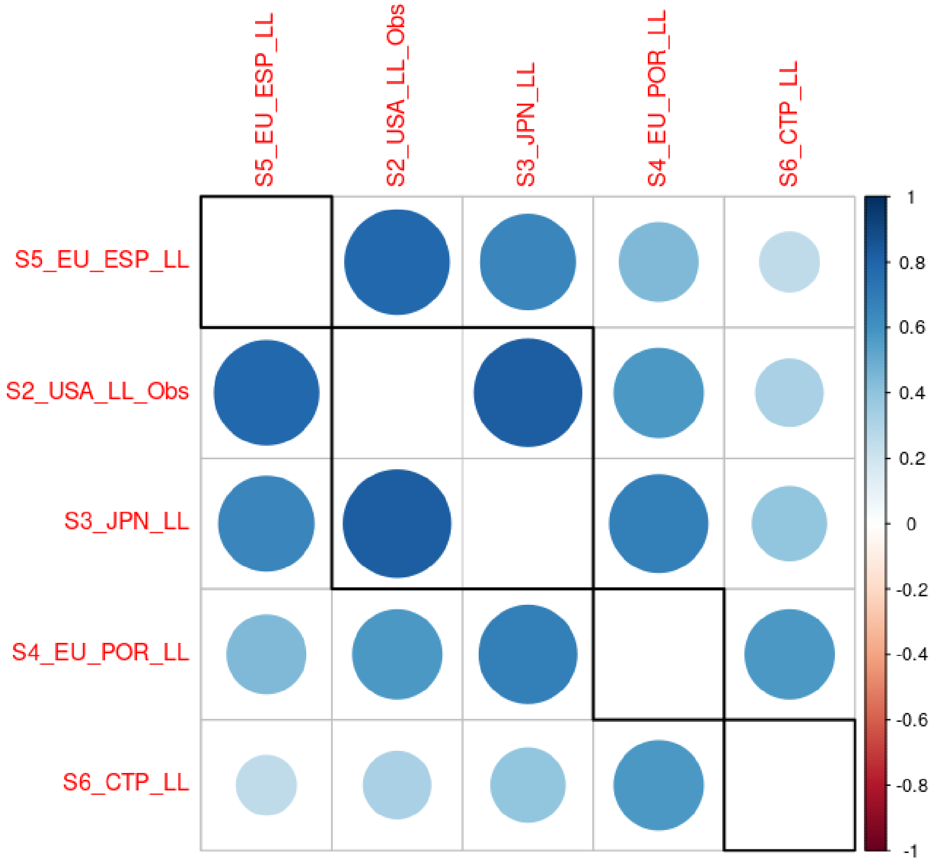


Figure 5. Pairwise scatter plots to examine correlations between Indices.



```
NULL
```

Figure 6. Plot of the correlation matrix for the CPUE indices, blue indicate a positive correlation and red negative. The order of the indices and the rectangular boxes are chosen based on a hierarchical cluster analysis using a set of dissimilarities for the indices.

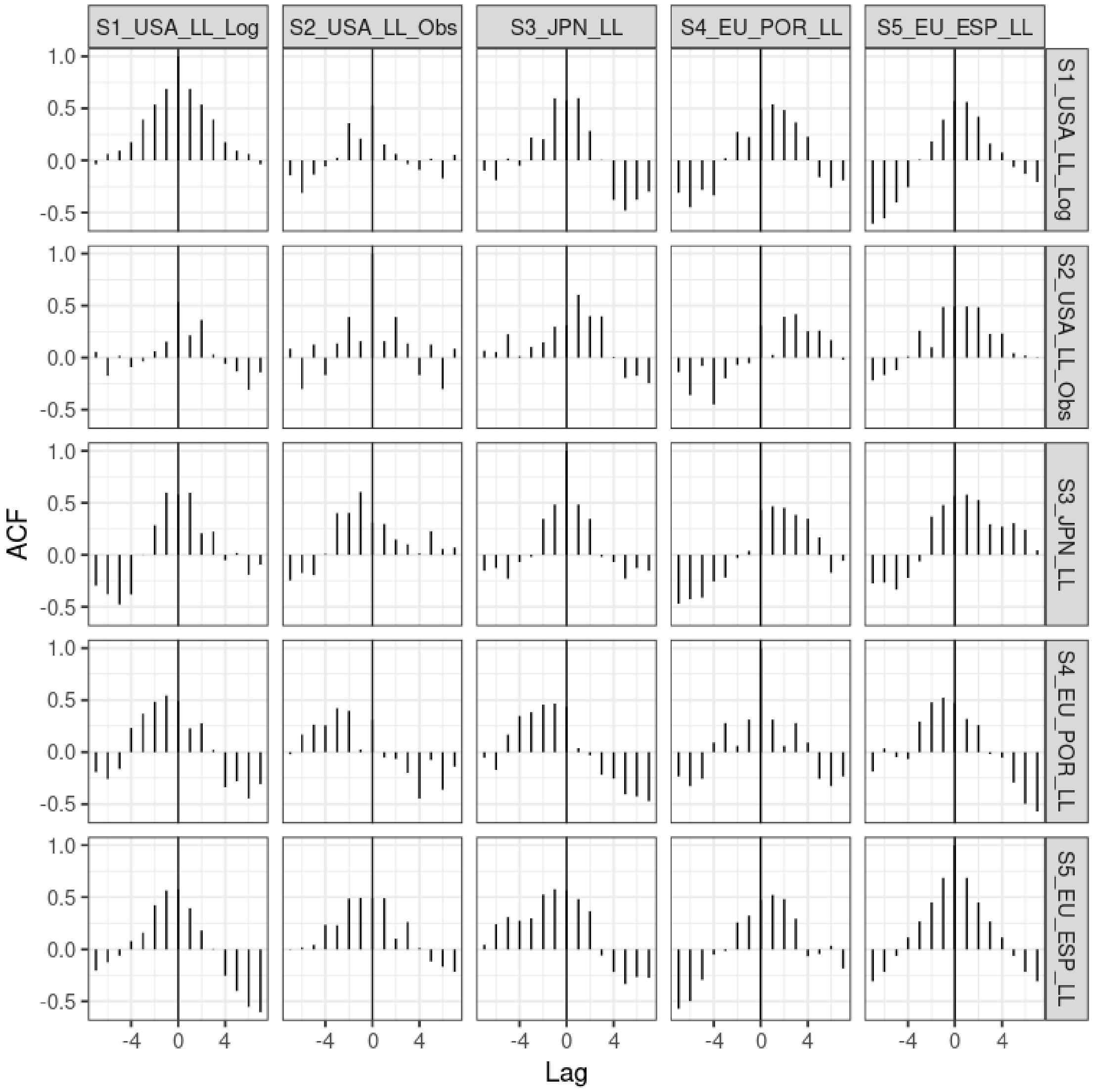


Figure 7. Cross correlations between indices, to identify potential lags due to year-class effects.

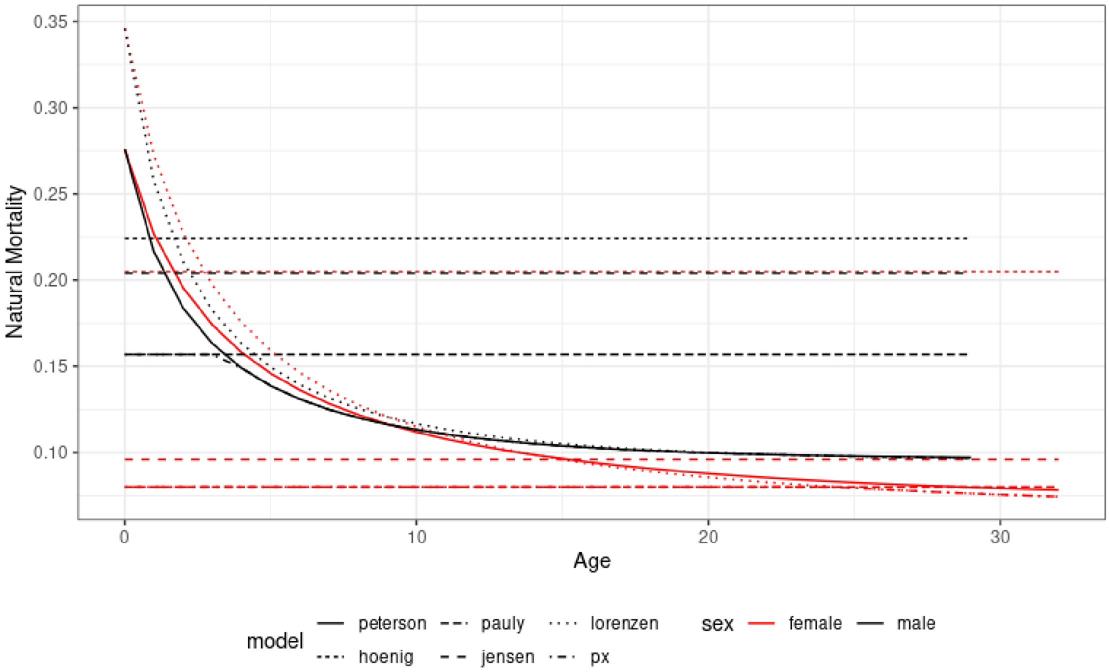


Figure 8. Natural mortality-at-age, showing hypotheses considered by Cortes (SCRS/2019/087).

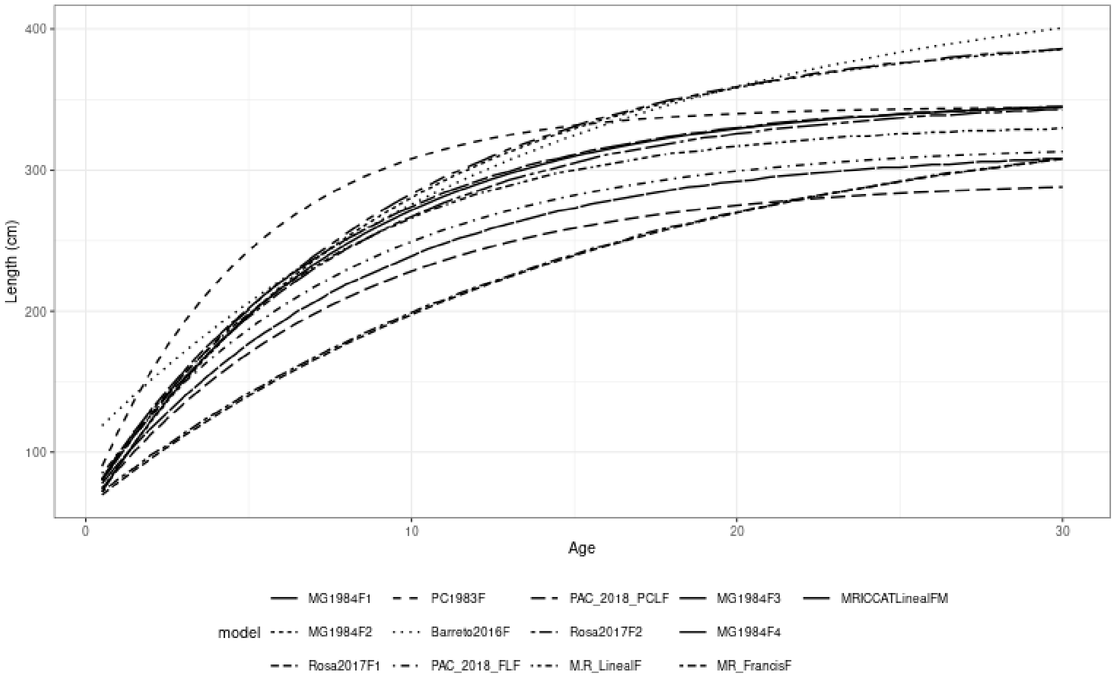


Figure 9. Length-at-age, hypotheses as summarised by Mejuto et al., (2021)

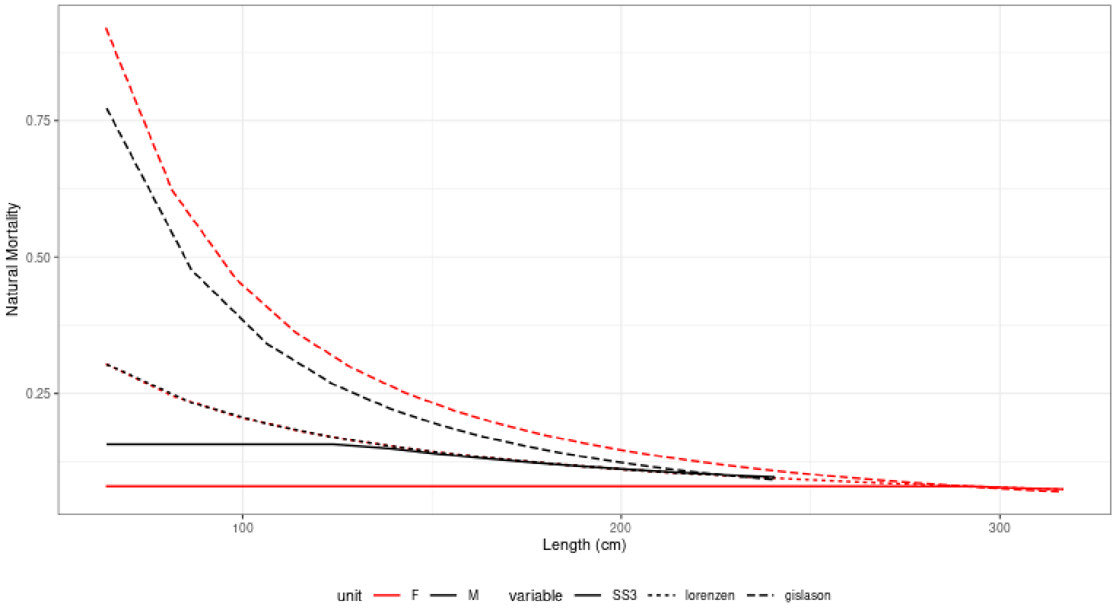


Figure 10. Natural mortality-at-length based on the assumptions used in the Stock Synthesis base case (run 1).

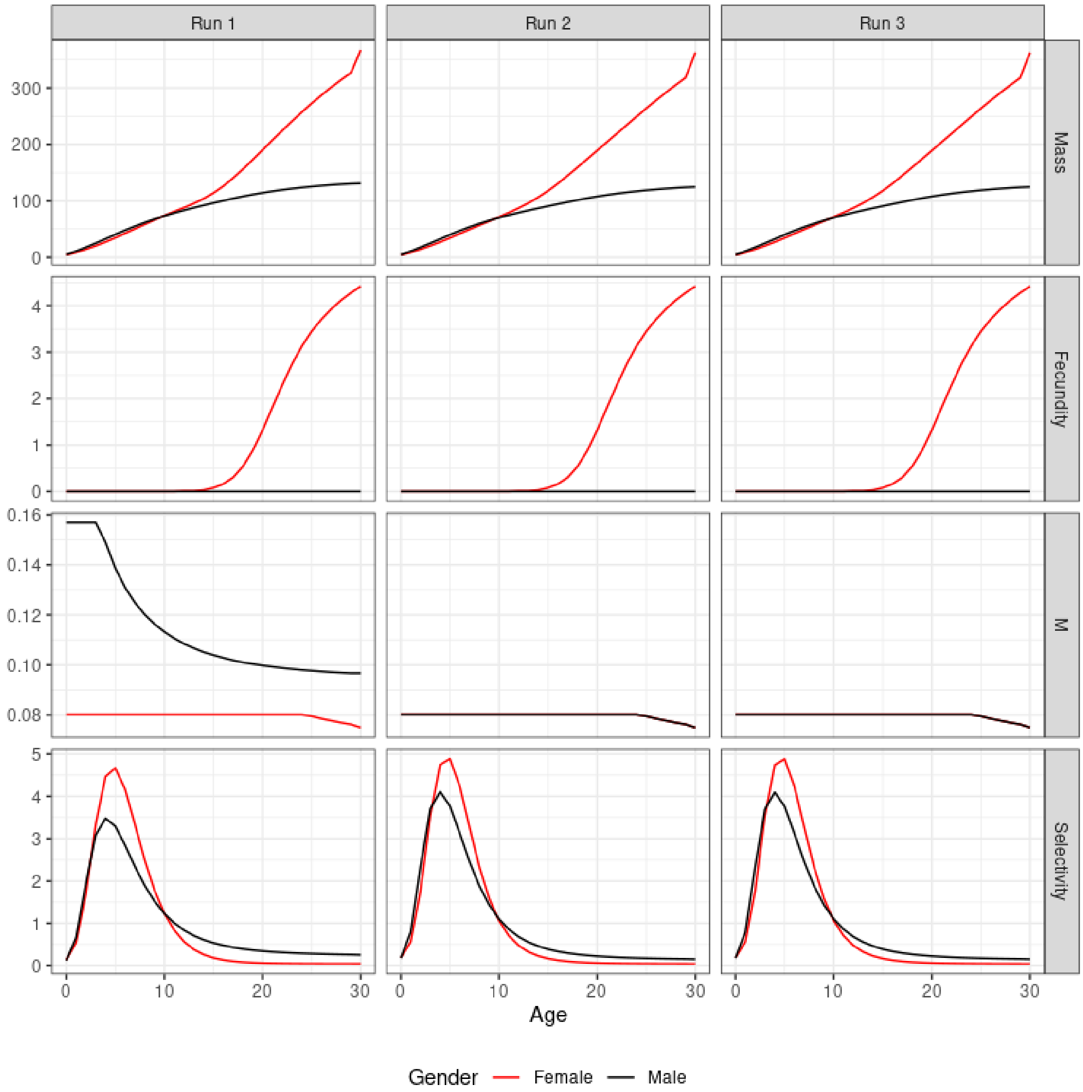


Figure 11. Vectors-at-age from the working group Stock Synthesis runs used to estimate per recruit reference points.

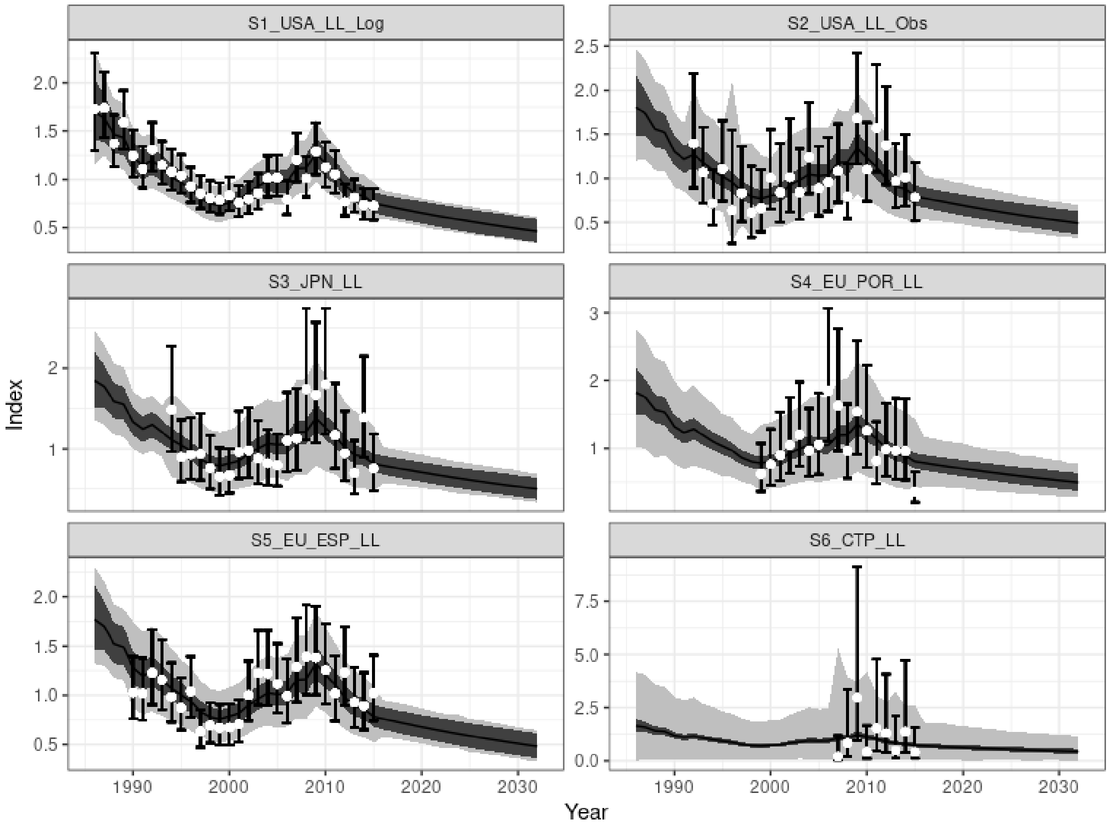


Figure 12. Fits to indices using JARA to examine trends, the ribbons show the common trend and the bars the observed indices with their 95% confidence intervals.

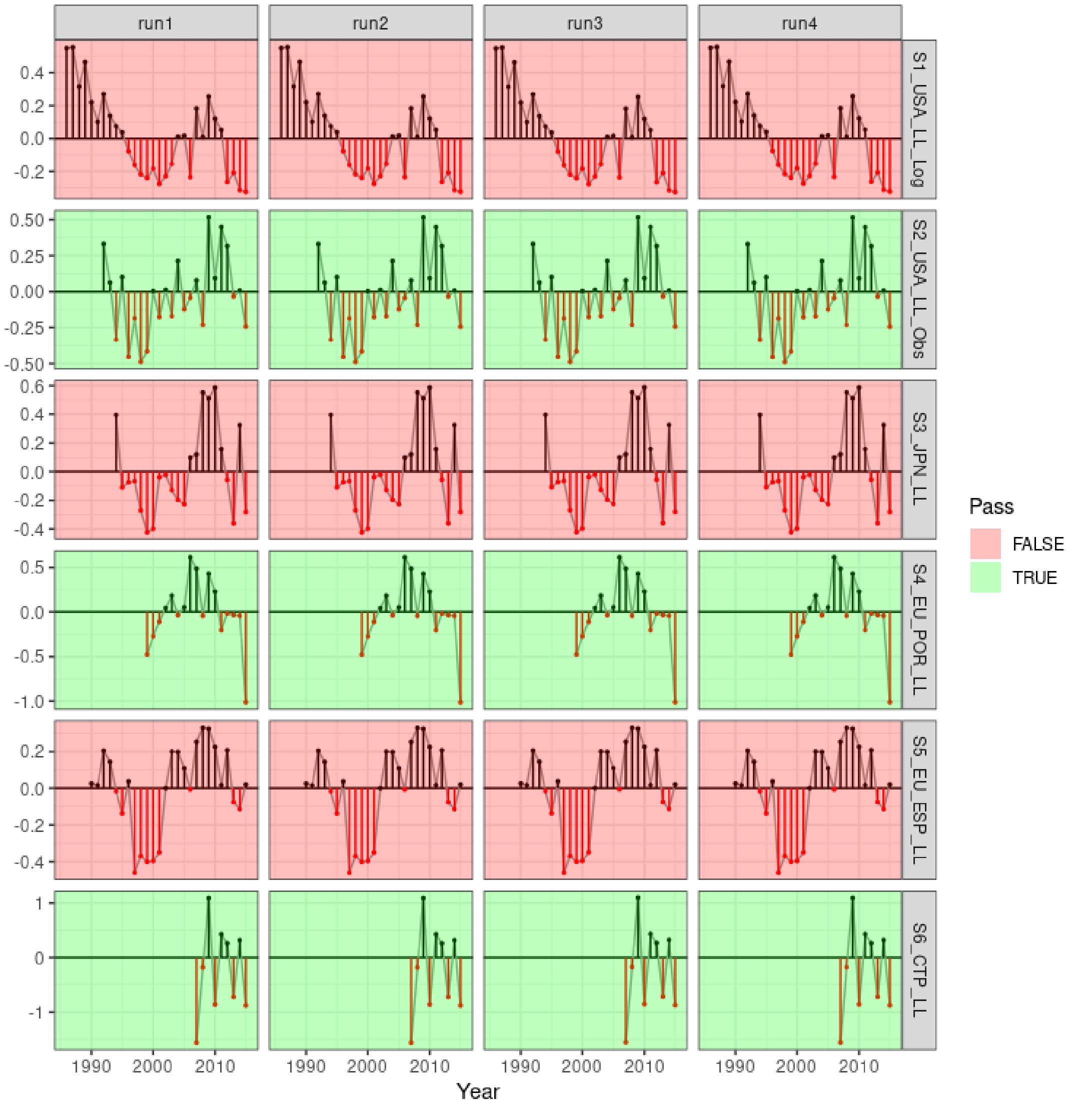


Figure 13. Runs tests for the trend analysis.

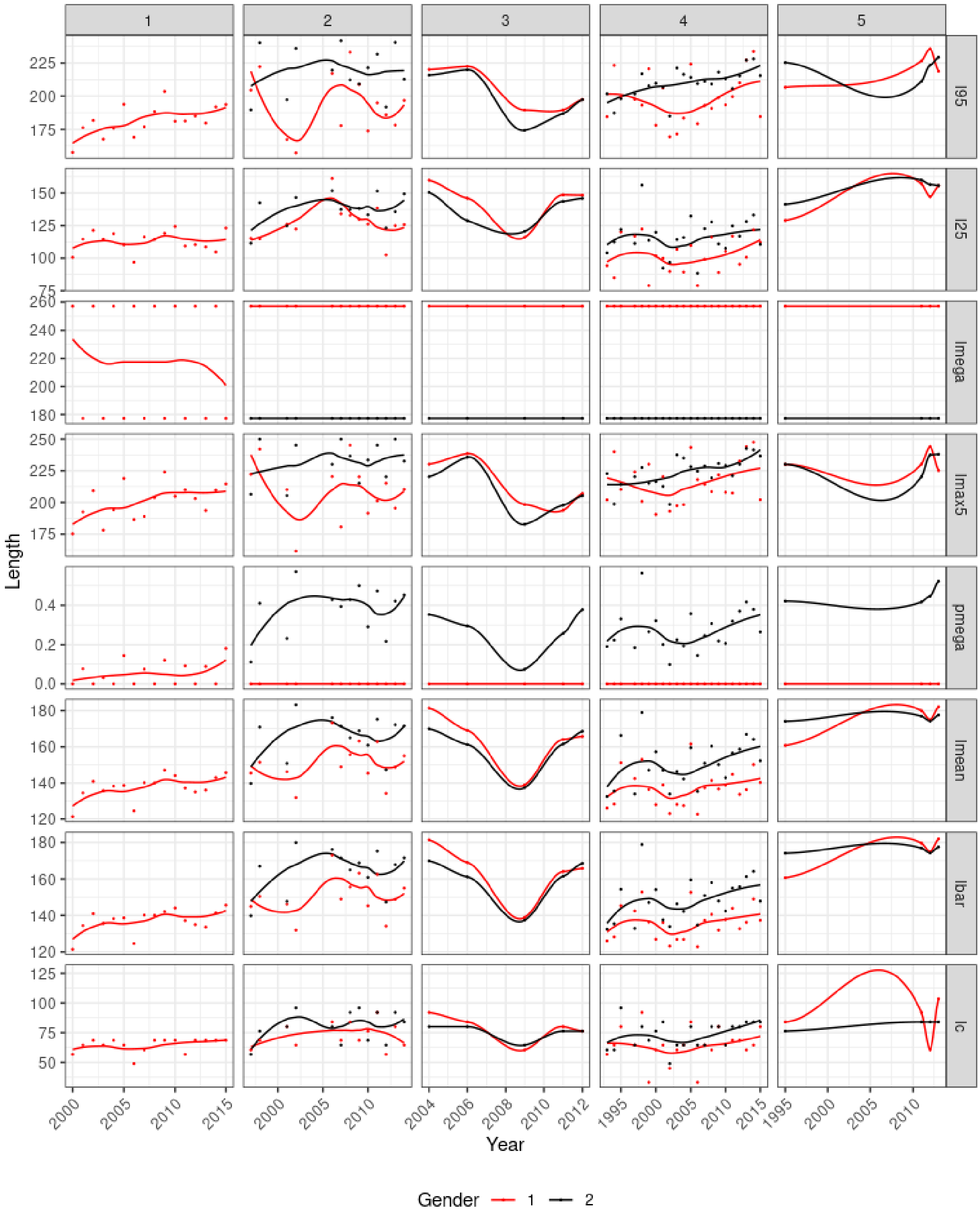


Figure 14. Length base indicators by fleet derived from the length composition data.

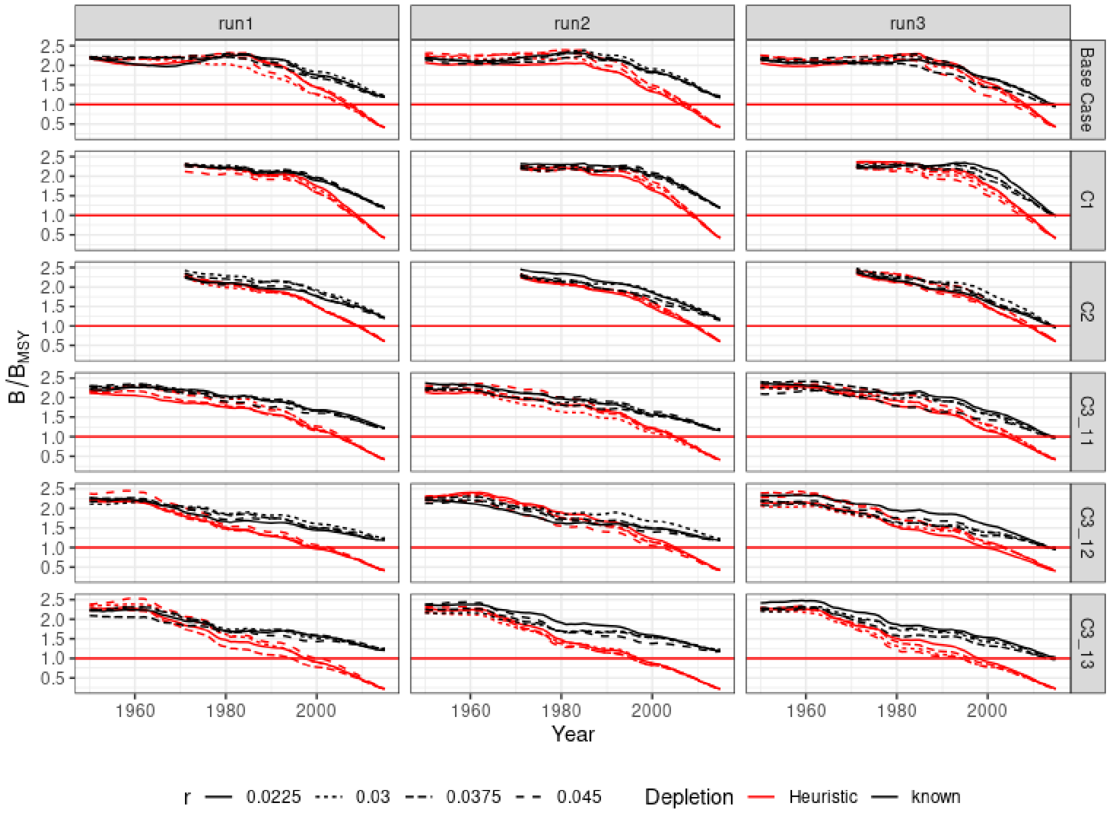


Figure 15 Trends in biomass relative to B_{MSY} from the catch only assessments for the alternative runs and priors for r. Depletion were either known, i.e. taken from the SS assessments or based on heuristics.

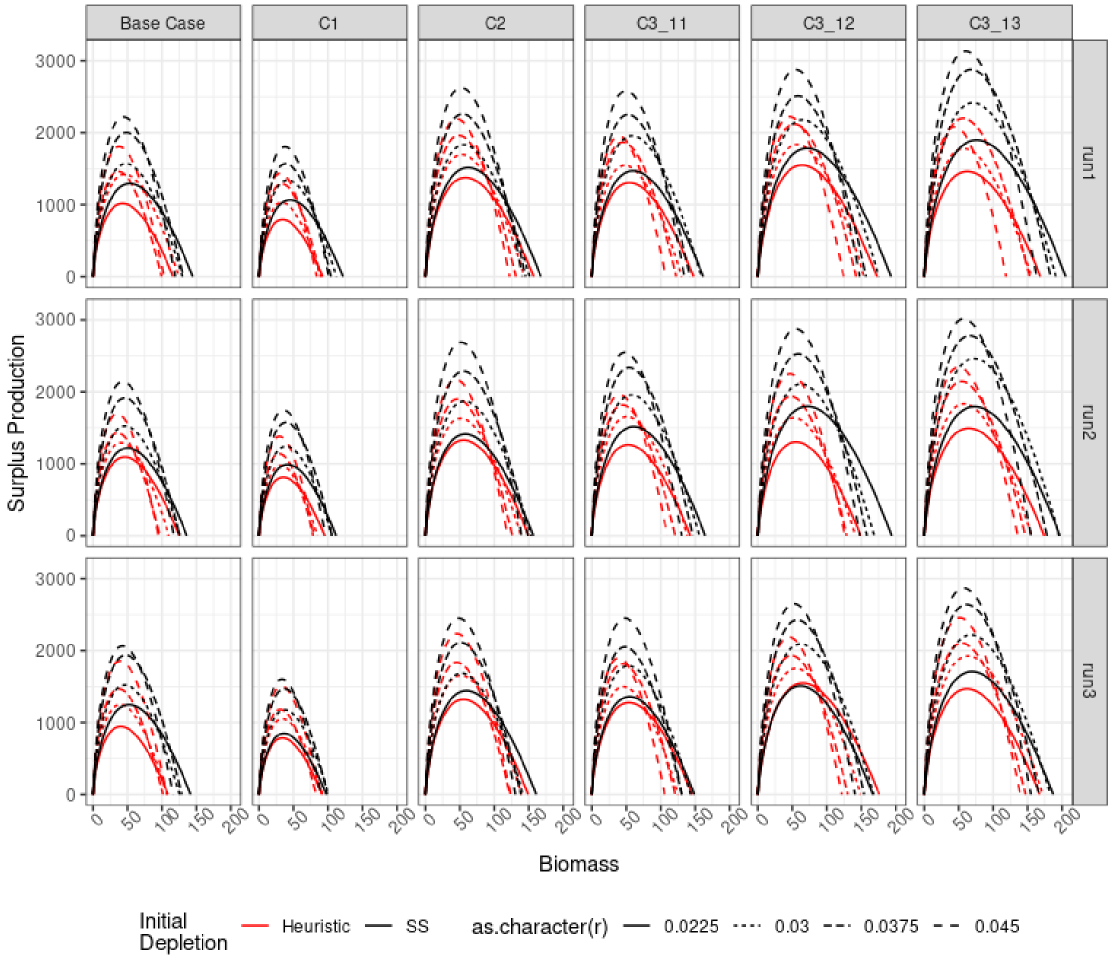


Figure 16. Production functions from catch only assessments.

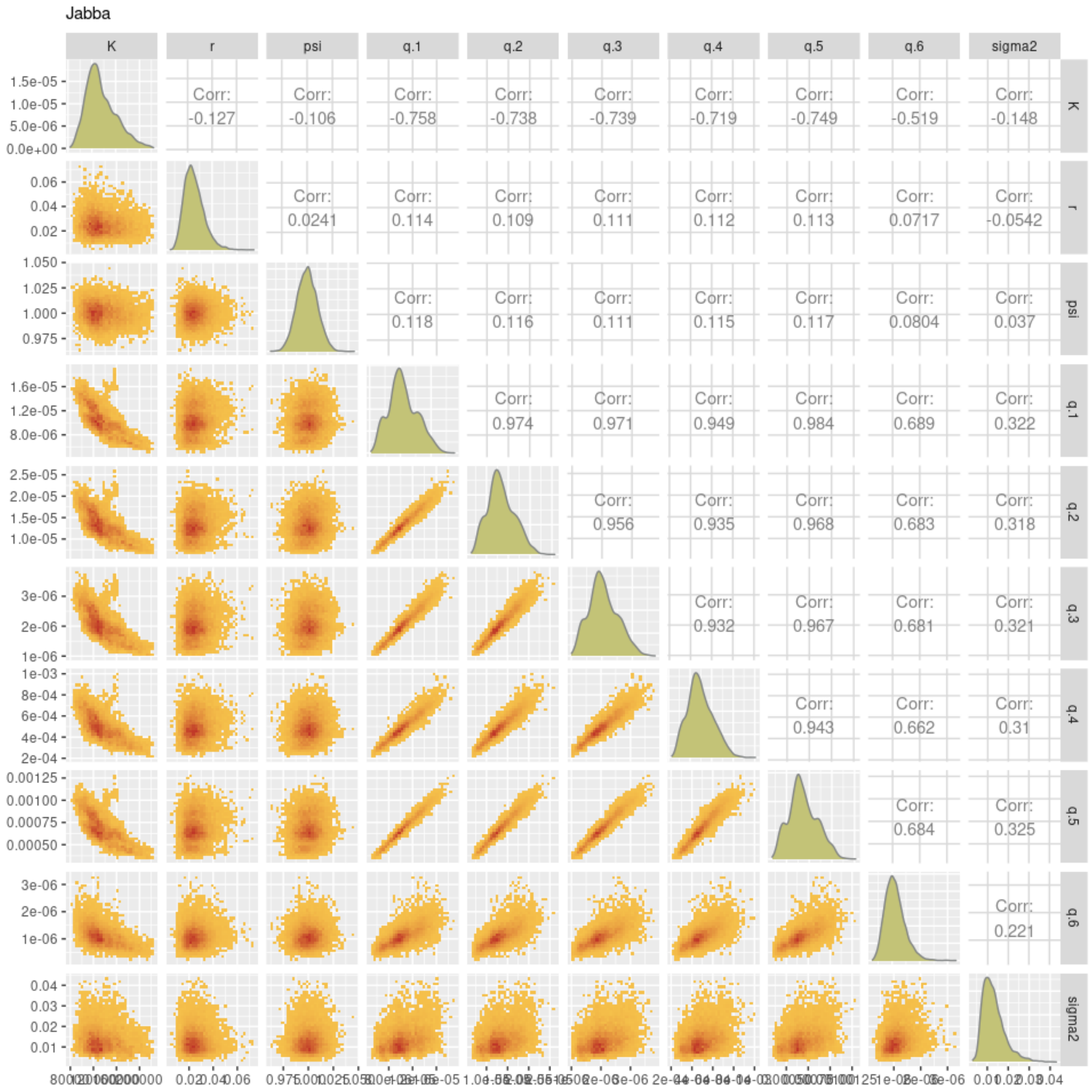


Figure 17. Parameter estimates and correlations for JABBA with SS catch scenario and r prior of 0.03.

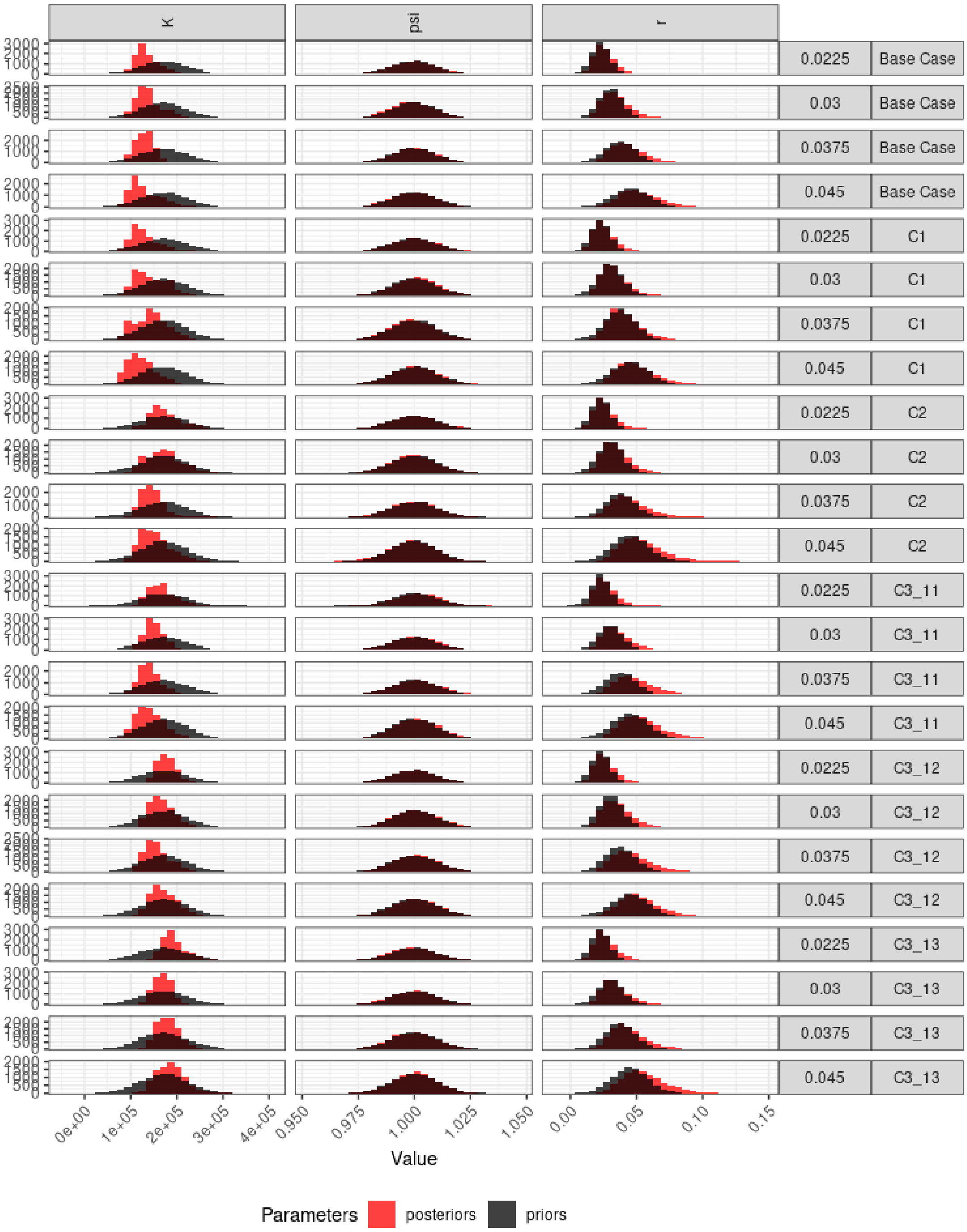


Figure 18. Posteriors and priors for JABBA assessments, by *r* and catch scenario.

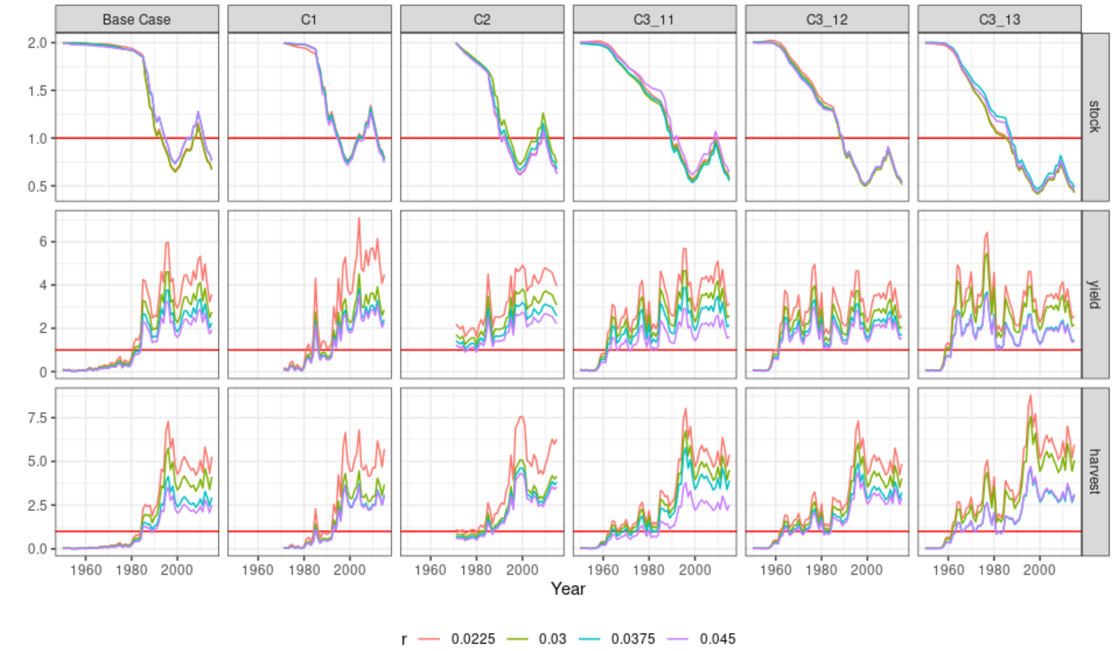


Figure 19. Time series relative to MSY benchmarks for JABBA scenarios.

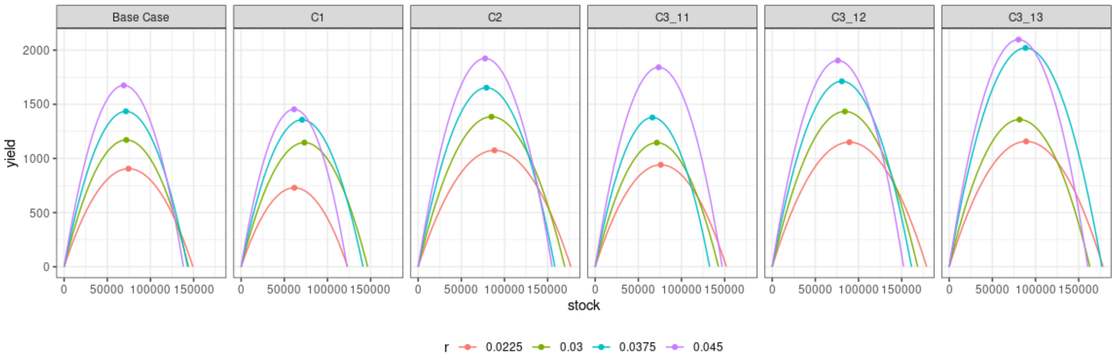


Figure 20. Production functions for the JABBA scenarios.

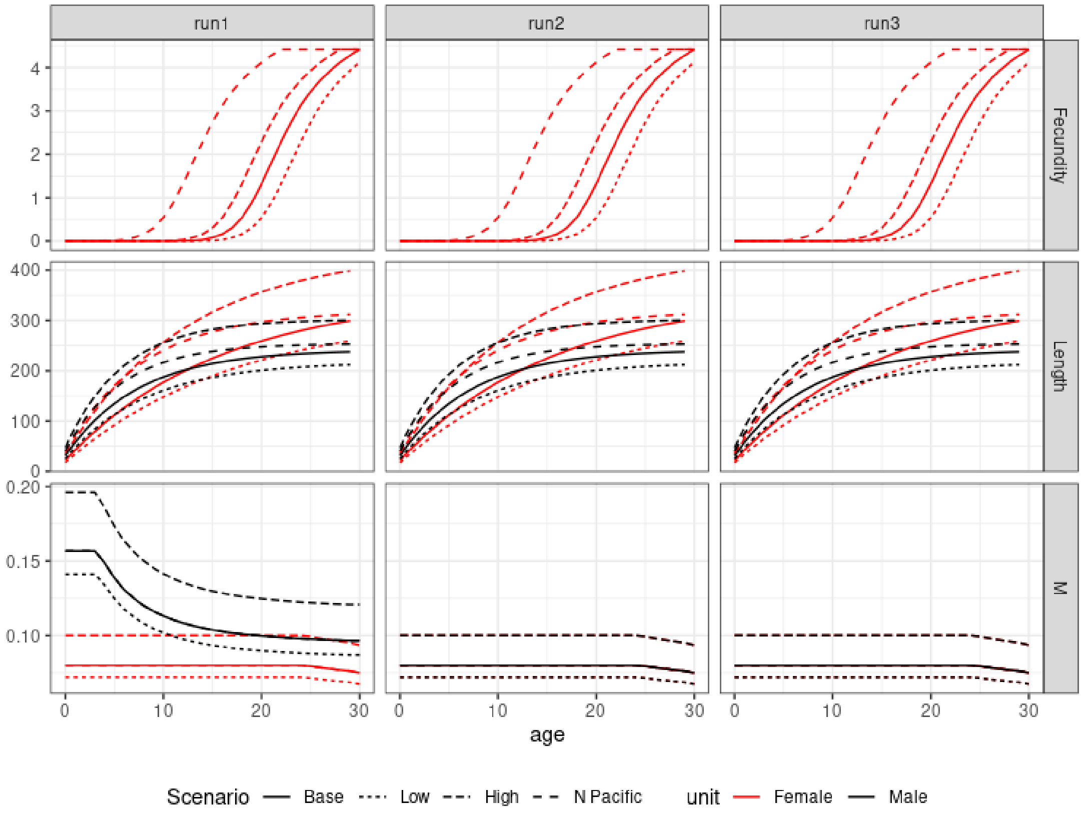


Figure 2.1 Scenarios for low and high productivity and North Pacific growth and maturity.

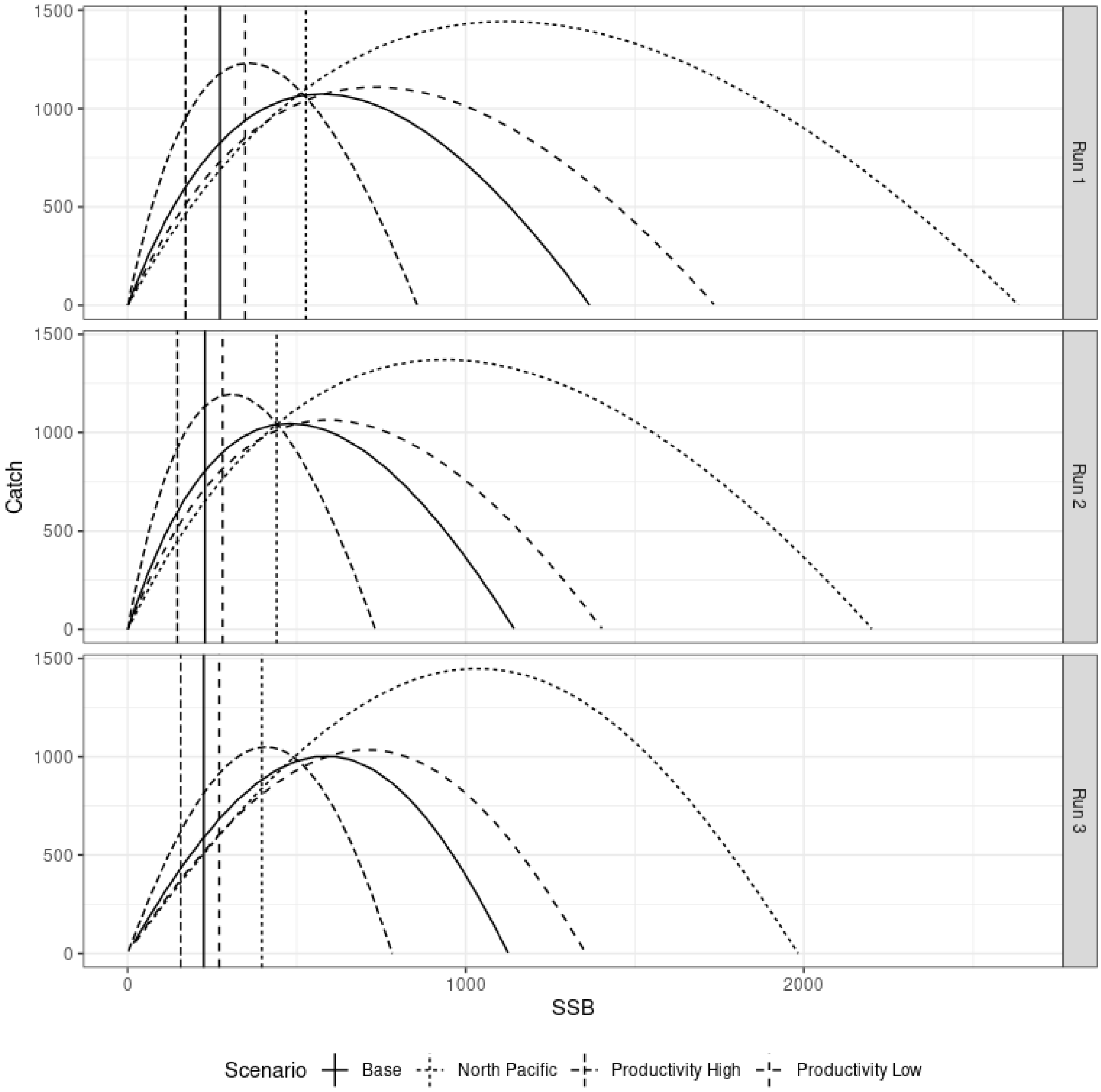


Figure 22. Production functions for the low and high productivity and North Pacific scenarios compared to the WG run, vertical lines are 20% of Virgin Biomass.

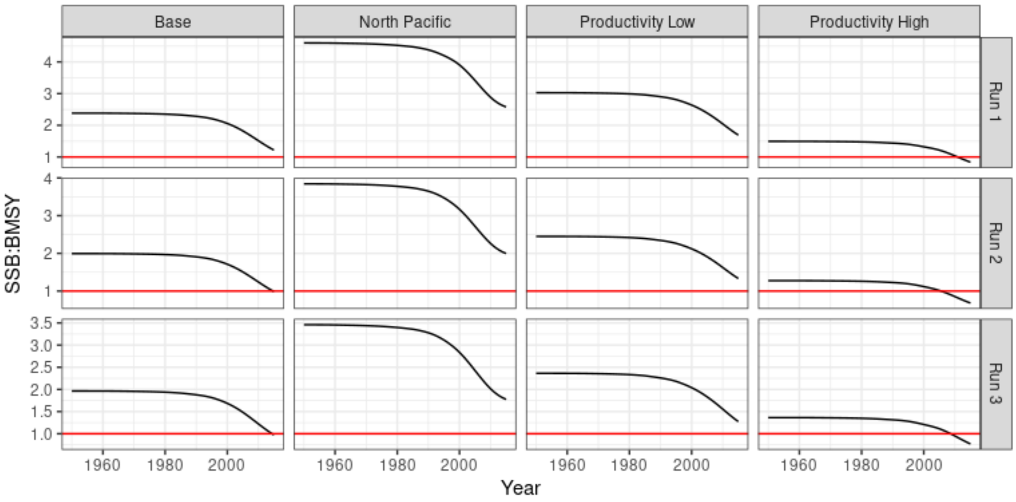


Figure 23. Time series of $SSB\colon B_{MSY}$ for the Stock Synthesis scenarios.

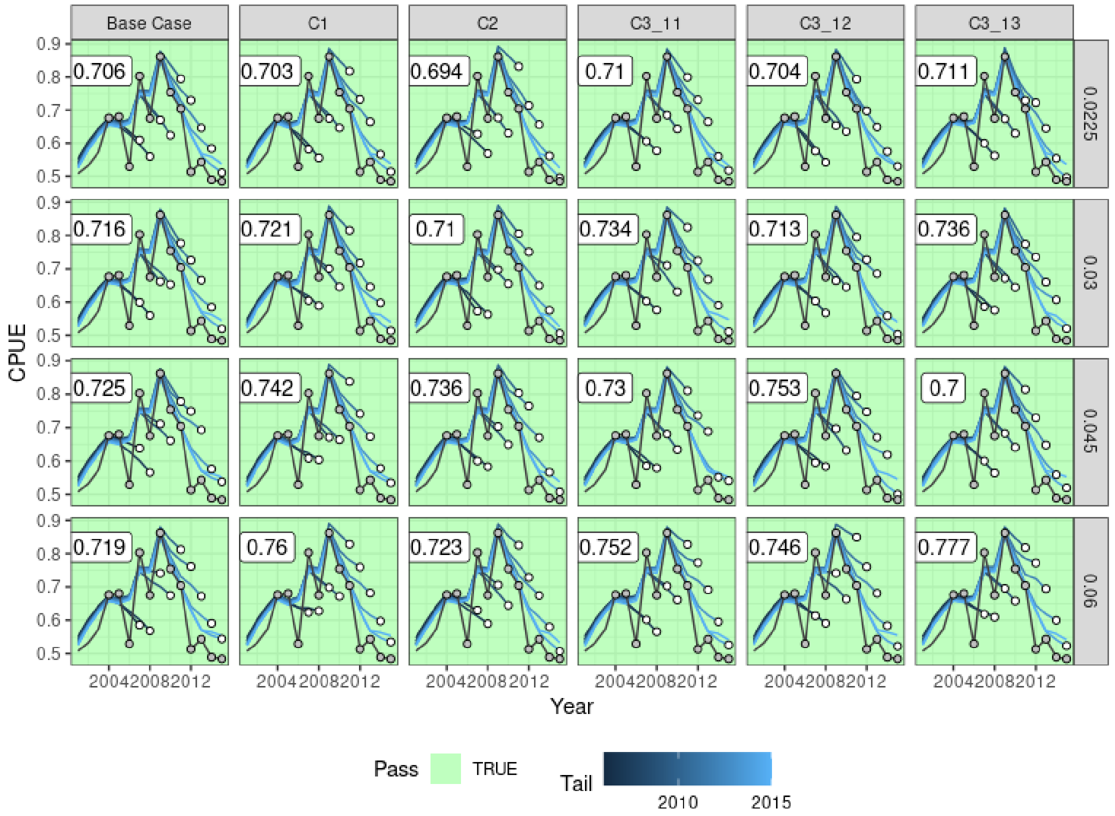


Figure 24 Hindcasts for biomass dynamic assessments S1_USA_LL_Log; black line and points are the observations, white points are the 1-step ahead estimates from the hindcast in the final year.

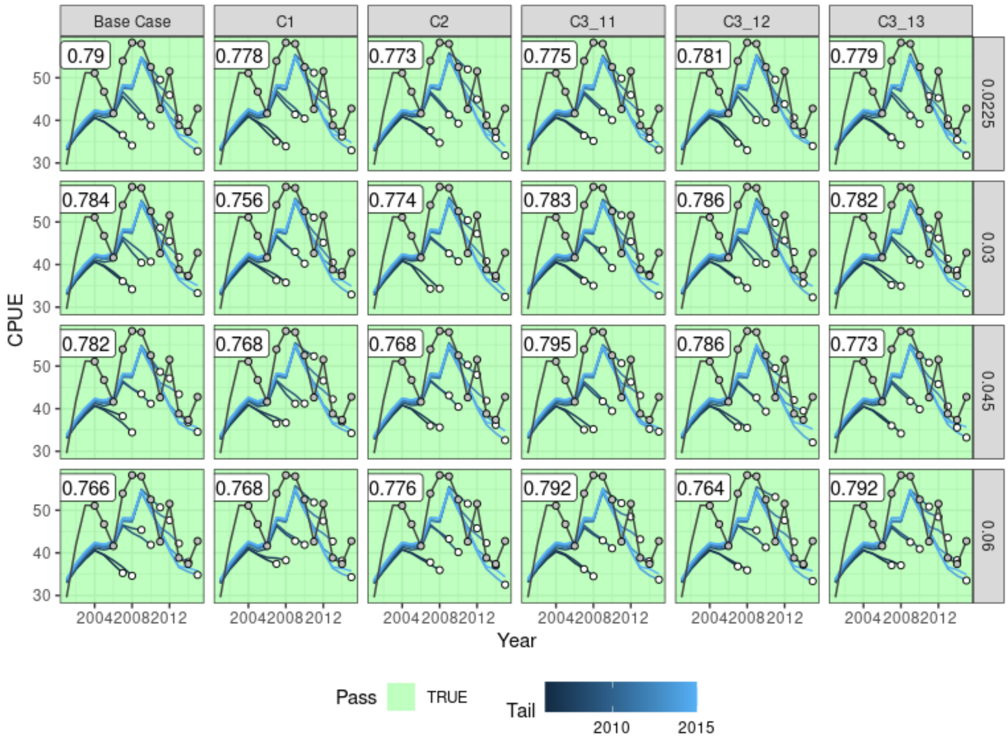


Figure 25. Hindcasts for biomass dynamic assessments S5_EU_ESP_LL.

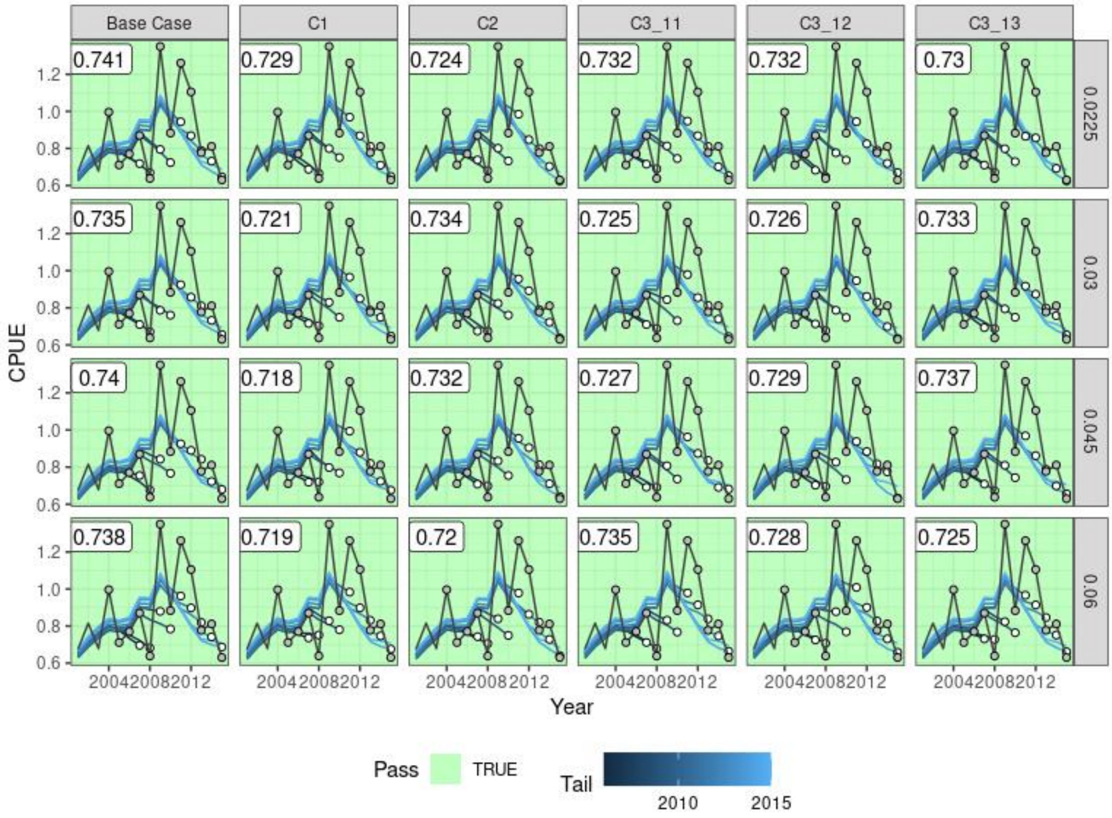


Figure 26. Hindcasts for biomass dynamic assessments S2_USA_LL_Obs.

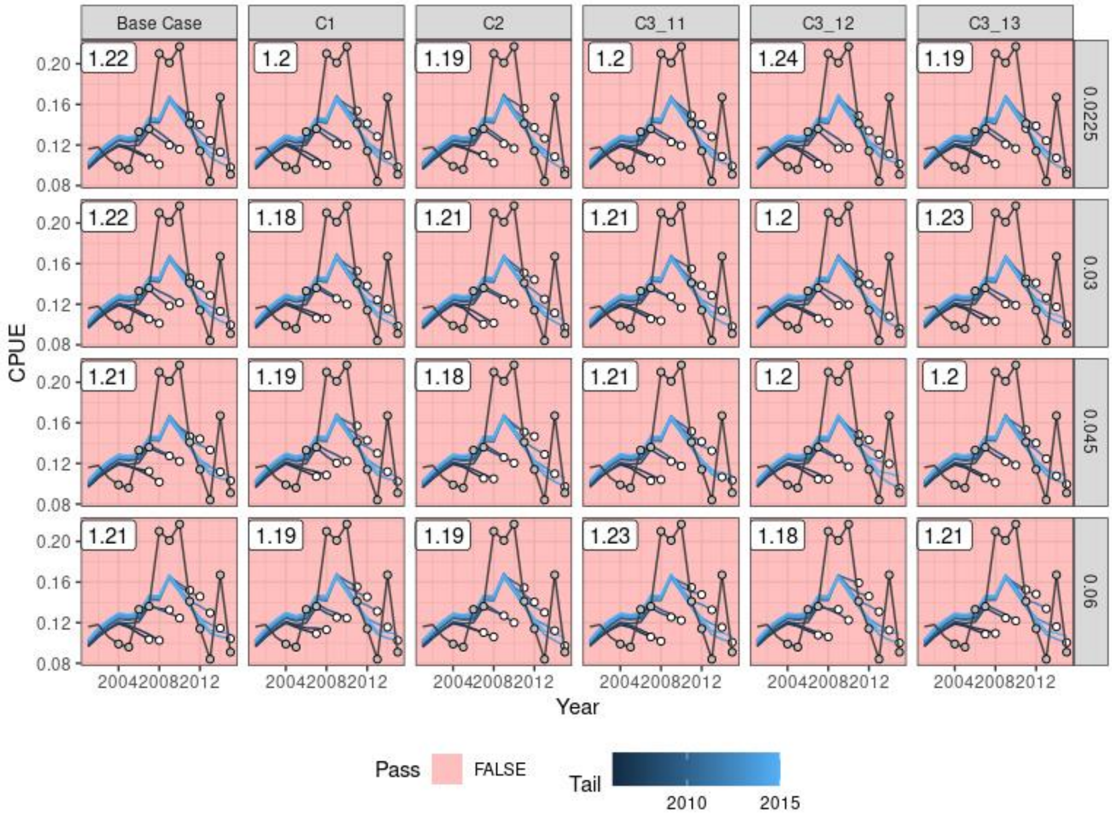


Figure 27. Hindcasts for biomass dynamic assessments S3_JPN_LL.

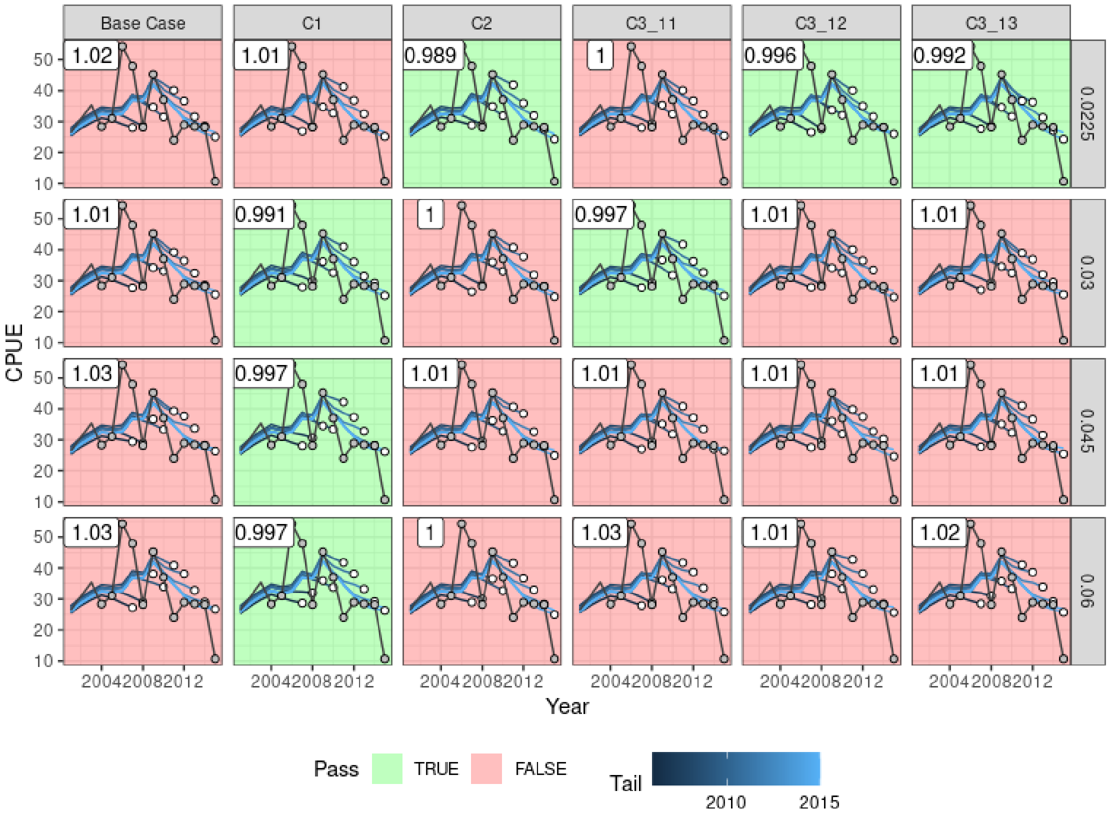


Figure 28. Hindcasts for biomass dynamic assessments S4_EU_POR_LL.

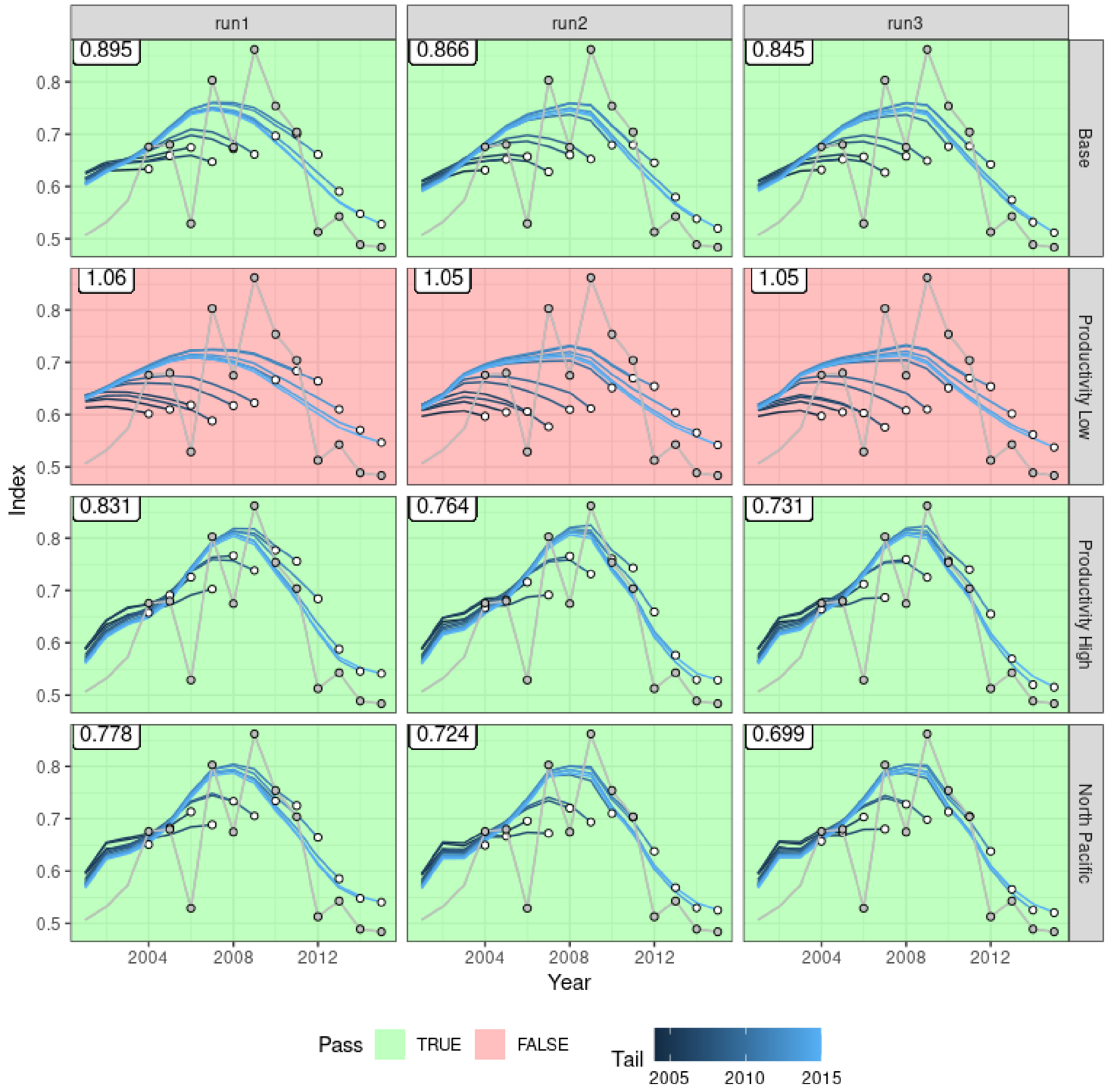


Figure 29. Hindcasts for Stock Synthesis assessments S1_USA_LL_Log.

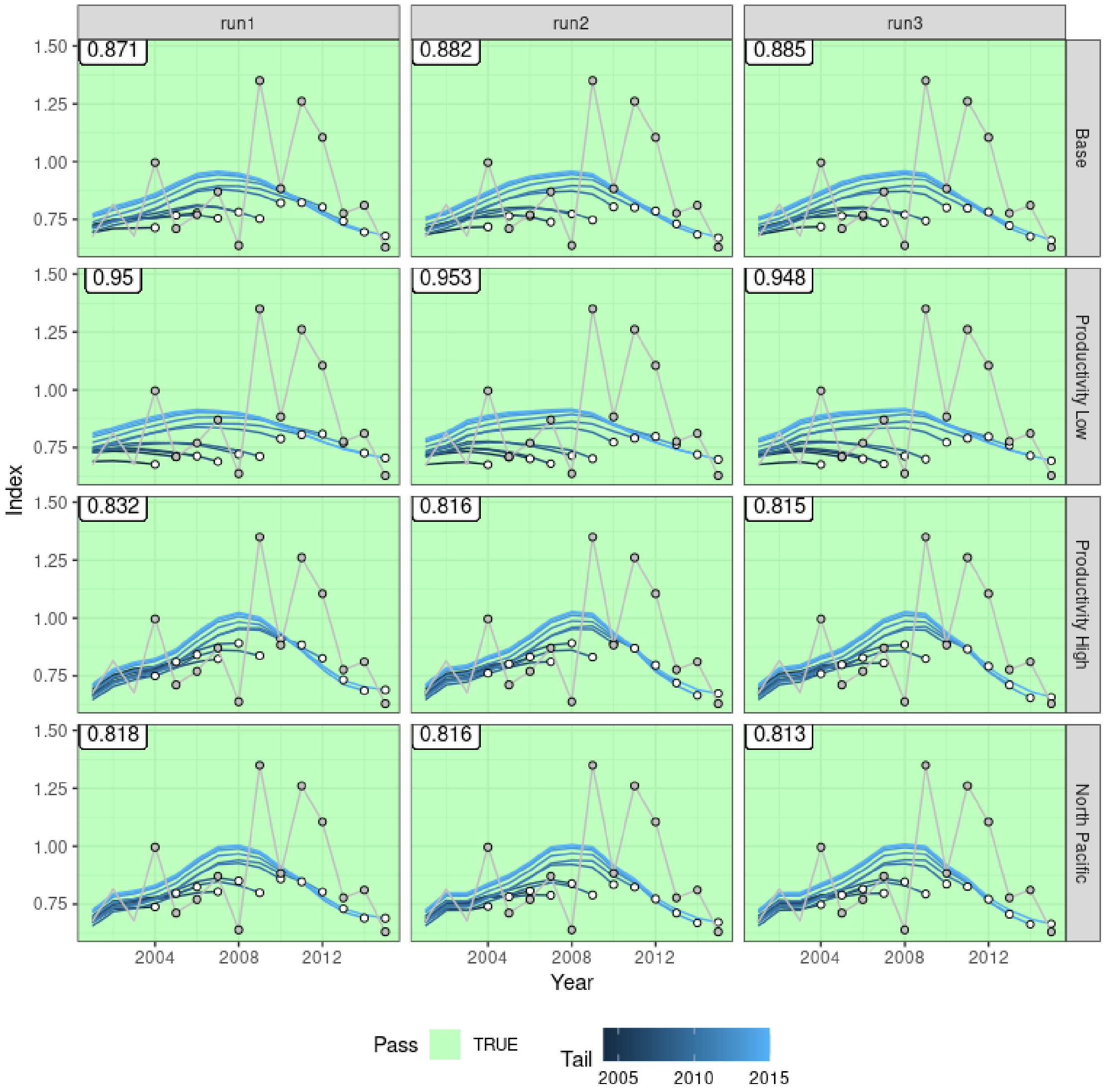


Figure 30. Hindcasts for Stock Synthesis assessments S2_USA_LL_Obs.

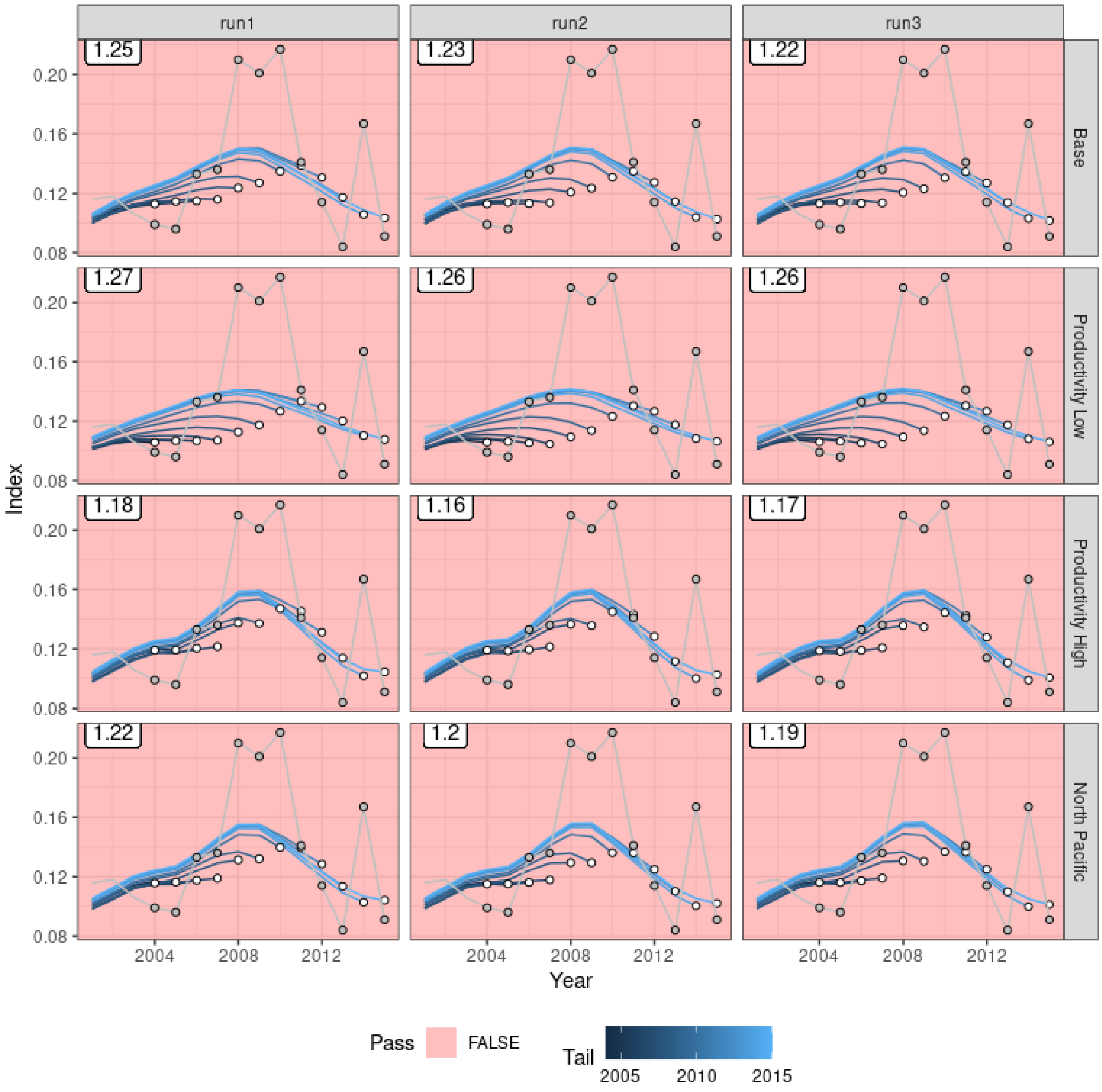


Figure 31. Hindcasts for Stock Synthesis assessments S3_JPN_LL.

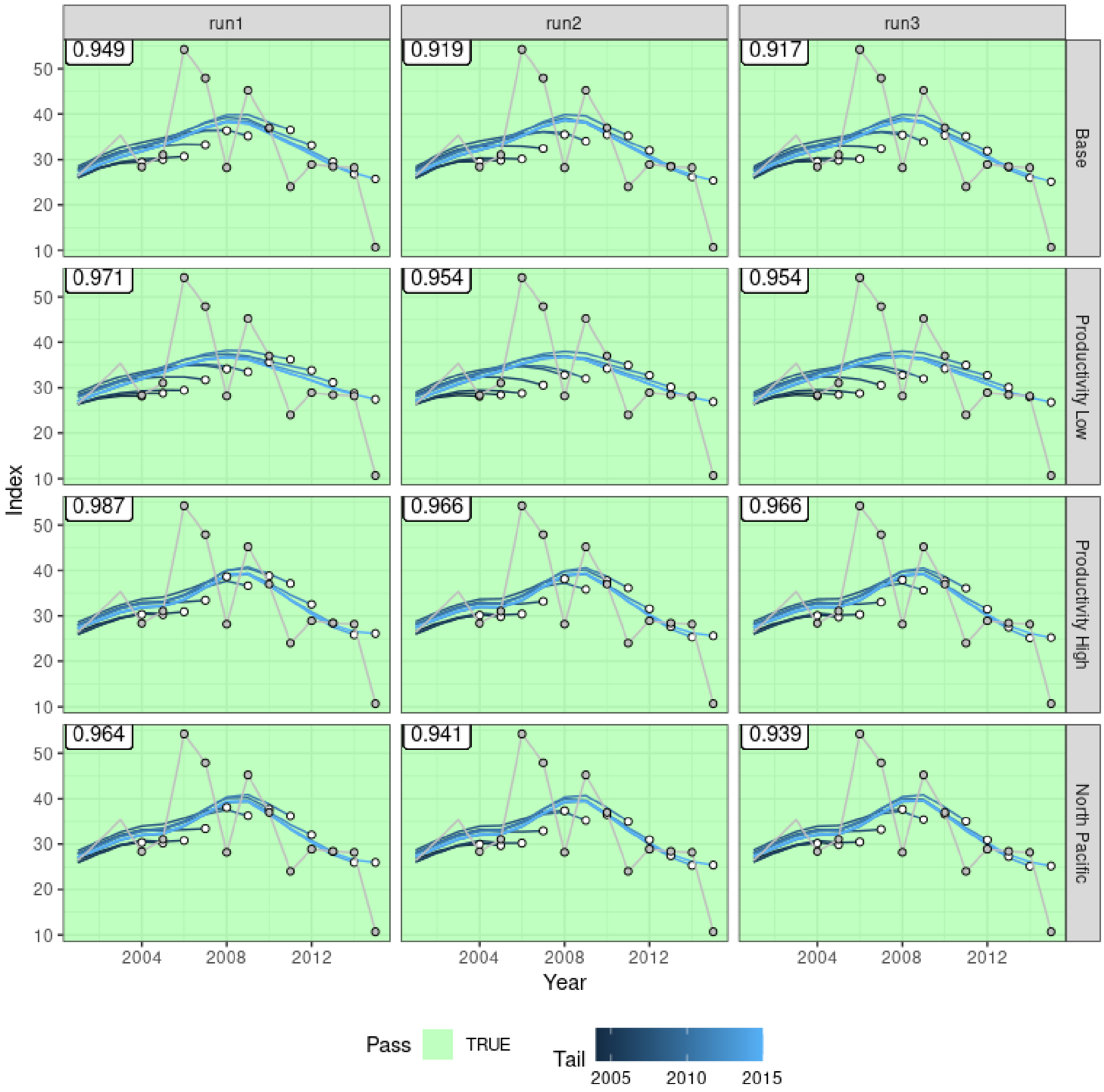


Figure 32. Hindcasts for Stock Synthesis assessments S4_EU_POR_LL.

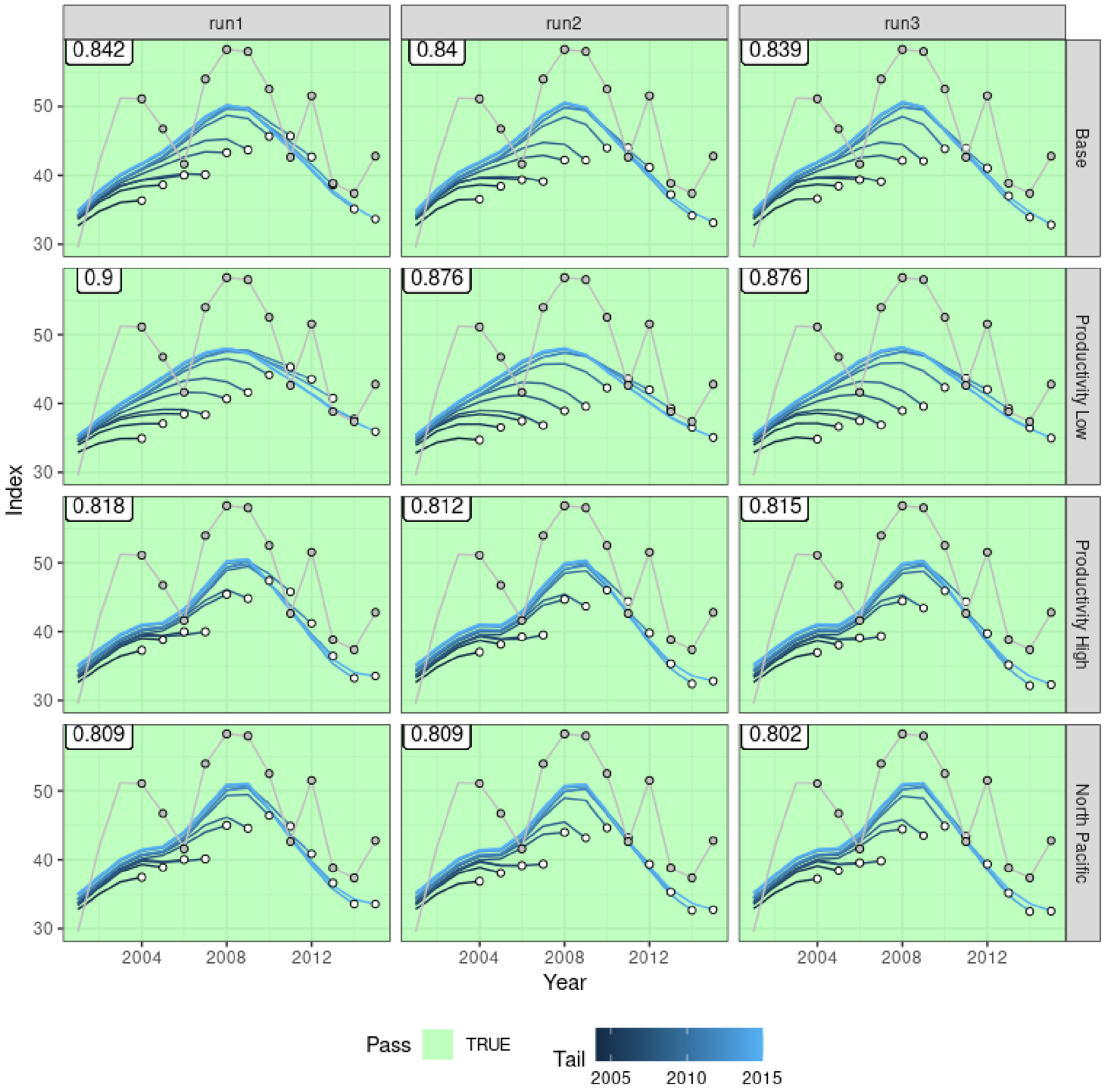


Figure 33. Hindcasts for Stock Synthesis assessments S5_EU_ESP_LL.

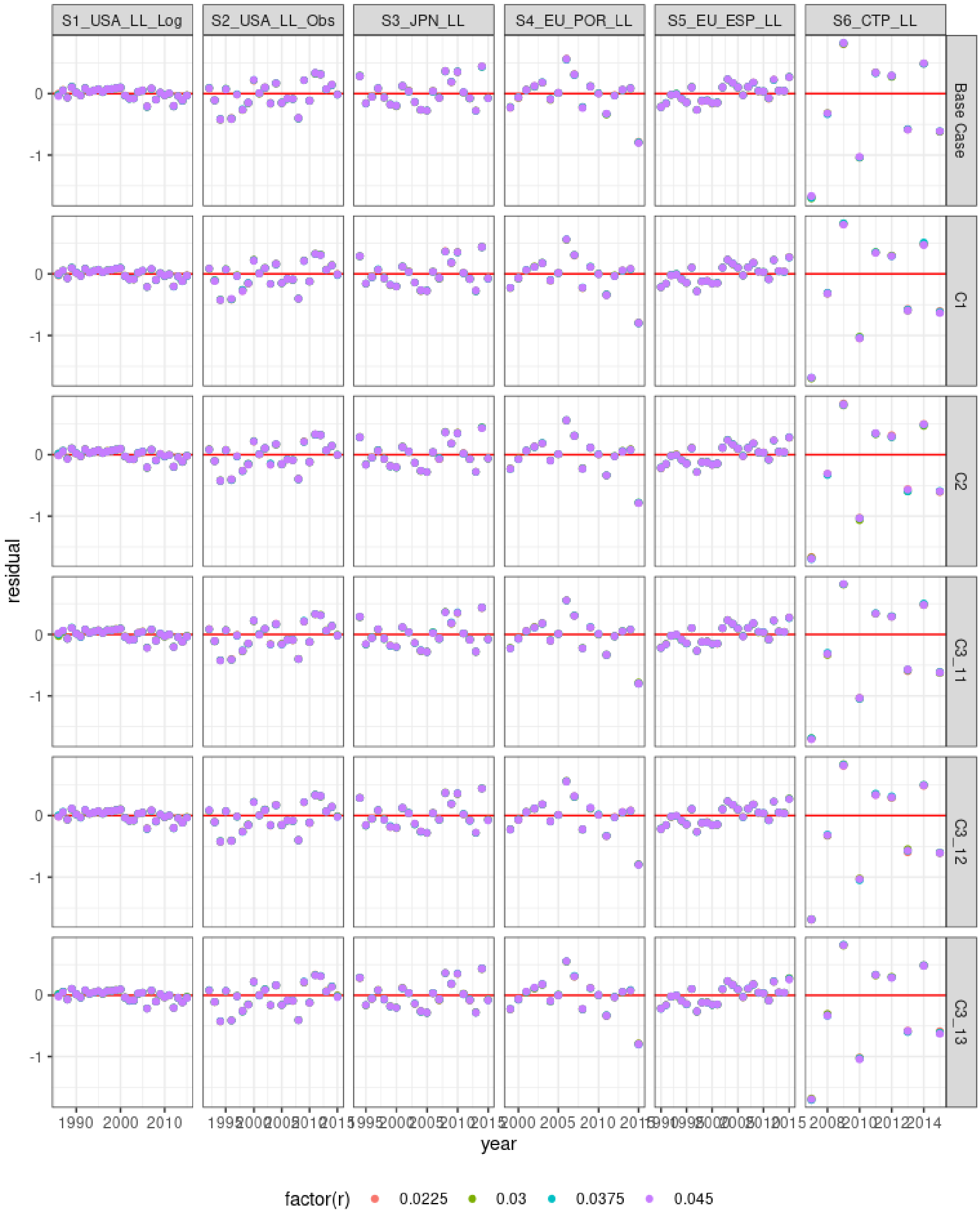


Figure 34. CPUE residuals for Jabba.

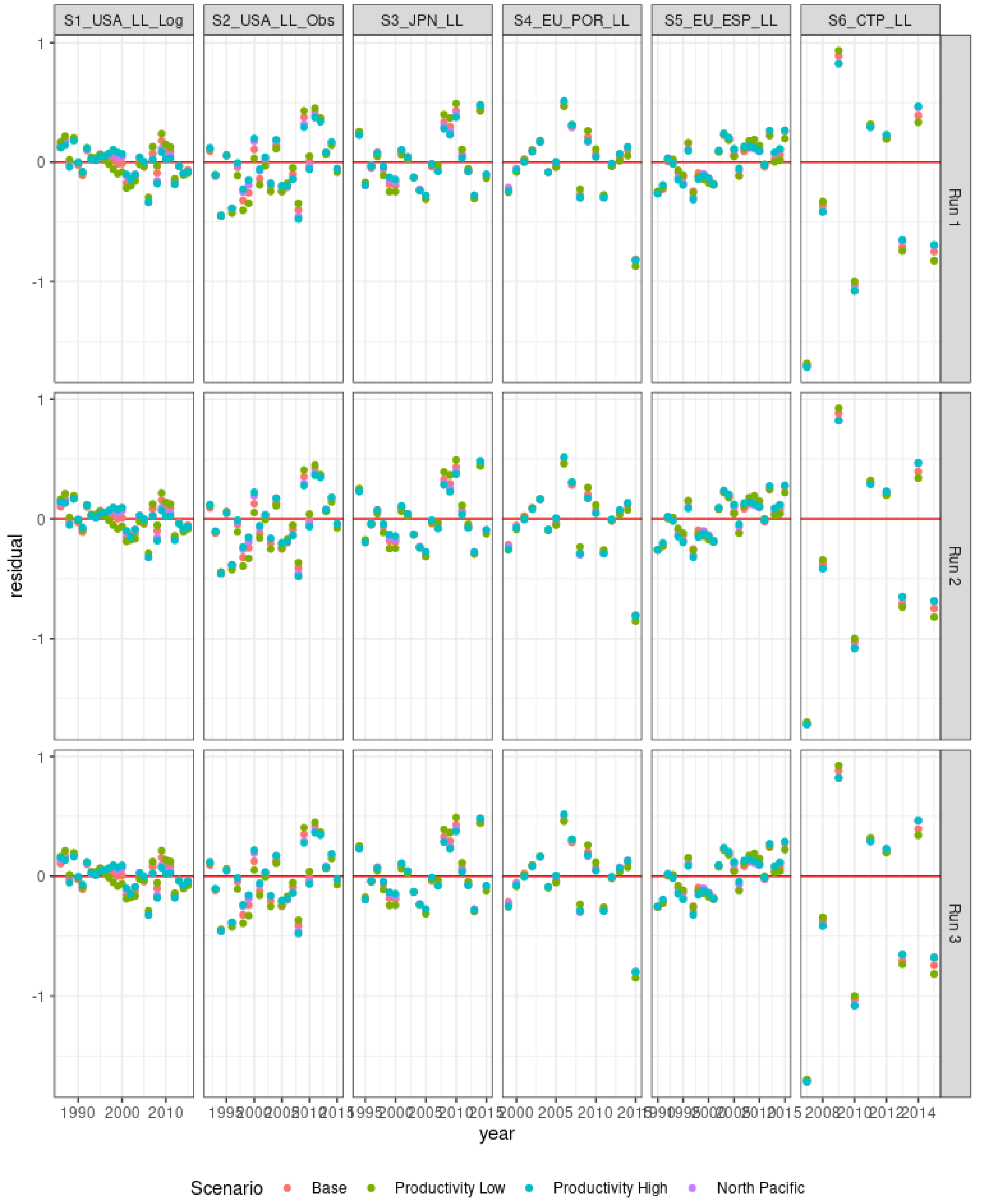


Figure 35. CPUE residuals for Stock Synthesis.

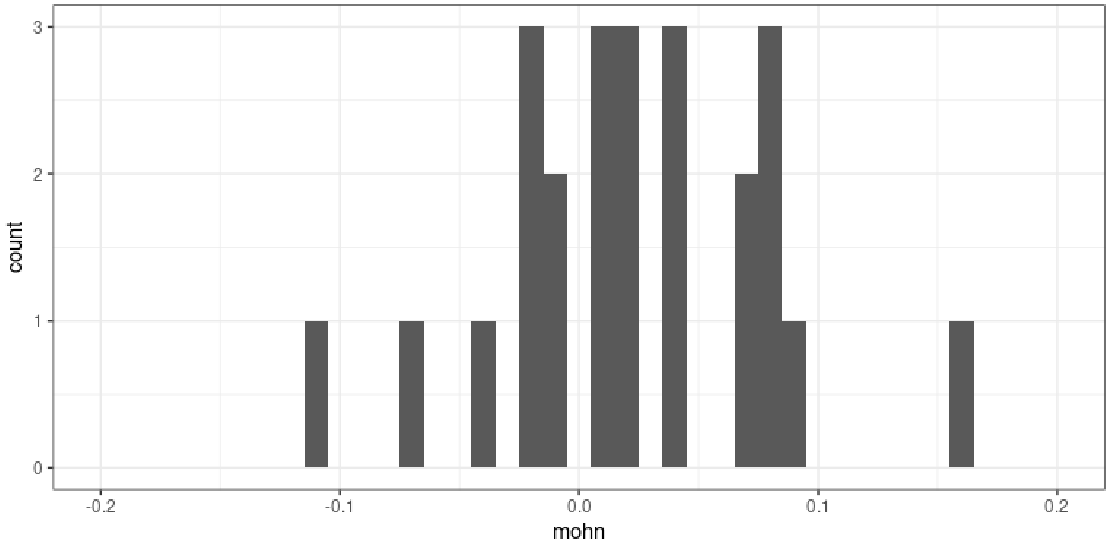


Figure 36. Mohn's ρ for JABBA, acceptable range is $[-0.15,2]$.

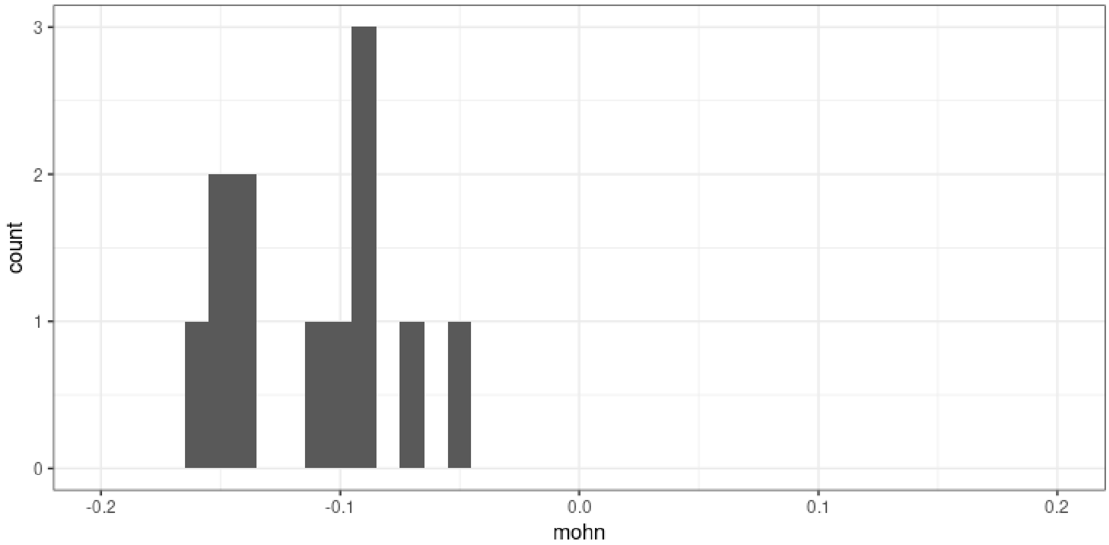


Figure 37. Mohn's ρ for Stock Synthesis, acceptable range is $[-0.15,2]$.

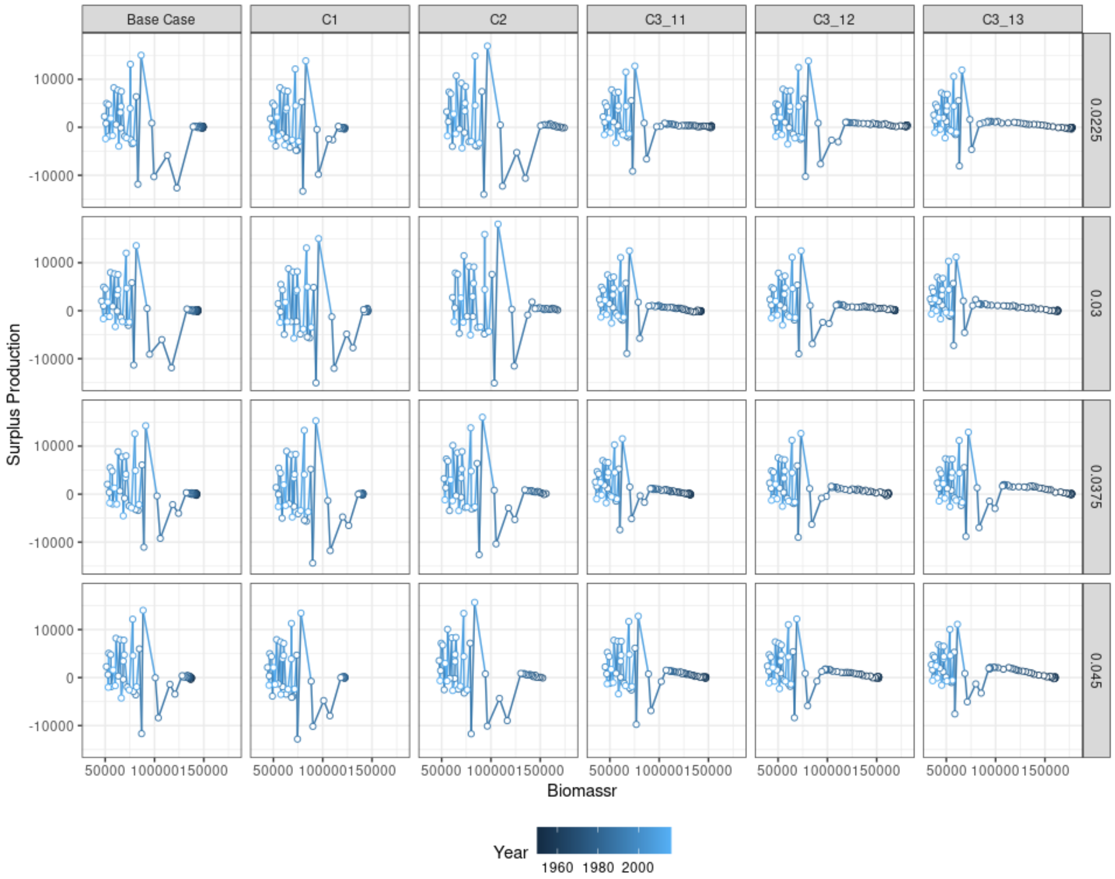


Figure 38. Surplus production for Jabba.

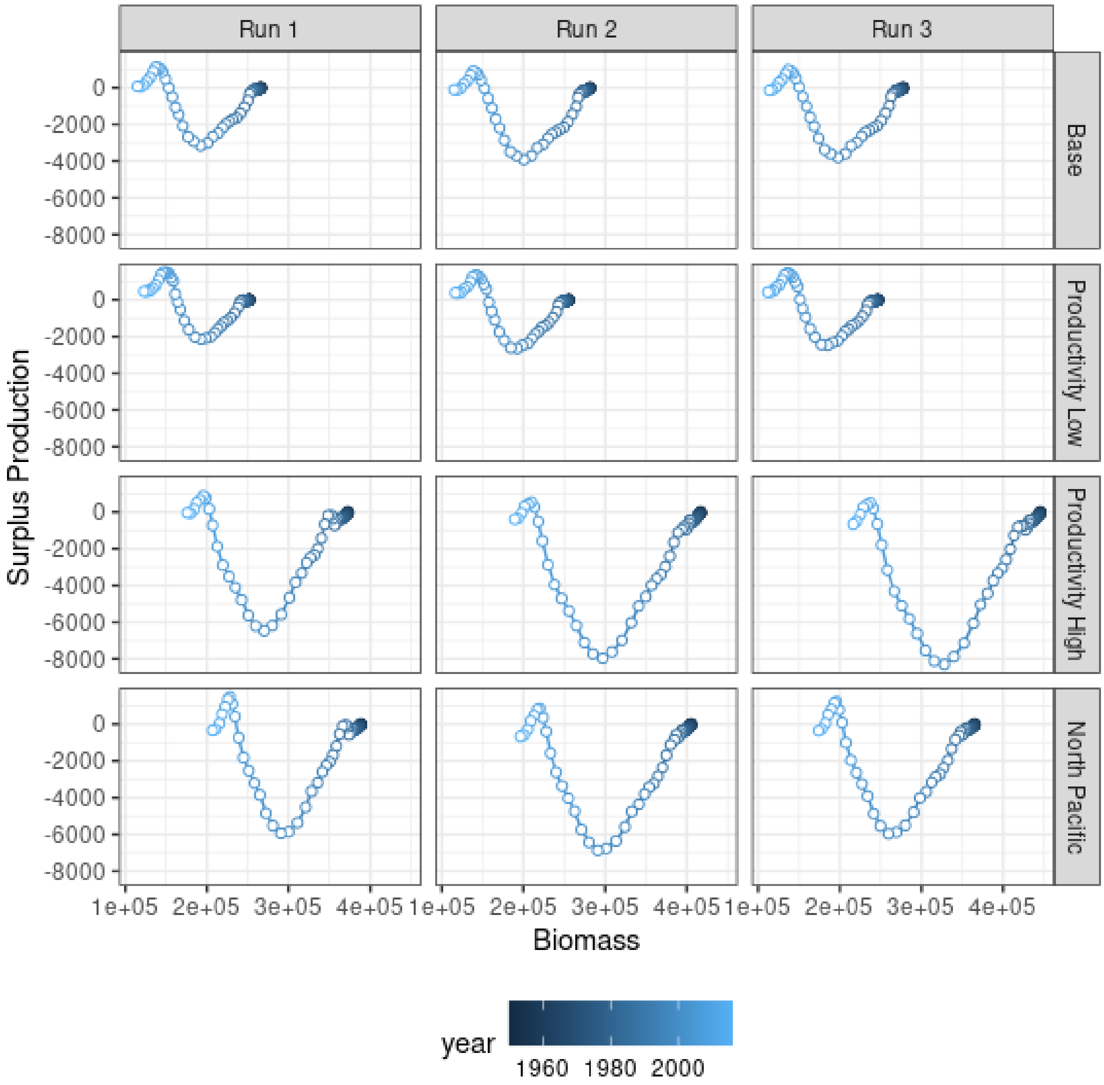


Figure 39. Surplus production for Stock Synthesis.

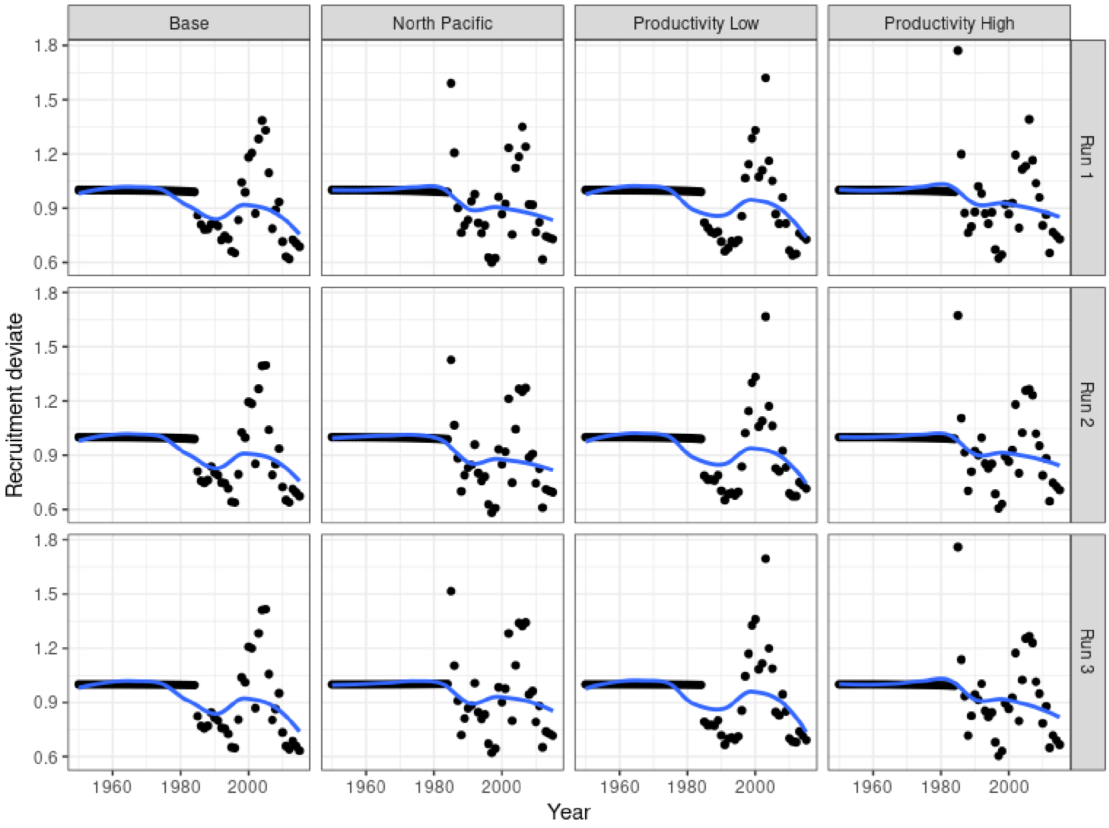


Figure 40. Recruitment deviates by scenario for Stock Synthesis.

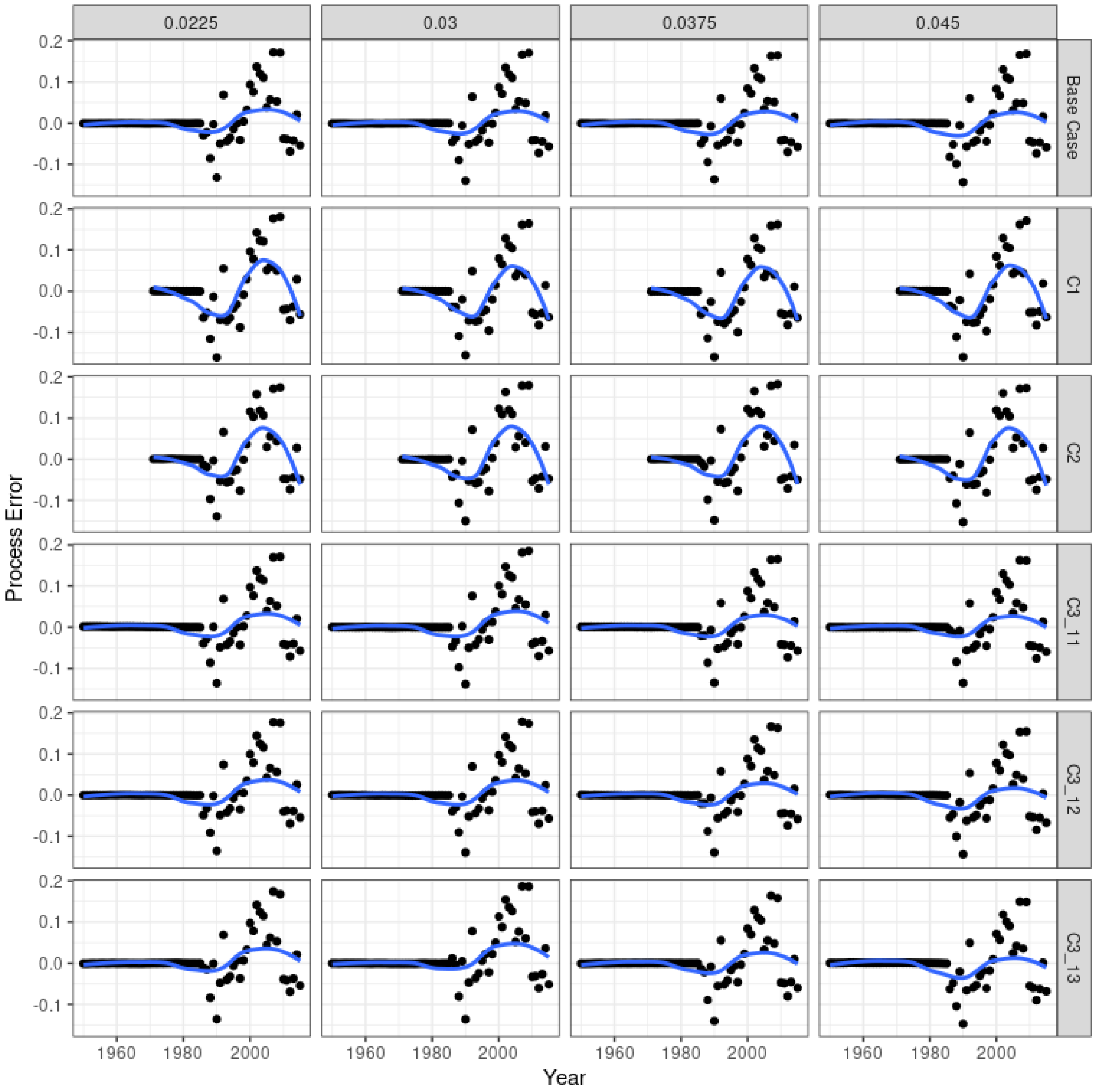


Figure 41. Processes error by scenario for JABBA.

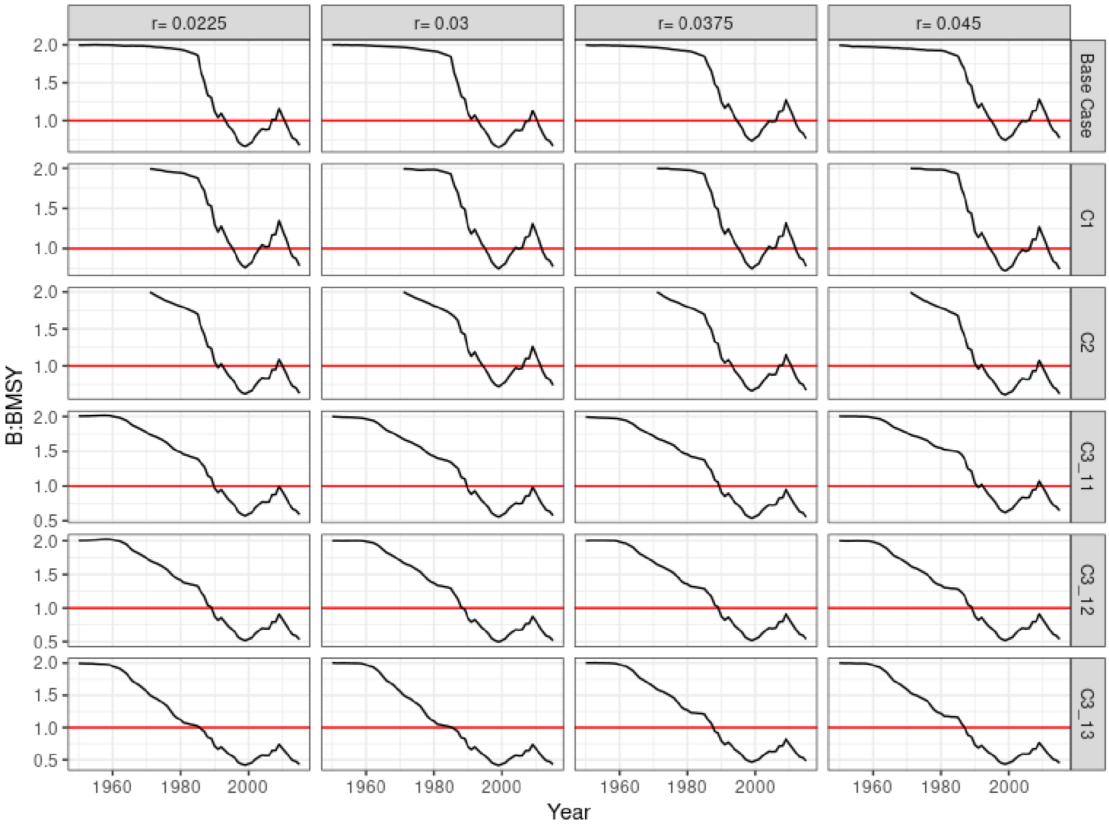


Figure 42. Time series relative to MSY benchmarks for JABBA scenarios.