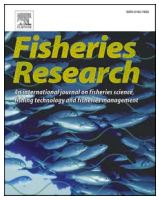
Fisheries Research 240 (2021) 105959

Contents lists available at ScienceDirect

Fisheries Research

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



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]

[a] NOAA Fisheries, Pacific Islands Fisheries Science Center, Honolulu, HI, United States
[b] Joint Research Centre (JRC), European Commission, Ispra, Varese, Italy
[c] NOAA Fisheries, Southeast Fisheries Science Center, Panama City Laboratory, Panama City, FL, United States
[d] School of Aquatic and Fisheries Sciences, University of Washington, Seattle, WA, United States
[e] Center for Environmental Policy, Imperial College London, London, United Kingdom
[f] Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Turistgatan, Lysekil, Sweden
[g] NOAA Fisheries, Southeast Fisheries Science Center, Miami, FL, United States
[h] Department of Marine Biosciences, Tokyo University of Marine Science and Technology, Konan, Minato, Tokyo, Japan
[i] Department of Environment, Forestry and Fisheries, Foretrust Building, Cape Town, South Africa
[j] NOAA Fisheries, Southwest Fisheries Science Center, La Jolla, CA, United States
[k] Inter-American Tropical Tuna Commission, La Jolla, CA, United States
[l] Center for the Advancement of Population, Assessment Methodology (CAPAM), United States
[m] NOAA Fisheries, Northwest Fisheries Science Center, Seattle, WA, United States
[n] Caelum Research Corporation Under Contract to National Marine Fisheries Service - Northwest Fisheries Science Center, Seattle, WA, United States

ARTICLE INFO

Handled by Simon Hoyle

Keywords:
Stock Synthesis
Model development
Hindcast
Profiling
Convergence
ASPM

ABSTRACT

Integrated analysis has increasingly been the preferred approach for conducting stock assessments and providing the basis for management advice for fish and invertebrate stocks around the world. Many decisions are required when developing integrated stock assessments. For example, the analyst needs to decide whether the model fits the data, if the optimization was successful, if estimates are consistent retrospectively, and if the model is suitable to predict future stock responses to fishing. This study provides practical guidelines for implementing selected diagnostic tools that can assist analysts in identifying problems with model specifications and alternatives that can be explored to minimize or eliminate such problems. Emphasis is placed on reviewing the implementation and interpretation of contemporary model diagnostic tools. We first describe each diagnostic approach and its utility. We then proceed by providing a "cookbook recipe" on how to implement each of the diagnostics, together with an interpretation of the results, using two worked examples of integrated stock assessments with Stock Synthesis. Further, we provide a conceptual flow chart that lays out a generic process of model development and selection using the presented model diagnostics. Based on this, we propose the following four properties as objective criteria for evaluating the plausibility of a model: (1) model convergence, (2) fit to the data, (3) model consistency, and (4) prediction skill. It would greatly benefit the stock assessment community if the next generation of stock assessment models could include the diagnostic tests presented in this study as a set of open source tools.

1. Introduction

Integrated analysis used for stock assessment combines several sources of data into a single model using a joint likelihood for the observed data (Fournier and Archibald, 1982; Maunder and Punt, 2013). For the assessment of exploited fish populations, these data may include records of landings, indices of abundance from research surveys, tagging data, and the composition of size classes and/or ages present in samples.

* Corresponding author.
E-mail addresses: felipe.carvalho@noaa.gov (F. Carvalho), henning.winker@ec.europa.eu (H. Winker).
[1] These authors contributed equally as lead authors to this work.

https://doi.org/10.1016/j.fishres.2021.105959
Received 4 August 2020; Received in revised form 19 March 2021; Accepted 22 March 2021
Available online 13 April 2021

Several general stock assessment software packages for implementing integrated analysis have been widely used around the world (Dichmont et al., 2016), including CASAL (Bull et al., 2005), MULTIFAN-CL (Fournier et al., 1998) and Stock Synthesis (Methot and Wetzel, 2013; https://github.com/nmfs-stock-synthesis/stock-synthesis).

Misspecification of key parameters or assumptions in integrated stock assessment models can strongly impact the estimates of quantities of management interest, such as stock depletion and biomass at maximum sustainable yield (Mangel et al., 2013). Model misspecifications can include incorrect specifications of important biological parameters, such as somatic growth (Minte-Vera et al., 2017), maturation (Thorson et al., 2019), or natural mortality (Lee et al., 2011). Model misspecifications can also arise due to incorrect specifications of selectivity functions (Ichinokawa et al., 2014; Punt et al., 2014; Vasilakopoulos et al., 2020) and variance parameters (Francis, 2011; Truesdell et al., 2017), or by not accounting for spatial stock structure (Goethel et al., 2011; Punt, 2019) or temporal variation in recruitment (Thorson et al., 2019), selectivity (Stewart and Monnahan, 2017), or any of the above listed biological processes. Failing to account for an important process can also lead to conflicting information among data sets (Francis, 2011; Ichinokawa et al., 2014) and retrospective and forecast bias (Brooks and Legault, 2016; Carvalho et al., 2017; Miller and Legault, 2017). Current solutions to data conflict include eliminating one of the conflicting data sources or, nearly equivalently, reducing the contribution of one of the conflicting data sources to the likelihood (i.e., data-weighting) when fitting the model (Maunder and Piner, 2017; Wang and Maunder, 2017). However, these approaches deal with the symptoms rather than the underlying causes of data conflicts (Wang et al., 2015) and leave intact model misspecifications that can affect estimates of management quantities. Thus, recognizing the source and impact of misspecified components is crucial for providing accurate advice to managers.

Tools to diagnose poor fits to the data and determine which data sources are in conflict can be used as starting places to identify model misspecification. Several diagnostics have been evaluated for their utility to identify poor fits to data and data conflicts within integrated stock assessment models (Carvalho et al., 2017; Lee et al., 2014; Maunder and Piner, 2017; Punt et al., 2014). Such model diagnostics range from graphical visualization and basic goodness-of-fit statistics to computationally intense techniques that can involve iterative refitting and profiling. Carvalho et al. (2017) tested several new and existing diagnostics (i.e., residual analysis, retrospective analysis, likelihood component profiling, age-structured production models −ASPMs, and catch curve analysis −CCA). They found that no single diagnostic worked well in all the evaluated cases and recommended the use of a selection of diagnostics (i.e., a diagnostic toolbox) to increase the ability to detect model misspecification while acknowledging that the use of multiple diagnostics may increase the probability that a diagnostic test results in a false positive.

Diagnostic tests are important in determining the robustness of estimates for management advice in integrated stock assessment models. For example, Maunder et al. (2020) developed a risk-based framework that assigns weights to models in an ensemble of candidate models, which involved the results of several diagnostics tests. In some cases, a simple fix within the assessment process can improve model diagnostics; in other cases, dedicated research studies are necessary to improve models outside the operational process (Eero et al., 2015; ICES, 2019). Maunder and Piner (2017) proposed a procedure based on diagnostic tests to guide the construction of stock assessment models and reduce model misspecification evidenced by data conflicts. Their procedure for model construction consisted of two components: (1) avoiding, diagnosing and fixing data conflicts, and (2) facilitating the interpretation of diagnostics results. Maunder and Piner (2017) also provided a flow chart to help users complete the various steps involved in model construction in an optimized sequence.

This paper first reviews the most recent developments of model diagnostic tools with a focus on integrated stock assessment models. We then proceed by providing "cookbook recipes" on how the diagnostic tools can be implemented and interpreted. We illustrate our cookbook recipes based on two worked examples of integrated stock assessments with Stock Synthesis: (1) North Atlantic shortfin mako (*Isurus oxyrinchus*) and (2) the Pacific hake (or Pacific whiting, *Merluccius productus*) off the west coast of the United States and Canada. Emphasis is placed on presenting a guide to produce graphics and supporting statistics to enable wider use of important diagnostic techniques, such as the runs test for residual analysis (Carvalho et al., 2017), model validation techniques using hindcast cross-validations (Kell et al., 2016), and deterministic ASPMs (Maunder and Piner, 2015). The model diagnostic process and recommended steps are discussed, and a conceptual flow chart for model development is provided. Although the diagnostics are presented for Stock Synthesis models, they apply to integrated, age-structured statistical catch-at-age/size population models that use multiple datasets and a variety of model structures.

2. Contemporary model diagnostic tools

In this section, we provide an overview of several current model diagnostics and how the analyst can use them when developing a stock assessment by following a flow chart. The diagnostics are grouped into the following four categories: convergence, goodness-of-fit, model consistency, and prediction skill.

Our process (Fig. 1) comprises a series of interconnected diagnostic tests that should be carried out to establish a base model (Carvalho et al., 2017) or an ensemble of candidate models (Maunder et al., 2020). In general, the process flows from top to bottom and shows when a diagnostic test is recommended (Fig. 1). The flow chart includes several 'detour' options to consider when model diagnostics do not show satisfactory results. In the process, we interpret a detour as a model exploration that might not necessarily lead to changes but typically includes some additional analysis to help justify modelling decisions. Although it sometimes becomes apparent that the model needs to be altered while engaging in a detour, the goal should be to explore alternative model formulations that could contribute to fixing the problems. We recognize that the model construction process varies depending on the analyst and the stock assessment, and thus, we sought to propose a flow chart that could work for the majority.

2.1. Convergence

There are several useful diagnostic checks for evaluating the convergence of a model, but when looked at in isolation, none of these convergence diagnostics alone may be sufficient to demonstrate convergence or the lack of it decisively. Therefore, model convergence should be assessed using several considerations. The first step is checking for parameters estimated at a bound, which can indicate problems with data or the assumed model structure. The second is checking that the final gradient (i.e., the terminal degree of descent of the objective function, which becomes lower or less steep, as the function approaches a minimum) is relatively small (e.g., $\leq$ 1.00E-04). A small final gradient is not an absolute requirement as our experience is that successful model outcomes can be obtained despite larger final gradients. The third is to determine whether the Hessian (i.e., the matrix of second derivatives of the log-likelihood concerning the parameters, from which the asymptotic standard error of the parameter estimates is derived) is positive definite. Parameters on bounds or with abnormal variance or covariance can prevent positive definite Hessian attainment. Models that are far from converged may also not attain a positive definite Hessian, but a positive definite Hessian is not, in itself, an indication of model convergence. Other convergence diagnostics include (i) examining the correlation matrix for highly correlated (e.g., $>$ 0.95) parameter pairs; and (ii) examining parameters for excessively high variance as an indication that they do not influence the fit to the data.

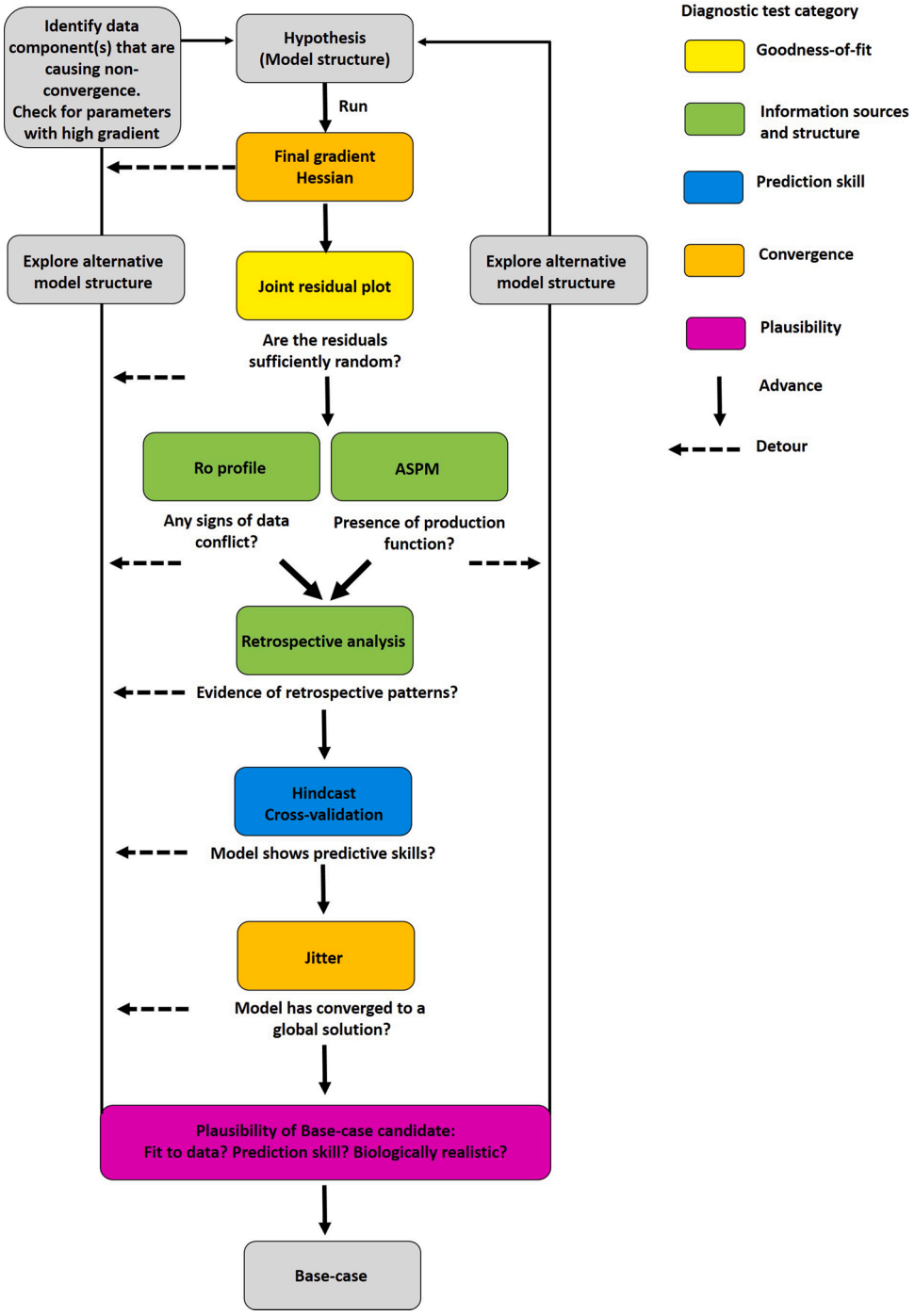


Fig. 1. Conceptual process flow chart illustrating a series of interconnected diagnostic tests recommended when developing a base model (or an ensemble of candidate models). The arrows represent broad guidance regarding the flow along the axis of diagnostic types from top to bottom. If diagnostic tests fail, a 'detour' path highlights the need for model exploration before advancing.

Highly correlated parameters and uninformed parameters can contribute to solutions that are spurious and possibly numerically unstable. In some instances, excessively high coefficients of variation (CVs) on estimated quantities (e.g., $>$ 100 %) can indicate insufficiently informative input data to estimate time-varying processes (e.g., recruitment or selectivity), if model convergence can even be attained in such over-parameterized situations (Methot and Wetzel, 2013).

Once individual model convergence has been established, 'jittering' the parameters' starting values and re-running the model is commonly used to evaluate whether the model has converged to a global solution rather than a local minimum. The primary check of jittering is to ensure that none of the randomly generated starting values of parameters results in a solution that has a smaller total negative log-likelihood than the reference model. However, the absence of a local minimum when running jittering is not a guarantee that the model is not stuck in a local minimum (Subbey, 2018). The jitters' magnitude should be done judiciously as extreme jitters could start the model search in an unrealistic place from which it cannot detect gradients pointing towards reasonable solutions.

2.2. Goodness-of-fit

Systematic misfit to data should be considered a sign of model misspecification. Unacceptable model fits (i.e., model estimates which do not match the data) can be detected by either the magnitude of the residuals being larger than implied by the observation error or the presence of trends in residuals (e.g., over time or age).

Plotting residuals is a simple method to observe trends, patterns, and variations in data fit over time (e.g., bias, drift, skewness, heavy tails, correlation with states or driving inputs, and heteroscedasticity). Technically, a random distribution of residuals will fall below or above the median 50 % of the time. However, analysts are also interested in whether the probability of being on either side of the median varies with time. The presence of temporal autocorrelation in residuals is evident by systematic drifts in the residual mean throughout time. The Wald and Wolfowitz (1940) runs test is a nonparametric hypothesis test for randomness in a data sequence that calculates the 2-sided *p*-value to estimate the number of runs (i.e., sequences of values of the same sign) above and below a reference value.

Another common goodness-of-fit statistic is the root mean square error (RMSE; Carvalho et al., 2017), which describes the standard deviation of residuals, such that

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t}\left(\hat{y}_t - y_t\right)^2}{n}} \quad (1)$$

where $\hat{y}_t$ is the predicted value at time step t, y_t is the observed value, and n is the number of observations. The RMSE can be interpreted as the standard deviation of the unexplained variance, an analog to the standard error. A relatively small RMSE ($\leq$ 0.3) indicates a reasonably precise model fit to relative abundance indices (Winker et al., 2018). However, to interpret the RMSE correctly, it is important to consider the observation error assumptions, mainly whether the observation error is an estimable quantity (e.g., estimable with additional variance) or fixed a priori (Winker et al., 2018). For example, if an abundance index is thought to be associated with a large sampling error, a fixed coefficient of variation (CV) larger than 0.3 may be assigned to that index a priori (Francis et al., 2003). In such a case, a small RMSE may point either towards a misspecified variance assumption or an over-fitted model. In general, we argue that the RMSE is not suitable to judge the goodness-of-fit across different time series in integrated assessments and should not be used for model selection purposes in isolation.

Winker et al. (2018) introduced a joint residual plot that incorporates several of the above features: lognormal residuals of abundance indices color-coded by fleet with combined RMSE; boxplots indicating the median and quantiles of all residuals available for any given year, with the area of each box indicating the strength of the discrepancy between abundance index (larger boxes indicate a higher degree of conflicting information); and a loess smoother through all residuals, which highlights systematically auto-correlated residual patterns. Here, we extended the implementation of this joint residual plot to the mean-length and mean-age residuals derived from observed and expected length- and age-composition data, respectively (Francis, 2011).

Similarly, the runs test can diagnose model misspecification using residuals from fits to abundance indices. It can also be applied to other data components in assessment models such as the mean-length residuals and mean-age residuals. In addition to the runs test, it is also recommended to look for patterns in residuals that may indicate the presence of non-random variation, for example, serially correlated residuals causing a systemic residual pattern (Punt et al., 2014) or obvious outliers. To objectively detect outliers, the three-sigma limit can be used to identify if any data point would be unlikely given a random process error in the observed residual distribution if it is further than three standard deviations away from the expected residual process average of zero (see details in Anhøj and Olesen, 2014).

2.3. Model consistency

2.3.1. Information sources and structure

A key model diagnostic developed to identify the influence of information sources on model estimates is the likelihood component profile (Ichinokawa et al., 2014; Lee et al., 2014; Wang et al., 2014). This diagnostic reports the likelihood over each data component across a particular parameter profile. The equilibrium recruitment parameter, R_0, is commonly profiled because it represents an ideal global scaling parameter given that unfished (virgin) recruitment is proportional to unfished biomass (Lee et al., 2014; Maunder and Piner, 2015; Wang et al., 2014). A profile of R_0 is conducted by sequentially fixing R_0 to a range of values and then examining the change in the total and data-component likelihoods. A relatively large change in negative log-likelihood units along the profile suggests a relatively informative data source for that particular model. Also, a difference in the location of the minimum negative log-likelihood along the profile between data sources might suggest either conflict in the data or model misspecification (or both). Profiling other scaling parameters (e.g., current biomass) or derived quantities should also be considered (Maunder and Starr, 2001).

The application of an Age-Structured Production Model (ASPM) diagnostic can detect misspecification of key systems-modeled processes that control the shape of the production function (Carvalho et al., 2017). This diagnostic evaluates whether the net effect between surplus production and observed catches alone could explain trends in the index of abundance versus a more complex model that uses annual deviations in recruitment to improve the fit to trends in the data. In the absence of information from length- or age-composition data (i.e., likelihood weighting of zero), all selectivity parameters in the ASPM are fixed to the estimated values from the fully integrated model (for further details, see section 3.2.2 ASPM Diagnostic). Maunder and Piner (2017) suggest that if the ASPM fits well to the indices of abundance that have good contrast (i.e., those that have declining as well as increasing trends), the production function is likely to drive the stock dynamics and the indices will provide information about absolute abundance (Minte-Vera et al., 2017). On the other hand, if there is not a good fit to the indices, then the catch data and the production function alone cannot explain the trajectories depicted in the indices of relative abundance. This can have several causes: the stock is recruitment-driven; the stock has not yet declined to the point at which catch is a major factor influencing abundance; the base-case model is misspecified because complex dynamics such as stock structure are being ignored, and thus, the signal in catches is lost; and the indices of relative abundance are not proportional to abundance. The ASPMdev is a variation of the ASPM diagnostic and designed to evaluate if composition data is needed to estimate the variability in recruitment (Minte-Vera et al., 2017). It involves fitting to indices of abundance while simultaneously estimating recruitment deviates in the absence of the composition data. Suppose the ASPMdev produces results substantially different from the fully integrated model and the ASPM. This would indicate that the composition data provide the primary source of information for estimating recruitment deviations.

2.3.2. Retrospective analysis

Retrospective analysis (Brooks and Legault, 2016; Carvalho et al., 2017; Hurtado-Ferro et al., 2015; Miller and Legault, 2017) is commonly used to check the consistency of model estimates, i.e., the invariance in spawning stock biomass (SSB) and fishing mortality (F) as the model is updated with new data in retrospect. The retrospective analysis involves sequentially removing observations from the terminal year (i.e., peels), fitting the model to the truncated series, and then comparing the relative difference between model estimates from the full-time series with the truncated time-series. The retrospective analysis focuses on the bias and accuracy of modeled quantities. The most commonly used statistic for

retrospective bias, rho (ρ_M), is obtained from Mohn (1999). In line with recent studies (Carvalho et al., 2017; Winker et al., 2018), we focus on the formulation proposed by Hurtado-Ferro et al. (2015) as mean relative error, of the form

$$\rho_M = \frac{1}{h}\sum_{t=1}^{h}\left(\frac{X_{T-t} - \widehat{X}_{T-t}}{\widehat{X}_{T-t}}\right) \quad (2)$$

where X is the quantity for which ρ_M is being calculated, $\widehat{X}$ is the corresponding estimate from the reference model that was fitted to the full dataset, T is the terminal year of the assessment, and h denotes the total number of time steps of sequentially removing years with data (hereafter referred to as retrospective peels). While it is straightforward to compare ρ_M among alternative model runs, deciding whether ρ_M of the 'best' model is acceptable or not, is subjective. A 'rule of thumb', proposed by Hurtado-Ferro et al. (2015), suggests values of ρ_M that fall outside (-0.15 to 0.20) for SSB for longer-lived species, or outside (-0.22 to 0.30) for shorter-lived species indicates an undesirable retrospective pattern. In addition, the direction of the retrospective bias has implications for characterizing risk associated with management advice. A positive ρ_M for SSB is of particular concern because it implies a systemic overestimation of biomass, which would lead to over-optimistic quota advice if not taken into consideration (Hurtado-Ferro et al., 2015).

2.4. Prediction skill

The model diagnostics introduced thus far evaluate how well the model fits all available observations and how consistent the modeled quantities are in retrospect. However, providing fisheries management advice requires predicting a stock's response to management and checking that predictions are consistent with future reality (Kell et al., 2016). The accuracy and precision of the predictions depend on the validity of the model, the information in the data, and how far ahead of time predictions are made.

An intuitive approach to assess potential forecast bias is to extend the retrospective analysis to conduct model-based hindcasts by adding the additional step of projecting quantities, such as SSB, over the truncated years (Brooks and Legault, 2016). The settings for these retrospective forecasts should be similar to the forecast settings used when conducting future projections, e.g., for alternative catch quota, only that the observed catches are used for the hindcast (Brooks and Legault, 2016). In age-structured integrated assessment models forecasts are typically forward-projections of the numbers- and catch-at-age matrices given assumptions about the expected recruitment (e.g., deterministic or short-term average) and other parameters that determine stock productivity and selectivity (Maunder et al., 2006; Johnson et al., 2016). The hindcast can estimate forecast bias by comparing the forecasted values to the reference model estimates (i.e., the assessment model that has zero peels) based on the most recent year. Forecast bias ρ_F can be computed as the average relative error analogous to the retrospective bias ρ_M (c.f. Eq. 2) and provides a measure of consistency with regards to updating the stock status estimates based on new data.

Retrospective forecasting is not suitable for validation, however, unless model estimates of latent quantities, such as SSB, could be known without error. To address this, Kell et al. (2016) proposed the hindcasting cross-validation technique (HCXval) where observations are compared to their predicted future values. The key concept behind the HCXval approach is 'prediction skill', which is defined as any measure of the accuracy of a forecasted value $(\widetilde{y}_t)$ to the actual observed value (y_t) that is not known by the model (Kell et al., 2021). The difference $\widetilde{y}_t - y_t$ is hereafter referred to as the 'prediction residual' (Michaelsen, 1987). The HCXval algorithm is similar to that used in the retrospective analysis. It requires the same procedure of peeling the observations and refitting the model to the truncated data series. Like retrospective forecasting, HCXval involves the additional steps of projecting forward (hindcasts). The difference is cross-validating the forecasts using the observations that were left out of the fit to the truncated time series in order to assess the model's prediction skill.

A robust statistic for evaluating prediction skill is the mean absolute scaled error (MASE; Hyndman and Koehler, 2006). 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 $y_t = y_{t-1}$, 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, such that:

$$MASE = \frac{\frac{1}{h}\sum_{t=T-h+1}^{T}(\widetilde{y}_t - y_t)}{\frac{1}{h}\sum_{t=T-h+1}^{T}|y_t - y_{t-1}|} \quad (3)$$

where $\widetilde{y}_t$ is the one-step-ahead forecast of the expected value for the observation at time t based on the model conditioned with data up to time t-1, and h denotes the number of hindcasting time steps for which forecasts $\widetilde{y}_t$ were made to compare with the observations y_t. A MASE score > 1 indicates that the average model forecasts are worse than a random walk. Conversely, a MASE score of 0.5 indicates that the model forecasts twice as accurately as a naïve baseline prediction; thus, the model has prediction skill.

To generate the MASE score for abundance indices, the predicted abundance index is calculated as the product of the fleet-specific vulnerable biomass trajectories and the estimated catchability coefficients q for the full model and for each of the reduced retrospective fits, including both observation and forecast time horizons. The prediction residuals are computed as the difference between the forecasts of $\log(y_t)$ for each retrospective model and the corresponding observation that was reserved for validation by omitting it from the fit. The MASE score can then be calculated for each index by scaling the MAE of the prediction residuals to the MAE of the baseline forecasts of $\log(y_{t-1})$ observations from the previous time step over the evaluation period. The above procedure can be applied to any observed or empirical quantity for which the expected value can be forecasted. In section *3.3. Prediction skill: Hindcast Cross-validation* we demonstrate how HCxval can be applied to indices (Kell et al., 2021) and also composition data based on prediction residuals of mean length- and age values.

3. Diagnostic cookbook and interpretation

In the following, we demonstrate the application of the above model diagnostics using outputs from two Stock Synthesis models. We chose Stock Synthesis for the integrated assessment modelling framework because it is widely used to perform assessments for fish stocks throughout the world. While Stock Synthesis was originally designed to fill the data moderate gap between biomass dynamic models on one side and Virtual Population Analysis (VPAs) and Statistical Catch-at-Age (SCAA) models on the other side, it now captures the entire spectrum from data-poor catch-only assessments (Cope, 2013; Wetzel and Punt, 2015) to data-rich situations with available age composition and abundance indices from multiple sources including research surveys. In particular, its use within its original realm of data-moderate stock assessments (where catch time series and abundance indices are available and catch composition data is limited or absent) has significantly increased in recent years, especially in tuna Regional Fishery Management Organizations (RFMOs). Various stocks of billfish and pelagic sharks that were exclusively assessed using Surplus Production Models and VPAs are now moving towards implementation in Stock Synthesis as additional data become available (e.g., Courtney et al., 2017; Wang

et al., 2015). Similarly, there has been a recent increase in the use of Stock Synthesis for benchmark assessments in Europe in place of the conventionally used VPA with extended survivor analysis or the state-space catch-at-age models such as SAM (ICES, 2019).

The visualization of model outputs and implementation of diagnostics for Stock Synthesis is facilitated by the R package r4ss (Taylor et al., 2021; github.com/r4ss/r4ss). For each technique, we point readers to relevant citations or source code. To enable Stock Synthesis users to reproduce the diagnostic plots presented here, we have implemented several functions in the new R package *ss3diags*, which is made available on github.com/JABBAmodel/ss3diags.

The first case study is based on the stock assessment for the North Atlantic shortfin mako shark (SMA; Courtney et al., 2020, 2017). The vast majority of SMA is caught by pelagic longline operations, but due to strong spatial structuring of size classes, the selectivity pattern differs among the fishing fleets operating in the different regions (Courtney et al., 2017). The population dynamics of SMA reveal an unusual combination of slow somatic growth, very late maturation, and steep dome-shaped selectivity. Fishing mortality predominantly occurs on sub-adults, whereas fishing mortality is expected to be low for larger adults, in particular for large, mature females (Winker et al., 2020). The SMA example represents a length-based age- and sex-structured multi-fleet model that is fit to five standardized catch-per-unit-effort (CPUE) indices. Fisheries-dependent length-composition data are assumed to be representative of the different selectivity patterns for the six major surface longline fishing fleets (Fig. 2). The lack of fisheries-independent abundance information and the absence of age-composition data are also typical characteristics of data availability for pelagic shark and tuna stocks.

The second case study is the most recent 2020 stock assessment of Pacific hake (HAKE; Grandin et al., 2020). The available data comprise a time series of total catches aggregated into a single fleet over the modeled period 1966–2019, an index of relative abundance and age-composition data from an acoustic biomass survey conducted over 1995–2019, age-composition data from the fishery for all years between 1975–2019, and weight-at-age and fecundity-at-age data are aggregated across all data sources. Prior to the start of industrial fishing operations in 1966, catches are assumed to be very small and thus not included in the model. Acoustic surveys were conducted once every three years over 1995–2001 and once every two years over 2001–2019, with an additional survey conducted in 2012. Empirical weight-at-age and fecundity-at-age data allow for time-varying growth without estimating time-varying parameters. Pacific hake appears to have low recruitment with occasional large recruitment events associated with high recruitment variability. The HAKE model represents an age-based, sex-aggregated integrated assessment built on information comparable to the 'data-rich' requirements of a conventional statistical catch-at-age model. Model features include estimating age-specific time-varying fisheries selectivity implemented as a random walk, year-specific ageing error, and Dirichlet-Multinomial weighting of age-composition data (Thorson et al., 2017).

3.1. Goodness-of-fit

To evaluate the overall model fit of the relative abundance indices and composition data, the joint-index residual plot was applied to the residuals from the fits to indices, mean length for SMA, and mean age for HAKE for multiple time series simultaneously. The code for this diagnostic plot was adapted from the JABBA R package (github.com/jabbamodel/JABBA) to Stock Synthesis output files and implemented as the plotting function ss3diags::SSplotJABBAres(), which provides the option to specify the type of data input.

Overall, the SMA joint-index residual plot indicated a good fit to the CPUE data with the RMSE around 30 % (Fig. 3; Winker et al., 2018). The boxes were small over time, except for the last year of CPUE data, 2015 (Fig. 3a). A loess-smoother indicated there appeared to be increased variability in the residuals of model fit to CPUE over time. Fit to the acoustic index in the HAKE model (Fig. 3b) included the estimation of an

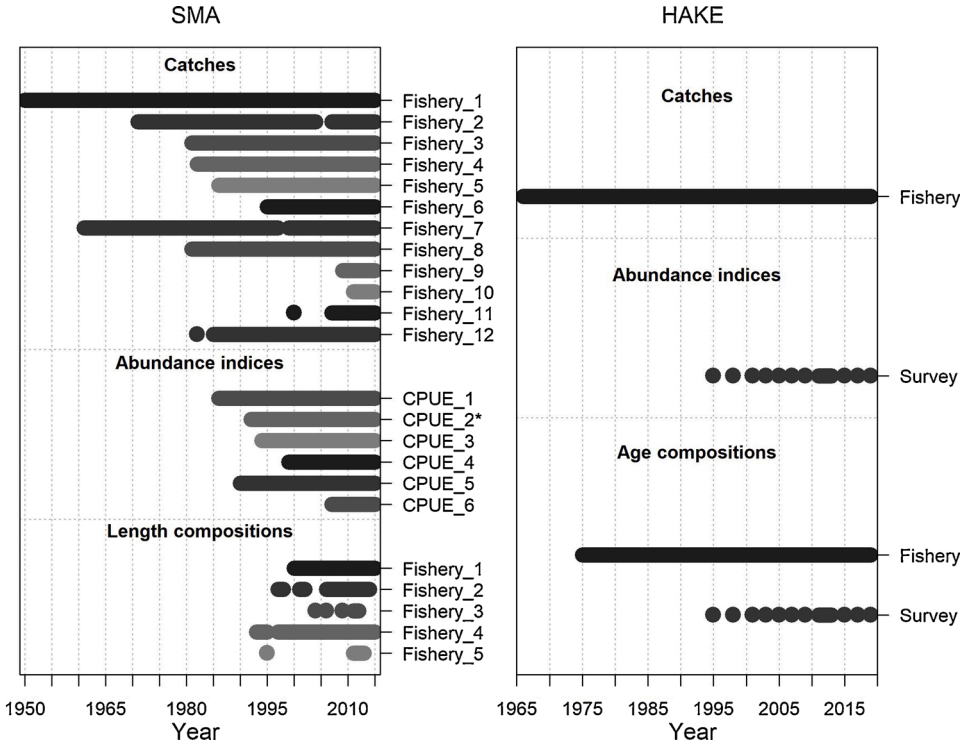


Fig. 2. Available temporal coverage and sources of catch, relative abundance, length-composition, and conditional age-at-length composition data used in the North Atlantic shortfin mako (SMA) model (left panel) and Pacific hake (HAKE) model (right panel) in Stock Synthesis. *CPUE_2 (US longline observer index) was not fitted by assigning zero weight to the SMA model's likelihood.

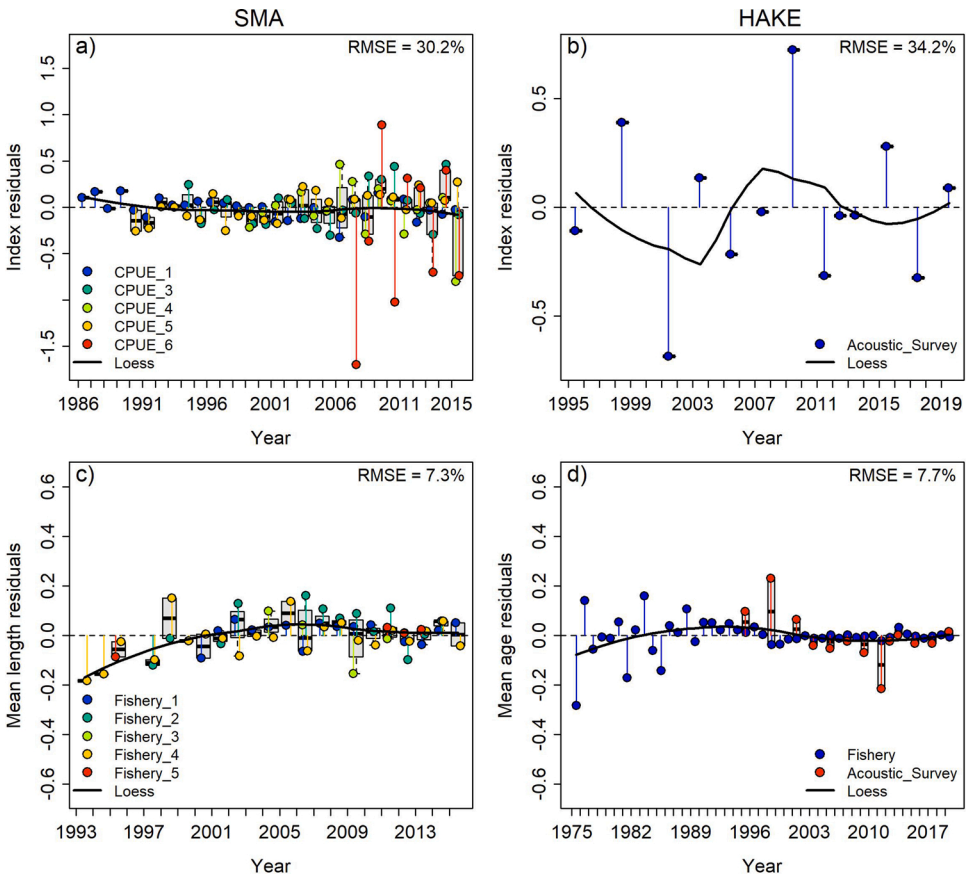


Fig. 3. Joint residual plots for (a) multiple CPUE fits from different fishing fleets from the North Atlantic shortfin mako (SMA) model color-coded by index and (b) a single acoustic survey abundance index from the Pacific hake (HAKE) model, (c) annual mean length estimates for multiple fishing fleets from the SMA model and (d) annual mean age estimates for surveys and the fishery from the HAKE model. Vertical lines with points show the residuals (in colors by index), and solid black lines show loess smoother through all residuals. Boxplots indicate the median and quantiles in cases where residuals from the multiple indices are available for any given year. Root-mean squared errors (RMSE) are included in the upper right-hand corner of each plot.

additional variance parameter, which resulted in relatively low precision (RMSE = 34.2 %). The RMSE for the joint residuals of the mean length estimates for SMA (Fig. 3c) and mean age for HAKE (Fig. 3d) were similar at around 7%. Observed mean age for HAKE indicated occasional conflicts between fisheries and survey data in 1998 and 2011 (Fig. 3d).

We developed the function ss3diags::SSplotRunstest() to visually denote passing (green) and failing (red) residual runs tests as judged by the p-values computed for each series (Carvalho et al., 2017), which can be applied to abundance index, mean-length residuals, and mean-age residuals by specifying the data type in the function.

There was no evidence ($p \geq 0.05$) to reject the hypothesis of randomly distributed residuals for all CPUE time series fit in the SMA model (Table 1; Fig. 4). Only the last data point of CPUE 3, which fell outside the three-sigma limit, may warrant additional evaluation of its influence on the final year's estimated stock abundance trend. The runs tests applied to the mean-length estimates from the five fishing fleets showed that only the mean-length residuals for Fishery 2 failed due to positive residuals over sequential years between 2002 and 2011 (Fig. 4d).

For HAKE, the residual series for the acoustic survey and the mean-age residuals for both the survey and the fishery passed the runs tests (Fig. 5). The early period 1975–1990 showed several larger residuals that fell outside the three-sigma limit for expected mean ages from the fishery (Fig. 5a). Time-varying selectivity is modeled only from 1991 onward leading to smaller residuals in mean ages for that later period. Considering that the sequence of these larger residuals are limited to the early years of the time series and appear to follow a random pattern, these outliers are likely to have little influence on the stock status estimates. By contrast, a sequence of positive or negative residuals falling outside the three-sigma limit would indicate potential model misspecification.

3.2. Model consistency

3.2.1. R_0 profile diagnostic

The R_0 likelihood component profiles for the SMA and HAKE models (Fig. 6) were developed using the function r4ss:SS_profile() (see https://github.com/jabbamodel/ss3diags for example R code).

For the SMA model, the gradient of the likelihood profile for the penalty on the recruitment deviations was greater than other data sources. The second strongest gradient in the log-likelihood profile was observed for the CPUE indices (Fig. 6a). The gradient of the likelihood profile supported by the length-composition data is lower than those supported by the penalty for the recruitment deviates and CPUE indices. Therefore, the length-composition data are the least informative data for the estimation of R_0. The minimum value along the R_0 profile for the penalty on the recruitment deviates was close to those from the CPUE data, which indicated no major conflict between these two likelihood components. On the other hand, the length-composition data showed a much higher minimum value along the R_0 profile compared to the other data sources. Among the CPUE indices, there was a relatively large change in the contribution to the likelihood over the profile from two of the time series, CPUE 1 and 5 (Fig. 6c). However, a difference in the minimum value along the R_0 profile was identified between these two indices, while a minimum value was not found for the other CPUE indices. Among the length-composition data, Fisheries 1 and 4 showed large changes in the contribution to the likelihood over the profile (Fig. 6e), with Fishery 1 showing a difference in the minimum value along the R_0 profile when compared to the other fisheries. These differences in the log-likelihood support of the minimum value indicate that there was also conflict among individual CPUE indices and length-composition data on the estimation of R_0, indicating that the maximum likelihood estimate of R_0 is somewhat balancing conflicting signals from multiple data sources.

Table 1
Summary statistics runs tests, retrospective analysis, retrospective forecasts, and hindcast cross-validation (HCxval) model diagnostics applied to the (a) North Atlantic shortfin mako (SMA) and (b) the Pacific hake (HAKE) Stock Synthesis models, where *n* denotes the number of observations to compute of the statistics.

Diagnostic	Quantity	Statistic	Value	*n*
a) SMA				
Runs Test	CPUE 1	p-value	0.069	30
Runs Test	CPUE 2*	p-value	0.717	24
Runs Test	CPUE 3	p-value	0.229	22
Runs Test	CPUE 4	p-value	0.406	17
Runs Test	CPUE 5	p-value	0.065	26
Runs Test	CPUE 6	p-value	0.87	9
Runs Test	Mean Length 1	p-value	0.127	16
Runs Test	Mean Length 2	p-value	0.04	13
Runs Test	Mean Length 3	p-value	0.331	5
Runs Test	Mean Length 4	p-value	0.806	22
Runs Test	Mean Length 5	p-value	0.159	4
Retrospective analysis	SSB	Mohn's Rho	0.059	5
Retrospective forecasts	SSB	Forecast bias	0.061	5
HCxval	CPUE 1	MASE	0.891	5
HCxval	CPUE 2*	MASE	0.862	5
HCxval	CPUE 3	MASE	0.463	5
HCxval	CPUE 4	MASE	0.936	5
HCxval	CPUE 5	MASE	0.763	5
HCxval	CPUE 6	MASE	0.534	5
HCxval	Mean Length 1	MASE	0.927	5
HCxval	Mean Length 2	MASE	0.650	4
HCxval	Mean Length 3	MASE	0.359	1
HCxval	Mean Length 4	MASE	0.636	5
HCxval	Mean Length 5	MASE	3.271	2
b) HAKE				
Runs Test	Survey Index	p-value	0.96	13
Runs Test	Survey mean age	p-value	0.093	13
Runs Test	Fishery mean age	p-value	0.234	45
Retrospective analysis	SSB	Mohn's Rho	−0.038	7
Retrospective forecasts	SSB	Forecast bias	−0.051	7
HCxval	Survey Index	MASE	1.065	4
HCxval	Survey mean age	MASE	0.356	4
HCxval	Fishery mean age	MASE	0.632	7

* CPUE 2 represents a subset of longline fleet observer data for the logbook based CPUE 1 index and was not used to fit the model (zero weight).

The age-composition data were more informative than the survey index for R_0 in the HAKE model (Fig. 6b). The log-likelihood profile for the penalty on the recruitment deviates attained a minimum at the upper R_0 range, which was opposite to the minimum in the log-likelihood profile for the age-composition data. Therefore, the resulting total log-likelihood profile could be interpreted as a trade-off between the recruitment penalty and achieving a good fit to the age-composition data. In the HAKE model, the recruitment standard deviation is fixed at 1.4, a relatively high value. This value was chosen to achieve consistency with the observed variability in the time series of recruitment deviation estimates (Grandin et al., 2020). Albeit of limited influence, the log-likelihood profile for the survey index appears to corroborate the total log-likelihood profile (Fig. 6d). Individual profiles for the age-composition data sources indicated no apparent conflicts between the survey and the fishery data. The age-composition time series for the fishery (1975–2019) is more informative about R_0 than the shorter survey age-composition data (Fig. 6f), even though the survey data are weighted higher by the estimated Dirichlet-Multinomial parameters.

3.2.2. ASPM diagnostic

We used the following workflow to compute the ASPM diagnostic (Minte-Vera et al., 2017; see github.com/jabbamodel/ss3diags for example R code): (1) run the integrated Stock Synthesis model, (2) fix the selectivity parameters at the maximum likelihood estimates (MLEs), (3) turn off estimation of all parameters except R_0 and the parameters representing the initial conditions (e.g., recruitment offset for the first time step of the model and initial fishing mortality parameters), (4) set the recruitment (and the initial age structure) deviates to zero (adjusting the bias-correction factor appropriately, see Methot and Taylor, 2011), and (5) fit the model to the indices of abundance only. Additionally, the model is run as for the ASPM but with recruitment deviates estimated (ASPMdev). Trends in relative SSB were then compared between the fully integrated stock assessment model, ASPM, and ASPMdev. Maunder et al. (2020) provide a flow chart to assign reliability weights to models based on combining the R_0 profile and the ASPM diagnostics.

The CPUE 1 index for the SMA model represented the longest CPUE time series in the model and was associated with the smallest RMSE of 11 % of all indices. The ASPM fit to CPUE 1 showed a consistent declining trend over time and an RMSE of 19.7 %. In contrast, the fit to the same index in the fully integrated SMA and ASPMdev (RMSE = 10.2 %) models were similar and were both associated with a more oscillatory pattern (Fig. 7a). All three models estimated similar trends for SSB. However, the SSB from ASPM was higher than SSB from the fully integrated model and ASPMdev (Fig. 7c). The differences between the CPUE fits from the fully integrated model and the ASPM for SMA are explained by the estimated recruitment deviations in the fully integrated model. The recruitment deviations allow for variability in age-0 recruitment and can be interpreted as the process error necessary to fit the observed trends in the CPUE data (Fig. 7e). By applying the ASPM diagnostic, it was possible to conclude that for SMA, the variability in recruitment must be taken into account to estimate both the trends in CPUE 1 and the absolute scaling of SSB. The ASPMdev indicated that the CPUE data and the catches contained information on temporal variability on recruitment.

For the HAKE model, the pattern in survey CPUE fit differed between the fully integrated model and ASPM (Fig. 7), resulting in RMSEs of 34.2 % and 39.4 %, respectively. The ASPM showed a decline at the beginning of the time series, followed by an increase and a flat trend over the most recent period that ends with a sharp decline in the last two years (Fig. 7b) that is likely due to the lack of information in the survey on young age classes. The ASPMdev had an RMSE of 0.3 % and fit the observed values almost perfectly, suggesting that the ASPMdev is overfitted. ASPM and ASPMdev estimates of SSB follow a different pattern from the fully integrated model, especially from the beginning of the time series to the early 1990s (Fig. 7d), suggesting that the population dynamics of HAKE are strongly driven by variation in recruitment. Yet, the highly inflated 95 % confidence intervals for the recruitment deviations estimates by the ASPMdev indicate that it is not possible to estimate recruitment deviations for HAKE without the age-composition data (Fig. 7f).

3.2.3. Retrospective analysis

The retrospective analysis was implemented in Stock Synthesis utilizing R and functions available in *r4ss* (see github.com/jabbamodel/ss3diags for example R code). The retrospective patterns were visualized using the function ss3diags::SSplotRetro(), which also routinely computes ρ_M and provides the option to illustrate hindcasts with one step ahead forecasts of SSB and to compute the associated forecast bias ρ_F. The retrospective diagnostic was implemented here for the SMA and HAKE models by sequentially eliminating the five and seven most recent years of data from the full stock assessment model, respectively. Miller and Legault (2017) found that estimates of ρ_M typically stabilized after five peels. For hake, the longer seven-year data peel was chosen to account for the bi-yearly survey index and age-composition updates.

For the SMA model, there was a consistently positive but small retrospective bias (Fig. 8a-c), with $\rho_M = 0.06$, falling well within the acceptable thresholds for long-lived species (Carvalho et al., 2017; Hurtado-Ferro et al., 2015). For the HAKE model, trends and scale in SSB were similar through the retrospective years (Fig. 8b). The small ρ_M of

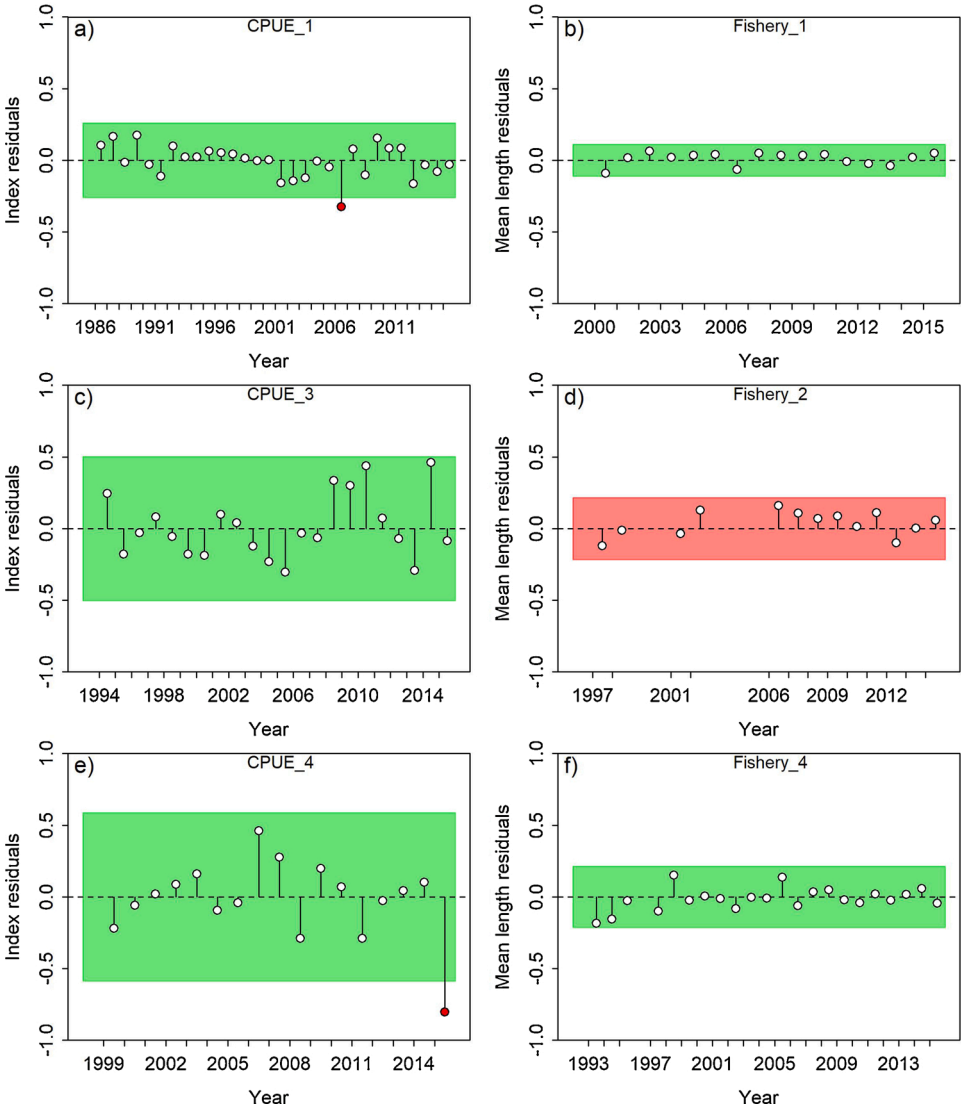


Fig. 4. Runs tests results illustrated for three catch-per-unit-effort (CPUE) fits (left panel: a, c, e) and three mean lengths of size composition data (right panel: b, d, f) from the North Atlantic shortfin mako (SMA) model. Green shading indicates no evidence ($p \geq 0.05$) and red shading evidence ($p < 0.05$) to reject the hypothesis of a randomly distributed time-series of residuals, respectively. The shaded (green/red) area spans three residual standard deviations to either side from zero, and the red points outside of the shading violate the 'three-sigma limit' for that series. The complete set of runs test results is presented in Table 1.

-0.04 for SSB and a random retrospective pattern indicates a consistently behaved model as sequential years of data are removed (Fig. 8d). For both the SMA and HAKE models, the one year forward projections of SSB are consistent with the estimated trend in reference models (Fig. 8). This also illustrates how the conventional retrospective procedure can be conceptually extended to hindcasting by implementing the additional step of a forecast (Legault and Brooks 2016). Here, the forecast bias ρ_F remained stable at 0.06 for SMA (Fig. 8c) and showed only a very slight increase to $\rho_F = -0.05$ when compared to the retrospective bias of $\rho_M =$ -0.04 for HAKE (Fig. 8d). We suggest that extending the conventional retrospective analysis by retrospective forecasts can be a useful tool when verifying that the modeled quantities are not only historically stable (i.e., retrospective ρ_M) but at the same time consistent between forward projections and subsequent updates with newly available data (i.e., retro forecasts ρ_F).

Both ρ_M and ρ_F are measures of an average bias across the years under evaluation. As such, they can lead to situations where large relative errors for individual years could cancel each other out, resulting in seemingly acceptable retrospective and forecast bias values, respectively. Therefore, we recommend checking if the retrospective peels and retrospective forecasts fall within the estimated 95 % confidence limits from the reference run. Here, this is the case for both SMA and HAKE, which confirms that the errors in SSB estimates resulting from additional years of data being removed or added to the models are consistent with estimated uncertainty.

3.3. Prediction skill: hindcast cross-validation

Implementing the HCxval diagnostic in Stock Synthesis required using the outputs produced for the retrospective routine generated by *r4ss* (see retrospective analysis section). The forecasts are based on the settings in 'forecast.ss', which are also evoked when conducting future projections with the same model, only that the observed catches are used for the retrospective forecasts. A desirable feature of Stock Synthesis is that the software also computes the expected values of the observational data (e.g., abundance indices, length- or age-composition data) based on the forward-projections of the numbers- and catch-at-age matrices. Therefore, there are no additional computationally intensive tasks needed if HCxval is conducted in conjunction with retrospective analysis. ss3diags::SSplotHCxval() produces novel HCxval diagnostic plots and computes the MASE scores for all indices of relative abundance, mean lengths, and mean ages, including observations that fall within the hindcasting evaluation period. To compute the observed forecasted mean lengths or mean ages from the composition data, we provide ss3diags::SSretroComps(), which builds on the function r4ss::SScompsTA1.8() to derive expected mean length and age based on the algorithms proposed by Francis (2011).

In the SMA model, all five fitted CPUE indices included at least one

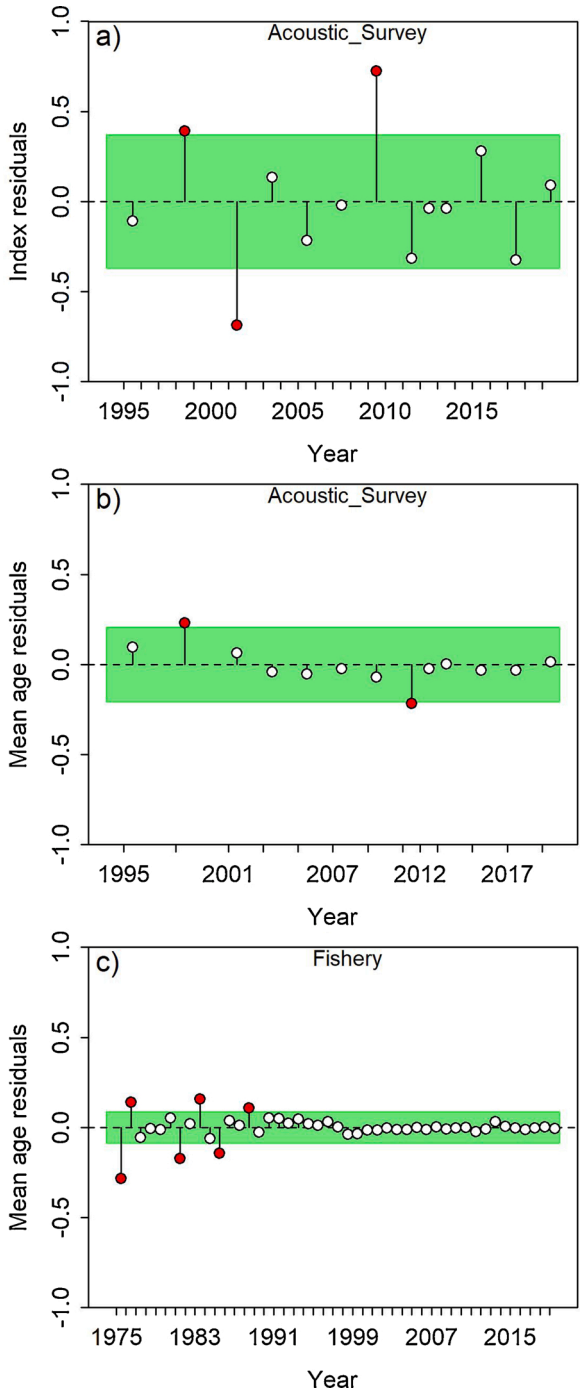


Fig. 5. Runs tests results for fits to (a) the acoustic biomass survey index and (b) annual mean age estimates for the survey, and (c) the fishery from the Pacific Hake (HAKE) model. Green shading indicates no evidence ($p \geq 0.05$) and red shading evidence ($p < 0.05$) to reject the hypothesis of a randomly distributed time-series of residuals, respectively. The shaded (green/red) area spans three residual standard deviations to either side from zero, and the red points outside of the shading violate the 'three-sigma limit' for that series. The complete set of runs test results is presented in Table 1.

observation that fell within the hindcast evaluation period 2010–2015 (Fig. 9). MASE scores < 1 indicated that the SMA model had a superior prediction skill than the naïve baseline forecast for all CPUE indices (Fig. 9a, c, e). The most accurate predictions were observed for CPUE 1 with a prediction residual MAE of 0.09. However, CPUE 1 also showed the least inter-annual variation among observations, associated with a small baseline MAE for the naïve predictions. Despite a higher prediction residual MAE of 0.130, the MASE score of CPUE 5 was slightly better than CPUE 1. In other words, less variable and thus more informative CPUE indices require higher prediction accuracy than noisy and less influential CPUE indices to 'pass' with a MASE < 1. Although the log-likelihood profile (Fig. 6) and ASPM diagnostics (Fig. 7) revealed that stock abundance was mostly informed by the CPUE data, the SMA model also indicated reasonably good prediction skill (MASE < 1) for mean lengths of four of the five fisheries (Fig. 9b, d, f). The only exception was Fishery 5 (MASE = 3.27), which comprised only four data points, of which two fell within the hindcasting horizon of the terminal 5 years.

In the HAKE model, the log-likelihood profile and ASPM diagnostics suggested that stock abundance trends and scale are predominantly informed by the age-composition data, but also indicated that the scale of SSB is sensitive to the assumption made for the penalty on the recruitment deviations. The MASE scores < 1 indicated an adequate prediction skill for the corresponding mean age estimates (Fig. 10). The mean-age estimates from the acoustic survey (MASE = 0.36) and the fishery (MASE = 0.63) can be seen as important data sources for validating that the HAKE model is consistent with the observed mean ages in retrospect (Fig. 10b-c). In contrast, the survey CPUE had a MASE score of 1.06, which suggests that the model's prediction skill for the bi-yearly survey index was low compared to the mean age estimates (Fig. 10a)

3.4. Convergence

The jitter test for global convergence was implemented in Stock Synthesis, utilizing the jitter feature described in detail within the Stock Synthesis manual (e.g., for version 3.30.15 see Methot et al., 2020). The jitter feature is implemented in R using a function in the *r4ss* package (see https://github.com/jabbamodel/ss3diags for example R code).

The final gradient of the SMA model was relatively small (e.g., < 1.00E-04), and the Hessian matrix for the parameter estimates was positive definite. Examination of parameter estimates indicated that some selectivity parameters were near their bounds, however, no parameters were estimated outside the reasonable minimum and maximum correlation thresholds (0.95 and 0.01, respectively). The 200 iterations of the jitter test for the SMA model (Fig. 11) resulted in 131 model runs that failed to converge, 44 model runs that converged at or close to the total likelihood estimate value of the base case model run (77 likelihood units), and five model runs with total likelihood values higher than 80. This demonstrates that the jittered model was sensitive to the initial values of the parameters. The specification of both bounds and priors on individual parameters, together with penalties, weights on associated likelihoods, and high correlations among parameters can all affect jitter convergence. Given that all converged model runs implemented within the jitter test resulted in total likelihood values equal to or greater than the base model, the jitter test did not provide evidence to reject the hypothesis that the base model parameter optimization converged to the global solution.

For the HAKE model, none of the estimated parameter values in the base case model were close to their specified bounds. The final gradient of the model was < 1.00E-04, and the Hessian matrix for the parameter estimates was positive definite. All of the 200 jitter model runs converged, with 196 model runs at the total negative likelihood estimate value of the base case model run (682 likelihood units), and 4 model runs had larger total negative likelihood values (Fig. 11). The jittered model was robust to initial values of the parameters and gave no evidence that the base case model converged to at local minimum of the objective function instead of the global minimum.

4. Discussion

Just like all models are wrong, but some are useful, it equally holds that all models are somewhat misspecified when fitting to empirical data (Francis, 2011; Maunder and Piner, 2017). By recognizing that it is

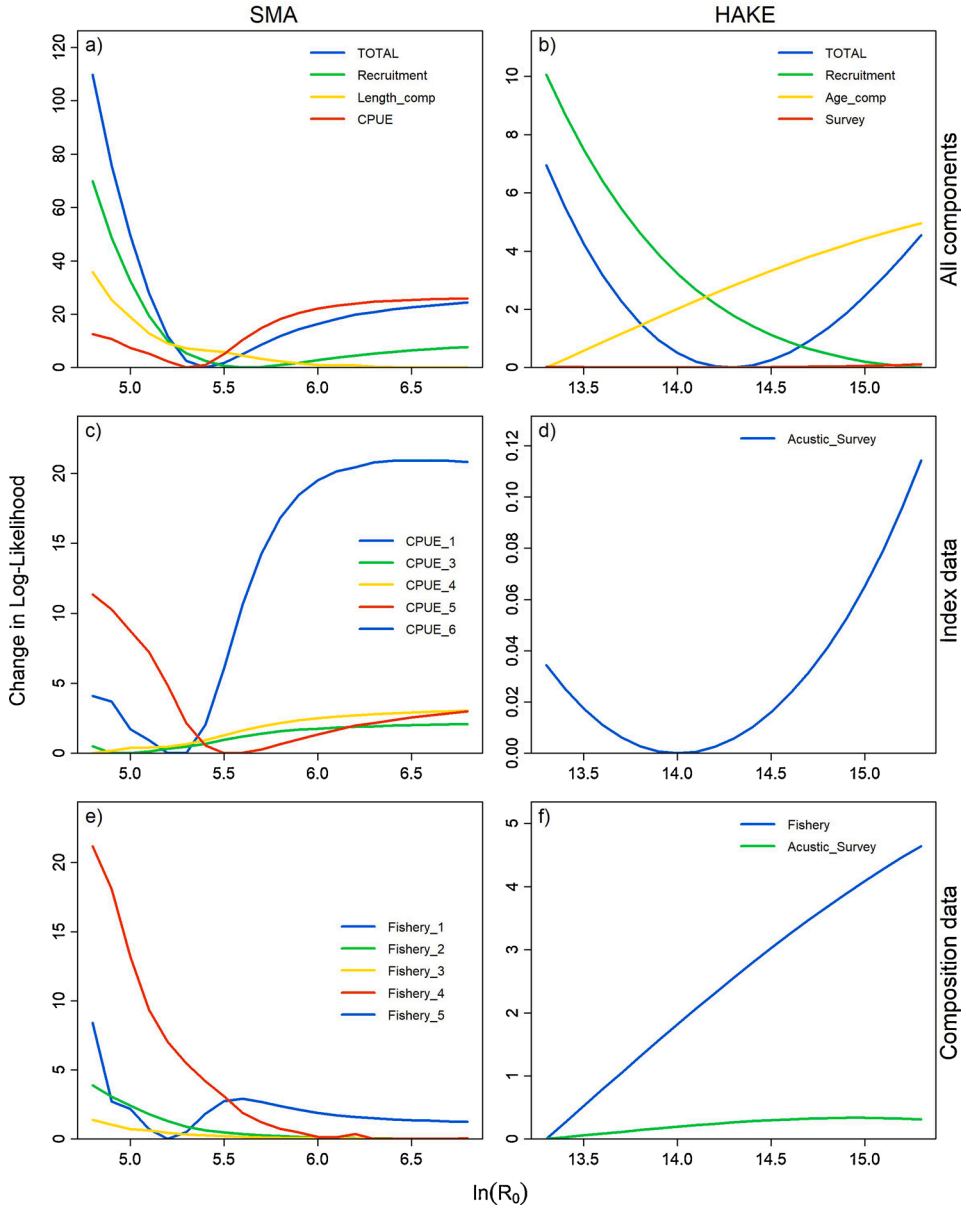


Fig. 6. Log-likelihood profiles for R_0 for the various data components included in the North Atlantic shortfin mako (SMA; left panels) and Pacific hake (HAKE; right panels) Stock Synthesis models, showing the contribution of (a) - (b) all data likelihood components, (c) - (d) among abundance indices and (e) - (f) among length- and age-composition data, respectively.

impossible to avoid some degree of model misspecification due to complexity and heterogeneity in the interplay between population dynamics and fisheries operation, we can start seeing a stock assessment for what it should be, a consistently evolving process to identify and improve models such that they are useful and provide more robust advice than others. First, we must ensure models converge adequately and best fit the data from a set of candidate models, and second, models must be validated before being used to forecast management advice.

In this paper, we demonstrate the application of contemporary model diagnostics ranging from convergence checks to model-free validation. The model diagnostics proposed here provide a set of tools to lay out the evidence in support of or against the candidate model(s) under consideration. Importantly, model diagnostics should enable the stock assessment analyst to lay out the main remaining concerns transparently. Addressing the remaining concerns may not always be feasible in the short term and may require improving the input data, revision of model structure, adding priors, and turning off estimation for poorly-informed parameters (i.e., regularization; Monnahan et al., 2019) or exploring sources of process error not included in the model (Walters et al., 2008).

In our proposed process, we first create a hypothesis and configure the model, including all the information known for that stock. The first stage of the flow chart evaluates model convergence. Some stock assessment models have thousands of parameters, so it is expected that some will have relatively high gradients (e.g., $>$ 1.00E-04), and some may be correlated. If this is the case, it is important first to identify which parameters have a high gradient and if there are correlations of potentially influential model parameters (e.g., between selectivity and growth) and then try to adjust from there. If a parameter has a high gradient or is highly correlated with other parameters, then fixing the parameter, using more informative priors, or loosening the bounds may form part of a necessary detour to investigate model properties. When examining model convergence problems, it is also important to determine if any specific data component(s) is causing the issues by inspecting the likelihood profile. Down-weighting or removing a data component can be considered for the final model if either of these

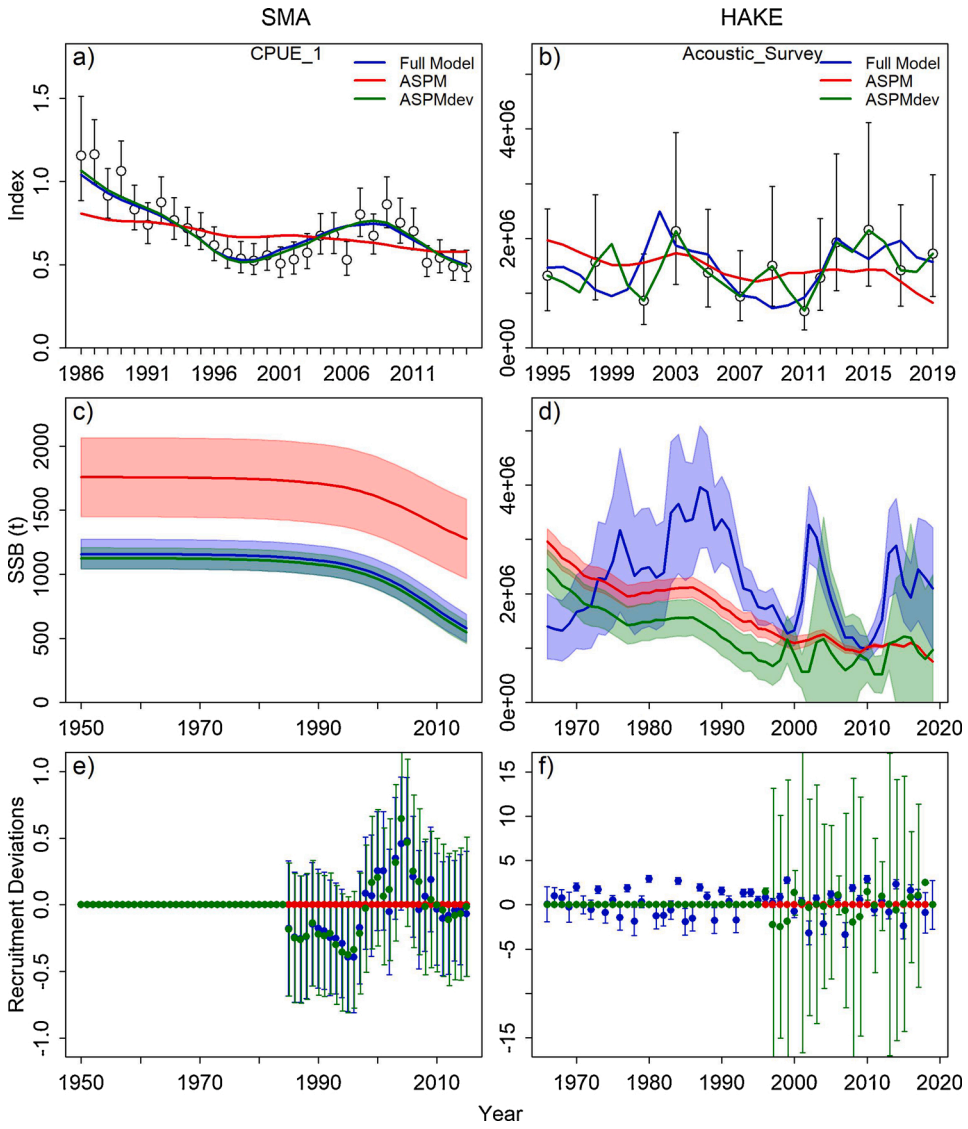


Fig. 7. Comparison between the fully integrated base-case and the deterministic Age-Structured-Production Model (ASPM) results for North Atlantic shortfin mako (SMA; left panels) and Pacific hake (HAKE; right panels), showing observed and predicted values for (a) the CPUE 1 index for SMA and (b) Survey index for HAKE, (c) – (d) spawning stock biomass trajectories relative to levels at MSY (SSB/SSB_{MSY}) and (e) – (f) recruitment deviation estimates.

options improves model convergence. If the initial runs produce meaningful results, we recommend checking the model fits and associated residual pattern as the second readily available diagnostic. There is little justification to ignore the evidence for poor fits to the data and non-random residual patterns by proceeding any further without first taking a detour. Lack of fit can be a sign of misspecification and an indication that an inappropriate model structure has been used. If the analyst decides to take a detour, initial explorations could include alternative parameterizations of key population dynamics processes, such as somatic growth or the spawning-recruitment relationship (Henríquez et al., 2016; Minte-Vera et al., 2017; Punt and Cope, 2019); adjusting the weight of the different data components in the likelihood (Francis, 2011; Wang et al., 2015); adding process complexity (e.g., time-varying selectivity (Stewart and Monnahan, 2017)) and time-varying catchability (Wilberg et al., 2009); regime shifts in recruitment (Haltuch and Punt, 2011; Johnson et al., 2016); or spatial structure (Goethel et al., 2011). It could be argued that, ideally, jitter runs should also be conducted at the beginning and during the model development process. However, in practice, time does not always allow to run this time-intensive diagnostic. Instead, users tend to verify model convergence because convergence will likely indicate no significant problems in a reasonably well-configured model. Thus, considering the jitter diagnostic's relatively long run times, we propose to reserve jitter diagnostic as the last step in the iterative model diagnostic process.

The R_0 profile has been widely used to identify data conflicts (Lee et al., 2014; Wang et al., 2015). Although R_0 is likely the most common parameter profiled over, the likelihood profile can also be a valuable diagnostic for any other estimable model parameter. Particularly, in cases where notoriously challenging parameters, such as natural mortality or the steepness of the spawning recruitment function, are estimated within the integrated model, the likelihood profile diagnostic is recommended to evaluate which data components are informative relative to the influence of the typically imposed priors or penalties. Even if parameters such as natural mortality or steepness are fixed, doing a likelihood profile of those parameters is still very useful, particularly, to measure the amount of information contained in the data and sensitivity (i.e., the consequences of using a fixed value) of the model results to the choice of those parameters. The difficulty with data conflicts arises because the source of the conflict may be an unknown misspecified process. In case the analyst decides to take a detour after inspecting the results from the R_0 profile, two options can be considered for further exploration; 1) eliminate or down-weight data, and 2)

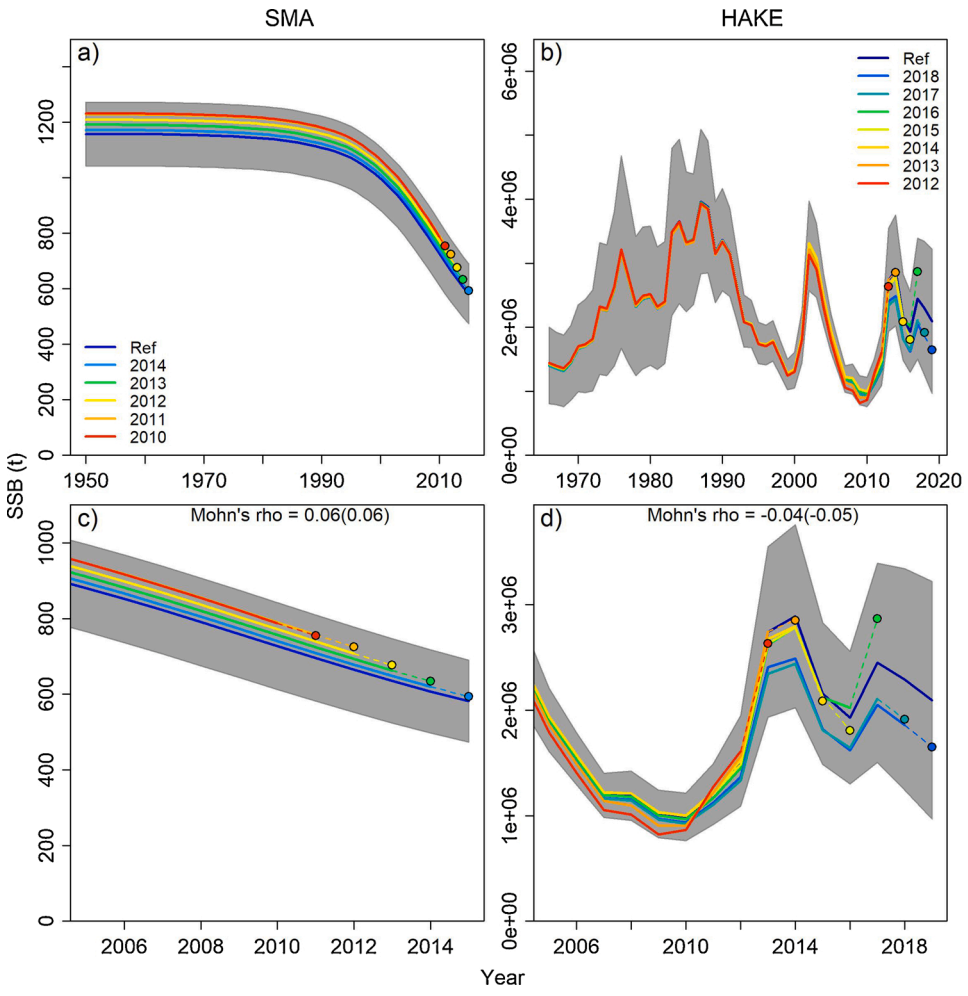


Fig. 8. Retrospective analysis of spawning stock biomass (SSB) estimates for North Atlantic shortfin mako (SMA; left panels) and Pacific hake (HAKE; right panels) models conducted by re-fitting the reference model (Ref) after removing five years of observations for SMA and seven years for HAKE, one year at a time sequentially. The retrospective results are shown for (a) – (b) the entire time series and (c) – (d) for the most recent years only. Mohn's rho statistic and the corresponding 'hindcast rho' values (in brackets) are printed at the top of the panels in (c) – (d). One-year-ahead projections denoted by color-coded dashed lines with terminal points are shown for each model. Grey shaded areas are the 95 % confidence intervals from the reference model.

include additional processes within the model structure (Wang and Maunder, 2017). However, any post-hoc approach for reweighting the data is unlikely to solve misspecification problems. It is important to note that exploring different weighting methods might produce different results, which can be biased in unknown ways because the model is still misspecified (Sharma et al., 2014).

In many cases, the fisheries-dependent input data are sampled in a biased way and are therefore not representative of the process they are meant to measure. A well-documented example is using non- or improperly standardized CPUE as a potentially biased index of abundance (Maunder and Punt, 2004). In addition to data processing, misreporting of catches, discards, or size composition data can commonly introduce data conflicts in stock assessment models. Therefore, it may be warranted to carefully re-evaluate the quality of input data to inform decisions on down-weighting or ultimately eliminate selected data sources. Alternatively, if the model misspecification includes models that are too simple to include the real complex processes that generated some of the data, changes to the model structure or temporal variation in model parameters can be modeled explicitly and help reduce or eliminate data conflict.

The ASPM diagnostic can evaluate data conflicts in information related to absolute abundance and abundance trends and detect misspecification in the population dynamics (e.g., steepness or natural mortality). It is important to find out early in the model construction process whether or not fishing affects the population (i.e., if the catches relative to surplus production caused observed trends in abundance). If, after inspecting the ASPM results, the analyst concludes that the model results strongly diverge from the expected changes in abundance given the catch (i.e., fishing does not affect the population), a detour is recommended to explore if the model is possibly misspecified or if there is evidence that stock dynamics are strongly driven by variations in recruitment (Minte-Vera et al., 2017). Based on the results of the detour, the analyst may consider developing an alternative model structure that does not rely on the catch and index to scale the model. Without finding an index that can inform the model about the absolute scale, the analyst needs to develop an alternative model that mimics a catch-curve analysis (i.e., fitting the model only to catch-composition data). A key component of such a model involves identifying which fleet or survey can provide the most information about the stock's productivity, which is needed to scale the model.

After advancing from the R_0 profile and ASPM diagnostics, the analyst reaches the diagnostics that provide insights about the invariance in important modeled quantities (e.g., SSB) by first looking into the rearview mirror with retrospective analysis (Hurtado-Ferro et al., 2015) and then evaluating the model's ability to forecast into the future by way of hindcasting cross-validation (Kell et al., 2016; Kell et al., 2021). Both diagnostics are useful to reveal systematic bias in the model estimation. Given that the variability in the retrospective bias, ρ_M depends on life history and that the statistic appears insensitive to the magnitude of F, Hurtado-Ferro et al. (2015) proposed a rule of thumb when determining whether a retrospective pattern should be addressed explicitly. However, ρ_M values smaller than those proposed should not be taken as confirmation that a given assessment does not present a retrospective pattern, as the choice of 90 % means that a 'false positive' can arise 10 % of the time. Retrospective analysis is widely used worldwide as a key diagnostic, and in Europe, it is often the key diagnostic for accepting or

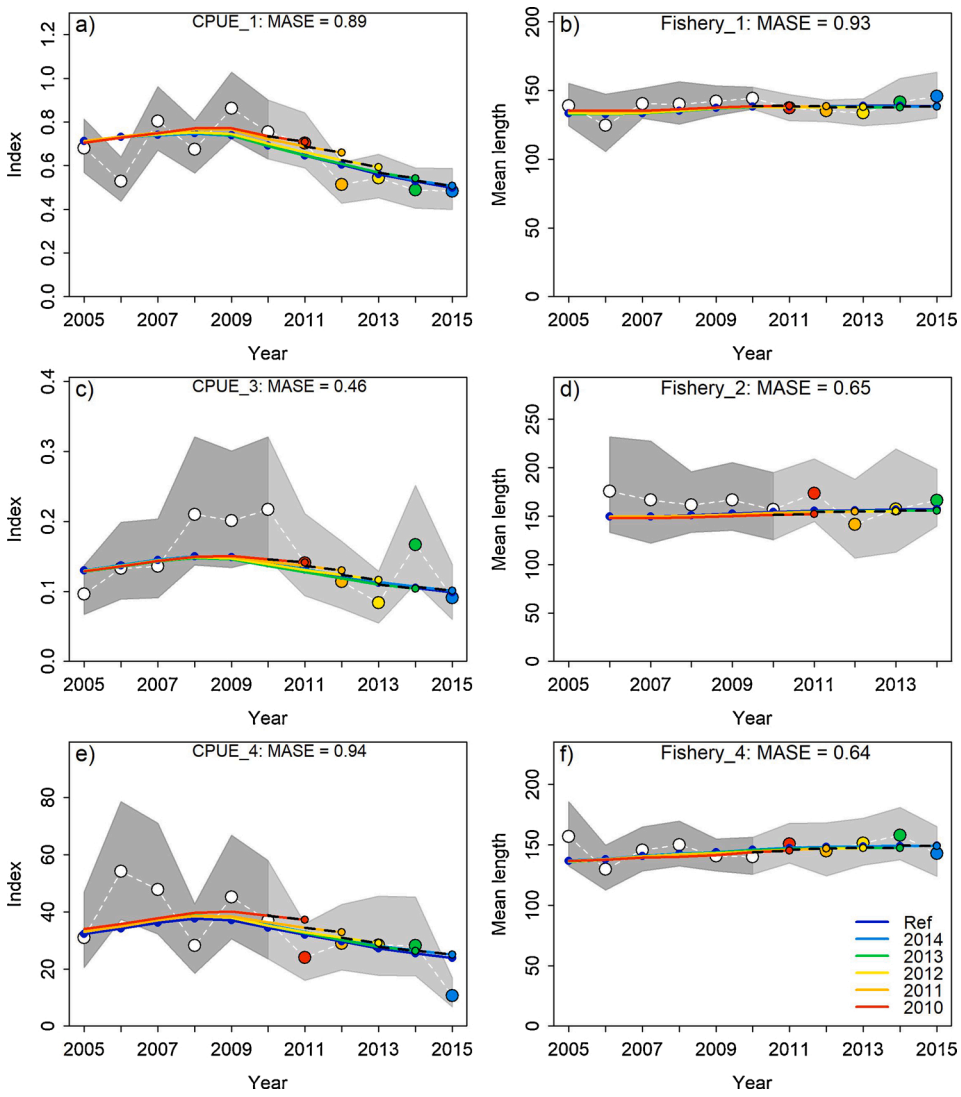


Fig. 9. Hindcasting cross-validation (HCxval) results for three catch-per-unit-effort (CPUE) fits (left panel: a, c, e) and three mean lengths of size composition data (right panel: b, d, f) from the North Atlantic shortfin mako (SMA) model, showing observed (large points connected with dashed line), fitted (solid lines) and one-year-ahead forecast values (small terminal points). HCxval was performed using one reference model (Ref) and five hindcast model runs (solid lines) relative to the expected catch-per-unit-effort (CPUE). The observations used for cross-validation are highlighted as color-coded solid circles with associated 95 % confidence intervals (light-gray shading). The model reference year refers to the endpoints of each one-year-ahead forecast and the corresponding observation (i.e., year of peel + 1). The mean absolute scaled error (MASE) score associated with each CPUE and size composition time series is denoted in each panel.

rejecting a model (ICES, 2019). A strong retrospective pattern indicates a problem with historical fits to data. The sources of a retrospective pattern can be anywhere in the time series. Therefore, when opting for a detour after inspecting the retrospective analysis results, it is recommended to explore alternative biology and fishery hypotheses. A closer inspection of the recruitment residuals is also advised, as retrospective patterns can be linked to the model's inability to capture components of the state-dependent dynamics of the ecosystem that appear to force the stock-recruitment relationship. However, simulation testing shows that data or model inconsistency may not always produce a retrospective pattern (Carvalho et al., 2017). This highlights the need of using hindcasting with cross-validation of observations to estimate prediction skill in combination with retrospective analysis, and why it should be routinely used as a diagnostic tool to evaluate the ability of a model to provide advice on future catches.

Model convergence, evaluating how well the model fits data, identifying data-conflicts, and evaluating model consistency in terms of retrospective and forecast bias have received much attention in fisheries science over the past decade. However, compared to other disciplines, such as oceanography and climate research, where model validation is an important prerequisite (e.g., Barnston et al., 2019; Keenlyside et al., 2008; Smith et al., 2010), key aspects of model validation have been mainly overlooked in fisheries science. Extending the stock assessment model diagnostic toolbox by including hindcast cross-validation techniques is building a bridge between current best practices in fisheries to what is already best practice in energy, oceanography, and climate research. This will ultimately increase confidence in the model-based scientific advice by stakeholders, managers, and policymakers.

The conceptual flow chart (Fig. 1) lays out a generic process of model development and selection using the presented model diagnostics. Based on this, we propose the following four properties as objective criteria for evaluating the plausibility of a model: (1) model convergence, (2) fit to the data, (3) model consistency, and (4) prediction skill. We recommend that none of these diagnostic criteria be interpreted in isolation or used as a definitive metric to accept or reject a model. For example, the analyst needs to decide whether the optimization was successful, the model fits the data, if the estimates are consistent when updated with new data (e.g., retrospective pattern), and if the model is not overfitted and able to make future predictions. These criteria are generic and, in principle, transferrable to any stock assessment model that provides an option for a forecast (Kell et al., 2021). For example, residuals run tests, retrospective analysis, and hindcast cross-validation are also available in the Bayesian state-space surplus production model 'JABBA' (Winker et al., 2018). In addition, stock specific plausibility criteria should be considered to evaluate if the assessment results are consistent with prior knowledge about the exploitation history and population biology (Maunder et al., 2020; Sharma et al., 2020; Thorson, 2020).

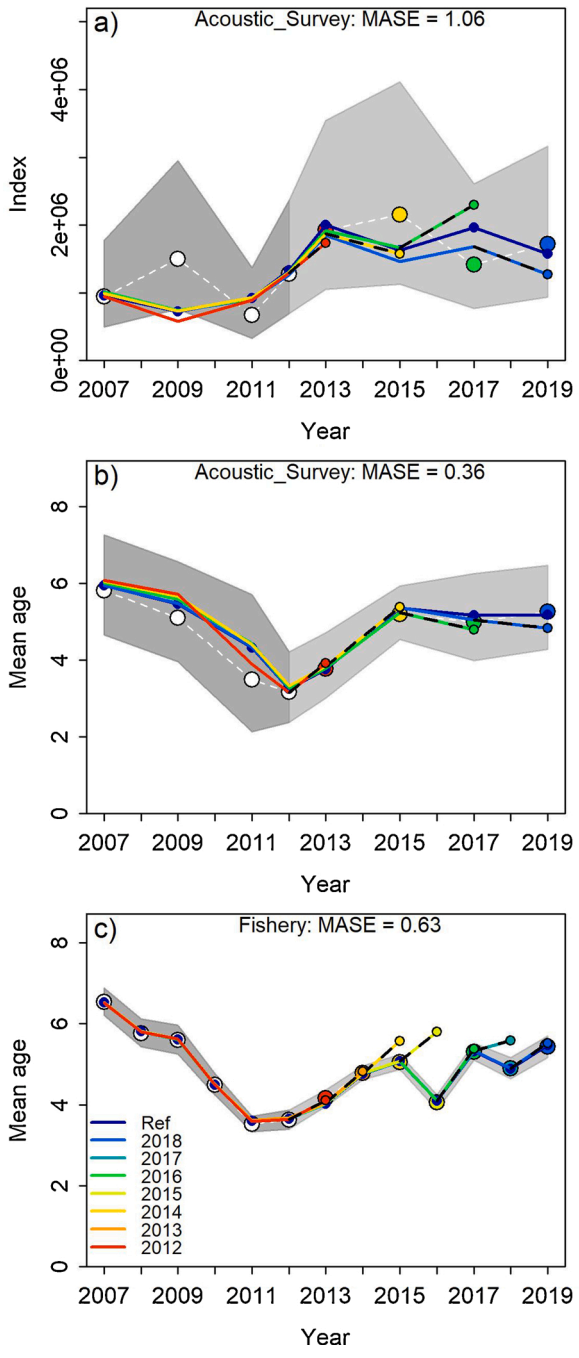


Fig. 10. Hindcasting cross-validation (HCxval) results for the fits to (a) the acoustic biomass survey index, and (b) annual mean age estimates from the survey, and (c) the fishery for the Pacific Hake (HAKE) model, showing observed (large points connected with dashed line), fitted (solid lines) and one-year-ahead forecast values (small terminal points). HCxval was performed using one reference model (Ref) and seven hindcast model runs (solid lines) relative to the expected survey index. The observations used for cross-validation are highlighted as color-coded solid circles with associated 95 % confidence intervals (light-gray shading). The model reference year refers to the endpoints of each one-year-ahead forecast and the corresponding observation (i.e., year of peel + 1). The mean absolute scaled error (MASE) score associated with the survey index and age-composition time series is denoted in each panel.

These four criteria are not limited to selecting a single base-case model but could also be used for objectively assigning weights to an ensemble of models (Maunder et al., 2020). They can also be of value in the process of developing Management Strategy Evaluation (MSE) frameworks, where integrated models are commonly used for conditioning Operating Models (OMs) to evaluate the performance of harvest control rules (Butterworth and Punt, 1999; Punt et al., 2015; Sharma et al., 2020). This often involves modelling the resource dynamics by fitting integrated assessment models to the available data based on some statistical criterion, such as a maximum likelihood (Hillary et al., 2015). The aim of conditioning is to discard OMs that do not fit the data satisfactorily and are consequently inconsistent with the observations and, therefore, implausible (Punt et al., 2015). So, when conditioning OMs, the intention is not to find a "best assessment" but a limited set of OMs with high plausibility (Sharma et al., 2020), which includes the most important uncertainties in the model structure, parameters, and data (Butterworth and Punt, 1999; Punt et al., 2015). The proposed plausibility criteria may be evaluated formally based on selected model diagnostic tests included in the toolbox and then ideally combined with expert judgment (e.g., Maunder et al., 2020) to weight performance statistics when integrating over results for different OMs or across a model ensemble.

The diagnostic toolbox presented here includes several promising, contemporary model diagnostic approaches, but these are far from exhaustive. Applications of Monte-Carlo Markov Chain (MCMC) approaches as model diagnostic tools have been rapidly evolving, in particular the process of "regularizing" of parameter penalties and priors in stock assessment models (i.e., to check that all parameters are identifiable; Monnahan et al., 2019) and posterior predictive checks and associated *p* values, which have become a standard approach for evaluating the goodness of fit for Bayesian models (Conn et al., 2018). One of their advantages is that different discrepancy measures can be used to check different components of the model, which is particularly useful for integrated models that use multiple data sets of various types. In posterior predictive checks, a potential challenge for stock assessment models is that this diagnostic can be conservative and, therefore, not overly sensitive to model misspecification (Conn et al., 2018). Also, contemporary integrated stock assessment models are complex and highly parameterized, and Bayesian inference is often not possible or impractical due to long computer processing time. Besbeas and Morgan (2014) developed a non-Bayesian approach based on posterior predictive checks that sample the model parameters from a multivariate normal distribution using the MLEs and variance-covariance matrix of the parameter estimates to sample the model parameters for simulating the data rather than sampling from the posterior distribution. Further work is needed to improve these methods so that they are practical for stock assessment models and are better at detecting and identifying model misspecification.

One approach to further evaluate the sensitivity and specificity of diagnostic tests would be to take a broad set of peer-reviewed stock assessments (e.g., all those implemented in Stock Synthesis and accepted for management advice through an independent review system) and use them as simulators to test a set of diagnostics under both correctly specified models and misspecified models. The misspecifications should include fixed parameter values (e.g., natural mortality, steepness), model structure (e.g., form of the selectivity curve), process variation (random and systematic), likelihood functions (e.g., size of the variance parameter, the structure of the likelihood function), observation models (e.g., non-proportionality between the index of abundance and population size, selectivity function) and other relevant aspects of the model. The results of this analysis could then be analyzed to determine which diagnostics detect which type of model misspecification, whether there are specifics of a diagnostic that can be associated with model misspecification, and whether combinations of diagnostics are associated with specific misspecifications.

Integrated models linking all data components via the estimated dynamics are both a blessing and a curse, as problems fitting to any particular data component may not necessarily result from a misspecified process directly linked to that component. The next generation of stock assessment models is likely to incorporate even more data types

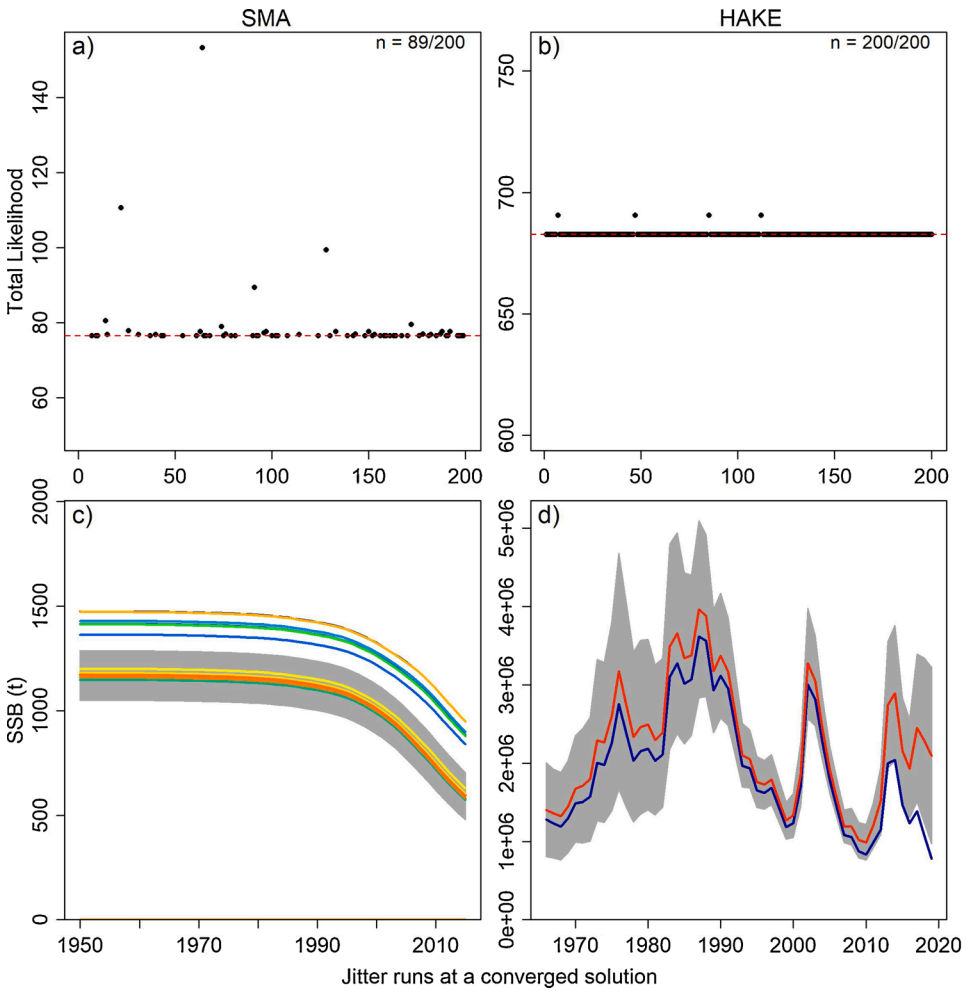


Fig. 11. The jitter diagnostic for global convergence conducted for the North Atlantic shortfin mako (SMA) and Pacific hake (HAKE) models. On top panels (a) – (b), solid black circles represent the total likelihood obtained from jittered model runs. The red horizontal dashed line represents the total likelihood value from the base-case model. Bottom panels (c) – (d) show the spawning stock biomass (SSB) from jittered model runs. Grey shaded areas are the 95 % confidence intervals from the base-case model.

and complex processes than the current models (Punt et al., 2020). These include spatial subpopulations, interactions with other species in the ecosystem, and influence from oceanographic processes (Punt et al., 2020). Problems associated with diagnosing the specific causes of poor model performance are likely to increase as these new complexities are explicitly considered. This will put even greater emphasis on the use of diagnostics (both existing and new) that can pinpoint the location of the misspecification.

In conclusion, we recommend that the next generation of stock assessment models should offer a routine diagnostic toolbox, as presented in this study. Automation of the toolbox across multiple operating systems and stock assessment software to produce warnings when diagnostics fail would also greatly facilitate their application. The diagnostic toolbox should be used to perform a comprehensive and systematic evaluation of the model(s) to determine which component(s) is misspecified and thus guide the analyst towards modifications and alternative model configurations that can be explored to minimize or eliminate such problems before the model is used for management advice.

Authorship contribution statement

FC and HW developed the concept of the paper; HW, FC, MC and LK developed the accompanying R package on GitHub; FC, HW, DC, MC, LK, MK, DY and IT conducted the diagnostic analyses; all authors contributed to writing and editing.

Declaration of Competing Interest

The authors have replied that they have no conflicts of interest to declare.

Acknowledgments

This work was carried out using data provided by the International Commission for the Conservation of Atlantic Tunas (ICCAT). The contents of this paper do not necessarily reflect the point of view of ICCAT and in no way anticipate the Commissions' future policy in this area. We thank The Center for the Advancement of Population Assessment Methodology (CAPAM) and New Zealand National Institute of Water and Atmospheric Research Ltd (NIWA) for hosting a technical workshop on the creation of frameworks for the next generation general stock assessment models in Wellington, New Zealand November 4-8, 2019. We also thank the NOAA staff who provided an internal review and the three anonymous reviewers for their suggestions and comments. The opinions expressed herein are those of the authors and do not necessarily reflect the view of NOAA or its sub-agencies.

References

Anhøj, J., Olesen, A.V., 2014. Run charts revisited: a simulation study of run chart rules for detection of non-random variation in health care processes. PLoS One 9, 1–13. https://doi.org/10.1371/journal.pone.0113825.

Barnston, A.G., Tippett, M.K., Ranganathan, M., L'Heureux, M.L., 2019. Deterministic skill of ENSO predictions from the north american multimodel ensemble. Clim. Dyn. 53, 7215–7234. https://doi.org/10.1007/s00382-017-3603-3.

Besbeas, P., Morgan, B.J.T., 2014. Goodness-of-fit of integrated population models using calibrated simulation. Methods Ecol. Evol. 5, 1373–1382. https://doi.org/10.1111/2041-210X.12279.

Brooks, E.N., Legault, C.M., 2016. Retrospective forecasting — evaluating performance of stock projections for New England groundfish stocks. Can. J. Fish. Aquat. Sci. 73, 935–950.

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H., 2005. CASAL (C++ Algorithmic Stock Assessment Laboratory): CASAL User Manual v2.07–2005/08/21 (No. Technical Report, 127). NIWA. http://www.niwascience.co.nz/ncfa/tools/casal.le.

Butterworth, D.S., Punt, A.E., 1999. Experiences in the evaluation and implementation of management procedures. ICES J. Mar. Sci. 56, 985–998. https://doi.org/10.1006/jmsc.1999.0532.

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.

Conn, P.B., Johnson, D.S., Williams, P.J., Melin, S.R., Hooten, M.B., 2018. A guide to Bayesian model checking for ecologists. Ecol. Monogr. 88, 526–542. https://doi.org/10.1002/ecm.1314.

Cope, J.M., 2013. Implementing a statistical catch-at-age model (Stock Synthesis) as a tool for deriving overfishing limits in data-limited situations. Fish. Res. 142, 3–14. https://doi.org/10.1016/j.fishres.2012.03.006.

Courtney, D., Carvalho, F., Winker, H., Kell, L., 2020. Examples of diagnostic methods implemented for previously completed North Atlantic shortfin mako Stock Synthesis model runs. Col. Vol. Sci. Pap. ICCAT 67, 173–234.

Courtney, D., Cortés, E., Zhang, X., 2017. Stock Synthesis (SS3) model runs conducted for North Atlantic shortfin mako. Collect. Vol. Sci. Pap. -ICCAT 74, 1759–1821. Available: https://www.iccat.int/en/pubs_CVSP.html (Accessed 4/2/2021).

Dichmont, C.M., Deng, R.A., Punt, A.E., Brodziak, J., Chang, Y.J., Cope, J.M., Ianelli, J. N., Legault, C.M., Methot, R.D., Porch, C.E., Prager, M.H., Shertzer, K.W., 2016. A reviewof stock assessment packages in the United States. Fish. Res. 183, 447–460. https://doi.org/10.1016/j.fishres.2016.07.001.

Eero, M., Hjelm, J., Behrens, J., Buchmann, K., Cardinale, M., Casini, M., Gasyukov, P., Holmgren, N., Horbowy, J., Hüssy, K., Kirkegaard, E., Kornilovs, G., Krumme, U., Köster, F.W., Oeberst, R., Plikshs, M., Radtke, K., Raid, T., Schmidt, J., Tomczak, M. T., Vinther, M., Zimmermann, C., Storr-Paulsen, M., 2015. Eastern Baltic cod in distress: biological changes and challenges for stock assessment. ICES J. Mar. Sci. 72, 2180–2186. https://doi.org/10.1093/icesjms/fsv109.

Fournier, D., Archibald, C.P., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39, 1195–1207. https://doi.org/10.1139/f82-157.

Fournier, D.A., Hampton, J., Sibert, J.R., 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat. Sci. 55, 2105–2116. https://doi.org/10.1139/cjfas-55-9-2105.

Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138. https://doi.org/10.1139/f2011-025.

Francis, R.I.C.C., Hurst, R.J., Renwick, J.A., 2003. Quantifying annual variation in catchability for commercial and research fishing. Fish. Bull. 101, 293–304. Available: https://spo.nmfs.noaa.gov/content/quantifying-annual-variation-catchability-commercial-and-research-fishing (Accessed 4/3/2021).

Goethel, D.R., Quinn, T.J., Cadrin, S.X., 2011. Incorporating spatial structure in stock assessment: movement modeling in marine fish population dynamics. Rev. Fish. Sci. Aquac. 19, 119–136. https://doi.org/10.1080/10641262.2011.557451.

Grandin, C.J., Johnson, K.F., Edwards, A.M., Berger, A.M., 2020. Status of the Pacific Hake (whiting) Stock in U.S. And Canadian Waters in 2020. Prepared by the Joint Technical Committee of the U.S. And Canada Pacific Hake/Whiting Agreement. National Marine Fisheries Service and Fisheries and Oceans Canada. Available: https://media.fisheries.noaa.gov/dam-migration/hake-assessment-2020-final.pdf (Accessed 4/2/2021).

Haltuch, M.A., Punt, A.E., 2011. The promises and pitfalls of including decadal-scale climate forcing of recruitment in groundfish stock assessment. Can. J. Fish. Aquat. Sci. 68, 912–926. https://doi.org/10.1139/f2011-030.

Henríquez, V., Licandeo, R., Cubillos, L.A., Cox, S.P., 2016. Interactions between ageing error and selectivity in statistical catch-at-age models: simulations and implications for assessment of the Chilean Patagonian toothfish fishery. ICES J. Mar. Sci. 73, 1074–1090. https://doi.org/10.1093/icesjms/fsv270.

Hillary, R.M., Preece, A.L., Davies, C.R., Kurota, H., Sakai, O., Itoh, T., Parma, A.M., Butterworth, D.S., Ianelli, J., Branch, T.A., 2015. A scientific alternative to moratoria for rebuilding depleted international tuna stocks. Fish Fish. Oxf. (Oxf) 17, 469–482. https://doi.org/10.1111/faf.12121.

Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., Licandeo, R., McGilliard, C.R., Monnahan, C.C., Muradian, M.L., Ono, K., Vert-Pre, K.A., Whitten, A.R., Punt, A.E., 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. Ices J. Mar. Sci. 72, 99–110. https://doi.org/10.1093/icesjms/fsu198.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

ICES, 2019. Benchmark Workshop on Baltic Cod Stocks (WKBALTCOD2). ICES Sci. Reports 1 (9), 1–310. https://doi.org/10.17895/ices.pub.4984.

Ichinokawa, M., Okamura, H., Takeuchi, Y., 2014. Data conflict caused by model misspecification of selectivity in an integrated stock assessment model and its potential effects on stock status estimation. Fish. Res. 158, 147–157. https://doi.org/10.1016/j.fishres.2014.02.003.

Johnson, K.F., Councill, E., Thorson, J.T., Brooks, E., Methot, R.D., Punt, A.E., 2016. Can autocorrelated recruitment be estimated using integrated assessment models and how does it affect population forecasts? Fish. Res. 183, 222–232. https://doi.org/10.1016/j.fishres.2016.06.004.

Keenlyside, N.S., Latif, M., Jungclaus, J., Kornblueh, L., Roeckner, E., 2008. Advancing decadal-scale climate prediction in the North Atlantic sector. Nature 453, 84–88. https://doi.org/10.1038/nature06921.

Kell, L.T., Kimoto, A., Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fish. Res. 183, 119–127. https://doi.org/10.1016/j.fishres.2016.05.017.

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., Fu, D., 2021. Validation of stock assessment methods: is it me or my model talking? *in press* ICES J. Mar. Sci.

Lee, H.-H., Maunder, M.N., Piner, K.R., Methot, R.D., 2011. Estimating natural mortality within a fisheries stock assessment model: An evaluation using simulation analysis based on twelve stock assessments. Fish. Res. 109, 89–94. https://doi.org/10.1016/j.fishres.2011.01.021.

Lee, H.-H., Piner, K.R., Methot, R.D., Maunder, M.N., 2014. Use of likelihood profiling over a global scaling parameter to structure the population dynamics model: An example using blue marlin in the Pacific Ocean. Fish. Res. 158, 138–146. https://doi.org/10.1016/j.fishres.2013.12.017.

Mangel, M., MacCall, A.D., Brodziak, J., Dick, E.J., Forrest, R.E., Pourzard, R., Ralston, S., 2013. A perspective on steepness, reference points, and stock assessment. Can. J. Fish. Aquat. Sci. 70, 930–940. https://doi.org/10.1139/cjfas-2012-0372.

Maunder, M.N., Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72, 7–18. https://doi.org/10.1093/icesjms/fsu015.

Maunder, M.N., Piner, K.R., 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fish. Res. 192, 16–27. https://doi.org/10.1016/j.fishres.2016.04.022.

Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. 70, 141–159. https://doi.org/10.1016/j.fishres.2004.08.002.

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74. https://doi.org/10.1016/j.fishres.2012.07.025.

Maunder, M.N., Starr, P.J., 2001. Bayesian assessment of the SNA1 snapper (Pagrus auratus) stock on the north - east coast of New Zealand. New Zeal. J. Mar. Freshw. Res. 35, 87–110. https://doi.org/10.1080/00288330.2001.9516980.

Maunder, M.N., Harley, S.J., Hampton, J., 2006. Including parameter uncertainty in forward projections of computationally intensive statistical population dynamic models. ICES J. Mar. Sci. 63, 969–979. http://icesjms.oxfordjournals.org/content/63/6/969.abstract.

Maunder, M.N., Xu, H., Lennert-Cody, C.E., Valero, J.L., Aires-da-Silva, A., Minte-Vera, 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.

Methot, R.D., Taylor, I.G., 2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. Can. J. Fish. Aquat. Sci. 68, 1744–1760.

Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99. https://doi.org/10.1016/j.fishres.2012.10.012.

Methot Jr., R.D., Wetzel, C.R., Taylor, I.G., Doering, K., 2020. Stock Synthesis User Manual Version 3.30.15. U.S. Department of Commerce. NOAA Processed Report NMFS-NWFSC-PR-2020-05. https://doi.org/10.25923/5wpn-qt71. https://vlab.ncep.noaa.gov/web/stock-synthesis.

Michaelsen, J., 1987. Cross-validation in statistical climate forecast models. J. Clim. Appl. Meteorol. 26, 1589–1600. https://doi.org/10.1175/1520-0450(1987)026<1589:CVISCF>2.0.CO;2.

Miller, T.J., Legault, C.M., 2017. Statistical behavior of retrospective patterns and their effects on estimation of stock and harvest status. Fish. Res. 186, 109–120. https://doi.org/10.1016/j.fishres.2016.08.002.

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

Mohn, R., 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56, 473–488. https://doi.org/10.1006/jmsc.1999.0481.

Monnahan, C.C., Branch, T.A., Thorson, J.T., Stewart, I.J., Szuwalski, C.S., 2019. Overcoming long Bayesian run times in integrated fisheries stock assessments. ICES J. Mar. Sci. 76, 1477–1488. https://doi.org/10.1093/icesjms/fsz059.

Punt, A.E., 2019. Spatial stock assessment methods: a viewpoint on current issues and assumptions. Fish. Res. 213, 132–143. https://doi.org/10.1016/j.fishres.2019.01.014.

Punt, A.E., Cope, J.M., 2019. Extending integrated stock assessments models to use non-depensatory three-parameter stock-recruitment relationships. Fish. Res. 217, 46–57. https://doi.org/10.1016/j.fishres.2017.07.007.

Punt, A.E., Hurtado-Ferro, F., Whitten, A.R., 2014. Model selection for selectivity in fisheries stock assessments. Fish. Res. 158, 124–134. https://doi.org/10.1016/j.fishres.2013.06.003.

Punt, A.E., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A.A., Haddon, M., 2015. Management strategy evaluation: best practices. Fish Fish. Oxf. (Oxf) 17, 303–334. https://doi.org/10.1111/faf.12104.

Punt, A.E., Dunn, A., Elvarsson, B.P., Hampton, J., Hoyle, S.D., Maunder, M.N., Methot, R.D., Nielsen, A., 2020. Essential features of the next-generation integrated fisheries stock assessment package: A perspective. Fish. Res. 229, 105617 https://doi.org/10.1016/j.fishres.2020.105617.

Sharma, R., Langley, A., Herrera, M., Greehan, J., Hyun, S.Y., 2014. Investigating the influence of length–frequency data on the stock assessment of Indian Ocean bigeye tuna. Fish. Res. 158, 50–62. https://doi.org/10.1016/j.fishres.2014.01.012.

Sharma, R., Levontin, P., Kitakado, T., Kell, L., Mosqueira, I., Kimoto, A., Scott, R., Minte-Vera, C., De Bruyn, P., Ye, Y., Kleineberg, J., Walton, J.L., Miller, S., Magnusson, A., 2020. Operating model design in tuna Regional Fishery Management Organizations: current practice, issues and implications. Fish Fish. 21, 940–961. https://doi.org/10.1111/faf.12480.

Smith, D.M., Eade, R., Dunstone, N.J., Fereday, D., Murphy, J.M., Pohlmann, H., Scaife, A.A., 2010. Skilful multi-year predictions of Atlantic hurricane frequency. Nat. Geosci. 3, 846–849. https://doi.org/10.1038/ngeo1004.

Stewart, I.J., Monnahan, C.C., 2017. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. Fish. Res. 192, 126–134. https://doi.org/10.1016/j.fishres.2016.06.018.

Subbey, S., 2018. Parameter estimation in stock assessment modelling: caveats with gradient-based algorithms. ICES J. Mar. Sci. 75, 1553–1559. https://doi.org/10.1093/icesjms/fsy044.

Taylor, I.G., Doering, K.L., Johnson, K.F., Wetzel, C.R., Stewart, I.J., 2021. Beyond visualizing catch-at-age models: lessons learned from the r4ss package about software to support stock assessments. Fish. Res. 239, 105924 https://doi.org/10.1016/j.fishres.2021.105924.

Thorson, J.T., 2020. Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model. Fish Fish. 21, 237–251. https://doi.org/10.1111/faf.12427.

Thorson, J.T., Johnson, K.F., Methot, R.D., Taylor, I.G., 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fish. Res. 192, 84–93. https://doi.org/10.1016/j.fishres.2016.06.005.

Thorson, J.T., Rudd, M.B., Winker, H., 2019. The case for estimating recruitment variation in data-moderate and data-poor age-structured models. Fish. Res. 217, 87–97. https://doi.org/10.1016/j.fishres.2018.07.007.

Truesdell, S.B., Bence, J.R., Syslo, J.M., Ebener, M.P., 2017. Estimating multinomial effective sample size in catch-at-age and catch-at-size models. Fish. Res. 192, 66–83. https://doi.org/10.1016/j.fishres.2016.11.003.

Vasilakopoulos, P., Jardim, E., Konrad, C., Rihan, D., Mannini, A., Pinto, C., Casey, J., Mosqueira, I., O'Neill, F.G., 2020. Selectivity metrics for fisheries management and advice. Fish Fish. 21, 621–638. https://doi.org/10.1111/faf.12451.

Wald, A., Wolfowitz, J., 1940. On a test whether two samples are from the same population. Ann. Math. Stat. 11, 147–162. http://www.jstor.org/stable/2235872.

Walters, C.J., Hilborn, R., Christensen, V., 2008. Surplus production dynamics in declining and recovering fish populations. Can. J. Fish. Aquat. Sci. 65, 2536–2551. https://doi.org/10.1139/F08-170.

Wang, S.-P., Maunder, M.N., 2017. Is down-weighting composition data adequate for dealing with model misspecification, or do we need to fix the model? Fish. Res. 192, 41–51.

Wang, S.-P., Maunder, M.N., Piner, K.R., Aires-da-Silva, A., Lee, H.-H., 2014. Evaluation of virgin recruitment profiling as a diagnostic for selectivity curve structure in integrated stock assessment models. Fish. Res. 158, 158–164. https://doi.org/10.1016/j.fishres.2013.12.009.

Wang, S.-P., Maunder, M.N., Nishida, T., Chen, Y.-R., 2015. Influence of model misspecification, temporal changes, and data weighting in stock assessment models: Application to swordfish (*Xiphias gladius*) in the Indian Ocean. Fish. Res. 166, 119–128. https://doi.org/10.1016/j.fishres.2014.08.004.

Wetzel, C.R., Punt, A.E., 2015. Evaluating the performance of data-moderate and catch-only assessment methods for U.S. west coast groundfish. Fish. Res. 171, 170–187. https://doi.org/10.1016/j.fishres.2015.06.005.

Wilberg, M.J., Thorson, J.T., Linton, B.C., Berkson, J., 2009. Incorporating time-varying catchability into population dynamic stock assessment models. Rev. Fish. Sci. Aquac. 18, 7–24. https://doi.org/10.1080/10641260903294647.

Winker, H., Carvalho, F., Kapur, M., 2018. JABBA: Just Another Bayesian Biomass Assessment. Fish. Res. 204, 275–288. https://doi.org/10.1016/j.fishres.2018.03.010.

Winker, H., Carvalho, F., Kerwath, S., 2020. Age-structured biomass dynamics of North Atlantic shortfin mako with implications for the interpretation of surplus production models. Col. Vol. Sci. Pap. ICCAT 76, 316–336. https://www.iccat.int/Documents/CVSP/CV076_2019/colvol76.html.