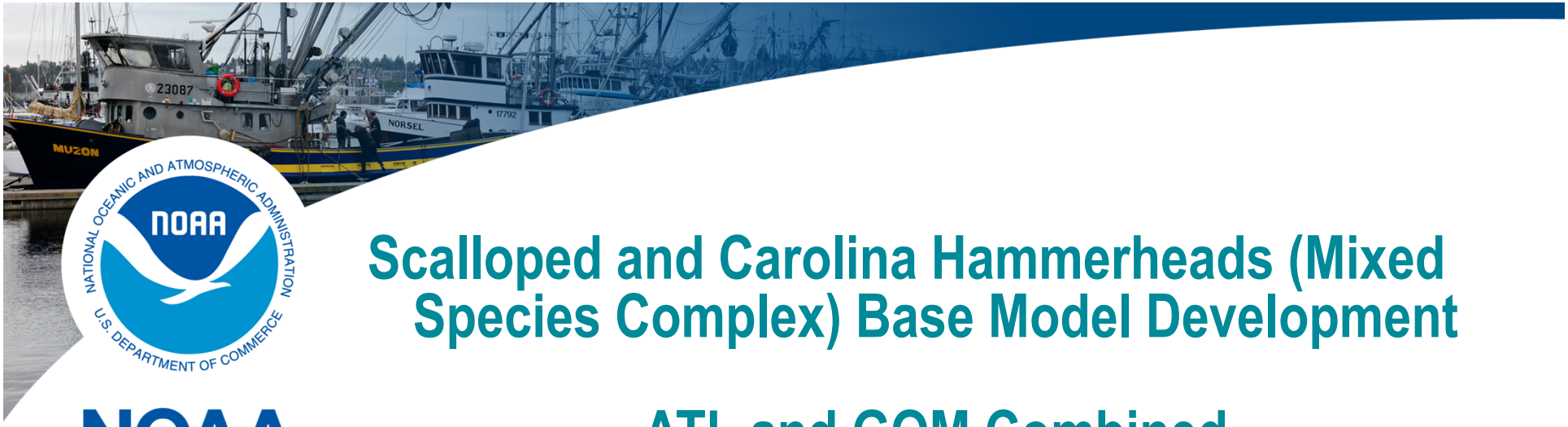


NOAA FISHERIES

SEFSC

Scalloped and Carolina Hammerheads (Mixed Species Complex) Base Model Development

ATL and GOM Combined Stock Synthesis Model Diagnostics

SEDAR 77 (Assessment Webinar IX)

March 21, 2023

Outline

Review Stock Synthesis model update(s) for Webinar 9

Multiple diagnostics implemented for Stock Synthesis model

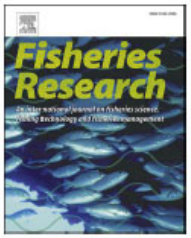
Diagnostics Adapted from Previous Examples

Fisheries Research 240 (2021) 105959

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



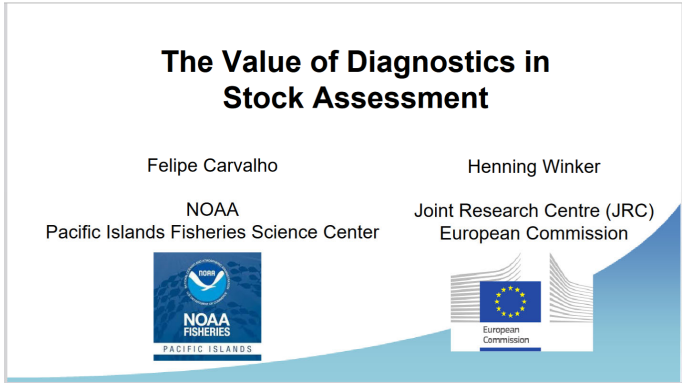
Check for updates

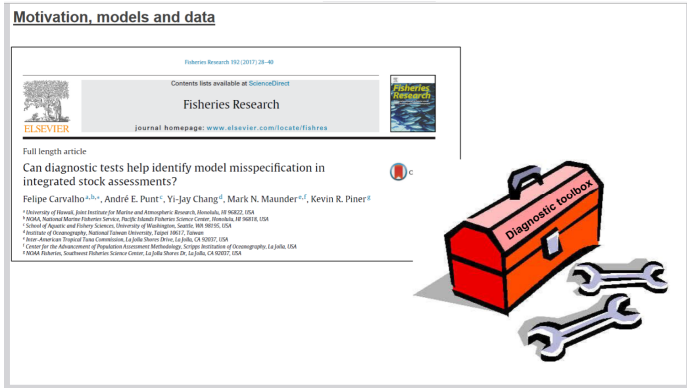
A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho [a,*,1], Henning Winker [b,1], Dean Courtney [c], Maia Kapur [d], Laurence Kell [e], Massimiliano Cardinale [f], Michael Schirripa [g], Toshihide Kitakado [h], Dawit Yemane [i], Kevin R. Piner [j], Mark N. Maunder [k,l], Ian Taylor [m], Chantel R. Wetzel [m], Kathryn Doering [n], Kelli F. Johnson [m], Richard D. Methot [m]

Diagnostics Implemented with r4ss and ss3diags

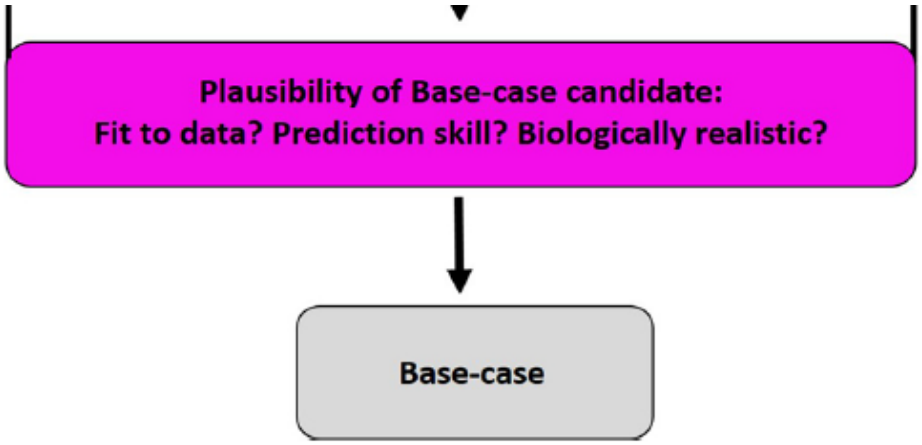
- Stock Synthesis model runs evaluated with r4ss
  - https://github.com/r4ss/r4ss
- Stock Synthesis diagnostics evaluated with ss3diags and r4ss
  - E.g., CAPAM Diagnostics Workshop 2022
    - https://github.com/PIFSCstockassessments/ss3diags
    - http://www.capamresearch.org/content/diagnostics-workshop-presentations
    - The Value of Diagnostics in Stock Assessment
    - www.capamresearch.org/sites/default/files/IATTC_Workshop_Final_Felipe.pdf

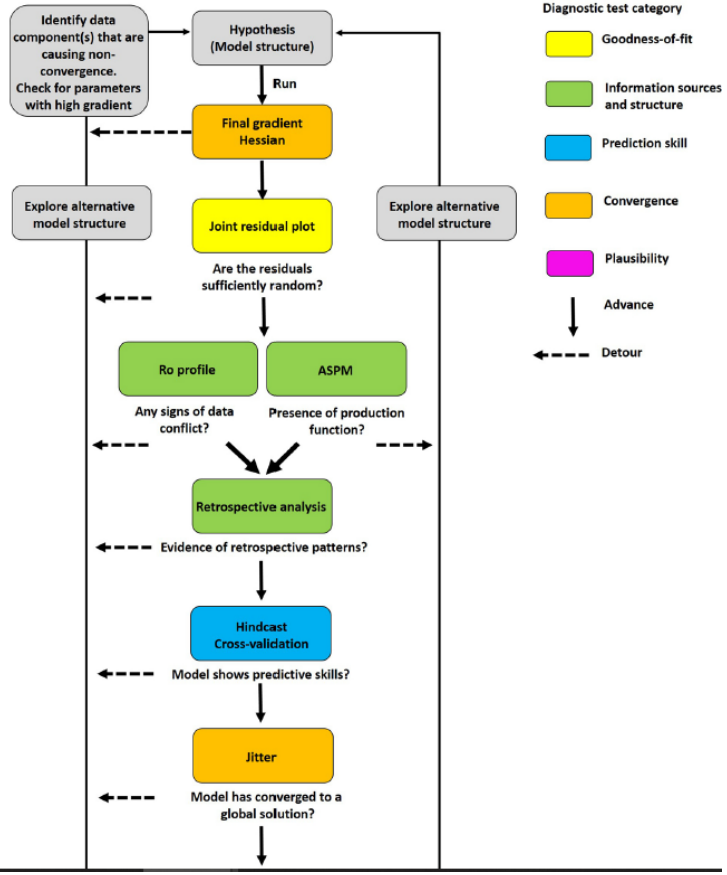




Multiple Diagnostics Evaluated

- Multiple diagnostics evaluated together can provide insight about model plausibility

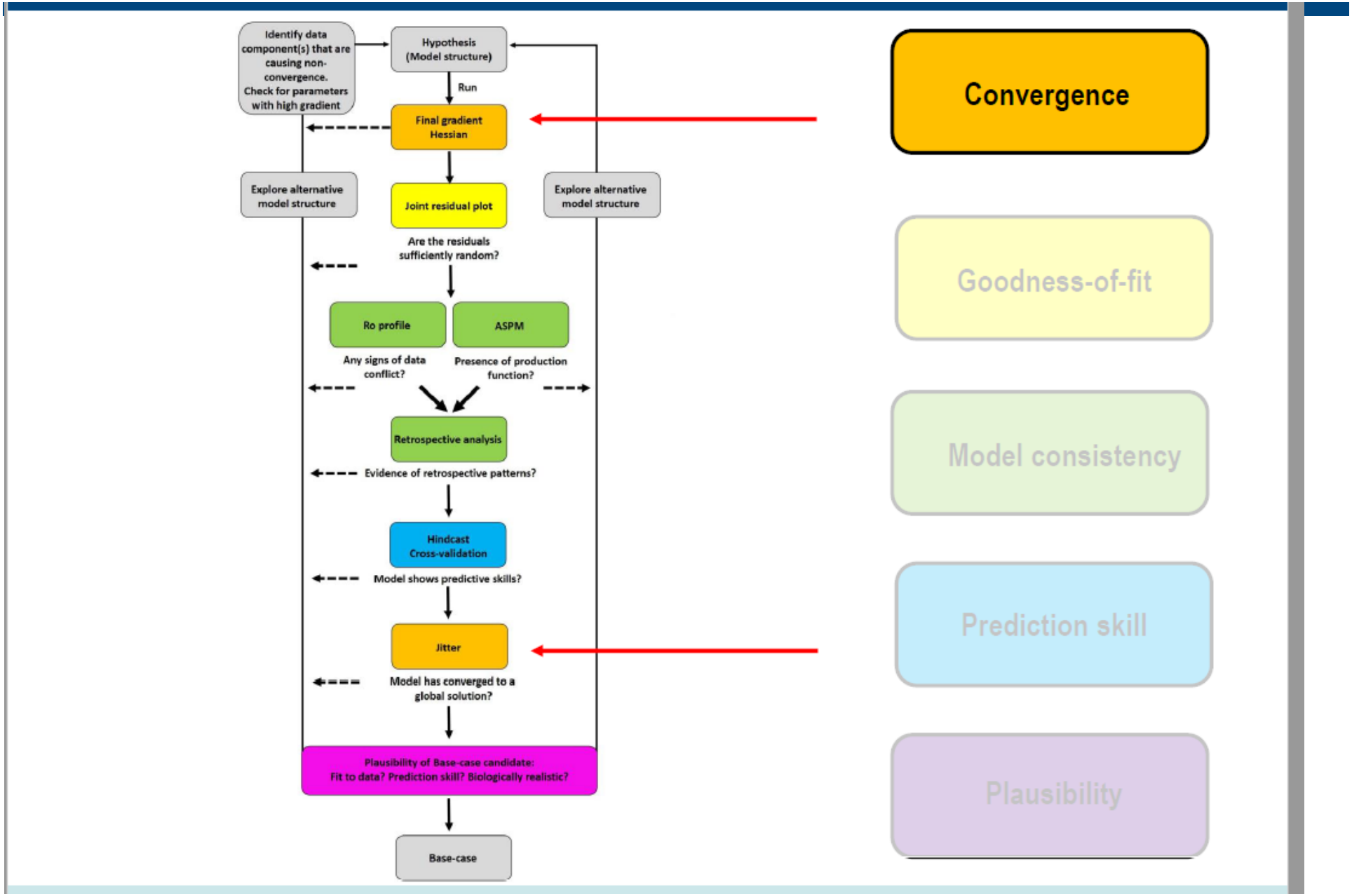




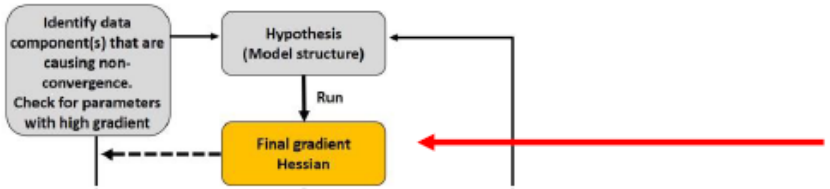
ICCAT_WGSAM 2021

SCRS/P/2021/022. A cookbook for using model diagnostics in integrated stock assessments (Carvalho et al.,)

SCRS/P/2021/020. Ensemble weighting and projections using model validation and prediction skill with ss3diags (Winker et al.,)



SCRS/P/2021/022. A cookbook for using model diagnostics in integrated stock assessments (Carvalho et al.,)



- Parameters estimated at a bound?
- Final gradient is relatively small (e.g., ≤ 1.00E-04)?
- Hessian positive definite?
- Highly correlated parameters?
- Parameters with high variance?

Convergence

Goodness-of-fit

Model consistency

Prediction skill

Plausibility

SCRS/P/2021/022. A cookbook for using model diagnostics in integrated stock assessments (Carvalho et al.,)

Jitter Analysis

- All jitter model runs resulted in total likelihood values equal to or greater than the continuity analysis model configuration (894 likelihood units within rounding error)
- The jitter test did not provide evidence to reject the hypothesis that the continuity analysis model configuration parameter optimization converged to the global solution

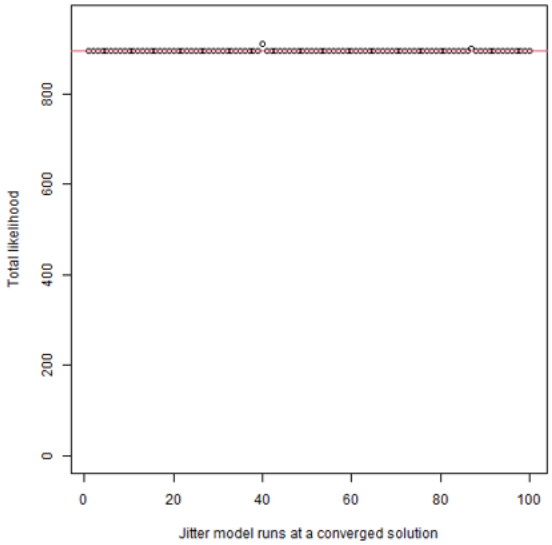


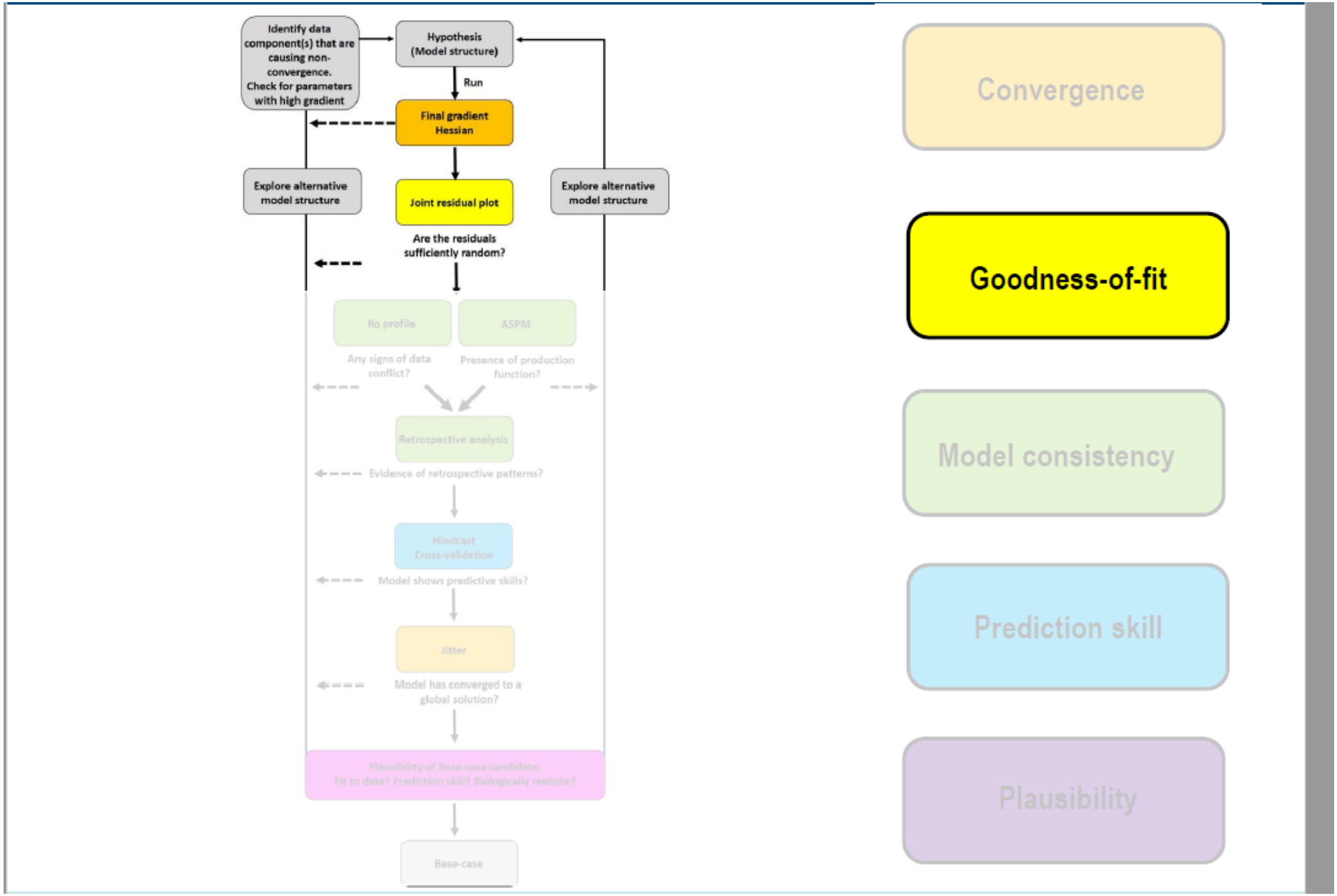
Table C.1. Jitter results for global convergence (100 iterations) obtained as described above for the Stock Synthesis (GOM + ATL) continuity analysis model configuration.

	Likelihood	Frequency
1	894.2[1]	4
2	894.3[2]	79
3	895.0	3
4	895.3	2
5	895.4	10
6	900.5	1
7	911.2	1
	Total	100

[1]Min 894.2
[2]Continuity analysis model configuration 894.3

Diagnostic-1 (Convergence and Jitter)

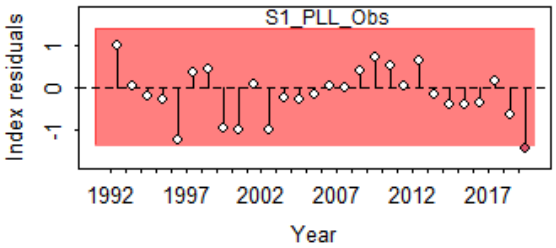
The model passed this diagnostic (except final gradient $3.6 \times 10^{-4} > 1.00e-04$)

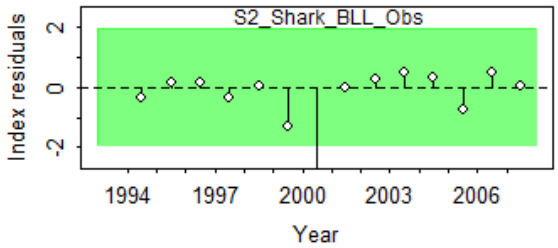


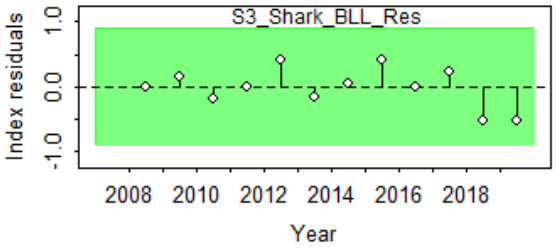
SCRS/P/2021/022. A cookbook for using model diagnostics in integrated stock assessments (Carvalho et al.,)

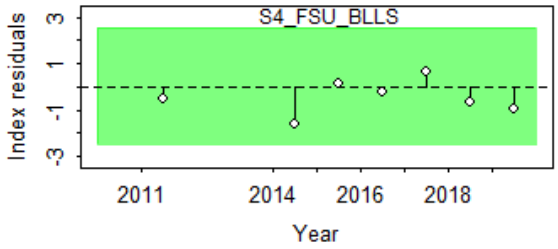
Runs test

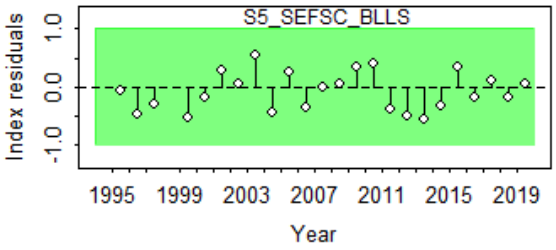
CPUE indices











Convergence

Goodness-of-fit

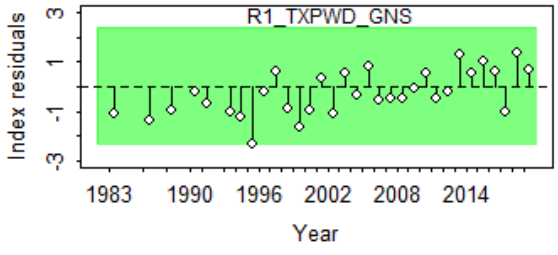
Model consistency

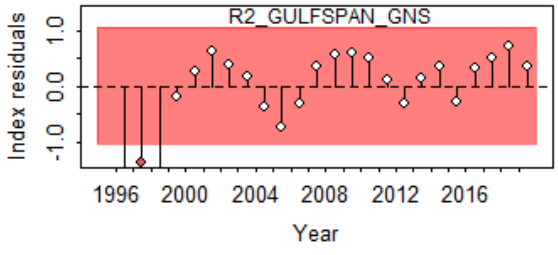
Prediction skill

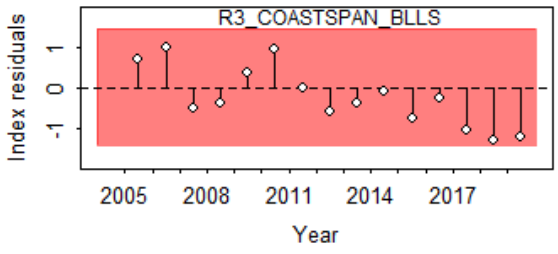
Plausibility

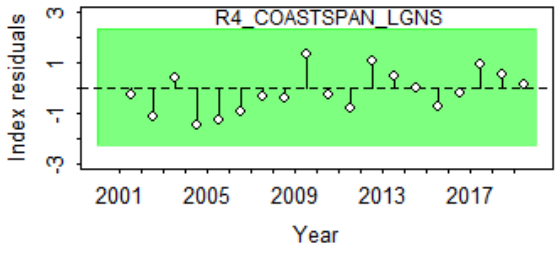
Runs test

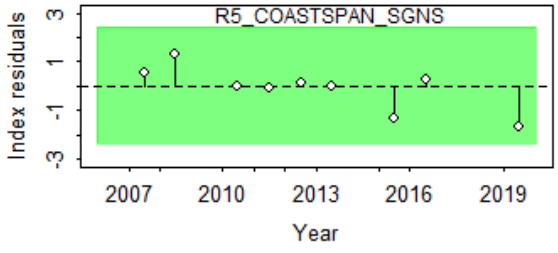
Age-0 CPUE indices











Convergence

Goodness-of-fit

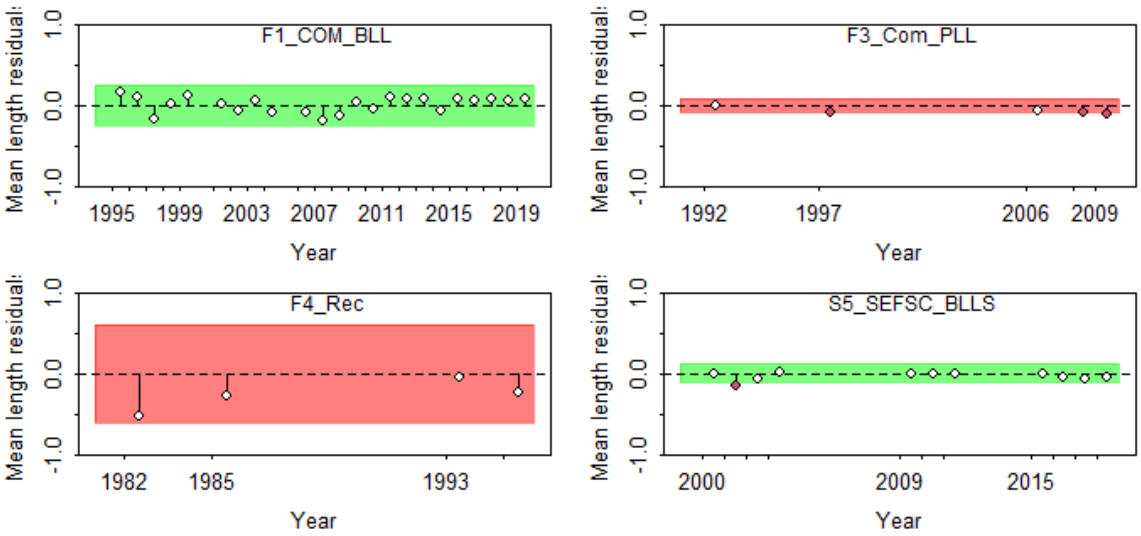
Model consistency

Prediction skill

Plausibility

Runs test

Mean length standardized residuals



Convergence

Goodness-of-fit

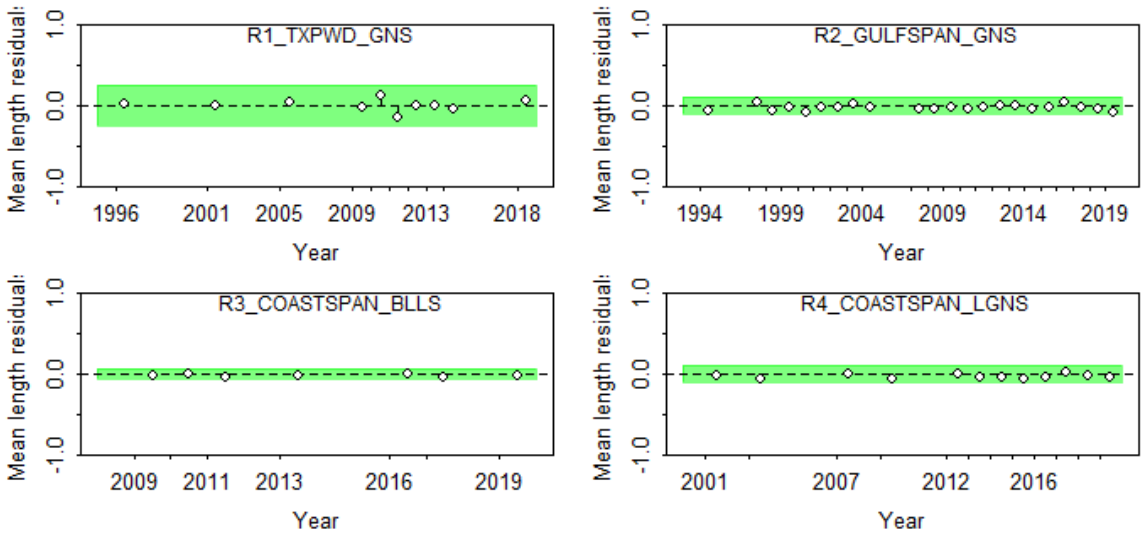
Model consistency

Prediction skill

Plausibility

Runs test

Age-0 mean length standardized residuals



Convergence

Goodness-of-fit

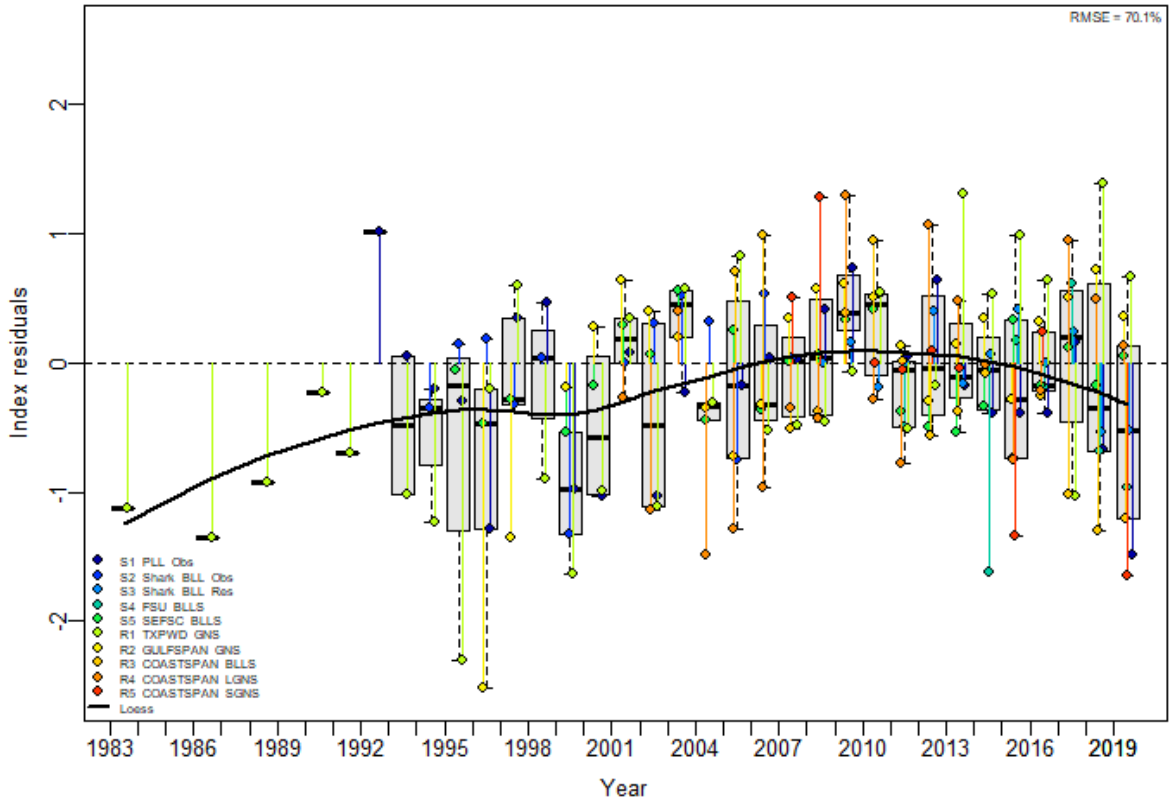
Model consistency

Prediction skill

Plausibility

Joint residual plots

CPUE time series (RMSE = 70.1%)



Convergence

Goodness-of-fit

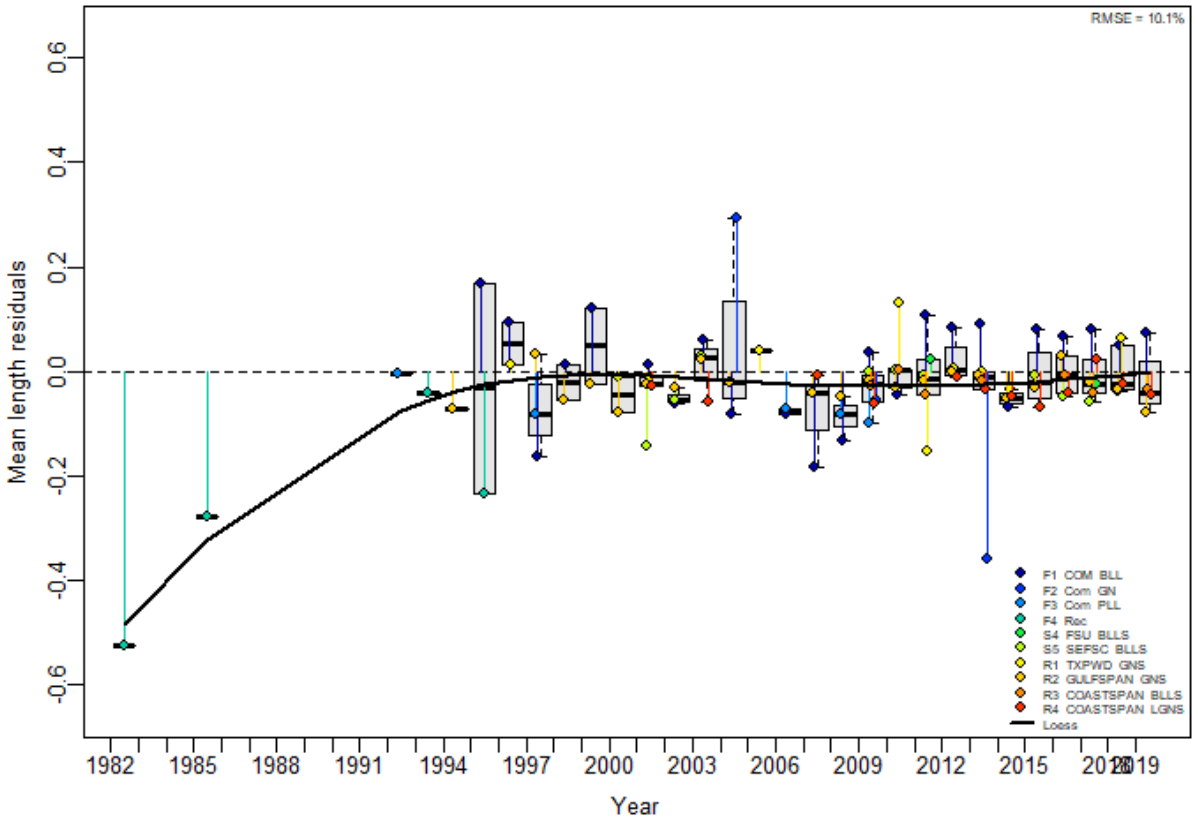
Model consistency

Prediction skill

Plausibility

Joint residual plots

Mean length time series (RMSE 10.1%)



Convergence

Goodness-of-fit

Model consistency

Prediction skill

Plausibility

Diagnostic-2 (Runs test of CPUE and mean length residuals)

The results for this diagnostic were mixed.

There was evidence ($p < 0.05$) to reject the hypothesis of randomly distributed residuals for one survey CPUE index (S1_PLL_Obs) and two age-0 recruitment CPUE indices (R2_GULFSPAN_GNS, R3_COASTSPAN_BLLS)

There was evidence ($p < 0.05$) to reject the hypothesis of randomly distributed residuals for two time series (F3_Com_PLL and F4_Rec)

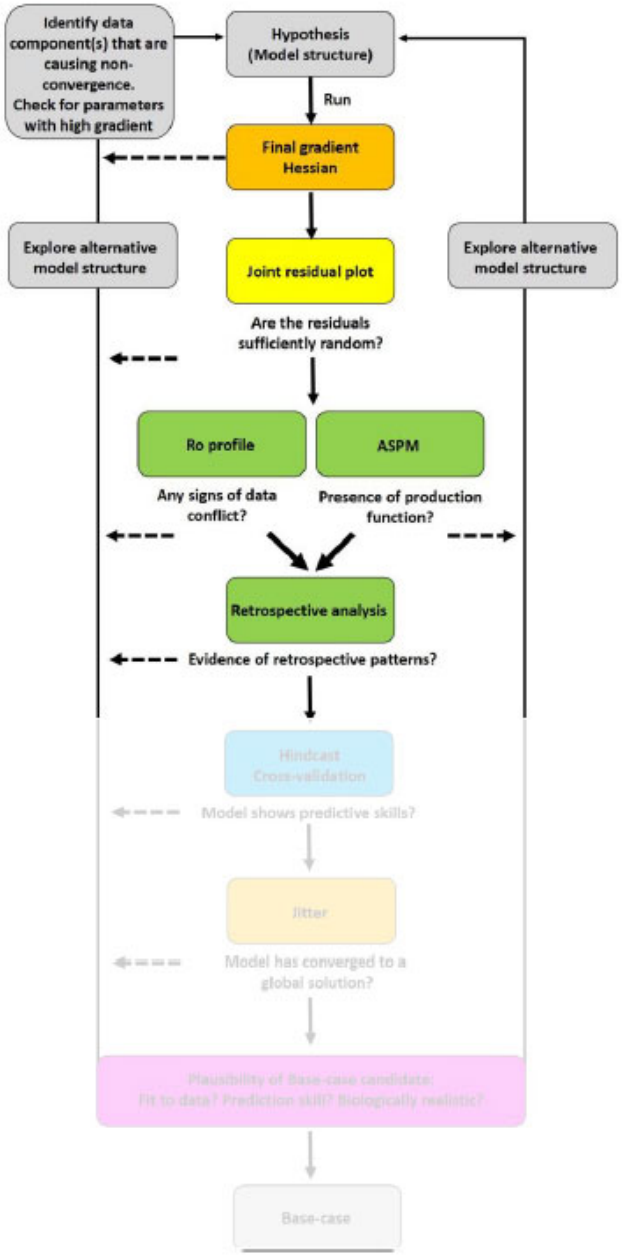
Diagnostic-3 (Joint residual plots and RMSE of CPUE and mean length)

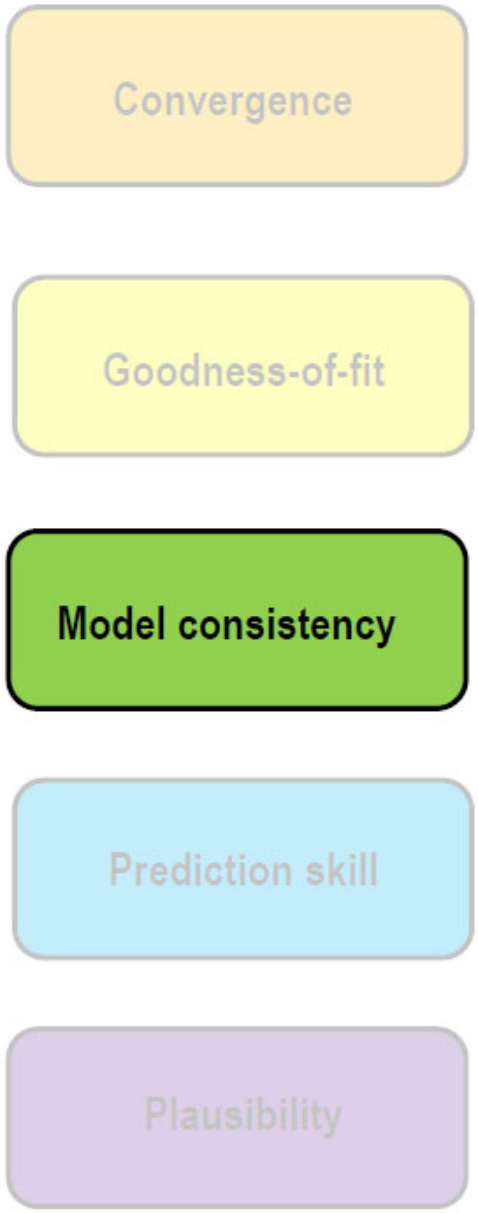
The results for this diagnostic were mixed.

The overall model fit to CPUE was imprecise (root mean square error of all residuals combined, RMSE >> 0.3)

In contrast the overall model fit to mean length was relatively more precise (RMSE < 0.3)

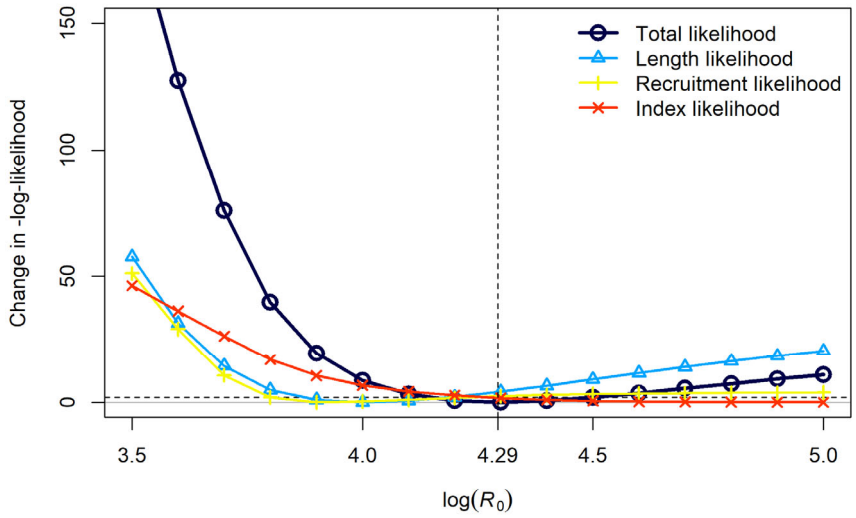
There were also trends in overall residuals for fits to CPUE and mean length, indicated by a loess smoother through all residuals, except for age-0 mean length time series

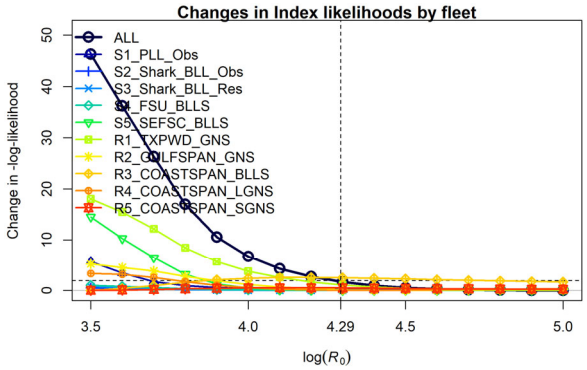


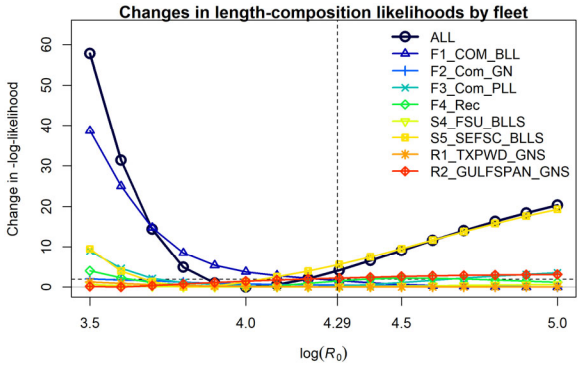


SCRS/P/2021/022. A cookbook for using model diagnostics in integrated stock assessments (Carvalho et al.,)

Log-likelihood profiles for R0







Convergence

Goodness-of-fit

Model consistency

Prediction skill

Plausibility

Diagnostic-4 (Log-likelihood component profiles for R_0)

The results for this diagnostic were mixed

Magnitude of the R_0 profiles indicated that estimation of the recruitment deviations, length composition, and CPUE were about equally informative within the likelihood

Relatively large changes in the magnitude of the R0 profiles for two CPUE time series (S5_SEFSC_BLLS, R1_TXPED_GNS) and two length compositions (F1_COM_BLL, S5_SEFSC_BLLS) indicated that these data sources were relatively more informative than the other data components included in the R0 profile

The location of the minimum negative log-likelihood along the R0 profile for length composition and recruitment were similar (about 4.0)

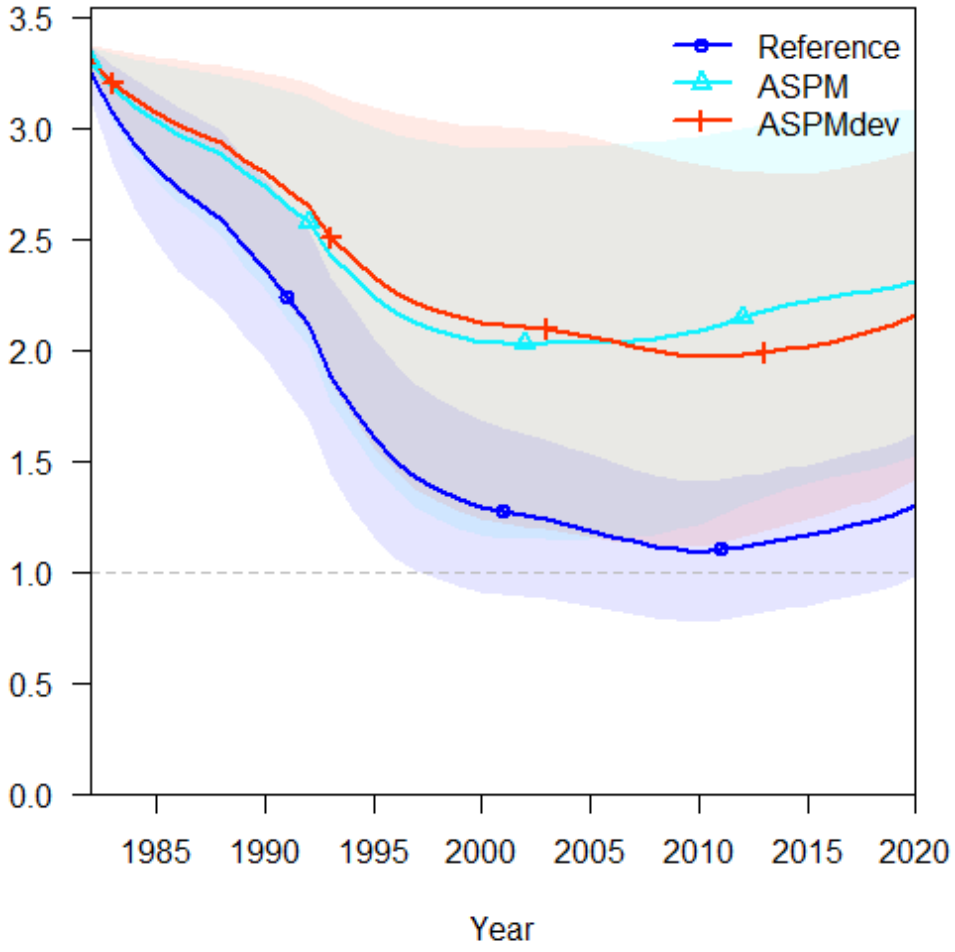
However, a minimum value was not identified for indices of relative abundance, indicating that 1) the scale of the population is driven by fit to length composition, and 2) there is conflict in the minimum likelihood for the Ro profile between data components

A flat profile likelihood or a profile likelihood with its minimum value occurring at a bound suggests that there is an inability to estimate the parameter from any of the data sets and that the parameter should potentially be fixed (Karp et al. 2022).

However, diagnosing which of many confounded model processes lead to the data conflicts is difficult even for stock assessments of targeted species. In particular, the R_0 likelihood component profile by itself performed poorly as a diagnostic to identify model misspecification in a simulation study (Carvalho et al. 2017)

ASPM

Spawning output relative to its MSY value



Convergence

Goodness-of-fit

Model consistency

Prediction skill

Plausibility

Diagnostic-5 (ASPM)

The results of this diagnostic were mixed

The large asymptotic 95% confidence intervals of relative spawning stock size for the ASPMs did not overlap the full integrated stock assessment model for recent years, indicating highly divergent results between the reference model and the ASPMs

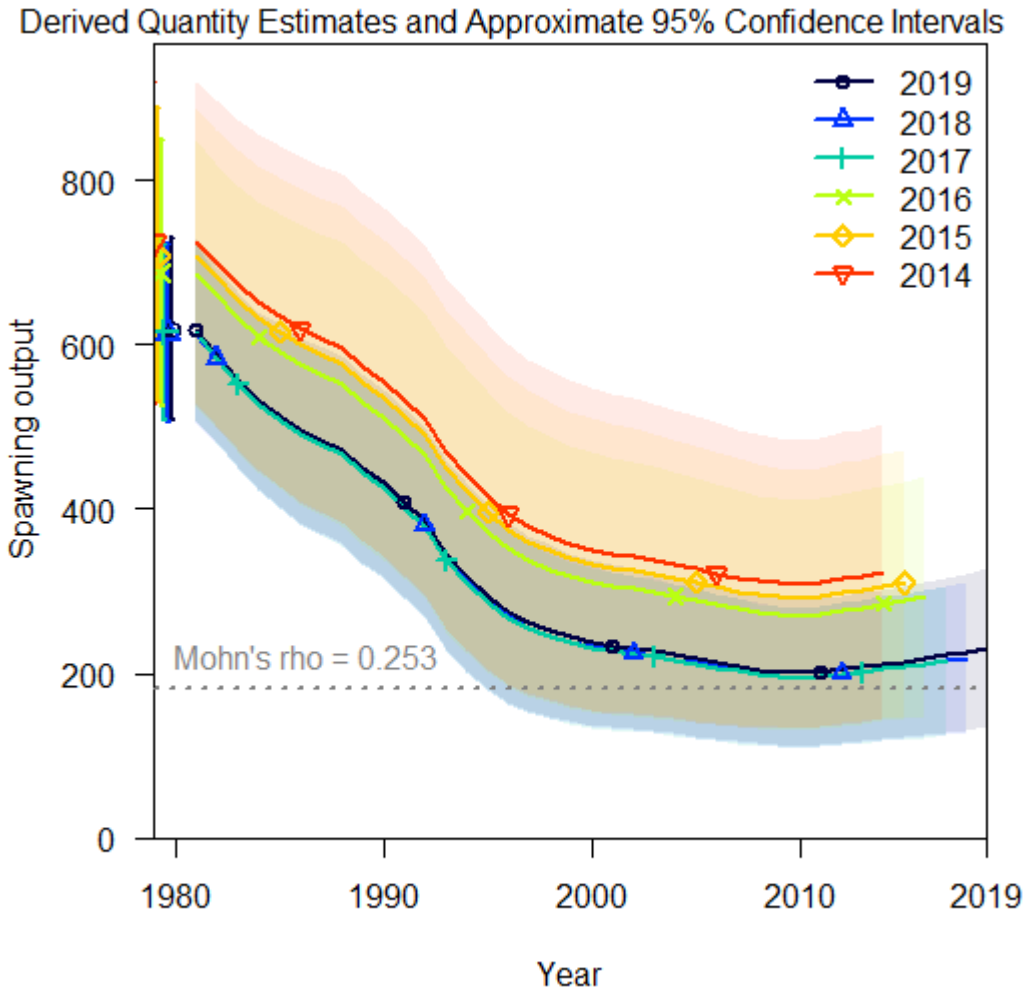
Consequently, the ASPM results indicate that the observed catches alone could not explain the trend in the indices of abundance and hence that the data available to the ASPM (i.e., the indices of abundance and the catch) did not provide enough information to estimate the scale of the population (e.g., see Punt 2023).

The differences observed between the full integrated stock assessment model compared to the ASPMs indicate that the fit to length composition data inform the estimated stock size.

As discussed in Minta-Vera et al. (2017), there is a trade-off within the fully integrated model between the fit to composition data (in general used to estimate recruitment) and the influence of fits to length composition on absolute abundance through a catch-curve type process.

The tradeoff was addressed in this assessment by right weighting the data following A Francis (2011) two-stage data weighting approach implemented in the base model configuration.

Retrospective patterns and Mohn's Rho test



Convergence

Goodness-of-fit

Model consistency

Prediction skill

Plausibility

Diagnostic-6 (Retrospective patterns and Mohn's Rho test)

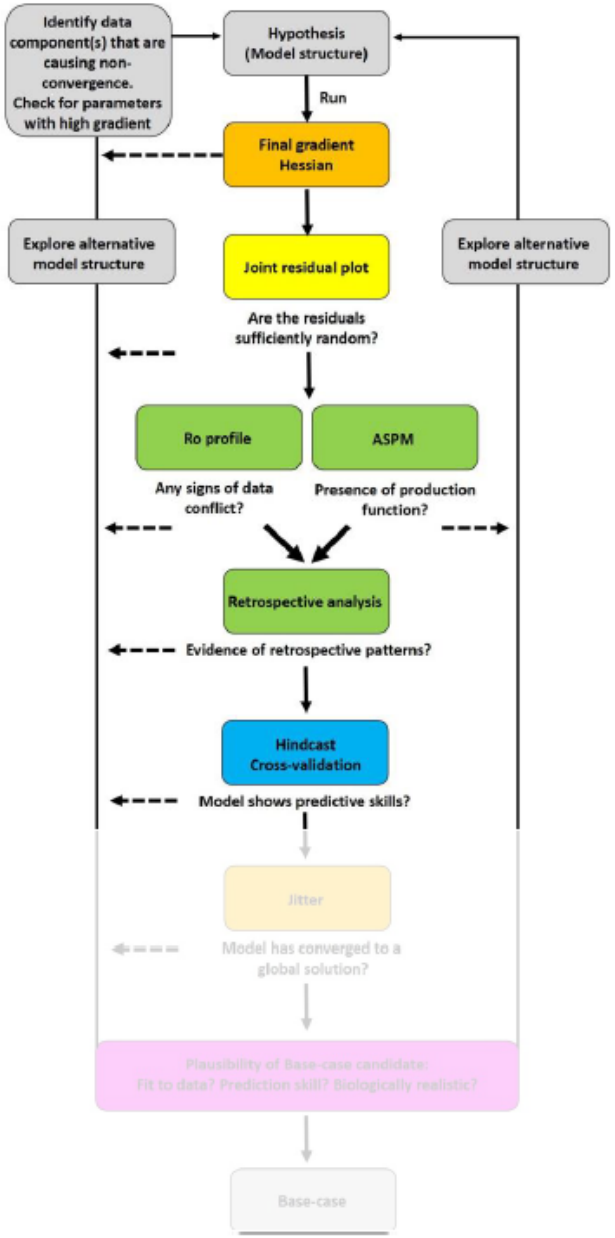
The model failed this diagnostic

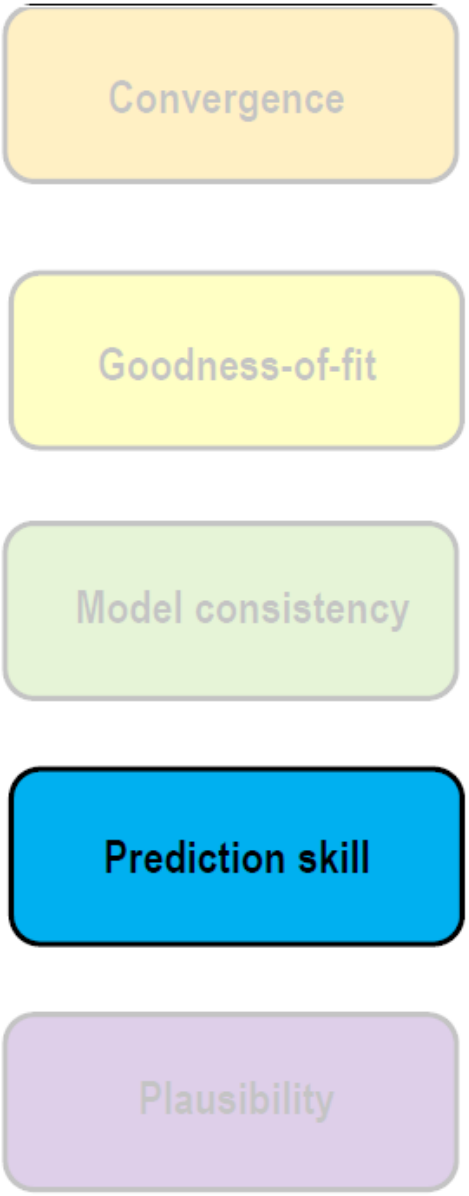
Mohn's rho was calculated for spawning biomass with a five year peel

The severity of the retrospective pattern was based on the range provided by Hurtado-Ferro et al. (2015), with values higher than 0.20 and lower than -0.15 used as an indication for problematic retrospective patterns

The model exhibited a retrospective pattern in recent years, with Mohn's rho values for spawning biomass (2.5) > 0.20

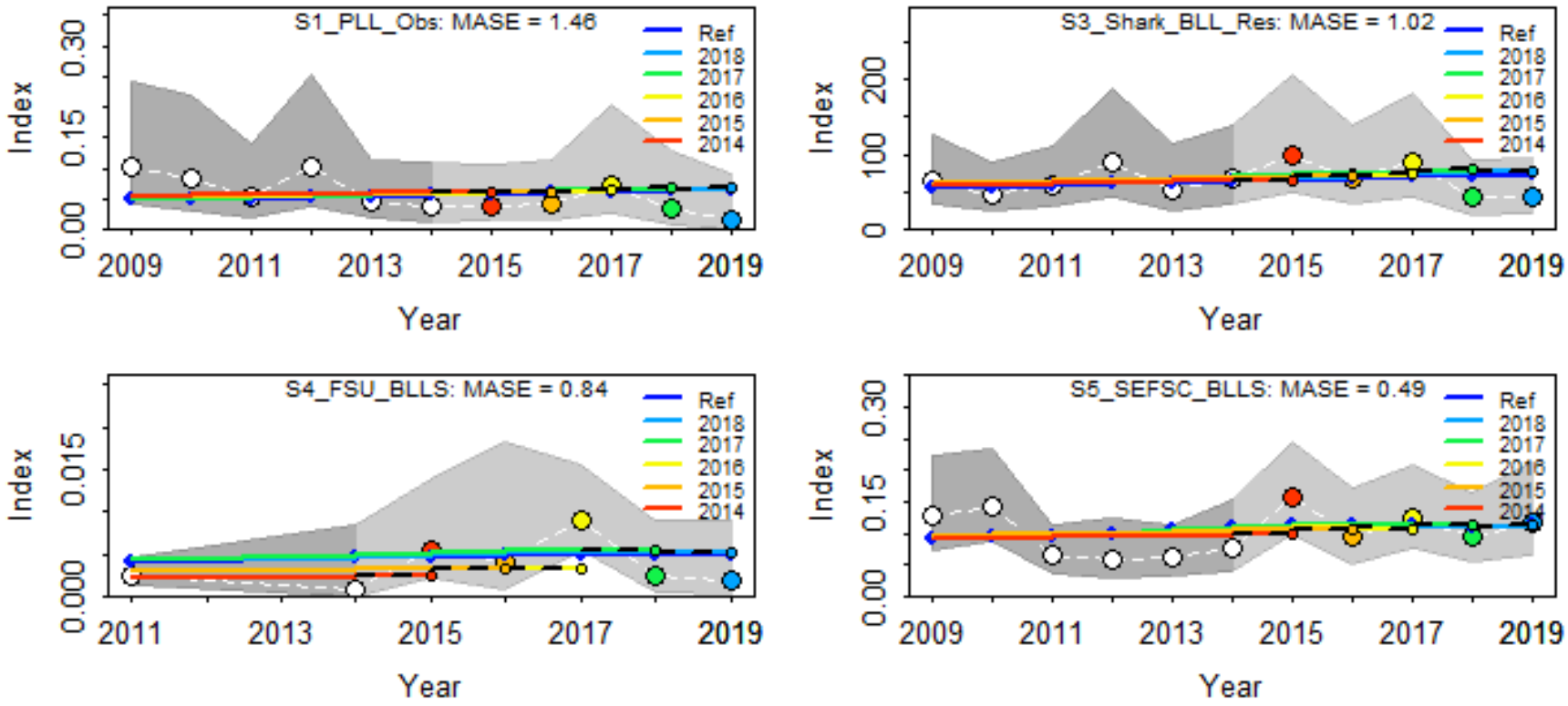
This result indicates that there is an apparent tendency to overestimate spawning biomass in recent years 2014, 2015, and 2016, but not 2017, and 2018





SCRS/P/2021/022. A cookbook for using model diagnostics in integrated stock assessments (Carvalho et al.,)

Hindcasting cross validation (HCxval) CPUE indices

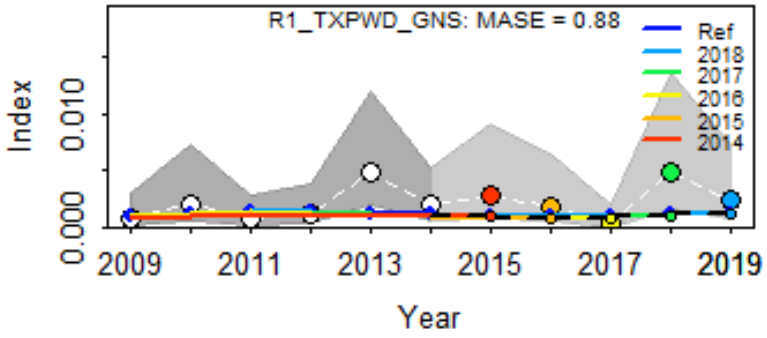


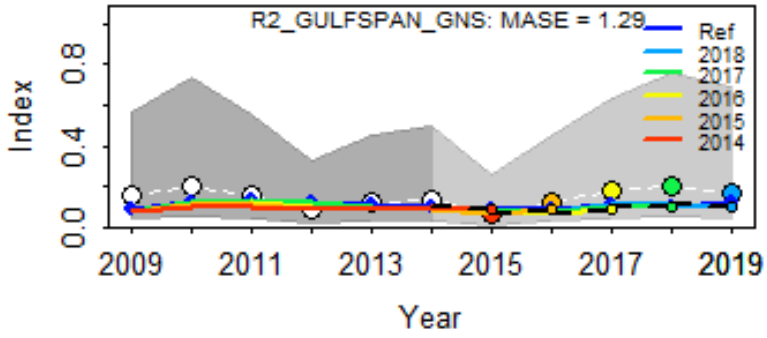
The most accurate CPUE index predictions were observed for S5_SEFSC_BLLS (MASE 0.5)

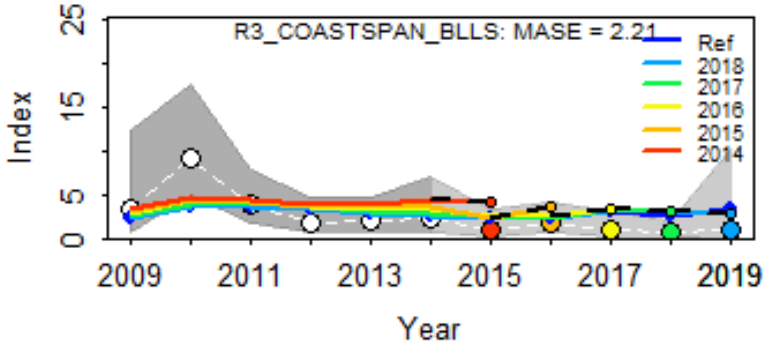
Four CPUE indices failed the diagnostic and four CPUE indices passed the diagnostic

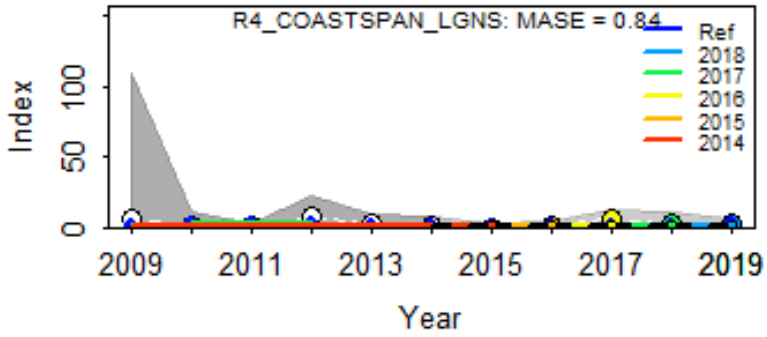
Predictions for CPUE time series were all relatively flat (neither increasing nor decreasing within the period 2014 – 2018

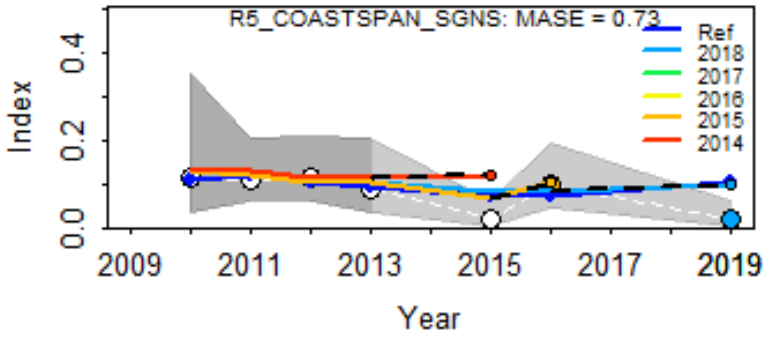
Hindcasting cross validation (HCxval) age-0 CPUE indices



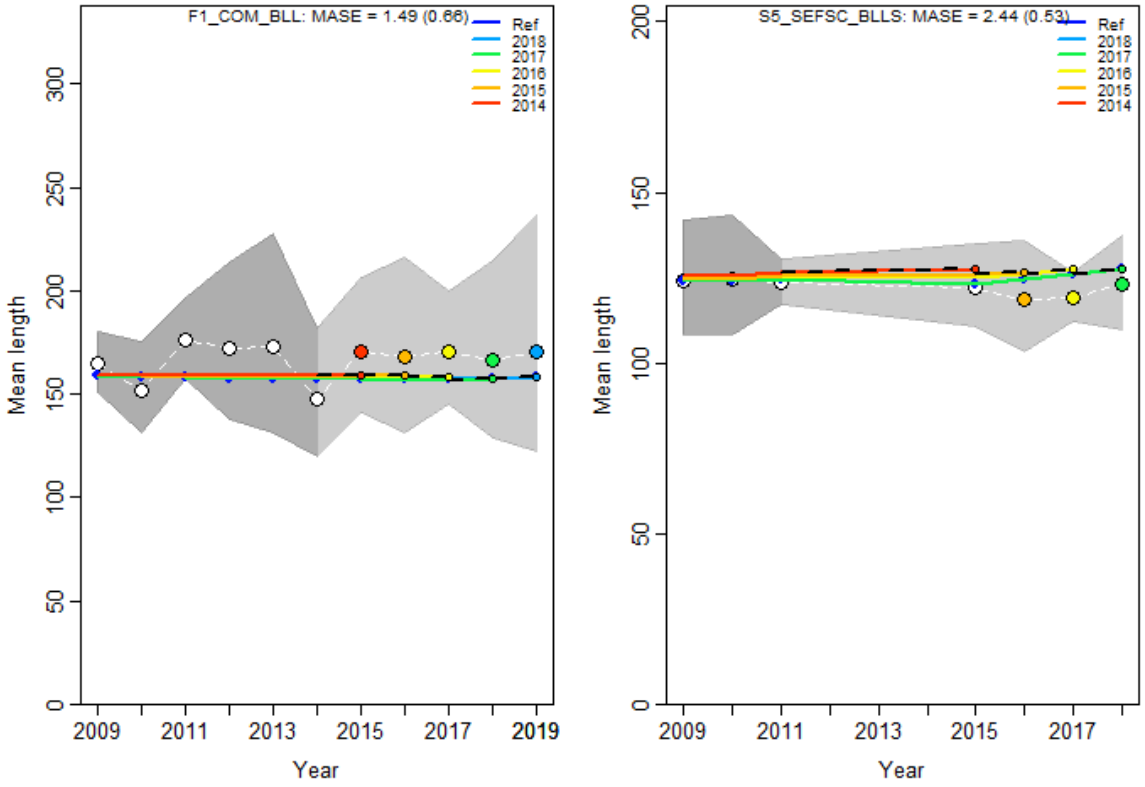








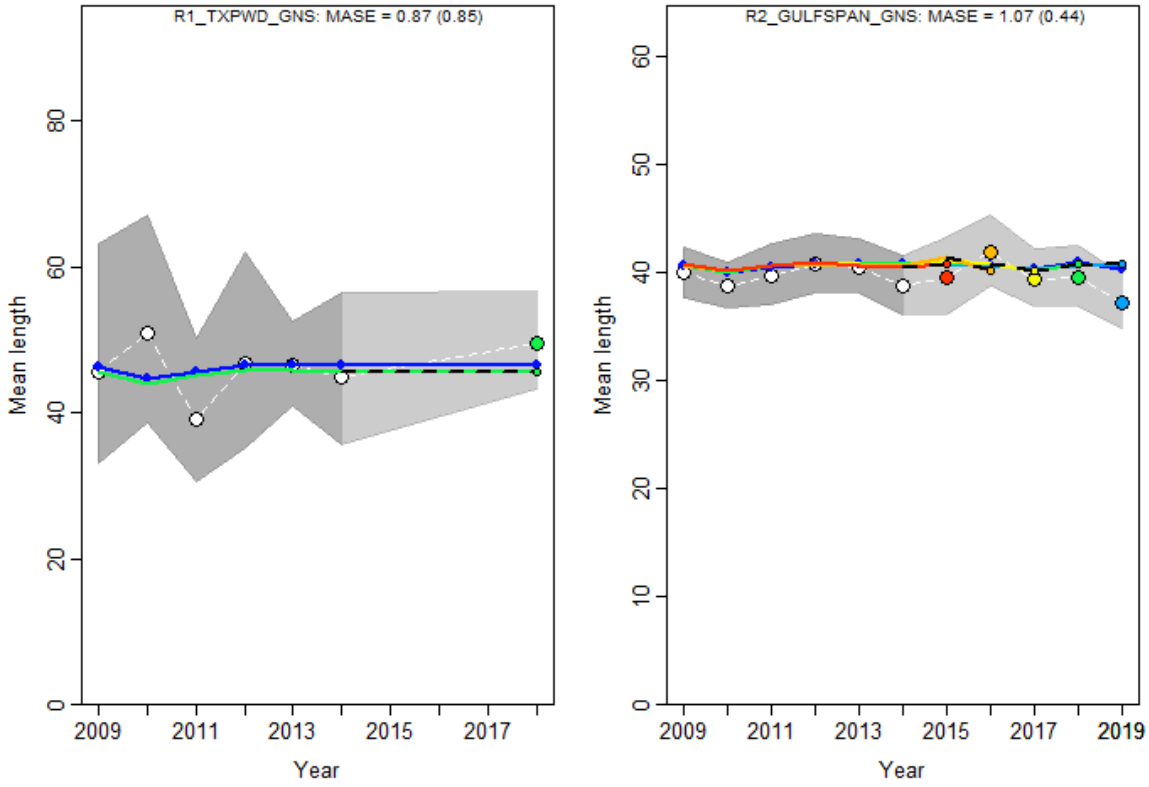
Hindcasting cross validation (HCxval) mean length time series



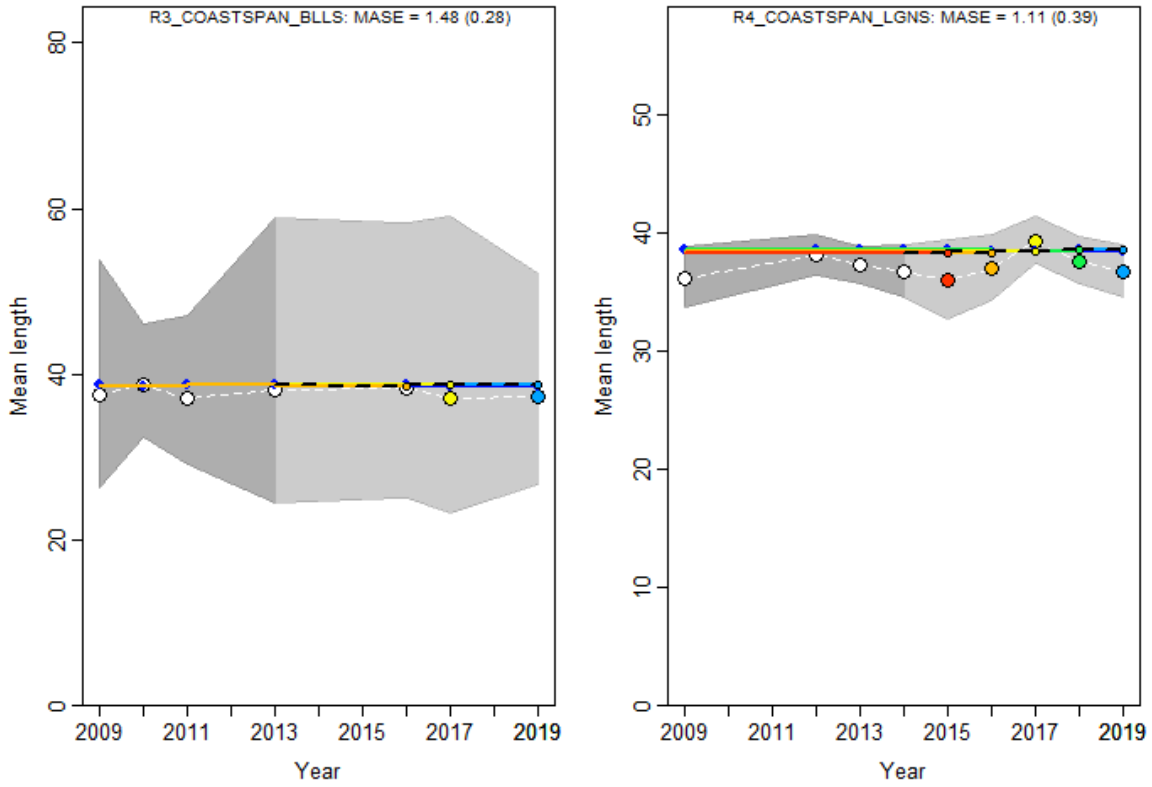
All three length composition time series with complete observations (n = 5) within the hindcast evaluation period 2014 – 2018 passed the HCXval diagnostic

R2_GULFSPAN_GNS and R4_COASTSPAN_LGNS (MASE.adj 0.4)

Hindcasting cross validation (HCxval) mean length time series



Hindcasting cross validation (HCxval) mean length time series



Diagnostic-7 (Hindcasting cross validation)

The results for this diagnostic were mixed for CPIE indices.

The model passed this diagnostic for mean length time series.

The hind-cast cross-validation diagnostic identified that four CPUE indices failed the diagnostic and four CPUE indices passed the diagnostic. CPUE indices which failed the diagnostic had poor prediction skill. An explanation may be that either the indices are not proportional to relative abundance or that there are processes that are not being accounted for in the model structure.

In the latter case fits to length composition may be driving trends in abundance. This interpretation is consistent with the R0 likelihood component profile, which indicated that the minimum R0 profile of the population is driven by fit to fit to length composition data and that there is conflict in the minimum likelihood for the Ro profile between data components.

This could be investigated further by considering a range of scenarios based on alternative datasets and model structures. Hindcasting could then be used to identify the best performing scenarios (e.g., choice of models and data which inform abundance from CPUE data and inform recruitment from length composition data) by comparing predictions with observations in the updated models with updated hind-cast cross-validation.

Diagnostic-1 (Convergence and Jitter)
The model passed this diagnostic (except final gradient $3.6x10^{-4}$ > 1.00e-04)

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

Diagnostic-3 (Joint residual plots and RMSE of CPUE and mean length)
The results for this diagnostic were mixed.

Diagnostic-4 (Log-likelihood component profiles for R_0)
The results for this diagnostic were mixed.

Diagnostic-5 (ASPM)
The results of this diagnostic were mixed.

Diagnostic-6 (Retrospective patterns and Mohn's Rho test)
The model failed this diagnostic.

Diagnostic-7 (Hindcasting cross validation)
The results for this diagnostic were mixed for CPIE indices.
The model passed this diagnostic for mean length time series.

Fisheries Research 261 (2023) 106642

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Those who fail to learn from history are condemned to repeat it: A perspective on current stock assessment good practices and the consequences of not following them

André E. Punt [a,b,*]

[a] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98185-5020, USA
[b] CSIRO Oceans and Atmosphere, GPO Box 1538, Hobart, TAS 7001, Australia

Overall, a model would be considered adequate for providing management advice if the optimization was successful, the model fits the data adequately (e.g., based on residual analysis), the model provides reliable estimates of trends and scale, the results of the model are consistent when updated with new data (e.g., retrospective analysis), and the model is able to make adequate future predictions (e.g., hindcasting) (Carvalho et al., 2021). It is generally best practice to apply a range of diagnostics. The diagnostics used most commonly are:

Convergence diagnostics.

Residual diagnostics.

Retrospective analysis.

Hindcast cross-validation.

Likelihood profiling.

Other diagnostics. Some diagnostics (e.g., the Age-structured Production Model, ASPM, diagnostic; … the catch curve diagnostic … ASPM diagnostic was developed to assess whether surplus production and observed catches alone could explain the trend in the index of abundance and hence whether the data (i.e., the indices of abundance) provide information on the scale of the population.…

Fisheries Research 261 (2023) 106642

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Those who fail to learn from history are condemned to repeat it: A perspective on current stock assessment good practices and the consequences of not following them

André E. Punt [a,b,*]

[a] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98185-5020, USA
[b] CSIRO Oceans and Atmosphere, GPO Box 1538, Hobart, TAS 7001, Australia

Overall, the ideal is to apply as many diagnostic analyses as possible, along with running sensitivity analyses to explore sensitivity even within a model that exhibits no obvious problems, recognizing that currently available diagnostics are not guaranteed to identify all problems or uncertainties. Carvalho et al. (2017) found that applying multiple diagnostics was likely to identify more problems, without a major increase in 'Type I error', i.e., incorrect reject of a correctly specified model. **Few assessments apply all of the above diagnostics and the minimum set would seem to be to evaluate convergence and model fit (as summarized using residuals) and to conduct a retrospective analysis and construct likelihood profiles.** The hindcast and the ASPM diagnostics can be used to better understand the "value" of the assessment (for example, is it any better than a simple AR-1 process) and its properties. **Weighting of alternative model configurations using diagnostics remains a research area unfortunately.**

Model Uncertainty with MVLN in ss3diags

SCRS/2019/145 Collect. Vol. Sci. Pap. ICCAT, 76(6): 725-739 (2020)

PROJECTIONS TO CREATE KOBE 2 STRATEGY MATRIX USING THE MULTIVARIATE LOG-NORMAL APPROXIMATION FOR ATLANTIC YELLOWFIN TUNA

John Walter[1] & Henning Winker[2]

SUMMARY

IOTC-2019-WPTT21-51

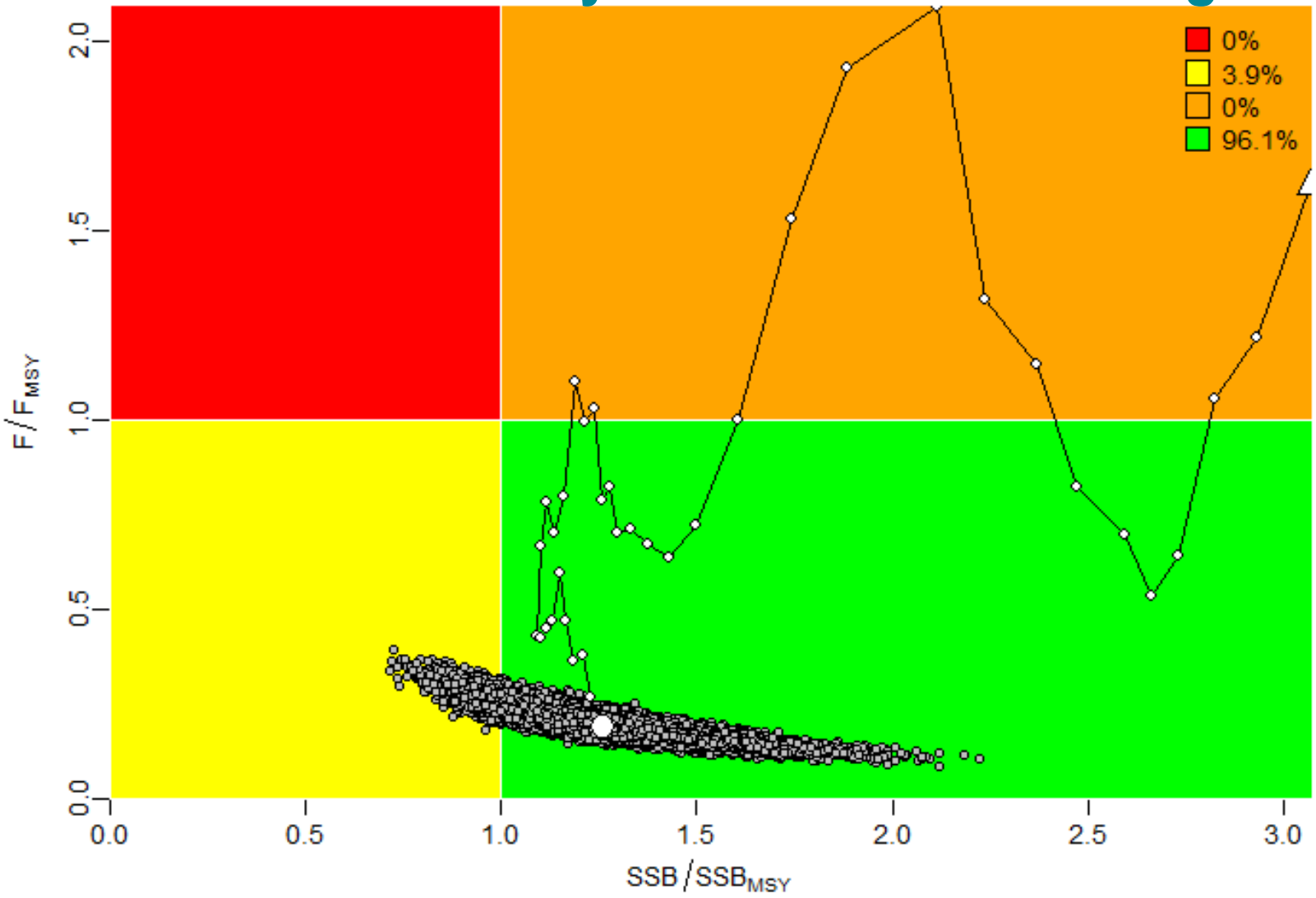
A multivariate lognormal Monte-Carlo approach for estimating structural uncertainty about the stock status and future projections for Indian Ocean Yellowfin tuna

Henning Winker[1,] John Walter[2], Massimiliano Cardinale[3], and Dan Fu[4]*

ICCAT_WGSAM 2021

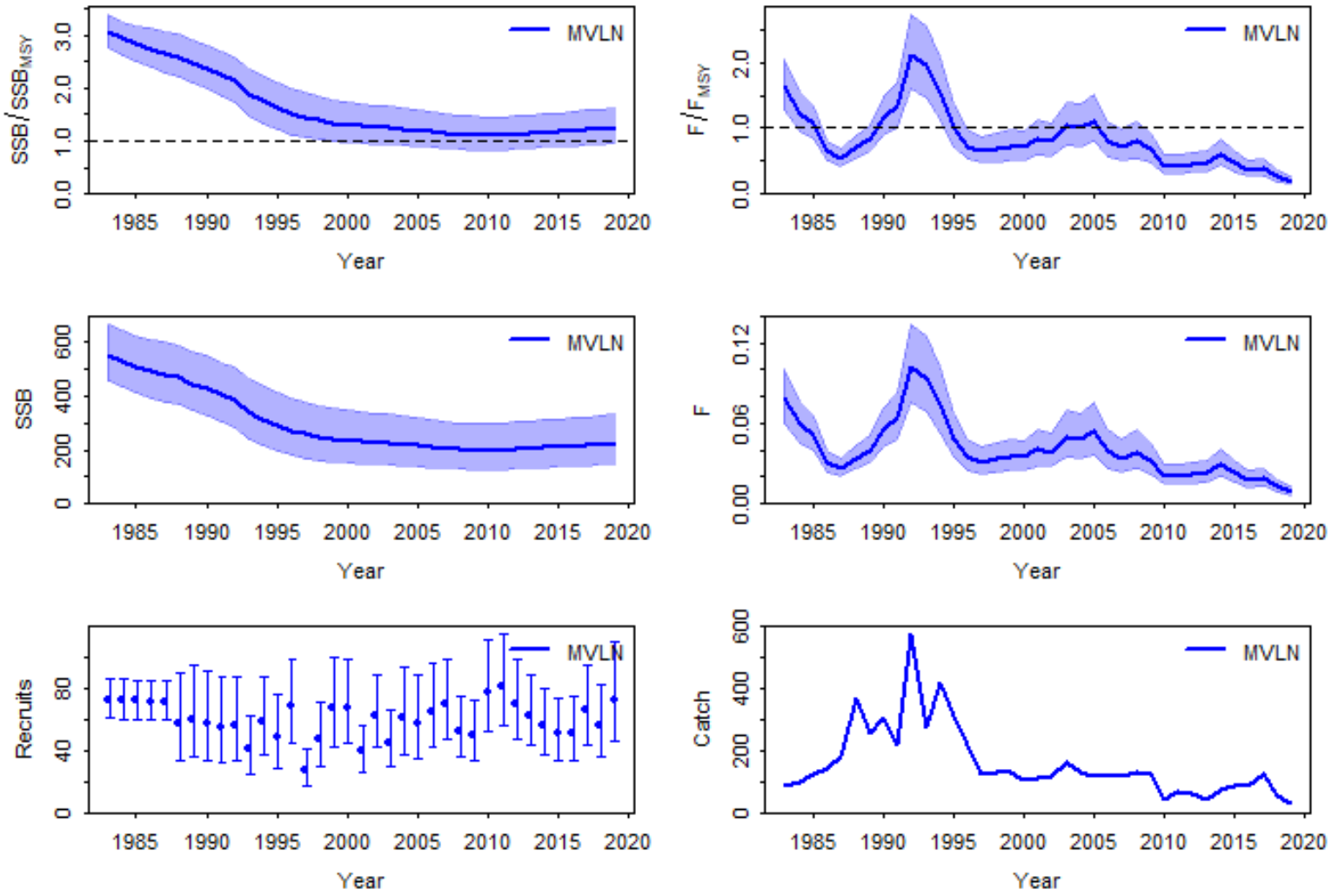
SCRS/P/2021/020. Ensemble weighting and projections using model validation and prediction skill with ss3diags (Winker et al.,)

Model Uncertainty with MVLN in ss3diags



mvln.SHH = SSdeltaMVLN(reference.SHH.output,mc=5000)

Model Uncertainty with MVLN in ss3diags



SSplotEnsemble(mvln.SHH$kb,ylabs=mvln.SHH$labels,add=T,verbose=F)

Carvalho, F., Punt, A. E., Chang, Y.-J., Maunder, M. N., and Piner, K. R. 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. 192:28–40. Available: https://doi.org/10.1016/j.fishres.2016.09.018 (3/7/2023).

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., and R. D. Methot. 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. 240:105959.
Available: https://doi.org/10.1016/j.fishres.2021.105959 (3/7/2023).

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

Karp, M. A., Kuriyama, P., Blackhart, K., Brodziak, J., Carvalho, F., Curti, K., Dick, E. J., Hanselman, D., Hennen, D., Ianelli, J., Sagarese, S., Shertzer, K., and I. Taylor. 2022. Common model diagnostics for fish stock assessments in the United States. NOAA Tech. Memo. NMFS-F/SPO-240, 28 p.

Minte-Vera, C. V., Maunder, M. N., Aires-da-Silva, A. M., Satoh, K., and K. Uosaki. 2017. Get the biology right, or use size-composition data at your own risk. Fish. Res. 192:114–125. Available: https://doi.org/10.1016/j.fishres.2017.01.014 (3/7/2023).

Punt, A. E. 2023. Those who fail to learn from history are condemned to repeat it: A perspective on current stock assessment good practices and the consequences of not following them. Fish. Res. 261:106642.
Available: https://doi.org/10.1016/j.fishres.2023.106642 (3/8/2023).