MODEL VALIDATION FOR THE SELECTION AND WEIGHTING OF STOCK ASSESSMENT SCENARIOS

Laurence T. Kell[1] Henning Winker[2]

SUMMARY

This worked example has been conducted in response to the Recommendation that the Shark Species Group, together with the Working Group on Stock Assessment Methods, should help develop guidelines for the selection, rejection, weighting and extension of stock assessment models when providing robust management advice. The blue shark assessment, in common with other ICCAT stock assessments, has to consider alternative often conflicting data sets, uncertain life history information, and auxiliary data sets such as length and tagging data. The Working Group on Stock Assessment Methods has therefore recommended that Species Groups should identify model uncertainties, biases and misspecifications, to be considered when specifying uncertainty grids to be considered. The Shark Species Group has also been asked to provide, "... options for a harvest control rule (HCR) with associated limit, target and threshold reference points for the management of blue shark in the ICCAT Convention area".

RÉSUMÉ

Cet exemple concret a été développé en réponse à la Recommandation visant à ce que le Groupe d'espèces sur les requins, conjointement avec le Groupe de travail sur les méthodes d'évaluation des stocks, contribuent à l'élaboration de directives pour la sélection, le rejet, la pondération et l'extension des modèles d'évaluation des stocks lors de la soumission d'un avis de gestion robuste. L'évaluation du stock de requin peau bleue, tout comme d'autres évaluations des stocks de l'ICCAT, doit examiner des jeux de données alternatifs souvent contradictoires, des informations incertaines sur le cycle vital et des jeux de données auxiliaires tels que les données de longueur et de marquage. Le Groupe de travail sur les méthodes d'évaluation des stocks a donc recommandé que le Groupe d'espèces identifie les incertitudes, les biais et les erreurs de spécification des modèles, qui seront étudiés lors de la spécification des grilles d'incertitude à examiner. Il a également été demandé au Groupe d'espèces sur les requins de soumettre « ... des options pour une règle de contrôle de l'exploitation (HCR) avec les points de référence limite, cible et seuil associés pour la gestion du requin peau bleue dans la zone de la Convention de l'ICCAT ».

RESUMEN

Este ejemplo práctico se ha realizado en respuesta a la Recomendación de que el Grupo de especies de tiburones, junto con el Grupo de trabajo sobre métodos de evaluación de stock, debería ayudar a desarrollar directrices para la selección, rechazo, ponderación y ampliación de los modelos de evaluación de stock a la hora de proporcionar un asesoramiento sólido en materia de ordenación. La evaluación del tiburón azul, al igual que otras evaluaciones de stock de ICCAT, tiene que considerar conjuntos de datos alternativos a menudo contradictorios, información incierta sobre el ciclo vital y conjuntos de datos auxiliares como datos de talla y marcado. Por lo tanto, el Grupo de trabajo sobre métodos de evaluación de stocks ha recomendado que los grupos de especies identifiquen las incertidumbres, sesgos y errores de especificación de los modelos, que deben tenerse en cuenta a la hora de especificar las matrices de incertidumbre que deben considerarse. También se ha pedido al Grupo de especies de tiburones que proporcione "... opciones para una norma de control de capturas (HCR) con niveles de referencia límite, objetivo y umbral asociados para la ordenación del tiburón azul en la zona del Convenio de ICCAT".

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

[2] Department of Aquatic Resources, Institute of Marine Research, Swedish University of Agricultural Sciences, Sweden

KEYWORDS

Mean Absolute Scaled Error, Prediction Skill, Robustness, Sensitivity Analysis, Uncertainty Analysis

1. Introduction

The blue shark assessment has to consider alternative data sets and review new life history information and tagging data. Therefore, this work has been conducted in response to the Recommendation in 2021 that the shark working group should *"Consider, together with the Working Group on Stock Assessment Methods, alternative stock assessment methods (as per Kell, 2021, other SCRS papers, and the fisheries literature)"* Following this recommendation the SCRS asked the author to help develop guidelines for the selection, rejection, weighting and extension of stock assessment models, for the upcoming blue shark assessment (ICCAT, 2022).

The shark working group has also been asked to provide, *"... options for a harvest control rule (HCR) with the associated limit, target and threshold reference points for the management of blue shark in the ICCAT Convention area."* Which also requires procedures for selection, rejection, the weighting of Operating Model (OM) scenarios when conducting Management Strategy Evaluation (MSE).

The Working Group on Stock Assessment Methods, therefore, recommended that SCRS meetings in preparation for stock assessment evaluations routinely include a presentation and discussion of the model and the diagnostics of the previous assessment being used to provide management advice. The presentations should identify model uncertainties, biases and misspecifications, which should be considered when specifying the uncertainty grid to be submitted at the subsequent stock assessment meeting.

The adoption of the Precautionary Approach requires providing advice that is robust to uncertainty. Therefore, when conducting a stock assessment, alternative model structures and data sets are commonly considered. The primary diagnostics used to compare models are to examine residual patterns to check goodness-of-fit and to conduct retrospective analysis to check the stability of estimates. However, residual patterns can be removed by adding more parameters than justified by the data, and retrospective patterns by ignoring the data. Therefore, neither alone can be used for validation, which requires assessing whether it is plausible that a system identical to the model generated the data (Hodge and Dewar, 1992).

A variety of diagnostics are available for assessing goodness of fit (e.g., Carvalaho *et al.* 2021). For example, indices of abundance are a primary contributor to the overall likelihood when fitting stock assessment models to data (Whitten *et al*., 2013), and the sum of squared errors (SSE) between observed and predicted indices in the log-space is often used as a fitness measure. SSE is problematic because complex models tend to have many parameters to allow flexibility, resulting in a low SSE due to overfitting by adding more parameters than can be justified by the data. This was a reason for the development of criteria such as AIC to aid in model selection. However, AIC needs to be performed on models with the same likelihood function and data, which is not the case if different hypotheses are modelled with alternative model structures and data sets.

Stock assessment models often have poor prediction performance (Patterson *et al*., 2001). For example, it is often observed at the Standing Committee on Research and Statistics (SCRS), when assessments are updated that the stock forecasts did not come to pass. Therefore, to provide advice that is robust to uncertainty, Management Strategy Evaluation (MSE) is used to develop Management Procedures (MPs) and HCR that are robust to uncertainty. To do this, alternative hypotheses and data sets are used for conditioning OMs. The OMs may be based on stock assessments and when conditioning OMs is necessary to ensure that the hypotheses represent plausible dynamics.

An MP is the combination of data collection schemes, the specific analyses applied to those data, such as a stock assessment to determine current status and also the pre-agreed method of standardisation of CPUE data, and the decision rules used to determine management actions based on the results of those analyses (Punt and Donovan 1999; Butterworth 2007; De Oliveira *et al*. 2008).

Retrospective forecasting has been used to compare models, however, this approach is not suitable for validation, as model estimates such as SSB which are latent quantities are not known without error. To address this, *Kell et al.* (2016) proposed hindcasting where actual observations (e.g. indices of abundance) are compared to their predicted future values. The key concept behind the approach is 'prediction skill', which is defined as any measure of the accuracy of a forecasted value to the actual observed value unknown by the model (Kell *et al*., 2021). The difference is hereafter referred to as the 'prediction residual' (Michaelsen, 1987).

We, therefore, use hindcasting to estimate prediction skill, a measure of the accuracy of a predicted value unknown by the model relative to its observed value to show to validate the Blue Shark JABBA assessment and scenario in order to accept and weight assessment models and hypotheses. We then discuss the extension of the approach to multimodel scenarios and conditioned OMs when developing HCRs.

2. Material and Methods

For the assessment the Bayesian State-Space Surplus Production Model framework, 'Just Another Bayesian Biomass Assessment' (JABBA, Winker *et al*., 2018) was used. JABBA has been applied in stock assessments of sharks, tuna, and billfishes around the world, and presents a unifying, flexible framework for biomass dynamic modelling, runs quickly, and generates reproducible stock status estimates and diagnostic tools. Specific emphasis has been placed on flexibility for specifying alternative scenarios, achieving high stability and improved convergence rates. The approach taken, however, is generic and can also be applied to other stock assessment models.

Biomass dynamic assessment models use as inputs total removals and indices of relative abundance. For example, catch-per-unit-of-effort (CPUE) is often the main piece of information used in fisheries stock assessments and is assumed to be proportional to abundance. However, CPUE are not observations, as they are generated by models to standardise them, by removing factors that are unrelated to abundance that may vary by year. It is important to understand and account for mechanisms that could affect catchability, which may be due to changes in fishing strategies, or stock distribution and productivity (de Bruyn and Schirripa, 2017). CPUE indices are generally brought by contracting parties to an assessment working group and provide insights into the fisheries and stocks. For example, some indices may have considerably wider geographical coverage than others, which likely affects their representativeness for the stock.

We therefore first explore the trends in and correlations between the CPUE submitted to the working group. We then run evaluate the prediction residuals for JABBA assessments based on different CPUE grouping i.e. i) all indices, ii) one-by-one, iii) leaving one out, and iv) clusters agreed at the Data Prep meeting.

For validation, observations or well-known quantities should be used. We, therefore, applied hindcasting to the time series of catch per unit effort (CPUE) used as indices of relative abundance. The algorithm is similar to that used in retrospective analysis, as it requires the same procedure of peeling observations from the end of the series and refitting the model to the truncated data set. Hindcasting involves the additional step of projecting forward, and an important difference is that the forecasts are for the observations.

2.1 Indices of Abundance

The data used are the indices of abundance summarised in **Figure 1**. To summarise trends, a GAM was fitted with an index factor for catchability and a common smoother was fitted to all the indices.

Abundance indices (annual values of the index and associated CV) are the standardised CPUEs, for Venezuela, Spain, Portugal, USA, Japan, Chinese Taipei and Morocco. The SPN, POR and MOR indices are in biomass, whereas the remaining indices are in numbers. This is not ideal for use in biomass dynamic models, where the population is in biomass and the indices used in fitting are assumed to be representative of exploitable biomass through an index-specific catchability factor (q).

At the data preparatory meeting, no index was identified as flawed, so the Group decided not to exclude any index. The suggestion was made to use a cluster analysis approach to formulate alternative states of nature and evaluate CPUE series using the resulting groups. The results of the cluster analysis for the North Atlantic suggest grouping Morocco and EU-Portugal together, and Japan, EU-Spain, and Chinese Taipei as another group, while waiting for the results of the new U.S. standardization. For Venezuela, it was noted that it was positively correlated with the Moroccan index, but the two indices only overlapped in a limited number of years. The Group agreed that it would explore using this index from Venezuela along with the updated U.S. index. It further recommended exploring the influence of this index on the assessment.

2.2 Model Settings

The intention at this stage is not to find a “best” assessment but to develop a generic approach for comparing indices, and propose hypotheses for testing, and an objective way to validate models. We, therefore, used common settings across all runs.

Model settings were the same as used in SCRS/2023/124. For the production function, a Schaefer form was used to correspond to B_{MSY} being half of Virgin Biomass ($B_{MSY}/K = 0.5$)

It was assumed that the catch values used as data inputs are “Reported Catch” values and the “True Catch” values may differ, i.e., “True Catch” ~ logNormal (median=Reported Catch, CV=0.1).

2.3 Hindcast

An objective of this study is to show how hindcasting and using multiple measures for prediction skill can help in the development of robust stock assessment advice frameworks. Using multiple measures to represent more than one attribute of prediction skill helps provide insight into the reliability of the measures and the error structure of the data (Kell *et al.,* 2016).

We use hindcasting to evaluate the prediction skill of series of catch per unit effort (CPUE) used as indices of stock abundance, across a range of stock assessment scenarios. Since time series of CPUE are often the most influential inputs to stock assessment models (Francis and Hilborn, 2011) it is important to be aware of the limitations of these data when fitting models with them.

A hindcast procedure was used where the indices of abundance are sequentially removed from the terminal year, i.e., peeled backwards from the model. In contrast, in a retrospective analysis, all observations for a year are peeled back, which means that quantities cannot be predicted for the years peeled back unless additional assumptions are made. The hindcast is a variant of cross-validation where, like retrospective analysis, recent data are removed, and the model is refitted with the remaining data. Known values (observations) or well-estimated historical values are then compared to model estimates. When observations are used for comparison, this is also referred to as model-free validation (Kell *et al*., 2016). In a hindcast, observations are removed from the terminal year and up to n years back, and then the missing observations are predicted by fitting to the remaining data for 1, 2, ... n steps ahead. Observations may be removed by series or fleets to evaluate data conflicts, time blocks to overcome serial correlations, or individually to estimate bias, as in the jackknife. No stock forecast or projection needs to be performed, and so there is no need to make assumptions about future parameters, as all parameters needed are estimated within the model. The hindcast may be conducted for individual data series or combinations of series and data types, for example, by fleet where both CPUE and length data are removed. This allows data conflicts to be explored.

2.4 Prediction Skill

The main aims of stock assessment models are to provide accurate estimates of state variables and predictions of the future. A variety of measures for accuracy are available, and **Table 1** shows a number of metrics. The JABBA package has a method that outputs the prediction and model residuals, model and hindcast estimates, and observations, allowing these metrics to be easily calculated. The intention is to create a similar function in the ss3diags package.

Where the error (e) is the difference between the observed state and the estimate or prediction $e = y - \hat{y}$, and where n is the sample length and h is the period for a forecast.

Metrics, such as MAE cannot be used to compare data at different scales unlike Mean Absolute Scaled Error MASE and Root Mean Squared Scaled Error (RMSSE). The mean absolute scaled error (MASE; Hyndman and Koehler, 2006) is a robust statistic for evaluating prediction skill. MASE builds on the principle of evaluating the prediction skill of a model relative to a naïve baseline prediction. A prediction is said to have 'skill' if it improves the model forecast compared to the baseline. A widely used baseline forecast for time series is the 'persistence algorithm' that takes the observation at the previous time step to predict the expected outcome at the next time step as a random walk of naïve in-sample predictions, e.g., tomorrow's weather will be the same as today's. The MASE score scales the mean absolute error (MAE) of forecasts (i.e., prediction residuals) to MAE of a naïve in-sample prediction.

MASE has the desirable properties of scale invariance, so it can compare forecasts across data sets with different scales, such as CPUE. It also has predictable behaviour, symmetry, interpretability, and asymptotic normality.

Unlike relative error, MASE does not skew its distribution even when the observed values are close to zero. It is easy to interpret, as a score of 0.5 indicates that the model forecasts are twice as accurate as a naïve baseline prediction. The Diebold-Mariano test (Diebold and Mariano, 1995) for one-step forecasts can also be used to test the statistical significance of the difference between two sets of forecasts.

3. Results

The indices are first compared before JABBA is fitted and the prediction residuals are used for validation.

3.1 Indices

A GAM with was fitted to all the indices, Tukey described this approach as residuals and reiteration: where removing a striking pattern allows more subtle patterns to be seen. Therefore, **Figure 2** plots the residuals, along with an indication of whether the run's test was passed. There appear to be similarities between the SPN and JPN series. There are also differences between the geographical coverage of the CPUE indices, and some may only cover a small area relative to the stock. Such factors need to be considered when developing hypotheses for testing.

Pairwise scatter plots (**Figure 3**) are then used to explore the correlations between the indices, and **Figure 4** shows the correlation matrix, blue indicates positive and red negative correlations. There appear to be 3 clusters por & mor, ven & usa1, ctp, usa2 & jpn, and conflicts e.g., spn & mor.

In a biomass dynamic model without age structure, it is assumed that the indices correspond to the same population components, i.e., no age effects, and no differences in spatial structure. Differences between quantities used for the index, i.e., biomass or numbers, will also impact trends and correlations, e.g. in a strong year-clas individuals will be subject to both mortality and growth in body mass meaning that a biomass index may increase when an index number is decreasing. If, however, only adults are caught then the difference between trends in numbers and biomass may be small.

Therefore, the cross-correlations are plotted in **Figure 5**. If two indices are mapping the same age classes, then the highest correlation will be at lag 0. If, however, two indices represent different age structures, then the highest correlation will be lagged (e.g., ven & jpn). While, if the two indices are mapping different stock components then correlations will be low (mor & ctp), and if an index does not track year-classes then its auto-correlations will be low (e.g. ctp).

3.2 JABBA Fits

After the comparison of the indices, JABBA was fitted using a common set of parameter settings across all scenarios. The aim was not to develop a best assessment but to show how comparisons can be made across scenarios, as a first step before considering model acceptance and weighting.

The JABBA runs are summarised in **Table 2**, **Table 3** presents the MASE and **Table 4**, the run tests

In **Table 2** a variety of summary statistics are presented, for example, the deviance information criterion (DIC) which has been extensively used for making Bayesian model selection. It is a Bayesian version of AIC and chooses a model that gives the smallest expected Kullback-Leibler divergence between the data generating process (DGP) and a predictive distribution asymptotically summary. The DIC varies widely depending on the choice of indices.

The prediction and model residuals are compared in **Figure 6,** for a five-year peel, none pass the runs tests. The patterns are quite different, and only MOR had MASE<1.

In summary, it looks like there are conflicts in the data and hence multiple scenarios may be warranted:

- *All*: If we run JABBA with all the indices then only MOR has prediction skill, the other indices are only adding noise or else their signal is being masked.
- *One by One*: When indices are run by themselves, JPN and SPN have prediction skill, as suggested above it looks like their signal is being masked by MOR.

- *ClDP1*: When JPN & SPN are removed, then POR, as well as MOR, have prediction skill.
- *ClDP2*: Now CTP has prediction skill.
- *ClTg1*: When JPN & SPN are removed then POR, as well as MOR, have prediction skill.
- *ClTg1*: compared to ClDP1 POR no longer has prediction skill, suggests that SPN is adding some conflict.

The runs tests show that if a single index is used in fitting, that there is no residual pattern, however, this does not always result in prediction skill as although CTP, POR and USA2 have no residual pattern MASE>1, which suggests overfitting.

3.4 JABBA Benchmarks and Reference Points

The distribution of benchmarks relative to reference points are also compared after JABBA is fitted.
The estimates of B/B_{MSY} and F/F_{MSY} in the final year are presented in **Figures 7** and **8** for scenarios All and ClDP1, and ClDP2. For Biomass, the "All" scenario gives a PDF that is between the scenario ClDP1 & ClDP2, while for F cluster ClDP1 is bimodal. This is likely due to non-convergence due to including MOR and VEN which are negatively correlated. These results demonstrate the problems of including conflicting indices in a single assessment and of combining assessments with different data sets and support the conclusion that multiple scenarios may be warranted.

4. Discussion

The analysis showed that the CPUE series are in conflict, causing convergence problems This is illustrated by **Figures 5 and 6,** which show the distributions of B/B_{MSY} and F/F_{MSY} for scenarios All and ClDP1, and ClDP2. For Biomass, the All scenario gives a PDF that is between the scenario ClDP1 & ClDP2 , while for F cluster ClDP1 is bimodal. This is likely due to including MOR and VEN which are negatively correlated. These results demonstrate the problems of including conflicting indices in a single assessment and combining assessments with different data sets.

The main input parameters of a stock assessment are often uncertain. This means that stock assessors are often faced with a range of model formulations and/or alternative management scenarios which should be scrutinised before decisions are made (Mannini *et al*, 2021). In this context, when discussing which could be the best model used in assessing stocks, Hilborn and Walters (1992) recalled an adage that "*the truth often lies at the intersection of competing lies*". This uncertainty in 'what is the best model?' necessitates a comparison of a range of alternative models. The use of structural uncertainty grids (Rice & Courtney, 2023) to conduct sensitivity analyses has become common practice in the tuna RFMOs. Since even moderate stocks like pelagic sharks typically have large uncertainties including catch and many inestimable parameters. This differs, from, and is often confused with uncertainty analysis using ensemble modelling where multiple models are developed and combined to provide estimates of uncertainty. Ensemble modelling can include model weighting and projections for alternative management options.

ICES are providing advice where instead of comparing multiple model outputs and selecting a single final model, an ensemble modelling approach (Dietterich, 2000) was used to present results with a quantitative criterion for weighting several model predictions. Ensemble methods provide a promising approach when decisions have to be made despite the presence of multiple and potentially conflicting estimates of stock status (Anderson *et al*. 2017). Ensemble models have been proven to be more accurate and less biased than the choice of an individual model, as they can effectively tease apart the conditions under which various model assumptions result in the most accurate predictions (Dietterich, 2000; Knutti *et al.,* 2009). In general, an ensemble approach will better encapsulate the variability and uncertainty of model predictions because instead of choosing a single set of fixed parameter values, you can explore a contrasting but plausible range of values. (Dietterich, 2000; Knutti *et al*, 2009). This is crucial when the reliability of single fixed parameters is in question. The objective when using an ensemble model is therefore to quantify the total uncertainty across all plausible models, where the structural uncertainty is likely to be much greater than the within-model uncertainty. For example, ensembles are often helpful because modellers need not decide on dome versus asymptotic fisheries selectivity (e.g., Sampson & Scott, 2012, FAO-GFCM, 2021), or whether to fix or estimate natural mortality (e.g., Johnson *et al.,* 2015). Moreover, ensemble forecasting has been proven to improve forecast accuracy, robustness in many fields, particularly in weather forecasting where the method originated (Wu and Levinson 2021).

4.1 Model weighting

Before running an ensemble model, and providing advice, each model needs to be assigned a weight. The need to weigh models based on available information is well recognised (Francis and Hilborn, 2011). However, the complexity of the stock assessment process may prevent strict statistical rigour from being applied. However, assigning the same weight (reliability) to all models could introduce biases in the management advice if some models are less likely than others or there is redundancy, i.e., multiple models with similar hypotheses are included at the expense of other equally plausible scenarios.

ICES (2022) developed an ensemble for Northern shrimp (Pandalus borealis), where the main uncertainty is in the specification of natural mortality. To assign weights to the various models, a system of discrete weight categories was used based on diagnostic scores (W(Diagnostics)) as weighting metrics (Maunder *table*., 2020) to judge the plausibility of each model based on its fit to the data. The W(Diagnostics) components were calculated based on a series of interconnected diagnostic tests as discussed by Carvalho *et al*. (2021)

$$(\boldsymbol{W}(\mathbf{Diags_1}) + \boldsymbol{W}(\mathbf{Diags}_2) \ldots + \boldsymbol{W}(\mathbf{Diags}_n))/\mathbf{Num\ of\ }\boldsymbol{W}(\mathbf{Diags})$$

Each W component was then assigned a value of 1 when the run passes the diagnostic test, and a 0 if it fails.

A similar approach could be used by ICCAT, tasks then is to agree on the tests to include and possible relative weights of the components. For example, do tests based on prediction and model residuals get the same weights.

Another issue is that are tests are also used for the development of scenarios, i.e. when failing a test requires exclusion or the development of an alternative scenario.

5. Conclusions

- The data are in conflict, and so different weighting schemes will give different results. MOR appears to be driving the assessment, or obscuring signals from the other indices.
- The next steps are to agree on scenarios based on a priori hypotheses and then define weights for model averaging when providing advice.
- The analysis should be re-conducted for SS3, i.e., do more information and associated assumptions.
- More complicated models may resolve the data conflicts? However, more assumptions require more validation.
- Prediction skill can be used to choose indices for use in either a model or model-free MP. However, the MP and indices used in an HCR must be common across OMs. I.e., you cannot pick and choose based on individual OMs. What if indices for potential use in an empirical HCR have different trends, averaging is likely to be wrong as the actual trend is more likely to be either or. A possibility is to run an MP with two HCRs based on each scenario, but then to take the most precautionary advice. This would also allow the value-of-information to be evaluated, as correctly identifying the unbiased index would permit higher yield.

Acknowledgements

This work was funded by the Spanish Surface Longline Fishing Sector and the Producer Organisations OPP BURELA, OPNAPA, ORPAGU and OPROMAR.

References

Akaike H. 1998. Information theory and an extension of the maximum likelihood principle. Selected Papers of Hirotugu Akaike. Springer, Berlin. 199pp.

Anderson S.C., Cooper A.B., Jensen O.P., Minto C., Thorson J.T., Walsh J.C., Afflerbach J., Dickey-Collas M., Kleisner K. M., Longo C., Osio G.C., Ovando D., Mosqueira I., Rosenberg A.A., Selig E.R. 2017. Improving estimates of population status and trend with superensemble models. Fish Fisheries,18: 732–741. Doi: https://doi.org/10.1111/faf.12200.

Butterworth D.S. 2007. Why a management procedure approach? Some positives and negatives. ICES Journal of Marine Science, 64:613-617. doi: 10.1093/icesjms/fsm003

Carvalho F., Winker H., Courtney D., Kapur M., Kell L., Cardinale M., Schirripa M., Kitakado T., Yemane D., Piner K.R., Maunder M.N., Taylor I., Wetzel C.R., Doering K., Johnson K.F., Methot R.D. 2021. A cookbook for using model diagnostics in integrated stock assessments, Fisheries Research, Volume 240, 2021, 105959, ISSN 0165-7836, https://doi.org/10.1016/j.fishres.2021.105959.

de Bruyn P. and Schirripa M.J. 2017. Tools To Guide The Selection Of CPUE Series – Revisited And Revised. SCRS/2017/081 Collect. Vol. Sci. Pap. ICCAT, 74(2): 404-409

Diebold F. and Mariano R. 1995. Comparing predictive accuracy. Journal Of Business And Economics Statistics, 20: 134–144.

Dietterich T.G. 2000. Ensemble methods in machine learning. In Multiple classifier systems (pp. 1–15). Berlin, Heidelberg: Springer.

De Oliveira J.A.A., Kell L.T., Punt A.E., Roel B.A., and Butterworth D.S. 2008. Managing without best predictions: the Management Strategy Evaluation framework. In Advances in Fisheries Science. 50 years on from Beverton and Holt. Edited by A. Payne, A. Cotter, T.Potter. Blackwell Publishing, Oxford. pp104-134

Hodges J. S. and Dewar J. A. 1992. Is it you or your model talking? A framework for model validation. Santa Monica, CA: Rand.

R.J. Hyndman, A.B. Koehler. 2006. Another look at measures of forecast accuracy. Int. J. Forecast., 22 (2006), pp. 679-688, 10.1016/j.ijforecast.2006.03.001

Kell L.T., Kimoto A. and Kitakado T. 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries research, 183, pp.119-127.

Kell L.T., Sharma R., Kitakado T., Winker H., Mosqueira I., Cardinale M. and Fu D. 2021. Validation of stock assessment methods: is it me or my model talking?. ICES Journal of Marine Science, 78(6), pp.2244-2255.

FAO-GFCM. 2021. Report of the Working Group on Stock Assessment of Demersal Species (WGSAD) – Benchmark session for the assessment of common sole in GSA 17, Scientific Advisory Committee on Fisheries (SAC). Online via Microsoft Teams, 12–16 April 2021.

Francis R.C. and Hilborn R. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences 68(6): 1124–1138. doi:10.1139/f2011-025.

Gunderson D.R. 1993. Surveys of fisheries resources. Wiley. New York. 248 pp.

Hilborn R., and Walters C. J. 1992. Quantitative Fish Stock Assessment. Choice, Dynamics and Uncertainty. New York: Chapman and Hall, 570.

ICCAT, 2022. REPORT OF THE 2022 ICCAT Intersessional Meeting Of The Sharks Species Group. Collect. Vol. Sci. Pap. ICCAT, 79(4): 61-132 (2022) 61

ICES. 2022. Benchmark workshop on Pandalus stocks (WKPRAWN). ICES Scientific Reports. 4:20. 249 pp. http://doi.org/10.17895/ices.pub.19714204

Knutti R., Furrer R., Tebaldi C., Cermak J., & Meehl G.A. 2009. Challenges in combining projections from multiple climate models. Journal of Climate,23,2739–2758.

Mannini A., Pinto C., Konrad C., Vasilakopoulos P. and Winker H. 2020. "The Elephant in the Room": Exploring Natural Mortality Uncertainty in Statistical Catch at Age Models. Front. Mar. Sci. 7:585654. doi: 10.3389/fmars.2020.585654.

Maunder M.N., Xu H., Lennert-Cody C.E., Valero J.L., Aires-da-Silva, A., MinteVera, C., 2020. Implementing Reference Point-based Fishery Harvest Control Rules Within a Probabilistic Framework That Considers Multiple Hypotheses (No. SAC-11- INF-F). Scientific Advisory Commitee, Inter-American Tropical Tuna Commission, San Diego.

Michaelsen J., 1987. Cross-validation in statistical climate forecast models. Journal of Applied Meteorology and Climatology, 26(11), pp.1589-1600.

Patterson K., Cook R., Darby C., Gavaris S., Kell L., Lewy P., Mesnil B., Punt A., Restrepo V., Skagen D.W. and Stefánsson G., 2001. Estimating uncertainty in fish stock assessment and forecasting. Fish and fisheries, 2(2), pp.125-157.

Punt A.E., and Donovan G.P. 1999. Developing management procedures that are robust to uncertainty: lessons from the International Whaling Commission. ICES Journal of Marine Science, 64:603-612. doi:10.1093/icesjms/fsm035.

Rice J. 2023 Comparison and analysis of North Atlantic CPUE; 2023 ICCAT BSH assessment. ICCAT Collect. Vol. Sci. Pap. ICCAT, 80(4): 345-352

Rice J. and Courtney D. 2023. Structural Uncertainty in RFMO Pelagic Shark Stock Assessments: Examples and Recommendations Resulting from Two Recent applications. SCRS/2023/051 (withdrawn)

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

Whitten A. R., Klaer N. L., Tuck G.N., Day R.W. 2013. Accounting for cohort-specific variable growth in fisheries stock assessments: a case study from south-eastern australia. Fisheries Research, 142: 27–36.

Wu H. and Levinson D. 2021. The ensemble approach to forecasting: A review and synthesis. https://doi.org/10.1016/j.trc.2021.103357

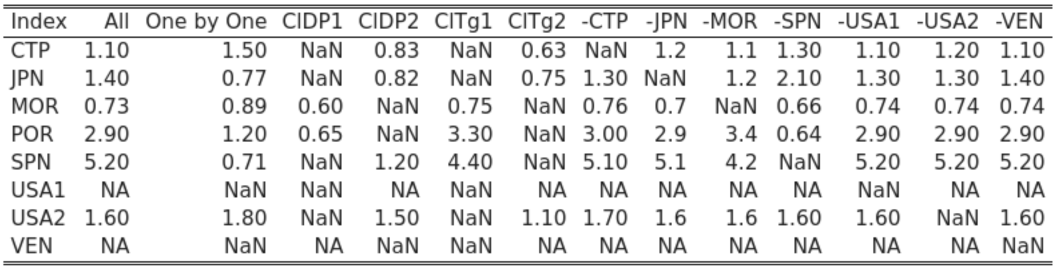
Table 1. Accuracy metrics

Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert e_i\rvert$
Mean Absolute Percentage Error (MAPE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert\frac{e_i}{y_i}\rvert$
Root Mean Squared Error (RMSSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
Mean Absolute Scaled Error (MASE)	$\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}\lvert y_t-\hat{y}_t\rvert}{\frac{1}{n-1}\sum_{t=2}^{n}\lvert y_t-y_{t-1}\rvert}$
Root Mean Squared Scaled Error (RMSSE)	$\sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(y_t-\hat{y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(y_t-y_{t-1})^2}}$

Table 2. Summary of fits scenario; -CTP etc show runs when a single index was omitted.

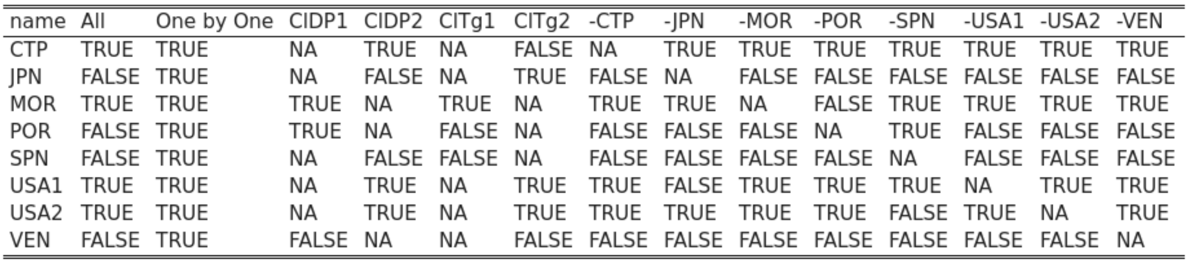
Statistic	All	CIDP1	CIDP2	CITg1	CITg2	CTP	JPN	MOR	POR	SPN	USA1	USA2	VEN	-CTP	-JPN	-MOR	-POR	-SPN	-USA1	-USA2	-VEN
N	160.0	57.00	98.0	62.0	93.00	15.00	28.00	12.0	25.00	25.0	23.0	7.00	20.00	140.0	130.0	140.0	130.0	130.0	130.0	150.0	140.0
DF	130.0	46.00	83.0	51.0	78.00	8.00	21.00	5.0	18.00	18.0	16.0	0.00	13.00	120.0	110.0	120.0	110.0	110.0	110.0	130.0	120.0
DIC	420.0	-100.00	6.8	270.0	-360.00	-400.00	-430.00	-270.0	-180.00	-180.0	-290.0	-390.00	-480.00	410.0	360.0	270.0	97.0	110.0	300.0	400.0	490.0
RMSE	53.0	71.00	43.0	21.0	65.00	110.00	12.00	42.0	6.70	4.9	40.0	Inf	120.00	48.0	59.0	54.0	58.0	59.0	56.0	54.0	39.0
SDNR	1.5	0.94	1.3	1.6	0.94	0.79	0.77	1.2	0.71	0.6	1.2	0.62	0.75	1.6	1.4	1.5	1.3	1.3	1.5	1.5	1.6
p	21.0	11.00	15.0	11.0	15.00	7.00	7.00	7.0	7.00	7.0	7.0	7.00	7.00	19.0	19.0	19.0	19.0	19.0	19.0	19.0	19.0

Table 3. MASE by scenario; -CTP etc show runs when a single index was omitted.



Index	All	One by One	CIDP1	CIDP2	CITg1	CITg2	-CTP	-JPN	-MOR	-SPN	-USA1	-USA2	-VEN
CTP	1.10	1.50	NaN	0.83	NaN	0.63	NaN	1.2	1.1	1.30	1.10	1.20	1.10
JPN	1.40	0.77	NaN	0.82	NaN	0.75	1.30	NaN	1.2	2.10	1.30	1.30	1.40
MOR	0.73	0.89	0.60	NaN	0.75	NaN	0.76	0.7	NaN	0.66	0.74	0.74	0.74
POR	2.90	1.20	0.65	NaN	3.30	NaN	3.00	2.9	3.4	0.64	2.90	2.90	2.90
SPN	5.20	0.71	NaN	1.20	4.40	NaN	5.10	5.1	4.2	NaN	5.20	5.20	5.20
USA1	NA	NaN	NaN	NA	NaN	NA	NA	NA	NA	NA	NaN	NA	NA
USA2	1.60	1.80	NaN	1.50	NaN	1.10	1.70	1.6	1.6	1.60	1.60	NaN	1.60
VEN	NA	NaN	NA	NaN	NaN	NA	NA	NA	NA	NA	NA	NA	NaN

Table 4. Summary of runs test, i.e. do the indices pass, by scenario; -CTP etc show runs when a single index was omitted.



name	All	One by One	CIDP1	CIDP2	CITg1	CITg2	-CTP	-JPN	-MOR	-POR	-SPN	-USA1	-USA2	-VEN
CTP	TRUE	TRUE	NA	TRUE	NA	FALSE	NA	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
JPN	FALSE	TRUE	NA	FALSE	NA	TRUE	FALSE	NA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
MOR	TRUE	TRUE	TRUE	NA	TRUE	NA	TRUE	TRUE	NA	FALSE	TRUE	TRUE	TRUE	TRUE
POR	FALSE	TRUE	TRUE	NA	FALSE	NA	FALSE	FALSE	FALSE	NA	TRUE	FALSE	FALSE	FALSE
SPN	FALSE	TRUE	NA	FALSE	FALSE	NA	FALSE	FALSE	FALSE	FALSE	NA	FALSE	FALSE	FALSE
USA1	TRUE	TRUE	NA	TRUE	NA	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	NA	TRUE	TRUE
USA2	TRUE	TRUE	NA	TRUE	NA	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	NA	TRUE
VEN	FALSE	TRUE	FALSE	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	NA

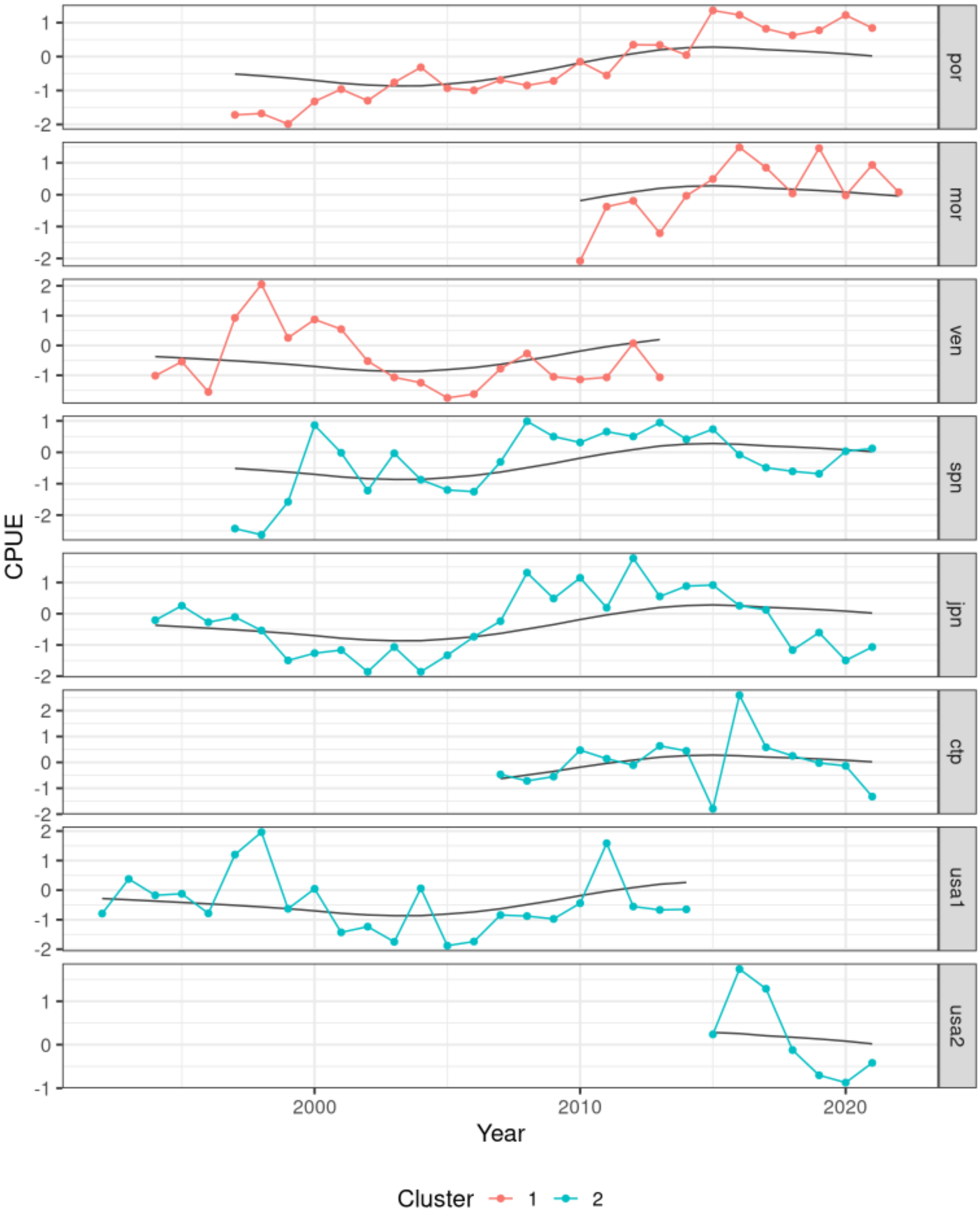


Figure 1. Time series of the CPUE indices, the continuous black line is a loess smother showing the average trend; i.e., fitted to year with series as a factor. Two clusters were chosen at the data prep meeting and are indicated.

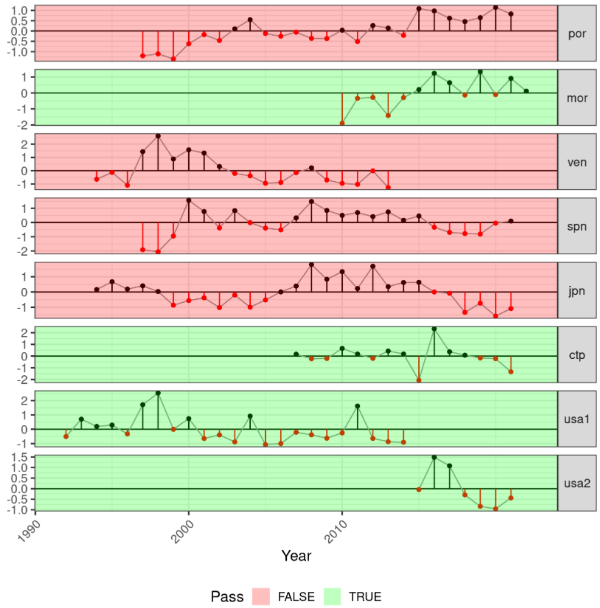


Figure 2. Residual runs tests for the fits to the loess smoother.

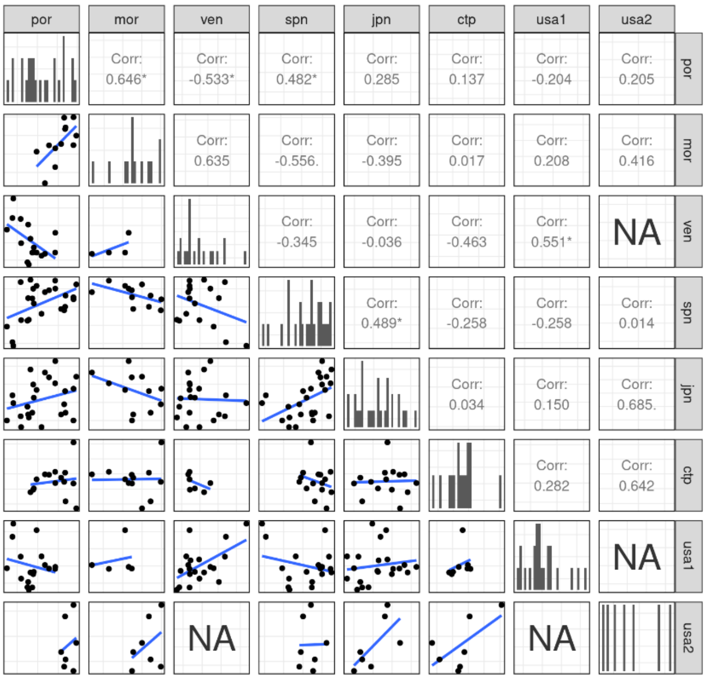


Figure 3. Pairwise CPUE plots showing the correlations between series.

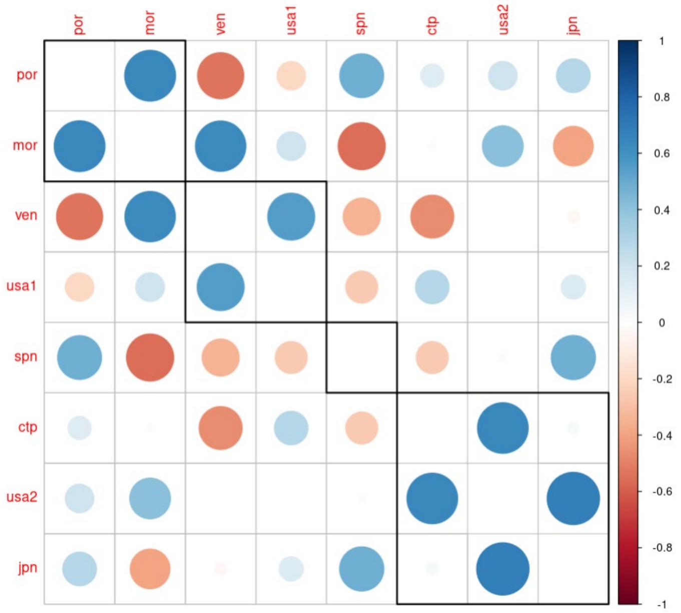


Figure 4. A plot of the correlation matrix for the CPUE indices, blue indicates a positive correlation and red negative. The order of the indices and the rectangular boxes are chosen based on a hierarchical cluster analysis using a set of dissimilarities for the indices.

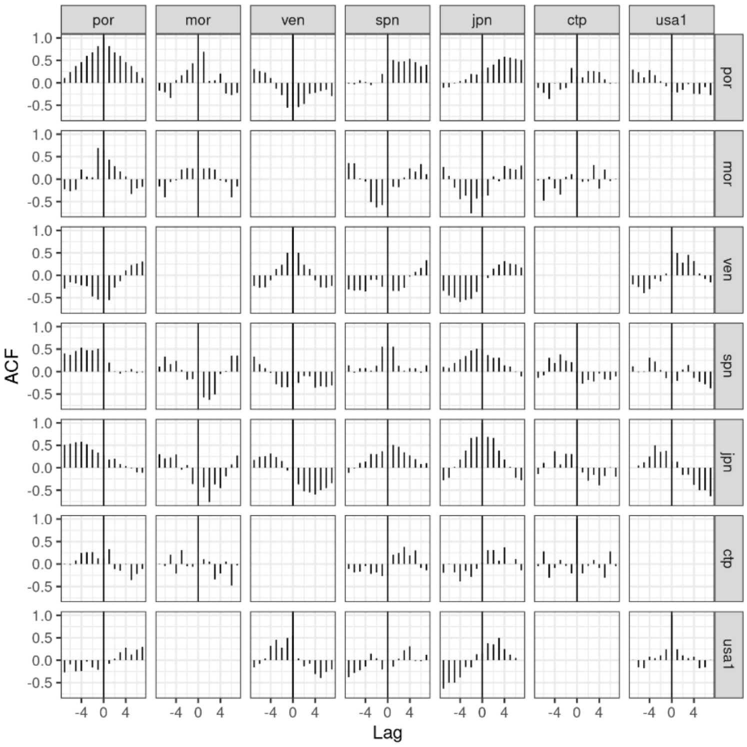


Figure 5. Cross correlations between indices.

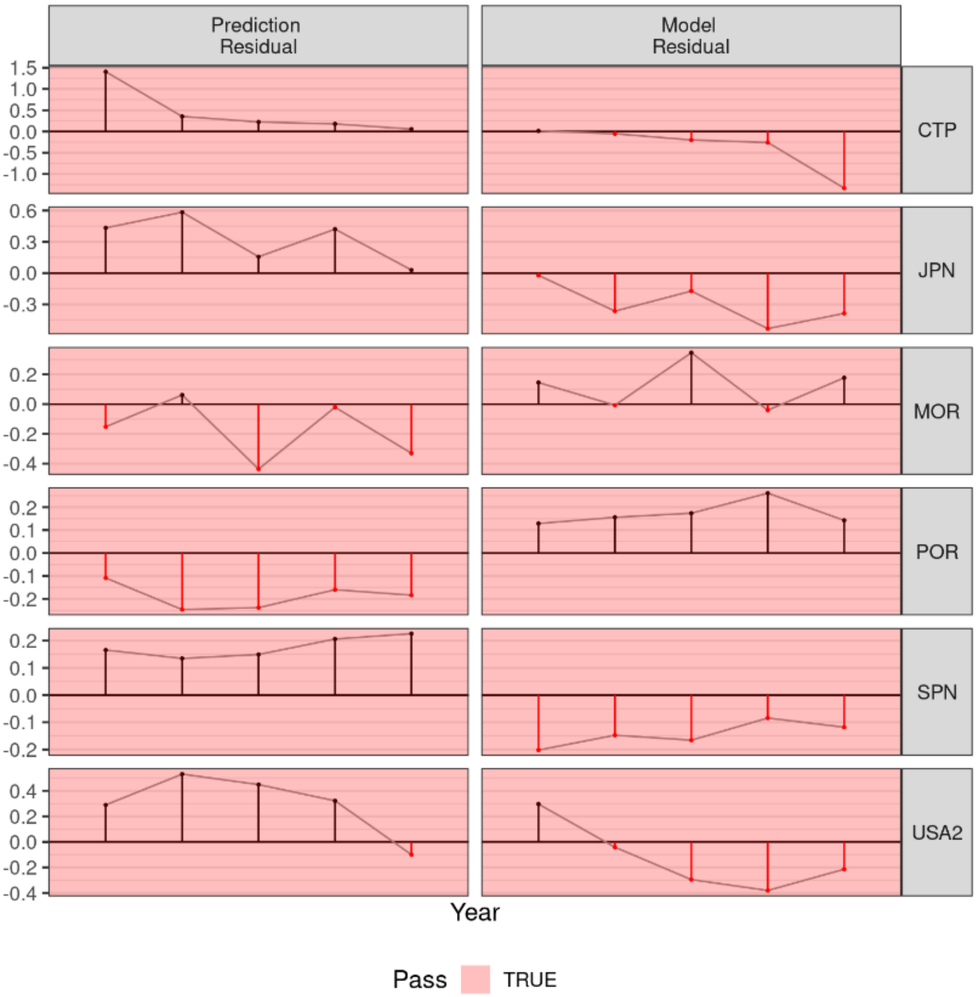


Figure 6. Comparison of runs tests for the scenario that used all indices.

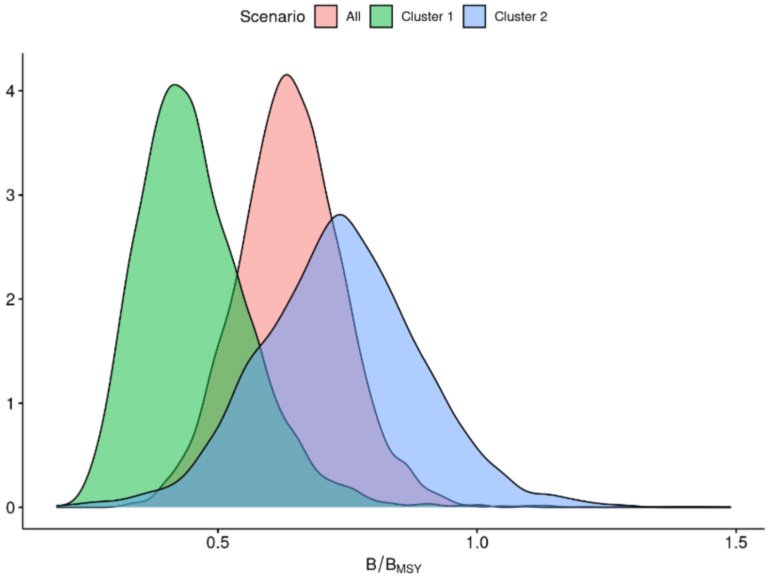


Figure 7. Distributions of B/B$_{MSY}$ for scenarios where either all indices or only cluster 1 or cluster 2 were included.

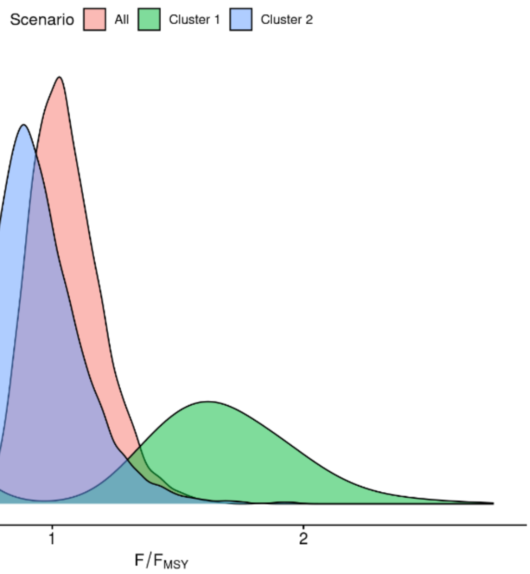


Figure 8. Distributions of F/F$_{MSY}$ for scenarios where either all indices or only cluster 1 or cluster 2 were included.