Fisheries Research xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Fisheries Research

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

Full length article

# Can diagnostic tests help identify model misspecification in integrated stock assessments?

Felipe Carvalho[a,b,*], André E. Punt[c], Yi-Jay Chang[d], Mark N. Maunder[e,f], Kevin R. Piner[g]

[a] University of Hawaii, Joint Institute for Marine and Atmospheric Research, Honolulu, HI 96822, USA
[b] NOAA, National Marine Fisheries Service, Pacific Islands Fisheries Science Center, Honolulu, HI 96818, USA
[c] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195, USA
[d] Institute of Oceanography, National Taiwan University, Taipei 10617, Taiwan
[e] Inter-American Tropical Tuna Commission, La Jolla Shores Drive, La Jolla, CA 92037, USA
[f] Center for the Advancement of Population Assessment Methodology, Scripps Institution of Oceanography, La Jolla, USA
[g] NOAA Fisheries, Southwest Fisheries Science Center, La Jolla Shores Dr, La Jolla, CA 92037, USA

ARTICLE INFO

*Article history:*
Received 14 June 2016
Received in revised form
15 September 2016
Accepted 16 September 2016
Handled by Prof. George A. Rose
Available online xxx

*Keywords:*
Integrated stock assessments
Simulation
Model diagnostics

ABSTRACT

A variety of data types can be included in contemporary integrated stock assessments to simultaneously provide information on all estimated parameters. Conflicts between data, which are often a symptom of model misspecification and evident as model misfit, can affect the estimates of important parameters and derived quantities. Unfortunately, there are few standard diagnostic tools available for integrated stock assessment models that can provide the analyst with all the information needed to determine if there is substantial model misspecification. In this study, we use simulation methods to evaluate the ability of commonly-used and recently-proposed diagnostic tests to detect model misspecification in the observation model process (i.e., the incorrect form for survey selectivity), systems dynamics (i.e., incorrect assumed values for steepness of the stock-recruitment relationship and natural mortality), and incorrect data weighting. The diagnostic tests evaluated here were: i) residuals analysis (SDNR and runs test); ii) retrospective analysis; iii) the $R_0$ likelihood component profile; iv) the age-structured production model (ASPM); and v) catch-curve analysis (CCA). The efficacy of the diagnostic tests depended on whether the misspecification was in the observation or systems dynamics model. Residual analyses were easily the best detector of misspecification of the observation model while the ASPM test was the only good diagnostic for detecting misspecification of system dynamics model. Retrospective analysis and the $R_0$ likelihood component profile infrequently detected misspecified models, and CCA had a high probability of rejecting correctly-specified models. Finally, applying multiple carefully selected diagnostics can increase the power to detect misspecification without substantially increasing the probability of falsely concluding there is misspecification when the model is correctly specified.



## 1. Introduction

The advantages of 'integrated' assessments are numerous, and include the ability to combine many data sources to estimate important population dynamics processes such as growth, natural mortality, fishing mortality and movement simultaneously (Doubleday, 1976; Fournier and Archibald, 1982; Maunder and Punt, 2013; Punt et al., 2013). This is made possible by summing the log-likelihoods from each data component (e.g., abundance indices, size-composition, tagging) into a single total log-likelihood. Another advantage of integrated assessments is that they allow the sensitivity to dataset choice to be evaluated and hence conflicts among datasets and model misspecification to be identified (Maunder and Punt, 2013). The Stock Synthesis (SS) assessment framework (Methot and Wetzel, 2013) is one of the most well-known examples of an integrated model, and has been applied in a wide variety of fish assessments globally (Wetzel and Punt, 2011; Methot and Wetzel, 2013).

However, simultaneously analyzing multiple data sources can lead to conflicts among the data sources, especially between size-composition data and indices of relative abundance (Francis, 2011; Ichinokawaa et al., 2014; Wang et al., 2014; Lee et al., 2014). Most recently, Maunder and Piner (2015) stated that conflicts between

* Corresponding author at: University of Hawaii, Joint Institute for Marine and Atmospheric Research, Honolulu, HI 96822, USA.
*E-mail address:* felipe.carvalho@noaa.gov (F. Carvalho).

http://dx.doi.org/10.1016/j.fishres.2016.09.018
0165-7836/

data sources arise due to: 1) random sampling error, 2) misspecification of the observation model (i.e., the model processes relating the population dynamics or states to data), and 3) misspecification of the system dynamics model (i.e., the population dynamics model). Analysts often down weight some of the data sources when confronted with conflicting data sources (e.g., Harle et al., 2015; Kell et al., 2014). However, this is not necessarily appropriate because it may not resolve the model misspecification (Wang et al., 2015). Deroba and Schueller (2013) and Lee et al. (2014) have shown that model misspecification can substantially bias assessment outcomes, affecting, in particular, parameter estimates, and determination of stock status. For example, assuming that the selectivity of a fishery is asymptotic when it is in fact dome-shaped can substantially bias estimates of absolute abundance (Wang et al., 2009). Alternative model structures can be explored to identify inconsistencies and hence form the basis to justify down weighting some data sources, as well as an indication of what component of the model structure is misspecified (Maunder and Piner, 2015). Francis (2011) recommends prioritizing indices of relative abundance, assuming that these data are representative of changes in stock abundance. However, age- and size-composition data can be more informative about the level of fishing mortality and biomass when the index is uninformative (i.e., there is no contrast in abundance levels) and/or is of poor quality (e.g., high sampling error or the index is not proportional to abundance). Although size- and age-composition data may provide substantial information on absolute abundance, the prioritization of indices of relative abundance is recommended because even slight model misspecifications can have a large impact on the information about absolute abundance contained in compositional data (Maunder and Piner, 2015; Lee et al., 2014; Wang et al., 2014).

There is little guidance and few objective criteria to determine how to best summarize the results of integrated assessments, determine if the model fits the data adequately, and if the model is well specified. Moreover, it is very difficult to easily evaluate convergence or identify problematic areas given the large number of estimable parameters in these assessments (Harley and Maunder, 2003). Applying classical model diagnostic tools in integrated stock assessments requires further investigation and possible refinement before good practice recommendations can be made. Some of the most common or recently proposed diagnostic tests to be used with integrated stock assessments include:

- *Residual analysis*. Analysis of residuals is perhaps the most common way to determine a model's goodness-of-fit (Cox and Snell, 1968). Residuals are examined for patterns to evaluate whether the model assumptions have been met (e.g., Wang et al., 2009). Many statistics exist to evaluate the residuals for desirable properties. One way is to calculate, for each abundance index, the standard deviation of the normalized (or standardized) residuals divided by the sampling (or assumed) standard deviation (SDNR) (Breen et al., 2003; Francis, 2011). The SDNR is a measure of the fit to the data that is independent of the number of data points. A relatively good model fit will be characterized by smaller residuals (i.e. close to zero) and a SDNR close to 1. Francis (2011) notes that it is also necessary to conduct a visual examination between observed and predicted values to be sure that the fit is good even when SDNR values are not much greater than 1. A non-random pattern of residuals may indicate that some heteroscedasticity is present, or there is some leftover serial correlation (serial correlation in sampling/observation error or model misspecification). Several well-known nonparametric tests for randomness in a time-series include: the runs test, the sign test, the runs up and down test, the Mann-Kendall test, and Bartel's rank test (Gibbons and Chakraborti, 1992). In this study, we used the runs test to evaluate whether residuals are random over time, because this test has been used to diagnose fits to indices and other data components in assessment models (e.g. SEDAR 40, 2015).
- *Retrospective analysis.* Retrospective analysis is another diagnostic approach widely used in stock assessment to evaluate the reliability of parameter and reference point estimates (Cadigan and Farrell, 2005; Hurtado-Ferro et al., 2014). Retrospective analysis involves fitting a stock assessment model to the full dataset, and the same model is then fitted to truncated datasets where the data for the most recent years have been sequentially removed. Retrospective analysis usually assumes that the estimates of historical abundance from the current assessment that uses all the data are more accurate than the estimates of "current" abundance from assessments that ignore recent data, therefore revealing possible bias of model predictions. In stock assessment, the "$e$; $\rho$"$e$; statistic proposed by Mohn (1999) is commonly used to evaluate the severity of retrospective patterns (Deroba, 2014). This statistic measures the average of relative difference between an estimated quantity from an assessment (e.g., biomass in final year) with a reduced time-series and the same quantity estimated from an assessment using the full time-series. According Hurtado-Ferro et al. (2014), retrospective patterns generally arise from two main causes: time-varying processes unaccounted for in the assessment (i.e., model misspecification), or incomplete data.
- $R_0$ *likelihood component profile.* Negative log-likelihoods of various data components for a profiled parameter (e.g., virgin recruitment) have been used as a diagnostic to evaluate the influence of each data component on estimates of model parameters and outputs (e.g., Maunder, 1998; Maunder and Starr, 2001; Francis, 2011; Lee et al., 2014; Ichinokawaa et al., 2014; Maunder and Piner, 2015). Wang et al. (2014) proposed an extension of $R_0$ (virgin recruitment) likelihood profiling to diagnose stock assessment models with misspecified selectivity. Their method consists of constructing a $R_0$ profile for data components simulated without error from a known stock assessment model. The $R_0$ profile from the known stock assessment model is assumed to represent the "true" information content of each data component. Any differences in subsequent models from the $R_0$ profile originated from the known stock assessment model are presumed to indicate conflict in the data or model misspecification. However, this diagnostic has not been used extensively or evaluated, and more research is needed before it can be recommended.
- *Age-structured production model*. In some integrated stock assessments the index of abundance provides almost no information on population scale. Consequently, the estimates of the model outputs rely almost completely on the size- and age-composition data and model structure. Maunder and Piner (2015) proposed a diagnostic tool that can be used to evaluate the information content of data about absolute abundance and assess whether the model is correctly specified. This diagnostic consists of comparing the results of an age-structured production model (ASPM) to those from a model estimating all of the model parameters and fitting to all the data (e.g., an integrated analysis). It is inferred that a production function is apparent in the data when the catch data explain indices with good contrast (e.g., declining and increasing trends), therefore providing evidence that the index is a reasonable proxy of stock trend. If the ASPM cannot mimic the index, then either the stock is recruitment-driven, catch levels have not been high enough to have a detectable impact on the population, the model is incorrect, or the index of relative abundance is uncertain or not proportional to abundance. Similar to the $R_0$ likelihood component profile, this diagnostic has only begun to be implemented, and its utility remains unknown.
- *Catch-curve analysis*. Most of the information on absolute abundance will come from the compositional data if the index of abundance provides little or no information on population scale.

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

It then becomes important to verify that the trend in the compositional data over time is consistent with the trend in the index of abundance. If this is the case, there should be more confidence in the estimates of trends in abundance. This can be interpreted as a diagnostic, because how consistent a data set is with the population model and the other data sets impacts how data sets are weighted. One of the simplest methods to evaluate the relationship between compositional data and fishing mortality is catch-curve analysis. Here, we propose and demonstrate for the first time how catch-curve analysis can potentially serve as a diagnostic to evaluate the consistency between trends in abundance over time from compositional data and an abundance index.

While diagnostic tests can be used to detect which model component has been misspecified, their reliability for this purpose is still unclear. In this study, we use simulation to evaluate the ability of commonly-used and recently-proposed diagnostic tests to detect model misspecification in integrated stock assessments. Three processes were misspecified: natural mortality, the steepness of the stock-recruitment relationship, and selectivity. The simulation analysis is based on the stock assessment for striped marlin (*Kajikia audax*) in the Western and Central North Pacific (WCNP) (Piner et al., 2013; Chang et al., 2015). We examined the estimation bias generated under each model misspecification and their resultant impact on the diagnostic tests. We also propose a range of diagnostic tests that can be used as defaults to determine whether a stock assessment model is misspecified.

## 2. Material and methods

### 2.1. Developing operating model

A simplified version of the 2015 stock assessment model for the WCNP striped marlin, implemented in SS was used in the simulation analysis as both the simulator and estimator. Three types of data were used in that model: fishery-specific catches, relative abundance indices, and length measurements. These data were compiled between 1975 and 2013. Available data sources and their temporal coverage are summarized in Fig. 1.

The operating model (OM) included three fisheries with catches from 1978 to 2013, denoted Fleet 1, Fleet 2, and Fleet 3. The observed total catches were input into the model seasonally (i.e., by calendar year and quarter) and in numbers (thousands of fish) for the three fisheries, and were assumed to be unbiased and relatively precise (Fig. 2a). Fleet 1 had three CPUE indices, while Fleets 2 and 3 had one CPUE index each (Fig. 2b). Selectivity was assumed to be asymptotic for Fleet 1 and 3, and dome-shaped for Fleet 2. The observed size-composition data from the original assessment were used for each fleet (Fig. 2c). The annual input effective sample sizes by fishery and year were assumed to be the associated total number of fish measured divided by 10 (Fig. 2d).

Biological, demographic, and fishery dynamic assumptions were taken from the original assessment (Table 1). Growth was modeled using a von Bertalanffy growth curve, recruitment was modeled using a Beverton-Holt stock-recruitment relationship, and the natural mortality rate ($M$) was age-specific, the steepness of the stock–recruitment relationship ($h$), and the extent of variation about the stock–recruitment relationship ($\sigma_R$) were pre-specified. The model started in 1975, and it was assumed that the combined fisheries were in equilibrium in 1975, with an equilibrium catch of 3000 mt.

Conditional age-at-length compositions were added to the OM, but were not available in the original assessment. These data were generated using the following process (Taylor and Methot, 2013):

1. The OM was used to calculate the expected conditional age-at-length compositions;
2. The size-compositions of fish to be aged were generated by sampling from a multinomial distribution with the chosen age sample size (50 individuals per season in this study) and proportions equal to the generated size-compositions; and
3. For each length bin with non-zero numbers of fish to be aged, the conditional age-at-length data were generated from a multinomial distribution with sample size equal to the value from step (ii) and proportions equal to the expected conditional age-at-length compositions from step (i). Conditional age-at-length compositions were generated for the time-steps for which size-composition data were available.

### 2.2. Data weighting

Data weighting is an important component of integrated stock assessment models. Consideration of the relative weighting of different data sources becomes even more important when the data appear to be in conflict. The goodness of fit to the relative abundance indices was prioritized rather than other data components, such as size-composition data when fitting to the original data to obtain the model parameters to use in the simulation analysis and set up the sample sizes to generate the data. This decision was based on the assumption that the relative abundance indices reflect a direct measure of population trend. The weighting method used for the CPUE indices and size-composition data followed the advice of Francis (2011) (Method TA1.8). For weighting the conditional age-at-length data we used the Francis-B approach described in Punt (2016). Iterative application of model fitting and reweighting occurred three times to explore the effects on successive estimates of the data weighting coefficient for each composition dataset. Weights from the first iteration were used for the results reported here because this iteration resulted in the smallest gradient for the objective function to be minimized among the three iterations of the model.

### 2.3. Simulation framework

The simulated data were generated using the parametric bootstrap feature of SS (Methot and Wetzel, 2013). The simulation framework consisted of two main parts: an OM (as described above), which determines the "true" population dynamics of the system from which data are sampled, and a separate estimation model (EM), which is fit to the data and provides estimates of quantities important for management. Critical inconsistencies between the OM and EM were expected to be minimal except for the intended model misspecifications in the simulation scenarios. In addition to the assumed correctly-specified model (CSM), the following misspecifications were introduced to the EM (based on the lack of information usually available to parameterize critical biological and fisheries processes in stock assessments):

1. all fleets were assumed to have asymptotic selectivity (EM_1);
2. steepness was fixed at 0.70 to reflect a less resilient stock (EM_2);
3. natural mortality was assumed to be constant for all ages and equal to 0.38 $yr^{-1}$ (EM_3); and
4. the weight assigned to the size- and age-composition data was increased by a factor of 10 (EM_4).

The simulation followed six general steps (Fig. 3):

1) the operating model was fit to the original dataset (see Table 1 for some of the parameters estimated in this process), and the

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

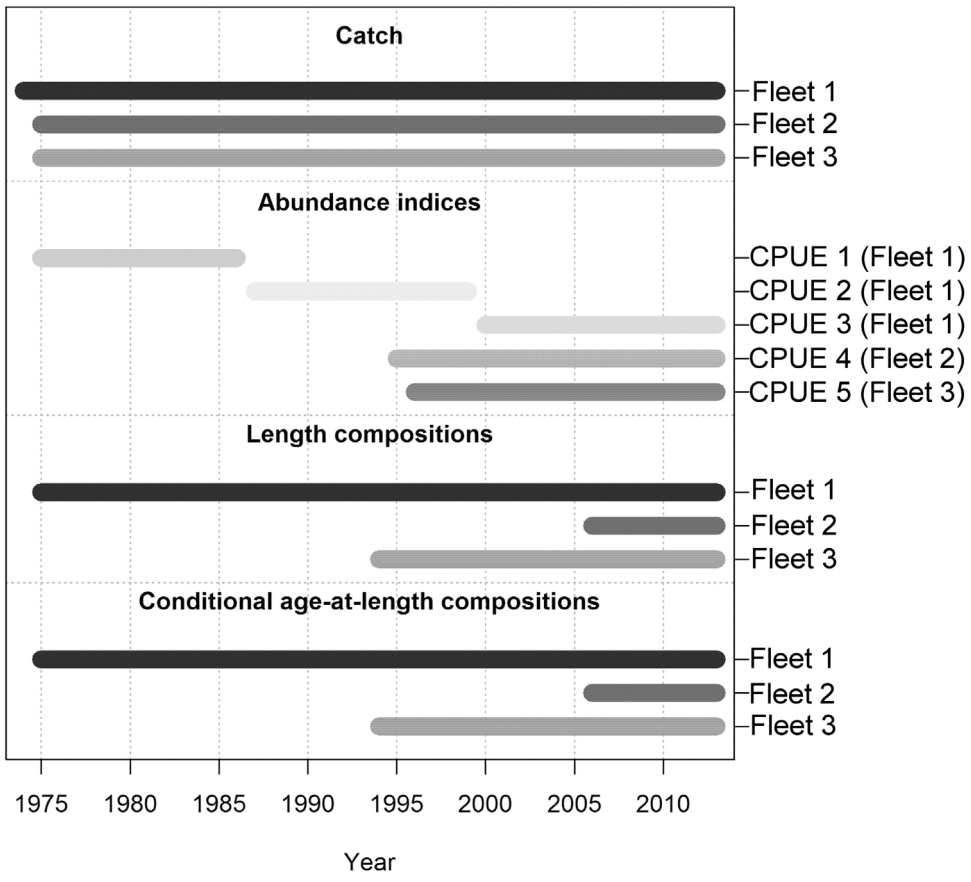


**Fig. 1.** Available temporal coverage and sources of catch, CPUE (abundance indices), size-composition, and conditional age-at-length data for the WCNP striped marlin stock assessment (Operating and Estimation Models).

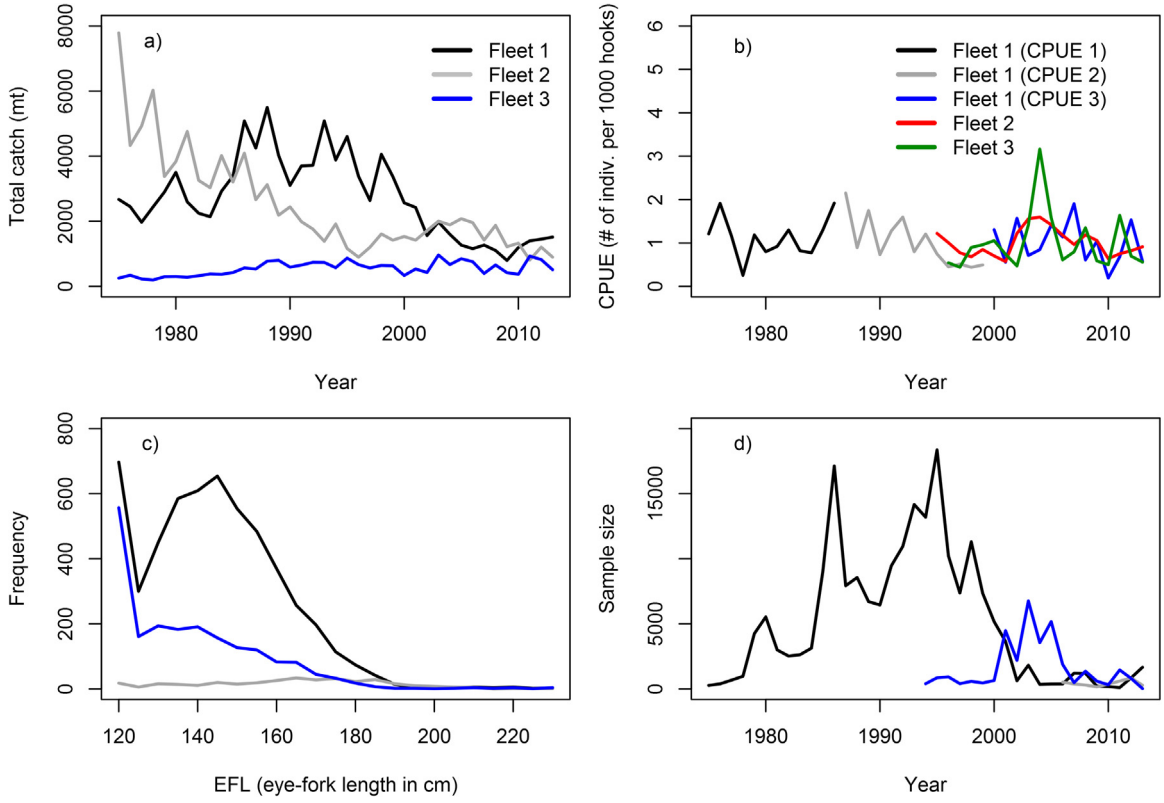


**Fig. 2.** Examples of data series for catch, CPUE, and length-frequency, and the sample sizes for size-frequency for WCNP striped marlin (Operating Model). The length-frequency data are aggregated across years to present overall patterns for each fishery.

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

**Table 1**
Key life history parameters and model structures used in the WCNP striped marlin stock assessment (Operating Model).

| Parameter | Value | Comments |
|---|---|---|
| Gender | Female only | |
| Natural mortality | 0.54 $yr^{-1}$ (age 0) 0.47 $yr^{-1}$ (age 1) 0.43 $yr^{-1}$ (age 2) 0.40 $yr^{-1}$ (age 3) 0.38 $yr^{-1}$ (age 4–15) | Age-specific natural mortality |
| Reference age (a1) | 0.3 yr | Fixed parameter |
| Maximum age (a2) | 15 yr | Fixed parameter |
| Length at a1 (L1) | 104 cm | Fixed parameter |
| Length at a2 (L2) | 214 cm | Fixed parameter |
| Growth rate (K) | 0.24 $yr^{-1}$ | Fixed parameter |
| CV of L1 | 0.14 | Fixed parameter |
| CV of L2 | 0.08 | Fixed parameter |
| Weight-at-length | $W = 4.68e{-}006 \times L^{3.16}$ | Fixed parameter |
| Size-at-50% Maturity | 177 cm | Fixed parameter |
| Slope of maturity ogive | $-0.064\,cm^{-1}$ | Fixed parameter |
| Fecundity | Proportional to spawning biomass | Fixed parameter |
| Spawning season | 2 | Model structure |
| Spawner-recruit relationship | Beverton-Holt | Model structure |
| Spawner-recruit steepness ($h$) | 0.87 | Fixed parameter |
| Log of Recruitment at virgin biomass log$R_0$ | 6.31642 | Estimated |
| Recruitment variability ($\sigma_R$) | 0.6 | Fixed parameter |
| Main recruitment deviations | 1975–2008 | Estimated |

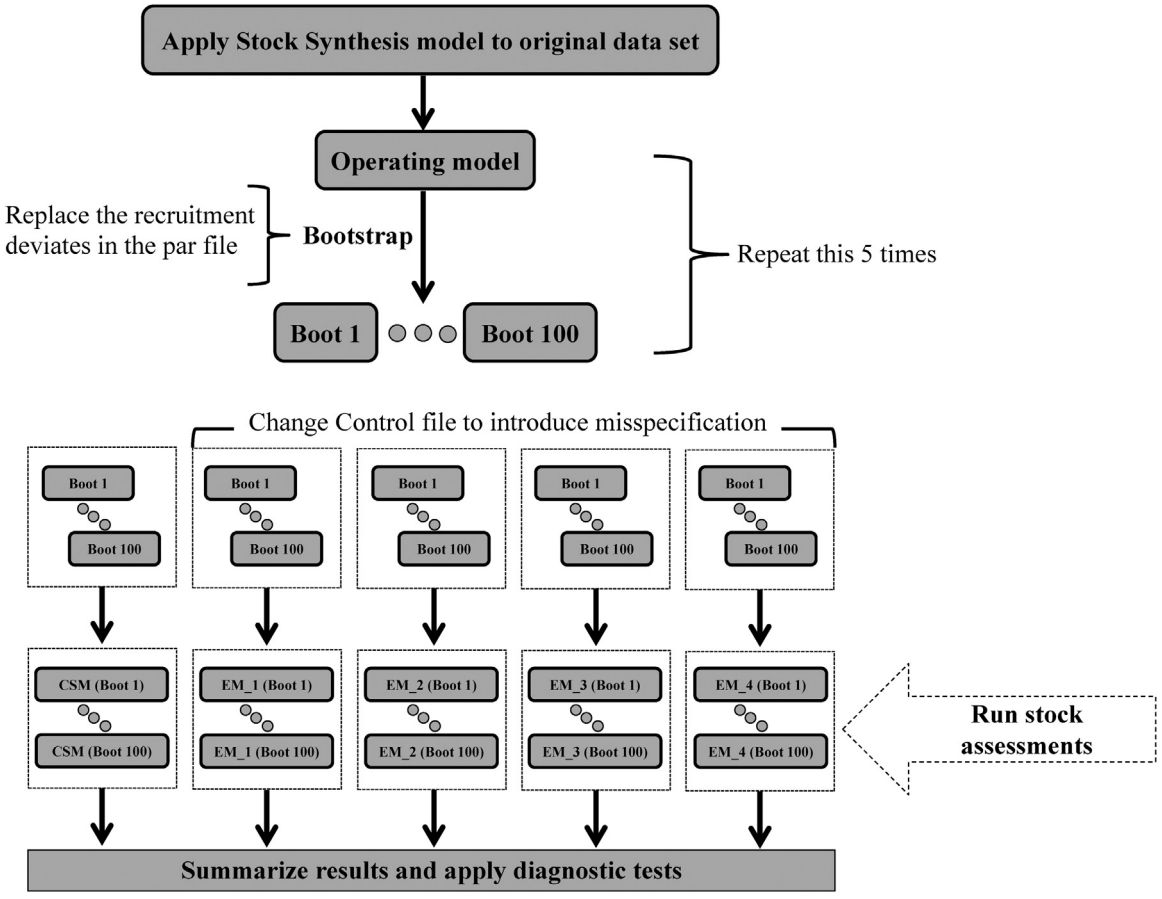


**Fig. 3.** General design of the simulation study.

resulting parameter estimates taken as the "true" values for the simulation;
2) random recruitment deviates were generated and the bootstrap procedures in SS were used to simulate 100 new datasets based on the OM;
3) the EM was fit to each data set with the same model as the OM to obtain CSM results (self-test; the same assessment platform, structural assumptions, and settings);
4) step 3 was repeated for each alternative EM;
5) estimates of relevant quantities from each EM were compared with their "true" values; and
6) the results of diagnostic tests results for the CSM (Step 3) were compared with those obtained by applying the diagnostic tests to the misspecified EMs (Step 4). This provides a form of 'Type I error' and "Type II error" evaluation, i.e. the probability of concluded that CSM is misspecified (which it is not) and the probability of detecting one of the misspecified models is indeed misspecified.

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

### 2.4. Assessing the impact of misspecification on assessment results

To quantify the impact of misspecification on assessment results, we calculate for each estimation method (i.e., EM_1, EM_2, EM_3, EM_4) the proportion of time that their estimates of a key management quantity (i.e., the ratio SSBterm/SSBinit) falls outside the 95% confidence interval (95% CI) of the same management quantity value from the CSM. Note that 5% of the cases will lead to misspecification being triggered even when the model is correctly specified (i.e., false positives). The impact of misspecification on estimation performance for all scenarios was also assessed by comparing the spawning biomass in the last year of the assessment as a ratio of the spawning biomass in the first year (SSBterm/SSBinit) estimated for each model with the "true" value. The bias and accuracy of the EMs were determined by calculating the median relative error (MRE) and the median absolute relative error (MARE) across simulations within a scenario as in Ono et al. (2015).

$$\mathrm{MRE} = \mathrm{median}\left(\frac{E_{(1)} - T_{(1)}}{T_{(1)}}, \ldots, \frac{E_{(100)} - T_{(100)}}{T_{(100)}}\right)$$

$$\mathrm{MARE} = \mathrm{median}\left(\left|\frac{E_{(1)} - T_{(1)}}{T_{(1)}}\right|, \ldots, \left|\frac{E_{(100)} - T_{(100)}}{T_{(100)}}\right|\right)$$

where $E$ is the estimated quantity of interest, $T$ the true value, and the subscript indicates the iteration number. Changes in model performance among scenarios can be evaluated by direct comparison of MRE and MARE values. As suggested by Ono et al. (2015), model parameters were considered accurately estimated when MARE was equal or below 16%, and have low bias when MRE was below ±4%.

### 2.5. Applying model diagnostic tests

The CSM and the alternative EMs were evaluated for lack of fit and resulting data conflicts using five diagnostic tests: i) analysis of residuals, ii) retrospective analysis, iii) $R_0$ likelihood component profile, iv) age-structured production model, and v) catch-curve analysis.

SDNR and a runs test were used to examine residual patterns. These analyses were made easier by the availability of the r4ss package (Taylor et al., 2011), which has been developed for SS to summarize and plot model results, manipulate files, and visualize model parameterizations. A strong non-random pattern in residuals may indicate model misspecification. To examine the randomness of the residuals, runs tests were conducted for CPUE indices, size-composition data, and conditional age-at-length data. To apply the runs tests to the size-composition data we used the Francis method (see equation 1.C in Table 2 of Punt, 2016), which calculates the standardized residuals of observed and model-predicted mean lengths by year (seasons combined). The runs test for the conditional age-at-length data were based on the Francis-B method (see equation 2.D in Table 2 of Punt, 2016), which calculates the standardized residuals of observed and model predicted mean conditional age-at-length by year. The runs tests were implemented using the function *runs.test* in the R package *tseries* (Trapletti, 2011). This function calculates the 2-sided $p$-value of the Wald-Wolfowitz runs test, which is a nonparametric statistical test that checks a randomness hypothesis for a data sequence. The SDNR tests were conducted only for the CPUE indices.

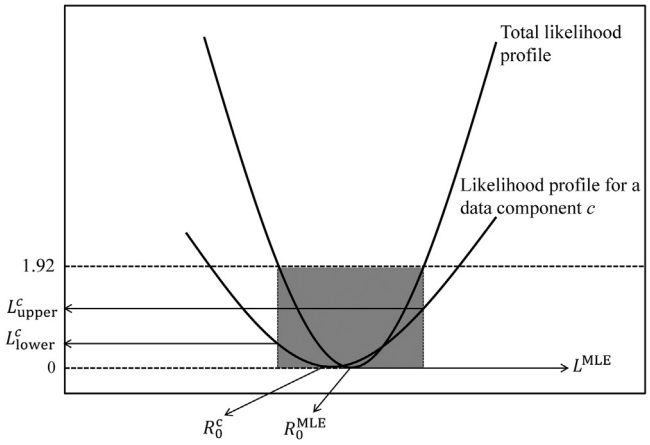


**Fig. 4.** Calculation of the $R_0$ component likelihood profile statistic. The grey area represents the 95% confidence interval for $R_0$.

A 10-year retrospective analysis was conducted on all assessment models by sequentially eliminating one year of data each time (model runs with fewer data are referred to as "peels"). Mohn's "$\rho$" was calculated for spawning stock biomass using the formulation proposed by Hurtado-Ferro et al. (2014).

$$\rho = \left(\frac{\overline{X_{Y-y,p} - X_{Y-y,\mathrm{ref}}}}{X_{Y-y,\mathrm{ref}}}\right)$$

where $X$ is the quantity for which Mohn's $\rho$ is being calculated, $Y$ the final year of the assessment period, $y$ the last year of a given "peel" $p$, and *ref* the reference peel, which is the most recent assessment. Note that this formulation is slightly different from that given by Mohn (1999), where instead of summing across peels, these are averaged.

The performance of the $R_0$ likelihood component profile diagnostic in detecting model misspecification was examined by calculating the $\phi$ statistic developed by Wang et al. (2014). We also calculated the number of simulations in which the estimates of $R_0$ from each data component occurred within the 95% confidence interval of the $R_0^{\mathrm{MLE}}$ (Fig. 4).

$$\phi = \begin{cases} \max\left[\left(L^c_{\mathrm{lower}} - L^{\mathrm{MLE,c}}\right), \left(L^c_{\mathrm{upper}} - L^{\mathrm{MLE,c}}\right)\right], & \text{if} R^c_0 \text{ is located within the95\% CI for} R^{\mathrm{MLE}}_0 \\ |L^c_{\mathrm{lower}} - L^c_{\mathrm{upper}}|, & \text{otherwise} \end{cases}$$

where $L^c_{\mathrm{lower}}$ and $L^c_{\mathrm{upper}}$ are the negative log-likelihoods for data component $c$ corresponding to the lower and upper boundaries of the 95% confidence interval for $R_0$, and $L^{\mathrm{MLE,c}}$ is the negative logarithm of the likelihood for data component $c$ corresponding to $R_0^{\mathrm{MLE}}$. The 95% confidence interval of the $R_0^{\mathrm{MLE}}$ was calculated based on likelihood-ratio test (i.e., 1.92 log-likelihood units from $L^{\mathrm{MLE}}$). A low value of $\phi$ for a data or penalty component indicates that it has a relatively small contribution to the estimation of $R_0$. If the estimate of $R_0$ for data component $c$ falls outside the 95% confidence interval of $R_0$based on the total likelihood, it might indicate conflict in the data or model misspecification. Two simulations are used to illustrate this method (Fig. 5). The examples illustrate the profiles of $R_0$ based on the total likelihood, the likelihoods for each data component, and for the penalty on the recruitment deviates. The first example (left panel) is simulation from the CSM scenario. In this example, the best fit values of $R_0$ for the abundance index data and the size-composition data were close to the maximum likelihood estimate of $R_0$ (890). This example illustrates how the changes in gradients of the likelihood profile for the abundance

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

**Table 2**
Median and 95% confidence intervals for $SSB_{term}$ and $SSB_{term}/SSB_{init}$ for the five scenarios for WCNP striped marlin. MRE and MARE performance metrics and the percentage of misspecified models identified.

| | $SSB_{term}$ in mt Median (95% CI) | $SSB_{term}/SSB_{init}$ Median (95% CI) | MRE (%) | MARE (%) | Misspecified (%) |
|---|---|---|---|---|---|
| Scenario | | | | | |
| CSM | 2181 (1791–2389) | 0.32 (0.27–0.35) | 0.4 | 4.3 | 5 |
| EM_1 | 1926 (1450–2481) | 0.28 (0.21–0.37) | −12.7 | 13.5 | 65 |
| EM_2 | 1840 (1407–2394) | 0.27 (0.21–0.35) | −16.5 | 16.3 | 74 |
| EM_3 | 2229 (1753–2719) | 0.33 (0.26–0.40) | 3.9 | 11.9 | 51 |
| EM_4 | 2184 (1738–2451) | 0.32 (0.26–0.36) | 0.5 | 5.3 | 7 |

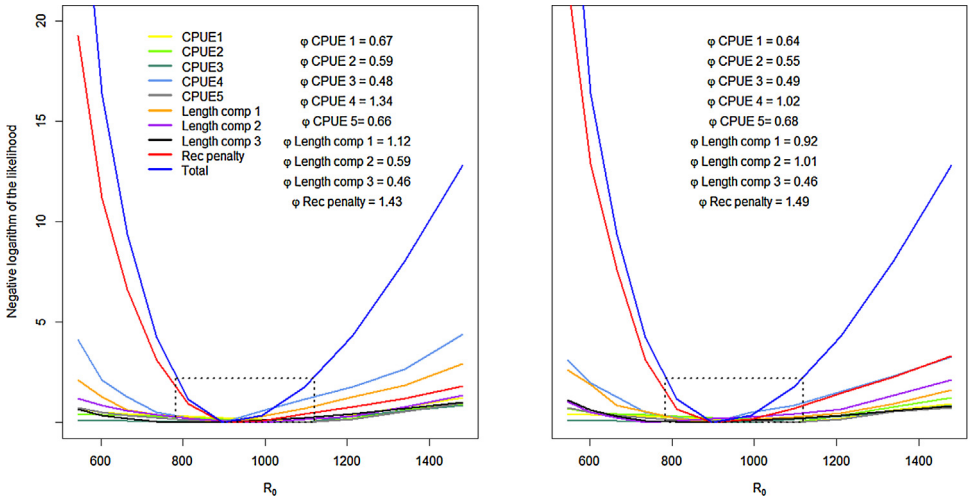


**Fig. 5.** Sample results showing $\varphi$ statistics and likelihood profiles for $R_0$ based on all data and on various data components for the CSM and EM_1, left and right panels respectively. The dashed rectangle represents the 95% confidence interval for $R_0$.

index data were similar over the range of $R_0$ when compared with those from the size-composition data. Also note that all the MLEs of $R_0$ based on the likelihood profile for each data component occurs within the 95% confidence interval of $R_0$ based on total likelihood. The second example (right panel) is one of the simulations for EM_1, and it shows the effects on the gradients of the likelihood profiles when selectivity is misspecified. Note how the recruitment deviates become more influential, while the influence of the index CPUE 4 and size-composition data from fleet 1 decrease. In this example, the MLEs of $R_0$ based on the size-composition data from fleet 2 lie outside the 95% confidence interval of $R_0$ based on total likelihood.

Originally, the $\phi$ statistic was designed specifically to identify data sets that are influential because they either have a likelihood component profile that has a steep gradient or they supported estimates of $R_0$ that are very different from those supported by the other data. However, Wang et al. (2014) did not provide criteria for the $\phi$ statistic that can be used to determine if a model is misspecified. In this study, we extend the use of the $\phi$ statistic, and test if it can be used to identify model misspecification. We argue that values greater than 2 might indicate misspecification because the $\phi$ statistic will be 2 if the component likelihood profile is equal to the total likelihood profile and is symmetrical, and greater than 2 if the shape is the same but the MLE for the component likelihood is different from the MLE for the total likelihood.

To perform the ASPM diagnostic test, we had to change the original model parameterization. SS can behave like an ASPM (Methot and Wetzel, 2013) when the parameters of the selectivity curve are fixed at those estimated from the fully integrated model (i.e., CSM, EM_1, EM_2, and EM_3), the annual recruitment deviates are not estimated (fixed at zero so that recruitment follows the stock–recruitment relationship), and the age- and size-composition data are not used for parameter estimation. The results from the ASPM should be similar to those from the fully integrated model if the size- and age-composition data are not informing absolute abundance or the trend in abundance and there is no strong pattern in recruitment. We implemented four ASPM scenarios to evaluate how the model misspecifications introduced in the EM affect the abundance estimates from the ASPM. The values used to fix the selectivity in the ASPM correspond to the bootstrap-specific estimates from the full model with similar parameterization. Selectivity for replicate 1 for the ASPM_CSM was thus set to the values from first replicate of the full CSM assessment. Natural mortality was assumed to be age-based in ASPM_CSM, ASPM_1, ASPM_2, and ASPM_4, and it was constant across ages and equal to 0.38 $yr^{-1}$ in ASPM_3.

The catch-curve analysis (CCA) was also implemented using SS. In this case the abundance index data are not used, natural mortality was age-based, and selectivity was estimated using the size-composition data. The CC model in SS uses catch to determine $F$, which is used with $M$ and other model parameters to estimate the numbers at age turned into length, and then fit to the observed proportions. Both $R_0$ and $R_{devs}$ were estimated, allowing composition data to directly influence the trend in absolute abundance over-time. We implemented one catch-curve analysis scenario (CCA_1) to evaluate how misspecification of selectivity affects the abundance estimates from the catch-curve analysis. In CCA_1, selectivity for all fleets was assumed to be asymptotic, as in EM_1.

## 3. Results

### 3.1. Management quantities and model performance

Before examining the results for the diagnostic tests, it is important to examine how well the CSM and the four EMs estimate

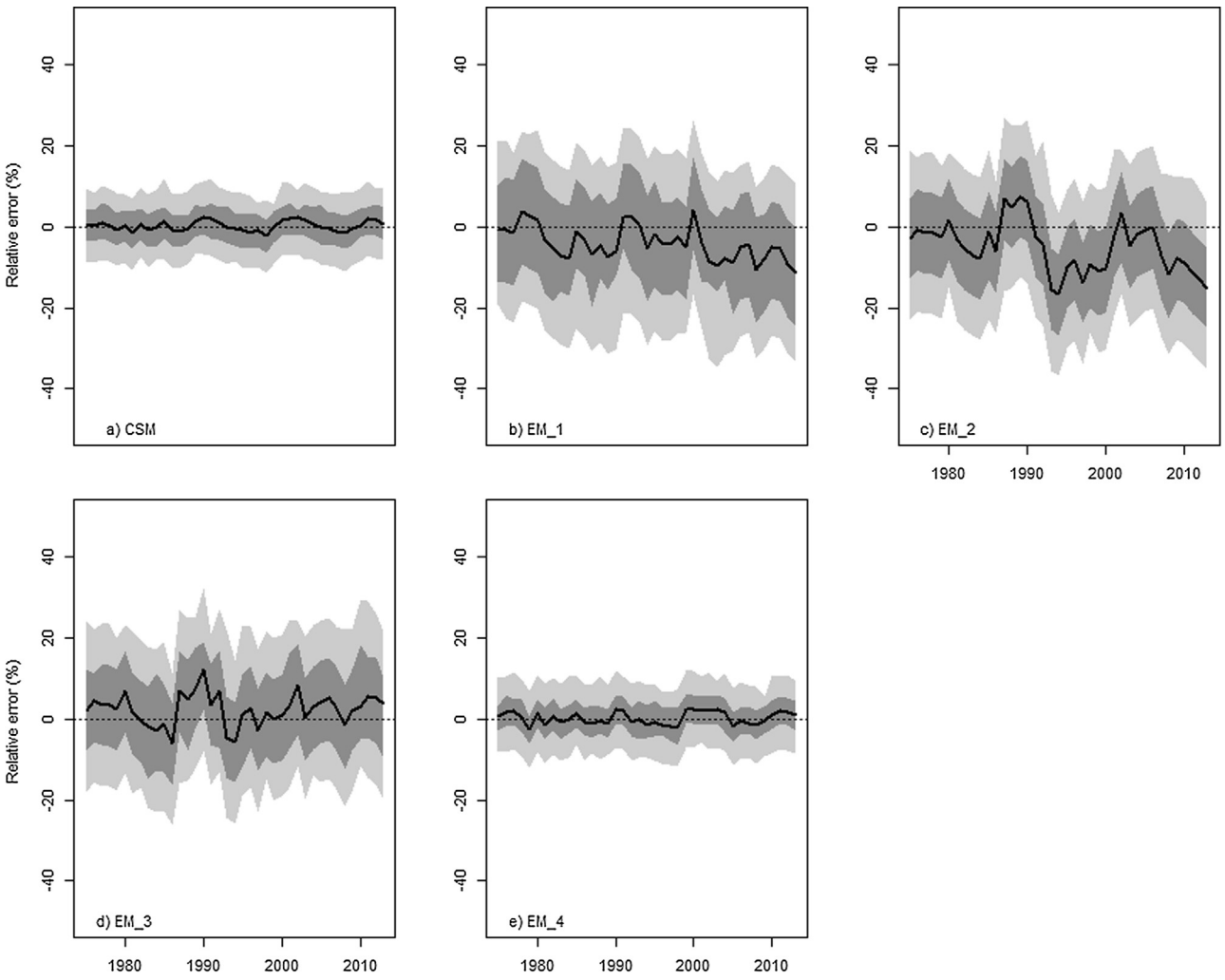


**Fig. 6.** Relative error distributions (median relative errors, and 50% and 90% intervals) for the time-trajectory of spawning stock biomass of WCNP striped marlin.

management quantities. Fig. 6 shows relative error distributions (median relative errors, and 50% and 90% simulation intervals) for the time trajectory of spawning stock biomass. The CSM is essentially unbiased (MRE = 0.4%) and precise (MARE = 4.3%), with some of the estimates slightly larger or smaller than the "true" value (Table 2). However, estimation performance is much poorer for the misspecified EM_1 and EM_2, with larger values for MRE and MARE compared to the CSM. The degradation in performance was associated with an increase in bias; the estimates of $SSB_{term}/SSB_{init}$ from EM_1 constantly underestimated (MRE = −12.7%), as well as from EM_2 (MRE = −16.5%). EM_3 was less biased (MRE = 3.9%) than EM_1 and EM_2. However, the MARE for EM_3 was higher (11.9%) than for the CSM. Over-weighting the size- and age-composition data (EM_4) had little impact on the MRE or MARE for $SSB_{term}/SSB_{init}$. The proportion of misspecified models (i.e., where the estimates of $SSB_{term}/SSB_{init}$ fell outside the 95% CI from the CSM) varied across scenarios. EM_2 had the highest proportion, followed by EM_1 and EM_3, while EM_4 had a much lower proportion than the other models (Table 2).

### 3.2. Diagnostic tests

#### 3.2.1. SDNR

Overall, the SDNR diagnostic test indicated that most of the models fit the CPUE indices adequately (Fig. 7). However, misspecification of selectivity for Fleet 2 (EM_1) resulted in a poor fit of the index for this fleet due to misspecified selectivity (Fig. 7b; CPUE 4), with the median and most of the interquartile range for SDNR for this index greater than 1. Furthermore, under EM_1 the proportion of models with SDNR values above 1 for CPUE 4 was very high (79%) compared to the other models. Overall, the proportion of SDNR values greater than 1 for EM_2 and EM_3 were similar, while for EM_4 the proportion was slightly larger than the CSM. These results do not necessarily indicate an unsatisfactory residual pattern for EM_1. However it is an indication that the introduced misspecifications in selectivity markedly impacted the CPUE fits, hence the SDNR values.

#### 3.2.2. Runs test

The runs test indicated that the residuals for CSM and EM_2, EM_3, and EM_4 for all five CPUE indices, the size-composition data, and the conditional age-at-length data were generally randomly positive and negative over the time series (Table 3). However, the number of simulations for which the residuals for CPUE_4 and the composition data for Fleet 2 were significantly not random was much larger for EM_1 than for the CSM.

#### 3.2.3. Retrospective analysis

Mohn's $\rho$ will be large, either positive or negative, when there is a consistent pattern of change in the peeled assessments relative to the full time series assessment. However, determining whether a given value of Mohn's $\rho$ indicates that an assessment exhibits a

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

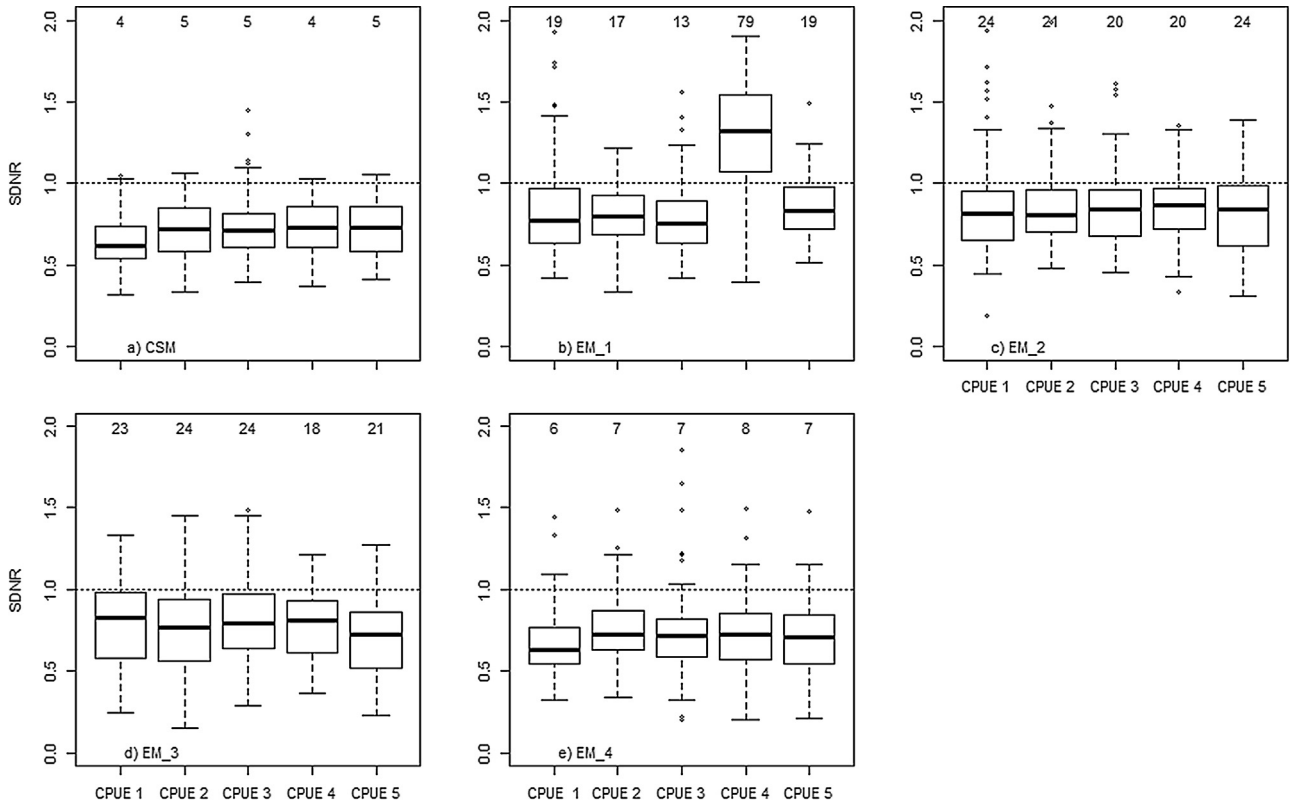


**Fig. 7.** Box-plots of SDNR values for the CPUE indices for the five scenarios. The box shows the interquartile range (IQR). The line inside the box shows the median. The two "whiskers" that extend from each box indicate the range of values that are outside of the IQR, but are close enough not to be considered outliers. Circles represent the outliers, which are observations with a distance of more than 1.5*IQR from the box. Numbers in the top are the proportion of models where SDNR values were above one.

**Table 3**
Percentage of the simulation runs which the hypothesis that the residuals are random can be rejected at $\alpha = 0.05$.

| Data component (Fleet) | CSM | EM_1 | EM_2 | EM_3 | EM_4 |
|---|---|---|---|---|---|
| CPUE 1 (1) | 4 | 3 | 4 | 5 | 5 |
| CPUE 2 (1) | 3 | 4 | 4 | 3 | 6 |
| CPUE 3 (1) | 5 | 4 | 5 | 6 | 4 |
| CPUE 4 (2) | 4 | 51 | 5 | 5 | 6 |
| CPUE 5 (3) | 5 | 3 | 6 | 4 | 5 |
| Size-composition (1) | 4 | 4 | 5 | 6 | 5 |
| Size-composition (2) | 5 | 39 | 9 | 7 | 7 |
| Size-composition (3) | 4 | 7 | 9 | 8 | 6 |
| Conditional age-at-length (1) | 5 | 6 | 7 | 9 | 6 |
| Conditional age-at-length (2) | 6 | 35 | 9 | 8 | 7 |
| Conditional age-at-length (3) | 6 | 8 | 7 | 8 | 9 |

retrospective pattern is subjective. We followed the rule of thumb proposed by Hurtado-Ferro et al. (2014), i.e., values of Mohn's $\rho$ that fall outside the range −0.15 and 0.20 can be interpreted as an indication of a retrospective pattern for long-lived species. Although Mohn's $\rho$ varied across simulations, the small median values of Mohn's $\rho$ for all models suggest negligible retrospective patterns (Fig. 8). CSM and EM_4 had the lowest median value for Mohn's $\rho$ among all scenarios, 0.005 and 0.009, respectively. However, stronger retrospective patterns are indicated for EM_1, EM_2, and EM_3, resulting in a slight increase in the median values of Mohn's $\rho$, as well as its variation compared to the CSM and EM_4. EM_2 had the largest proportion of simulations (15%) where Mohn's $\rho$ fell outside the range −0.15–0.2, followed by EM_1 (12%), and EM_3 (11%). None of the simulations for the CSM and EM_4 resulted in values for Mohn's $\rho$ outside that range. Given the almost near to zero values for Mohn's $\rho$ for most models, the retrospective analysis diagnostic test generally appeared to be unreliable at detecting the model misspecifications we introduced.

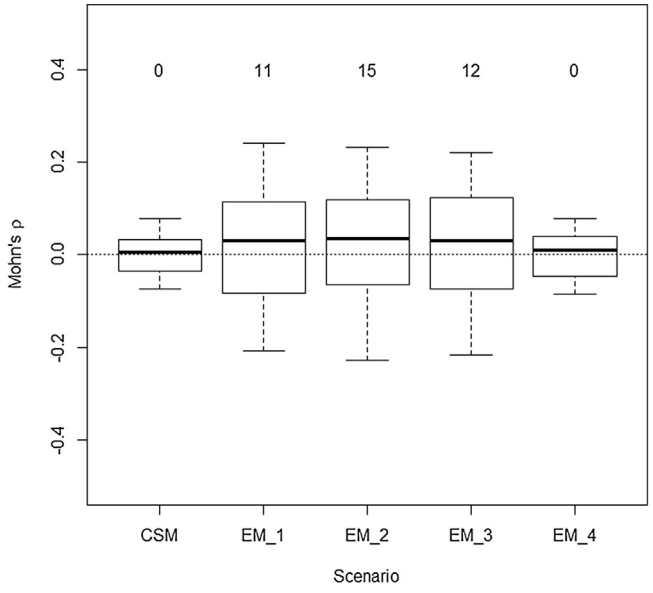


**Fig. 8.** Box-plots of Mohn's $\rho$ values for stock spawning biomass for each of the five scenarios for WCNP striped marlin. Numbers in the top are the proportion of models where Mohn's $\rho$ values fell outside the range −0.15 to 0.2.

### 3.2.4. $R_0$ *likelihood component profile*

Overall, the $R_0$ component likelihood profile statistic for the CSM and misspecified EMs shows that the penalty on the recruitment deviates has the largest influence on the estimation of $R_0$, and that

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

**Table 4**
Mean and standard deviation (over simulations) of the $R_0$ component likelihood profile statistic $\phi$ based on various data components for each of the four EMs. The grey shaded area represents the value for all fleets combined.

| Data component (Fleet) | CSM Mean (± SD) | EM_1 Mean (± SD) | EM_2 Mean (± SD) | EM_3 Mean (± SD) |
|---|---|---|---|---|
| **Index** | 0.75 (0.19) | 0.70 (0.15) | 0.73 (0.17) | 0.72(0.18) |
| CPUE 1 (1) | 0.65 (0.14) | 0.63 (0.11) | 0.61 (0.12) | 0.60 (0.14) |
| CPUE 2 (1) | 0.61 (0.13) | 0.58 (0.10) | 0.60 (0.13) | 0.63 (0.15) |
| CPUE 3 (1) | 0.49 (0.09) | 0.49 (0.11) | 0.52 (0.15) | 0.48 (0.13) |
| CPUE 4 (2) | 1.39 (0.27) | 1.09 (0.31) | 1.10 (0.32) | 1.13 (0.35) |
| CPUE 5 (3) | 0.67 (0.13) | 0.67 (0.15) | 0.66 (0.14) | 0.69 (0.17) |
| **Size-composition data** | 0.70 (0.16) | 0.67 (0.14) | 0.68 (0.16) | 0.67 (0.17) |
| Size-composition (1) | 1.09 (0.22) | 0.95 (0.21) | 1.03 (0.23) | 0.94 (0.25) |
| Size-composition (2) | 0.58 (0.13) | 0.66 (0.15) | 0.55 (0.19) | 0.61 (0.14) |
| Size-composition (3) | 0.46 (0.09) | 0.49 (0.11) | 0.43 (0.13) | 0.48 (0.12) |
| **Rec penalty** | 1.44 (0.29) | 1.49 (0.31) | 1.47 (0.37) | 1.48 (0.34) |

**Table 5**
Percentage of simulation runs in which the estimates of $R_0$ from each data component occurs outside the 95% confidence interval of the $R_0^{MLE}$. The number of simulations with the ϕ statistic value larger than 2 is shown inside the parenthesis. The grey shaded area represents the value for all fleets combined.

| Data component (Fleet) | CSM | EM_1 | EM_2 | EM_3 |
|---|---|---|---|---|
| **Index** | 0.03 (2) | 0.05 (1) | 0.04 (1) | 0.05 (1) |
| CPUE 1 (1) | 0.02 (1) | 0.06 (0) | 0.04 (1) | 0.05 (0) |
| CPUE 2 (1) | 0.04 (1) | 0.05 (0) | 0.03 (0) | 0.02 (0) |
| CPUE 3 (1) | 0.02 (0) | 0.08 (0) | 0.05 (0) | 0.03 (0) |
| CPUE 4 (2) | 0.03 (0) | 0.04 (0) | 0.05 (0) | 0.05 (0) |
| CPUE 5 (3) | 0.04 (0) | 0.02 (1) | 0.03 (0) | 0.05 (1) |
| **Size-composition data** | 0.04 (1) | 0.04 (2) | 0.04 (1) | 0.05 (0) |
| Size-composition (1) | 0.04 (1) | 0.05 (1) | 0.05 (1) | 0.06 (0) |
| Size-composition (2) | 0.06 (0) | 0.03 (0) | 0.05 (0) | 0.04 (0) |
| Size-composition (3) | 0.02 (0) | 0.04 (1) | 0.02 (0) | 0.05 (0) |
| **Rec penalty** | 0 (0) | 0 (0) | 0 (0) | 0 (0) |

the abundance index data are slightly more informative than the size-composition data for estimating $R_0$ (Table 4). However, the influence of the penalty on the recruitment deviates on the estimation of $R_0$ increases slightly while the abundance index and the size-composition data become less influential when the selectivity pattern for Fleet_2 is misspecified (EM_1). The misspecifications introduced to EM_2 and EM_3 also affected the $R_0$ profile statistics for each data component, with the influence of the penalty on the recruitment deviates on the estimation of $R_0$ increasing and that of the abundance index and the size-composition data decreasing.

Although the values of the $R_0$ component likelihood profile statistic changed in our simulations, the order of influence of each data component did not. If the $R_0$ component likelihood profile diagnostic were to correctly identify a misspecified model, we would have expected to see a shift in the order. Although the values of the influence statistic did change, they failed to differ enough to fall outside of the 95% CI range of the $R_0^{MLE}$, indicating that they were not statistically different. Furthermore, the number of simulations where the value of the influence statistic $\phi$ was greater than 2 was very low for all scenarios (Table 5).

### 3.2.5. ASPM

Whether $SSB_{term}/SSB_{init}$ from ASPM for an EM fell outside the (asymptotic) 95% CI of its corresponding fully-integrated model was used as a trigger for the ASPM diagnostic test. For example, we calculated the proportion of simulations under scenario ASPM_1 where the estimates of $SSB_{term}/SSB_{init}$ fell outside 95% CI of this same management quantity from the fully integrated EM_1 scenario. A production relationship was evident in the assessment models, with some ASPM results leading to similar estimates of $SSB_{term}$ and $SSB_{term}/SSB_{init}$ to their correspondent fully integrated model (Table 6). Four percent of the ASPM_CSM and ASPM_4 simulations were identified as misspecified. Only 9% of ASPM_1 simulations were assessed as misspecified. In contrast, 86% and 87% respectively of the ASPM_2 and ASPM_3 simulations were identified as misspecified. ASPM_2 led to much smaller median estimates for $SSB_{term}$ and $SSB_{term}/SSB_{init}$, while ASPM_3 led to much larger median estimates for these management quantities.

**Table 6**
Median and 95% confidence intervals for $SSB_{term}$ and $SSB_{term}/SSB_{init}$ for ASPMs and catch-curve analysis for WCNP striped marlin, and the percentage of misspecified models found.

| | $SSB_{term}$ in mt Median (95% CI) | $SSB_{term}/SSB_{init}$ Median (95% CI) | Identified to be misspecified (%) |
|---|---|---|---|
| Scenario | | | |
| ASPM_CSM | 2241 (1790–2603) | 0.33 (0.27–0.39) | 4 |
| ASPM_1 | 2475 (1912–3125) | 0.36 (0.28–0.43) | 9 |
| ASPM_2 | 1312 (1083–1826) | 0.19 (0.14–0.26) | 86 |
| ASPM_3 | 2675 (2273–3357) | 0.40 (0.36–0.48) | 87 |
| ASPM_4 | 2275 (1802–2659) | 0.33 (0.28–0.39) | 4 |
| CCA | 3259 (2420–4018) | 0.48 (0.34–0.62) | 91 |
| CCA_1 | 2658 (2215–3284) | 0.39 (0.32–0.53) | 92 |

### 3.2.6. CCA

A large proportion (91%) of simulations for the CCA_CSM was identified as misspecified, with much larger estimates of management quantities of $SSB_{term}$ and $SSB_{term}/SSB_{init}$ than the CSM (Table 6). Similar results were also found in CCA_1, with 92% of simulations identified as misspecified.

### 3.2.7. Reliability of diagnostic tests

The power of a diagnostic test to detect model misspecification in an individual model varied by the type of misspecification (Table 7). For diagnostics with multiple components (SDNR and runs test), a misspecification is defined as at least one component failing to pass that diagnostic. The probability of falsely rejecting a correctly specified model was low (less than 6%) for all diagnostic tests except the CCA (91% of false positives). To detect misspecifi-

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

**Table 7**
Percentage of models identified as misspecified by each diagnostic test under different scenarios.

| Diagnostic | CSM (%) | EM_1 (%) | EM_2 (%) | EM_3 (%) | EM_4 (%) |
|---|---|---|---|---|---|
| SDNR | 5 | 79 | 24 | 24 | 6 |
| Runs test | 6 | 51 | 9 | 9 | 9 |
| ASPM | 4 | 9 | 86 | 87 | 4 |
| Retrospective analysis | 0 | 11 | 15 | 12 | 0 |
| $R_o$ Likelihood component profile | 4 | 5 | 4 | 5 | – |
| CCA | 91 | 92 | – | – | – |

**Table 8**
Percentage of models identified as misspecified by at least one, at least two, or at least three, out of the five diagnostic tests under different scenarios. The CCA diagnostic is exclude due to high probability of Type I error.

| | CSM (%) | EM_1 (%) | EM_2 (%) | EM_3 (%) | EM_4 (%) |
|---|---|---|---|---|---|
| At least one diagnostic | 7 | 88 | 90 | 91 | 9 |
| At least two diagnostics | 4 | 43 | 22 | 21 | 6 |
| At least three diagnostics | 4 | 8 | 7 | 8 | 4 |

cation on selectivity pattern, the SDNR showed good power, while the runs test and ASPM showed moderate and low power, respectively. However, the inverse happens when the misspecification was related to the system dynamics ($h$ and $M$), where SDNR and the runs test showed lower power, and ASPM showed good power. The retrospective analysis,$R_0$ likelihood component profile, and CCA had low rates of detection of misspecified models.

When all diagnostic tests are considered together (excluding the CCA test due to its high false positive rates), the power to detect model misspecification improves without a substantial increase in the probability of incorrectly rejecting a correctly specified model (Table 8). When the criteria for rejecting a model as correctly specified is a failure of at least one of the diagnostic tests, nearly 90% of most misspecified models (EM_1, EM_2, EM_3) are detected with no real increase in the probability of a false detection. If the criterion for rejecting is changed to at least 2 diagnostics failing, power drops by more than half. The power to detect incorrectly weighted data (EM_4) was low for all diagnostics used separately or together.

## 4. Discussion

A major goal of model diagnostics is to have high power to reject misspecified models and a low probability of incorrectly rejecting correctly specified models. No individual diagnostic was sufficient to ensure high power of detecting all forms of misspecification tested. However, applying multiple diagnostic tests did increase the power to detect misspecification. This is important because the type of misspecification will not be known in real applications. Perhaps more importantly, the application of the multiple diagnostics tested (excluding the CCA diagnostic) did not dramatically increase the probability of a false detection even when the criterion for deciding there was model misspecification was only one diagnostic of the set being triggered.

There appear to be differences in the efficacy of the various diagnostic tests depending on whether the misspecification is in the observation or systems dynamics model. Residual analyses were easily the best detector of misspecification of the observation model, while the ASPM was the only good diagnostic for misspecification of system dynamics model. Note that in this case selectivity is used for both the observation process (sampling the size-compositions) and the system dynamics (the size of fish removed from the population) because the index is based on fishery CPUE not a survey, but the main influence is assumed to be derived from the observation process. Residual analysis is focused on lack of fit, which in our case was associated with misspecification the selection pattern. Some lack of fit is expected when the selectivity is misspecified because it is very likely that the model predictions would not be able to match the composition data. Using SDNR and runs test together as a diagnostic test provides a way to assess both the magnitude of residual variation as well as the presence of serial correlation of residuals. Although both SDNR and the run tests had small chances of false positives, SDNR seems to be more effective than the runs test in identifying misspecification of selectivity. It is also important to note that many time-series analyses (including for residuals) require relatively long and stationary time-series, and do not accept missing values. While model residuals from integrated models may be stationary, it is rare to encounter time-series of fisheries data that are long and complete, especially for size-composition data. The ASPM appears to detect misspecification of processes related to the underlying production function in the model, such as natural mortality. However, it is important to consider that it may be unlikely that it will have the same detection capability for a stock that does not have an elucidated production function, such as short-lived species and those that have highly-variable recruitment, particularly when fishing intensity is low. A potential benefit of the ASPM diagnostic is that when the changes in the index of abundance can be explained solely by catch and the production function, it may be seen as validating the index itself because a connection between catch and an abundance index would be unlikely to occur at random.

None of the diagnostic tests that were applicable to EM_4 had a high power to detect the incorrect weighting. The incorrect data weight had little effect on model results besides increasing the precision assumed for some data components because all the model processes were correctly specified. The resulting estimated population dynamics were also quite similar to the correctly specified model, with the differences between those two models caused by the random noise generated in the bootstrapping procedure. Thus, it is not surprising that our diagnostics failed to detect this misspecification. Although not included in this study, SDNR diagnostic test on the composition data for which the data weighting was changed may have been more powerful in detecting the misspecified weighting. Data weighting misspecification is expected to be more influential when combined with system dynamics or observation model misspecification and may be easier to detect by the diagnostic tests. However, this was not investigated in our analysis. The insensitivity of estimates to data weighting when a model is correctly specified and the sensitivity when the model is misspecified implies that data weighting sensitivity could be used to identify model misspecification. Further investigation into diagnostic tests based on data weighting or modification of the $R_0$ component likelihood profile diagnostic might be fruitful.

Several diagnostic methods tested performed poorly. The $R_0$ component likelihood profile diagnostic as proposed by Wang et al.

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

(2014) had very little power to detect any form of misspecification. The statistic was designed to identify a data set that is problematic in the sense that it either conflicted with the other data (the MLE from the likelihood component is different from the MLE from the total likelihood) or had little information (the confidence interval from the component likelihood is wide). A different statistic might be more appropriate and further research is needed. However, the use of profiling may be better when used in the manner described by Lee et al. (2014), which allows users to quantify the contribution of each data component to the estimates of absolute population scale. Retrospective analysis also performed poorly, even though this is one of the most used diagnostics. One of the potential causes for such poor performance is the choice of the misspecifications introduced in our simulations, as some types of misspecifications do not produce retrospective patterns (Legault, 2009). For example, in our study we did not account for time-varying biological parameters and selectivity, which according to Hurtado-Ferro et al. (2014) is one of the main causes of high variability in the magnitude of Mohn's $\rho$ statistic. This paper introduced the CCA diagnostic, but because of its high probability of rejecting correctly specified models needs further development and testing before it can be recommended for use.

Although simulation studies are very useful for testing analytical approaches, several caveats need to be mentioned. Our simulations were based on a single stock and its associated data. It is unknown if results would be similar for stocks with different life history strategies, data types, or different responses to fishing. In addition, simulation studies use highly simplified systems and real world examples will almost certainly deal with much more complexity in both fishery (e.g., time varying selection) and biological (e.g., spatial dynamics) structure. A key simplification was controlling for only a single misspecification, while real world examples will likely contain multiple misspecified processes. Finally, we did not test all diagnostics and many other potential candidates most certainly exist (e.g., Piner et al., 2011; Besbeas and Morgan 2014).

## 5. Conclusions

Despite the limitations to the study, it is clear that multiple diagnostic tests need to be applied when evaluating model reliability. Some diagnostic tests may be better at detecting misspecification in observation model processes and others at finding misspecification of the systems dynamics model. Researchers should carefully consider the suite of diagnostics to be used to insure that all relevant model processes can be explored. The ASPM test (Maunder and Piner, 2015) appears to have promise in detecting systems dynamic misspecification, residual analysis in detecting observation model misspecification, and retrospective analysis in detecting unmodeled temporal variation (Hurtado-Ferro et al., 2014), and these diagnostics should be applied routinely. Other diagnostics presented here and elsewhere should also be considered, but require further development. Because it is likely that the properties of each of the diagnostic will change with differences between dynamic models and the systems they represent, further research into model diagnostics is warranted.

## Acknowledgements

Part of this research was carried out while Felipe Carvalho was a visiting scientist at the Center for the Advancement of Population Assessment Methodology (CAPAM) under funding from NOAA. The authors would like to thank the members of the billfish working group of the International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean (ISC) for their contributions to the 2015 striped marlin stock assessment. In addition, we thank reviewers, both known and anonymous, for their helpful critique and suggestions on improving the manuscript. This research addresses the data weighting component of the good practices guide to stock assessment program of CAPAM. This work was partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative agreement N°. NA10OAR4320148.

## References

Besbeas, P., Morgan, B.J.T., 2014. Goodness-of-fit of integrated population models using calibrated simulation. Methods Ecol. Evol. 5, 1373–1382.
Breen, P.A., Kim, S.W., Andrew, N.L., 2003. A length-based Bayesian stock assessment model for the New Zealand abalone *Haliotis iris*. Mar. Freshw. Res. 54 (5), 619–634.
Cadigan, N.G., Farrell, P.J., 2005. Local influence diagnostics for the retrospective problem in sequential population analysis. ICES J. Mar. Sci. 62, 256–265.
Chang, Y.J., Langseth, B., Yau, A., Brodziak, J., 2015. Stock Assessment Update for Striped Marlin (Kajikia audax) in the Western and Central North Pacific Ocean through 2013, ISC/15/BILLWG-2/01.
Cox, D.R., Snell, E.J., 1968. A general definition of residuals. J. R. Stat. Soc. Ser. B 30, 248–275.
Deroba, J.J., Schueller, A.M., 2013. Performance of stock assessments with misspecified age- and time-varying natural mortality. Fish. Res. 146, 27–40.
Deroba, J.J., 2014. Evaluating the consequences of adjusting fish stock assessment estimates of biomass for retrospective patterns using Mohn's Rho. N. Am. J. Fish. Manage. 34, 380–390.
Doubleday, W.G., 1976. A least squares approach to analyzing catch at age data. Res. Bull. Int. Comm. Northwest Atl. Fish. 12, 69–81.
Fournier, D., Archibald, C.P., 1982. A general-theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39, 1195–1207.
Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138.
Gibbons, J.D., Chakraborti, S., 1992. Nonparametric Statistical Inference, third edition, Marcel Dekker, Inc., New York.
Harle, S., Davies, N., Hampton, J., McKechnie, S., 2015. Stock assessment of bigeye tuna in the western and central Pacific ocean. WCPFC-SC10, https://www.wcpfc.int/node/18975.
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., 2014. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES J. Mar. Sci. 72, 99–110.
Ichinokawaa, 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.
Kell, L.T., De Bruyn, P., Maunder, M.N., Piner, K.R., Taylor, I.G., 2014. Likelihood component profiling as a data exploratory tool for north Atlantic albacore. Collect. Vol. Sci. Pap. ICCAT 70 (3), 1288–1293.
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.
Legault, C.M., 2009. Report of the Retrospective Working Group, January 14–16, 2008, Woods Hole, Mass. NEFSC Reference Doc. 09-01.
Maunder, M.N., Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72 (1), 7–18.
Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74.
Maunder, M.N., Starr, P.J., 2001. Bayesian aassessment of the SNA1 snapper (*Pagrus auratus*) stock on the northeast coast of New Zealand. N. Z. J. Mar. Freshw. Res. 35, 87–110.
Maunder, M.N., 1998. Integration of Tagging and Population Dynamics Models in Fisheries Stock Assessment. PhD Thesis. University of Washington.
Methot, R.D., Wetzel, C., 2013. Stock Synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.
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.
Ono, K., Licandeo, R.R., Muradian, M.L., Cunningham, C.J., Anderson, S.C., Hurtado-Ferro, F., Johnson, K., McGilliard, C., Monnahan, C.C., Szuwalski, C.S., Valero, J.L., Vert-Pre, K.A., Whitten, A.R., Punt, A.E., 2015. The importance of length and age composition data in statistical catch-at-age model for marine species. ICES J. Mar Sci. 72, 31–43.
Piner, K.R., Lee, H.H., Maunder, M.N., Methot, R.D., 2011. A simulation-based method to determine model misspecificaton: examples using natural mortality and population dynamics models. Mar. Coast. Fish. 3, 336–343.
Piner, K., Lee, H.H., Kimoto, A., Taylor, I., Kanaiwa, M., Sun, C.L., 2013. Population dynamics and status of striped marlin (*Kajikia audax*) in the western and central northern Pacific Ocean. Mar. Freshw. Res. 64, 108–118.
Punt, A.E., Huang, T.-C., Maunder, M.N., 2013. Review of integrated size-structured models for stock assessment of hard-to-age crustacean and mollusc species. ICES J. Mar. Sci. 70, 16–33.

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018

Punt, A.E., 2015. Some insights into data weighting in integrated stock assessments. Fish. Res., http://dx.doi.org/10.1016/j.fishres.2015.12.006.
SEDAR 40, 2015. Atlantic Menhaden Stock Assessment Report. SEDAR, North Charleston, SC643.
Taylor, I.G., Methot, R.D., 2013. Hiding or dead? A computationally efficient model of 830 selective fisheries mortality. Fish. Res. 142, 75–85.
Taylor, I.G., Stewart, I.J., Hicks, A., Garrison, T.M., Punt, A.E., Wallace, J.R., Wetzel, C.R., 2011. r4ss: R code for Stock Synthesis. R package version 1.16. http://R-Forge.R-project.org/projects/r4ss/.
Trapletti, A., 2011. tseries: Time series analysis and computational finance. R package version 0. 10-25. http://CRAN.R-project.org/package=tseries.
Wang, S.P., Chen, Y.R., Maunder, N.M., Nishida, T., 2009. Preliminary application of an age–structured assessment model to swordfish (Xiphias gladius) in the Indian Ocean. IOTC-WPB-2009-11, http://www.iotc.org/sites/default/files/documents/proceedings/2009/wpb/IOTC-2009-WPB-11.pdf.
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.
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.
Wetzel, C.R., Punt, A.E., 2011. Performance of a fisheries catch-at-age model (stock synthesis) in data-limited situations. Mar. Freshw. Res. 62, 927–936.

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018