ICES Journal of Marine Science

ICES Journal of Marine Science (2021), 78(6), 2244–2255. https://doi.org/10.1093/icesjms/fsab104

Original Article

Validation of stock assessment methods: is it me or my model talking?

Laurence T. Kell [1,*], Rishi Sharma[2], Toshihide Kitakado[3], Henning Winker[4], Iago Mosqueira [5], Massimiliano Cardinale [6], and Dan Fu[7]

[1]*Centre for Environmental Policy, Imperial College London, Weeks Building, 16-18 Princes Gardens, London SW7 1NE, UK*
[2]*Food and Agricultural Organization, Fishery and Aquaculture Policy and Resources Division, Rome, Lazio 00153, Italy*
[3]*Department of Marine Biosciences, Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato, Tokyo 108-8477, Japan*
[4]*Joint Research Centre (JRC), European Commission, TP 051, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy*
[5]*Wageningen Marine Research, Haringkade 1, 1976CP IJmuiden, the Netherlands*
[6]*Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, SE-453 30 Lysekil, Sweden*
[7]*Indian Ocean Tuna Commission, Le Chantier Mall, Po Box 1011, Victoria, Seychelles*

[*]*Corresponding author: tel: +44 7517071190; e-mail: laurie@kell.es*

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

Received 22 July 2020; revised 10 March 2021; accepted 7 May 2021; advance access publication 11 June 2021.

The adoption of the Precautionary Approach requires providing advice that is robust to uncertainty. Therefore, when conducting stock assessment alternative, model structures and data sets are commonly considered. The primary diagnostics used to compare models are to examine residuals patterns to check goodness-of-fit and to conduct retrospective analysis to check the stability of estimates. However, residual patterns can be removed by adding more parameters than justified by the data, and retrospective patterns removed by ignoring the data. Therefore, neither alone can be used for validation, which requires assessing whether it is plausible that a system identical to the model generated the data. Therefore, we use hindcasting to estimate prediction skill, a measure of the accuracy of a predicted value unknown by the model relative to its observed value, to explore model misspecification and data conflicts. We compare alternative model structures based on integrated statistical and Bayesian state-space biomass dynamic models using, as an example, Indian Ocean yellowfin tuna. Validation is not a binary process (i.e. pass or fail) but a continuum; therefore, we discuss the use of prediction skill to identify alternative hypotheses, weight ensemble models and agree on reference sets of operating models when conducting Management Strategy Evaluation.

Keywords: diagnostics, hindcast, prediction skill, retrospective analysis, stock assessment, validation

Introduction

Fisheries management relies upon stock assessment to provide advice. There are various definitions of stock assessment (e.g. Hilborn, 2003; Cadrin and Dickey-Collas, 2014), and our preference is for "the description of the characteristics of a 'stock' so that its biological reaction to being exploited can be rationally predicted and the predictions tested" (Sidney Holt, pers. comm.). The reasoning for this is because it explicitly recognizes that the main aim of a stock assessment is to provide the basis for long-term sustainable management. Stock assessment, therefore, requires making and validating probabilistic estimates of stock status and forecasts of the consequences of different management actions.

The adoption of the Precautionary Approach to fisheries management (PA; FAO, 1996) requires a formal consideration of uncertainty, which is increasingly being addressed by conducting stock assessment using alternative modelling frameworks conditioned on a variety of assumptions and data sets. This requires practices for re-

ducing subjectivity when deciding whether to accept an assessment (Punt *et al.*, 2020). Current literature on the comparison of stock assessment methods primarily focuses on how well models fit observational data (e.g. Deroba *et al.*, 2015), and diagnostic tests for explaining bias in model estimates of parameters and derived quantities (Carvalho *et al.*, 2021).

It can be challenging, however, to use traditional diagnostics based on model residuals and likelihoods such as Akaike's Information Criteria (AIC; Akaike, 1998) to compare models. For example, indices of abundance are a primary contributor to the overall likelihood when fitting stock assessment models to data (Whitten *et al.*, 2013), and the sum of squared errors (SSE) between observed and predicted indices in the log-space is often used as a fitness measure. SSE is problematic because complex models tend to have many parameters to allow flexibility, resulting in a low SSE due to overfitting by adding more parameters than can be justified by the data. Therefore, criteria such as AIC have been developed to aid in model selection. However, AIC needs to be performed on models with the same likelihood function and data, which is not the case if different hypotheses are modelled with alternative model structures and data sets.

Historical performance is also no indicator of how well a model may perform in the future, which needs to be evaluated if a model is to provide credible and robust advice. This is of particular importance for stock assessment models where the quantities of interest (i.e. fishing mortality and spawning stock biomass) are not directly observable unlike in weather forecasting and there is often insufficient data to allow some of it to be kept back for testing as in machine learning applications. A diagnostic tool to check the potential future stability of stock assessment models is retrospective analysis (Mohn, 1999). The procedure involves sequentially removing all data from the most recent period (i.e. peeling), refitting the model, and then comparing terminal year estimates of spawning stock biomass (SSB) and fishing mortality (F) to the full model. Retrospective analysis is widely used to evaluate the stability of model outputs, and in Europe is often the key diagnostic for accepting or rejecting a model (ICES, 2019). Retrospective analysis has been extended to include stock forecasts, where the terminal year estimates are projected for assumptions about future catches, recruitment, biological parameters, and the vulnerability of the stock to fishing (e.g. Brooks and Legault, 2016). However, stability and a reduction in variance can be achieved at the expense of bias by shrinking terminal estimates towards recent historical values. It is impossible to validate a model if bias is unknown, as is the case for unobservable quantities, such as SSB and F (Hodges and Dewar, 1992); since in such cases, the simplest way to remove a retrospective pattern is to ignore the data.

An alternative approach is to compare model estimates to observations. This is commonly used in many fields when known or closely estimated values for past events are used to evaluate how well model outputs match known results (Balmaseda *et al.*, 1995; Jin *et al.*, 2008; Weigel *et al.*, 2008). The comparison of model outputs to observations not used in fitting is referred to as "predictive validation" or "cross-validation", and when the observations are peeled back from the terminal year, this is known as "hindcasting". Removing observations allows models to be compared using prediction skill (Glickman and Zenk, 2000), a measure of a predictor's accuracy compared to its observed value unknown by the model, using metrics such as correlation, relative error, mean absolute scaled error (MASE), and bias.

Model validation increases confidence in the outputs of a model, leads to an increase in trust amongst the public, stake and assetholders and policymakers (Saltelli *et al.*, 2020), and can identify model limitations that should be addressed in future research. In this paper, we validate models using prediction skill by peeling back observations from the final year in the assessment and making predictions of the removed values using a hindcast procedure. We are not proposing the hindcast and prediction skill as the only diagnostic tool used in stock assessment but as a key tool for the assessment toolbox (Carvalho *et al.*, 2021). The hindcast procedure can be applied to many fields, e.g. climate and energy modelling (Kell *et al.*, 2020).

Material and methods

As a worked example, we compare three model families used to assess Indian Ocean yellowfin tuna stock (IOTC, 2019), namely a full integrated statistical model (SS; Methot and Wetzel, 2013), a deterministic age-structured production model (ASPM; Maunder and Piner, 2015), and a Bayesian state-space biomass dynamic model (JABBA; Winker *et al.*, 2018). Both the SS and ASPM models were based on a seasonal structure with four regions, JABBA in comparison, had an annual time step with no spatial structure.

After a model structure is agreed upon, it is crucial to validate the model to assesses whether it is plausible that a system identical to the model generated the data (Thygesen *et al.*, 2017). The ambition of validation is not to prove that a model is correct, but to check that it can not be falsified with the available data. This a different question from asking if the model is fit for a given purpose, which depends on the model's intended use. For example, to evaluate whether an assessment model is robust despite being misspecified, Management Strategy Evaluation (MSE; Punt and Donovan, 2007) can be conducted. See Sharma *et al.*, (2020) for a review of current practice in the Tuna Regional Fisheries Management Organisations (tRFMOs). Validation is not a binary process, i.e. identifying whether a model is valid or invalid, as there is a continuum between these two extremes. Therefore, a primary objective of validation is not to select a "best assessment" but to identify if models are overfitted and how they can be extended or modified to better describe the dynamics.

Model validation, therefore, serves a complimentary purpose to model selection and hypothesis testing. Model selection searches for the most suitable model within a specified family; hypothesis testing examines how to reduce the model structure, while model validation examines if it should be modified or extended. For models to be valid, they must satisfy four prerequisites (Hodges and Dewar, 1992). Namely, the situation modelled must: (i) be observable and measurable; (ii) be possible to collect sufficient informative data about it; (iii) exhibit constancy of structure in time, and (iv) exhibit constancy across variations in conditions not specified in the model.

The first two prerequisites should be straight forward; however, many stock assessments, particularly for highly migratory stocks like yellowfin tuna fished in areas beyond national jurisdiction, rely on fishery-dependent data rather than direct scientific observations. The use of fishery-dependent data is a concern since there is evidence that commercial catch per unit effort (CPUE) is likely to remain high while abundance declines (Harley

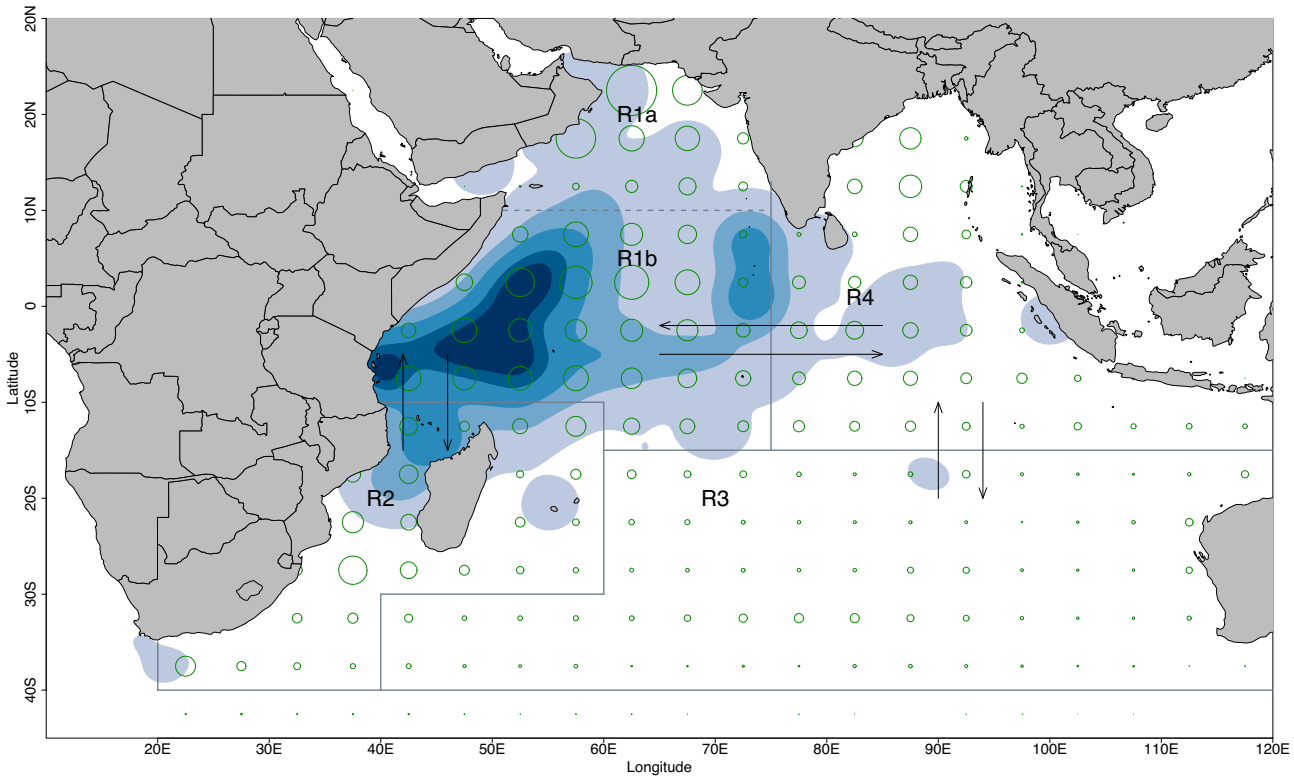

Figure 1. Spatial stratification of the Indian Ocean for the four region assessment model (R1a and R1b were treated as a single model region R1, but were retained for the fleet definition). The black arrows represent the configuration of the movement parametrization. Density contours represent of the dispersal of tag releases and subsequent recaptures from Indian Ocean Regional tuna tagging programme. Green circles represent the distribution of catches from the longline fishery aggregated by 5° longitude and latitude for 1980–2017 (max. = 133770 t).

et al., 2001). Prerequisite (iii) ensures that the model has prediction skill for the same conditions under which the validation tests were conducted. Prerequisite (iv) ensures that the model will still be valid under conditions that differ from those in the validation tests.

Material

Yellowfin tuna supports one of the largest tuna fisheries in the Indian Ocean, with catches currently exceeding 400000 t annually. The stock is harvested by various gears, from small-scale artisanal fisheries to large gill netters, industrial longliners, and purse seiners (Fiorellato *et al.*, 2019). There are regional differences in the stock and fisheries (Figure 1); and the western tropical region (Region 1) is considered the core area of the stock's distribution.

The majority of data available for assessing the stock is fishery-dependent. These include time series of the total catch, seasonal CPUE based on the long-line fisheries (Hoyle and Langley, 2020), samples of length compositions, tagging recaptures, and environmental data. CPUE are the primary source of information on abundance and are based on a composite long-line index, spatially stratified by region, from the main distant water fleets.

Indices in each region are standardized using generalized linear models that accounted for differences in targeting practices and catchability amongst fleets, based on gear configurations and species composition (Hoyle and Langley, 2020). The reason for this is because tuna long-line fishing strategies have changed over time. In the assessment, the CPUE indices across regions were linked by a common catchability coefficient, thus improving the model's ability to estimate regional biomass distribution. This required the calculation of arbitrary regional scaling factors related to a reference fleet's catch rates.

The length composition data are considered sufficient to provide reasonable estimates of fishery selectivity and recruitment trends but not stock abundance trends. Regional environmental indices (current and sea temperature) allow seasonal and temporal variations to be incorporated in fish movement estimation. Tag release and recovery data collected from the main phase of the Indian Ocean large-scale tuna tagging programme inform estimates of mortality, abundance, and movement.

Assessment models

Model development has focused on the spatial structure to account for differences in regional exploitation patterns; and non-stationarity in selectivity and catchability and seasonal movements have been found to resolve data conflicts (Urtizberea *et al.*, 2019). Although a fully integrated statistical model is used to develop the base case, other models are also used. These include an ASPM-R (Maunder and Piner, 2015) and a Bayesian state-space biomass dynamic model (JABBA; Winker *et al.*, 2018).

Stock Synthesis (SS; Methot and Wetzel, 2013) is used to conduct the base case assessment and implements an age and spatially structured model that reflects the complex population and fishery dynamics of the stock. The most recent assessment established a base case as a reference model for diagnostics and sce-

narios to capture various uncertainties (Fu *et al.*, 2018). The assessment indicates that the stock has declined substantially since 2012; and SSB in 2017 is now estimated to be close to the historical lowest level. The stock is estimated to be overfished; and the IOTC has implemented a rebuilding plan to reduce overall fishing pressure.

SS provides a flexible framework for conducting stock assessment and Maunder and Piner (2015) proposed a deterministic implementation in SS of an ASPM as a diagnostic of processes that control the expected dynamics through a production function (Carvalho *et al.*, 2017). Selectivity in ASPM is parametrized based on that estimated by a "full" SS model. The model is then refitted to the abundance indices without the size composition contributing to the likelihood. Recruitment deviations can either be estimated (as in our example) or set to zero. This enables an evaluation of whether the observed catches alone can not explain trends in the index of abundance. If the ASPM can fit the indices of abundance well, then a production function is likely to exist (i.e. the dynamics are driven by density-dependent processes), and the indices provide information about absolute abundance. If the fit is poor, then the expected surplus production and observed catches alone can not explain the indices' trends. This can have several causes, namely the (i) stock dynamics are recruitment-driven, (ii) stock has not yet declined to the point at which catch is a major factor influencing abundance; (iii) indices of relative abundance are not proportional to abundance; (iv) model is incorrectly specified, or (v) data are biased. While a production function was evident in the fit, the overall fit to the indices of abundance in 3 of the 4 areas was poor, and hence, we used recruitment deviates to help capture the trends in abundance by area (see Minte-Vera *et al.*, 2017). In this study, we implemented ASPM with estimated recruitment deviates (ASPM-R).

An alternative to an integrated assessment is to use a biomass dynamic model based on an explicit production function. This requires estimation and fixing of fewer parameters and does not use the length composition. We used the R package JABBA as it provides a unifying, flexible framework for state-space biomass dynamic modelling, runs quickly, and generates reproducible stock status estimates (Winker *et al.*, 2018). A Pella Tomlinson production function (Pella and Tomlinson, 1969) was assumed as this allows the shape of the production function to be varied. Allowing alternative assumptions about productivity, stock status, and reference points to be evaluated. JABBA does not account for spatial dynamics, and in this analysis, priors of production function parameters were based on the SS base case.

Assessment hypotheses

The base case is spatially disaggregated into two tropical regions (R1 and R4) and two austral subtropical regions (R2 and R3). The tropics encompass the main year-round fisheries, while the long-line fisheries occur more seasonally in the austral regions (Langley, 2015), reciprocal movement is assumed to occur between adjacent regions. The base case assumes a quarterly time step to approximate the continuous recruitment and rapid growth seen in the yellowfin stock. The population comprised 28 quarterly age-classes with an assumed unexploited equilibrium initial state in each region. Twenty-five fisheries are defined based on fishing gear, region, time period, fishing mode, and vessel type. Fisheries were modelled, allowing flexibility in selectivity (e.g. cubic spline or double normal), whereas long-line selectivity was constrained to be fully selective for the older ages.

Recruitment occurs in the two equatorial regions with temporal deviates in the regional distribution and is assumed to follow a Beverton and Holt stock-recruitment relationship. Growth is parametrized using age-specific deviates on the k growth parameter to mimic the non-von Bertalanffy growth of juvenile and adults' near-linear growth. Natural mortality varies by age, with the relative trend in age-specific natural mortality based on Pacific Ocean yellowfin (Maunder and Aires-da Silva, 2012).

Hindcast

Validation requires that the system be observable and measurable. So observations should be used unless model estimates are known to be very close to their true values. For example, when conducting a retrospective analysis, a reduction in mean squared error (a measure of variance) of model estimates can be achieved by shrinkage. However, the bias is difficult to quantify in model-based quantities, and therefore, the absence of retrospective patterns while reassuring is not sufficient for validation. For this reason, validation should be conducted using prediction skill based on observations. Therefore, we used a hindcast procedure where the indices of abundance are sequentially removed from the terminal year, i.e. peeled backwards from the model. In contrast, in a retrospective analysis, all observations for a year are peeled back, which means that quantities can not be predicted for the years peeled back unless additional assumptions are made.

The hindcast is a variant of cross-validation where, like retrospective analysis, recent data are removed, and the model refitted with the remaining data. Known values (observations) or well estimated historical values are then compared to model estimates. When observations are used for comparison, this is also referred to as model-free validation (Kell *et al.*, 2016). In a hindcast, observations are removed from the terminal year and up to n years back, and then the missing observations are predicted by fitting to the remaining data for 1, 2, ... n steps ahead. Observations may be removed by series or fleets to evaluate data conflicts, time blocks to overcome serial correlations, or individually to estimate bias as in the jackknife. No stock forecast or projection needs to be performed, and so there is no need to make assumptions about future parameters as all parameters needed are estimated within the model. The hindcast may be conducted for individual data series or combinations of series and data types, for example, by fleet where both CPUE and length data are removed. This allows data conflicts to be explored. Theoretically, the projection period is to the end of the historical time period (Brooks and Legault, 2016); However, in practice, a step size of one or several years ahead (the horizon h) is chosen for the hindcast when removing observations removed from the model fit. This should reflect the time horizon required for robust management advice, considering typical process stochasticity in fishery population dynamics and observation uncertainty. Assessment cycles are typically for three years in most tuna Regional Fisheries Management Organisations and so a horizon of three years was also used.

In this study, only CPUE observations were removed, catch and length composition remained in the model. Thus, all model fits had the same terminal year and differed only in the length of the CPUE time series. Therefore, the implemented procedure is similar to a jackknife in that we remove points using a peel and then "predict" missing values as part of the fitting process. Time series of pseudo

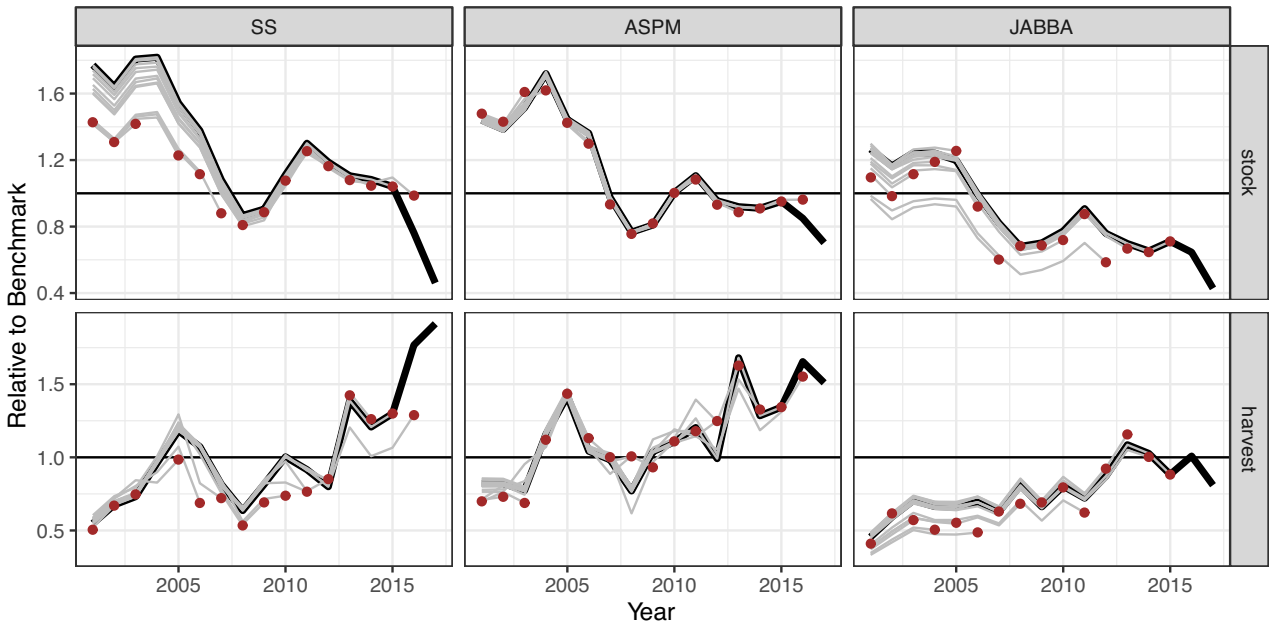

Figure 2. Hindcasts for one-step ahead, the thick solid line represent the model estimates for *stock* and *harvest* relative to *MSY* benchmarks, based on SSB and instantaneous fishing mortality for SS and ASPM-R and total biomass and harvest rate for JABBA. The points indicate the terminal years of the assessments peeled for the CPUE.

data (i.e. data that are artificially generated to test a program or procedure) were generated from estimates of vulnerable biomass and catchability (q). Prediction residuals (e) were then computed as the difference between the predictions and the observations. It is possible to perform the hindcast by peeling other data, e.g. the length or age compositions (see Carvalho *et al.*, 2021).

Relative error

In a retrospective analysis, Mohn's ρ (Mohn, 1999), is commonly used as a measure of relative error for model-based estimates. We used a variant, where we scaled by the mean, so the metric is not affected by the peel's length or the number of steps ahead.

$$\rho_M = \frac{1}{n}\sum_{t=T-n}^{T-1}\frac{\hat{y}_{(1:t),t} - \hat{y}_{(1:T),t}}{\hat{y}_{(1:T),t}}, \quad (1)$$

where n is the number of time steps that the peel is performed for, t is the time for which the missing value estimates, T is the terminal year in the CPUE series, and $\hat{y}$ denotes a model-based quantity, which in this case was *SSB*. The value with suffix $\hat{y}_{(1:T)|t}$ means a value estimated at time t from the full series running from time 1 to T, and $\hat{y}_{(1:t),t}$ is the value estimated using the data window from 1 to $t(\leq T)$. The data window is only applicable to the CPUE data window, as the catch and length composition data remain unchanged.

ρ_M is an average of the relative differences at the final time of each window and is a measure of relative retrospective 'bias' (scale-free) in a statistical sense. The metric tends to be applied not on the log but the original scale because both positive and negative directions are equivalent. ρ can be estimated for different horizons

$$\rho_M = \frac{1}{n-h+1}\sum_{t=T-n}^{T-h}\frac{\hat{y}_{(1:t)|t+h} - \hat{y}_{(1:T)|t+h}}{\hat{y}_{(1:T)|t+h}}. \quad (2)$$

There is no upper limit for reference values that are low relative to the alternative, while in the reverse case, the error cannot exceed 1.0. Therefore, it is usual to use a lower bound of −0.15 and an upper bound of 0.20 to identify acceptable performance for long-lived species (Hurtado-Ferro *et al.*, 2015) in practice. For values near or equal to 0, e.g. stocks where exploitation or stock size is low, small absolute differences can result in large relative differences. This may result in assessments being rejected when needed the most, e.g. during the development of recovery plans, when both stock biomass and fishing mortality may be low.

Prediction skill

Prediction skill compares an observation at time t (y_t) to an prediction of that observation made h time steps previously ($\hat{y}_{t|t-h}$). As a metric, we use the MASE, as it is a robust and easy to interpret statistic (Hyndman and Koehler, 2006). The MASE compares prediction error (e_t) for a prediction horizon of h

$$e_t = y_t - \hat{y}_{t|t-h} \quad (3)$$

to a benchmark forecast corresponding to a naïve forecast equal to the last observed value

$$\hat{y}_{t|t-h} = y_{t-h}. \quad (4)$$

For a peel of n and a horizon of h years

$$MASE = \frac{\frac{1}{n+1}\sum_{t=T-n}^{T}\left|y_t - \hat{y}_{t|t-h}\right|}{\frac{1}{n+1+h}\sum_{t=T-n-h}^{T}\left|y_t - y_{t-h}\right|}. \quad (5)$$

The MASE has the desirable properties of scale invariance, so it can compare forecasts across data sets with different scales and has predictable behaviour, symmetry, interpretability, and asymptotic normality. Unlike relative error, MASE does not skew its distribution even when the observed values are close to zero. It is easy to interpret as a score of 0.5 indicates that the model forecasts are twice

Table 1. Summary of Mohn's rho (ρ_M) statistics for relative stock status estimates from retrospective analysis (ρ_{M_r}) and hindcasts with one and three-step ahead projections (ρ_{M_p}) using the full Stock Synthesis (SS) reference case, the corresponding Age-Structure Equilibrium Model (ASPM-R) and Bayesian state-space biomass dynamics model JABBA.

Quantity	Method	1-step ahead (ρ_M)	3-step ahead (ρ_{M_p})
SSB/SSB_{MSY}	SS	0.03	0.32
SSB/SSB_{MSY}	ASPM-R	0.06	− 0.09
B/B_{MSY}	JABBA	− 0.03	− 0.21
F/F_{MSY}	SS	− 0.16	− 0.24
F/F_{MSY}	ASPM-R	0.01	0.08
U/U_{MSY}	JABBA	− 0.12	− 0.09

as accurate as a naïve baseline prediction. The Diebold-Mariano test (Diebold and Mariano, 1995) for one-step forecasts can also be used to test the statistical significance of the difference between two sets of forecasts, i.e. by comparing the prediction $y_t - \hat{y}_t$ to a random walk $y_t - y_{t-1}$.

Results

The one-step ahead and the three-step ahead estimates for model-based quantities are presented in Figure 2 and summarized in Table 1. These show estimates of stock size (*SSB* for age-based and biomass (*B*) for length-based methods) and exploitation level (fishing mortality (*F*) or harvest rate (*U*)), relative to their maximum sustainable yield (*MSY*) reference points.

No retrospective pattern is seen for ASPM-R, either for the one (Figure 2) or three-step (Figure 3) ahead forecasts. For SS, a negative bias is seen in SSB/SSB_{MSY} for one-step ahead, while for three-steps ahead, *SSB* is overestimated in the recent period little change is seen in the estimates of F/F_{MSY}. JABBA shows a negative pattern for harvest rate, which increases for three-step ahead forecast. In the case of JABBA, although exploitation is less than the *MSY* level stock biomass still declines below B_{MSY}, which implies that process error is driving the dynamics. When considering retrospective patterns, ρ has to be in the range $[-0.15, 0.2]$ for an assessment to be accepted (ICES, 2019). For the one-step ahead, all the assessments apart from the SS estimates of *F* pass this test. When the 3-year projection is considered, however, only ASPM-R shows acceptable performance.

The results from the model-free hindcasts (based on CPUE) are shown in Figures 4 and 5 for the one- and three-year ahead predictions respectively; the background colour indicates whether MASE ≤ 1. Table 2 summarizes the MASE values. For SS and JABBA, prediction skill is poor for CPUE indices 2 and 4, while the ASPM-R performs poorly for Region 2. Prediction skill deteriorates for the three-step ahead projections, particularly for SS and JABBA; although for ASPM-R, CPUE indices for Regions 1 and 3 still have good prediction skill.

The model and prediction residuals for future periods from one to five steps ahead are summarized in Figure 6 pooling across all CPUE indices. SS becomes increasingly imprecise and biased as the prediction horizon is increased. Although JABBA is more precise, it is still biased.

Discussion

Of the three structurally different model families used to assess Indian Ocean yellowfin, it was found that the model with the best prediction skill was the ASPM-R. Despite integrating all the available data, the SS base case assessment's poor performance is likely due to large sampling error in the length compositions, introducing noise rather than information about year-class strength. This was confirmed by an additional run performed as a check where the length data were down-weighted (effective sample size was 0.00005), where results were very similar to those of the ASPM-R.

Results for the ASPM-R suggest that a deterministic age-structured surplus production and observed catches could explain the trends in the indices of abundance and that the base case assessment model is incorrectly specified as it assigns too much weight to spurious signals in the length composition data. The Bayesian state-space biomass dynamic model, by contrast, produced reasonable performance metrics for the core fishing area (Region 1) and the south-western Indian Ocean (Region 3), but could not predict the diverging trends in CPUE for the Eastern Indian Ocean (Regions 3 and 4). Therefore, it appears that it is important for this stock to model both the age structure and spatial dynamics, while the quality of length samples needs to be improved.

There are many aspects of resource dynamics and productivity about which there is little information in stock assessment data sets (e.g. Lee *et al.* 2011, 2012; Jiao *et al.*, 2012; Simon *et al.*, 2012; Mangel *et al.*, 2013; Pepin and Marshall, 2015; Cury *et al.*, 2014). Therefore hypotheses about different plausible states of nature are increasingly represented by alternative model structures, fixed parameters, and weighting of data components (Sharma *et al.*, 2020). Thus, as well as methods for identifying uncertainties and agreeing on scenarios (Leach *et al.*, 2014), There is a need for methods to weigh, reject, extend models to include alternative hypotheses, and evaluate the value-of-information associated with different data sets.

However, model selection based on methods like AIC is only suitable for comparing frameworks based on the same input data. There is also a danger, with diagnostics based on the inspection of residuals, of "hypothesis fishing" or "p-hacking" (Head *et al.*, 2015; Wasserstein and Lazar, 2016), i.e. finding a pretext for excluding an index or adding extra parameters to improve the fit. On the other hand, if multiple true hypotheses are tested, some will likely be rejected falsely. Thus, it is valuable to reserve part of the data for hindcasting based on model-free validation so that a pattern's significance is not tested on the same data set that suggested the pattern (Arlot *et al.*, 2010).

The accuracy and precision of projections depend on the information in the data and determine how far ahead we can predict and the model's potential uses. If a model can not be validated, other uses include using a model as part of a feedback management system and scenario modelling when conducting MSE. An example of a model that can not be validated is a catch only model where if catch observations are removed can not be run. However, it can be simulation tested as part of an MP since feedback means that there is no need for prediction skill.

For example, backtesting is a form of hindcasting used in financial risk modelling to assess a trading or investment strategy's performance. This requires simulating past conditions, which is simple with the hindcast. MSE can be conducted as part of the backtest to allow the impact of feedback on historical catches and stock status to be evaluated. Although it is possible to find a strategy that would have worked well in the past, this is no guarantee that it will

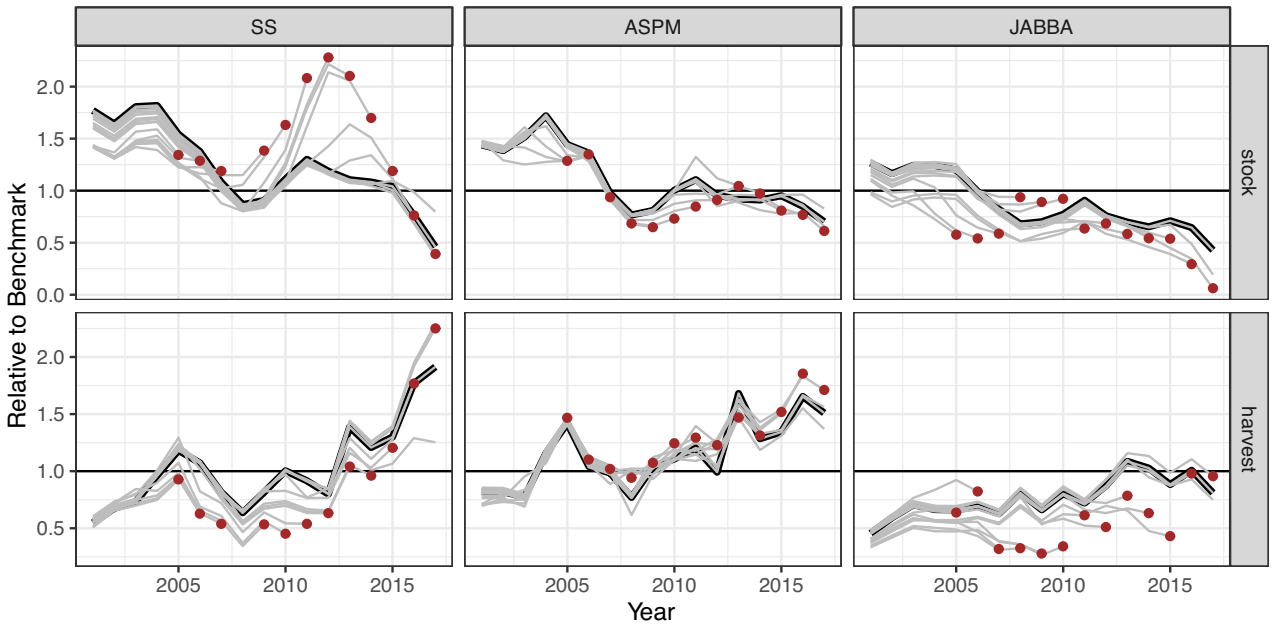

Figure 3. Hindcasts for three-step ahead, the thick solid line represent the model estimates for *stock* and *harvest* relative to *MSY* benchmarks, based on SSB and instantaneous fishing mortality for SS and ASPM-R and total biomass and harvest rate for JABBA. The points indicate the terminal years of the assessments peeled for the CPUE.

Table 2. MASE used for model-free validation of the full Stock Synthesis (SS) reference case, the corresponding Age-Structure Equilibrium Model (ASPM-R), and Bayesian state-space biomass dynamics model JABBA based on individual CPUE observations by region and quarter. The MASE values are shown for hindcasts with made with 1-year ahead and 3-year ahead projections

Region	Quarter	1 year			3 year		
		SS	ASPM-R	JABBA	SS	ASPM-R	JABBA
1	1	0.94	0.53	0.77	1.04	0.56	0.98
1	2	0.63	0.67	0.96	0.80	0.44	0.95
1	3	1.23	1.17	0.85	1.33	0.99	1.08
1	4	0.82	0.45	0.74	1.13	0.58	1.23
2	1	1.52	1.73	2.41	1.91	1.48	1.51
2	2	2.11	2.10	1.56	2.29	1.83	2.03
2	3	0.95	0.82	0.83	1.50	1.10	1.05
2	4	1.33	1.65	1.26	1.60	1.97	1.63
3	1	0.86	0.57	0.88	1.65	0.71	0.68
3	2	0.92	0.81	0.92	1.43	0.96	1.36
3	3	0.93	0.76	1.22	1.53	0.86	1.34
3	4	0.85	0.91	0.99	1.03	0.75	1.07
4	1	2.10	0.76	3.00	4.02	0.96	2.99
4	2	0.91	0.63	1.14	1.74	0.66	1.28
4	3	2.07	0.93	1.92	3.45	1.04	2.13
4	4	3.29	1.09	3.90	5.79	1.28	5.12

work well in the future. Therefore, although a backtest MSE is useful, particularly as it allows stakeholders to see the consequences of a different strategy, it is not sufficient to ensure the robustness of candidate future strategies. Despite this limitation, backtesting may provide insights that are unavailable when models and strategies are tested on simulated data alone.

Another promising field of applications is to use skill-based weighting for multimodel ensembles. A prediction skill score can be used to assign more weight to the better performing models objectively and has been found to improve forecasts (e.g. Casanova and Ahrens, 2009).

Conclusions

The hindcast is an important tool to achieve the aim of this paper, which was to support the definition of stock assessment "as

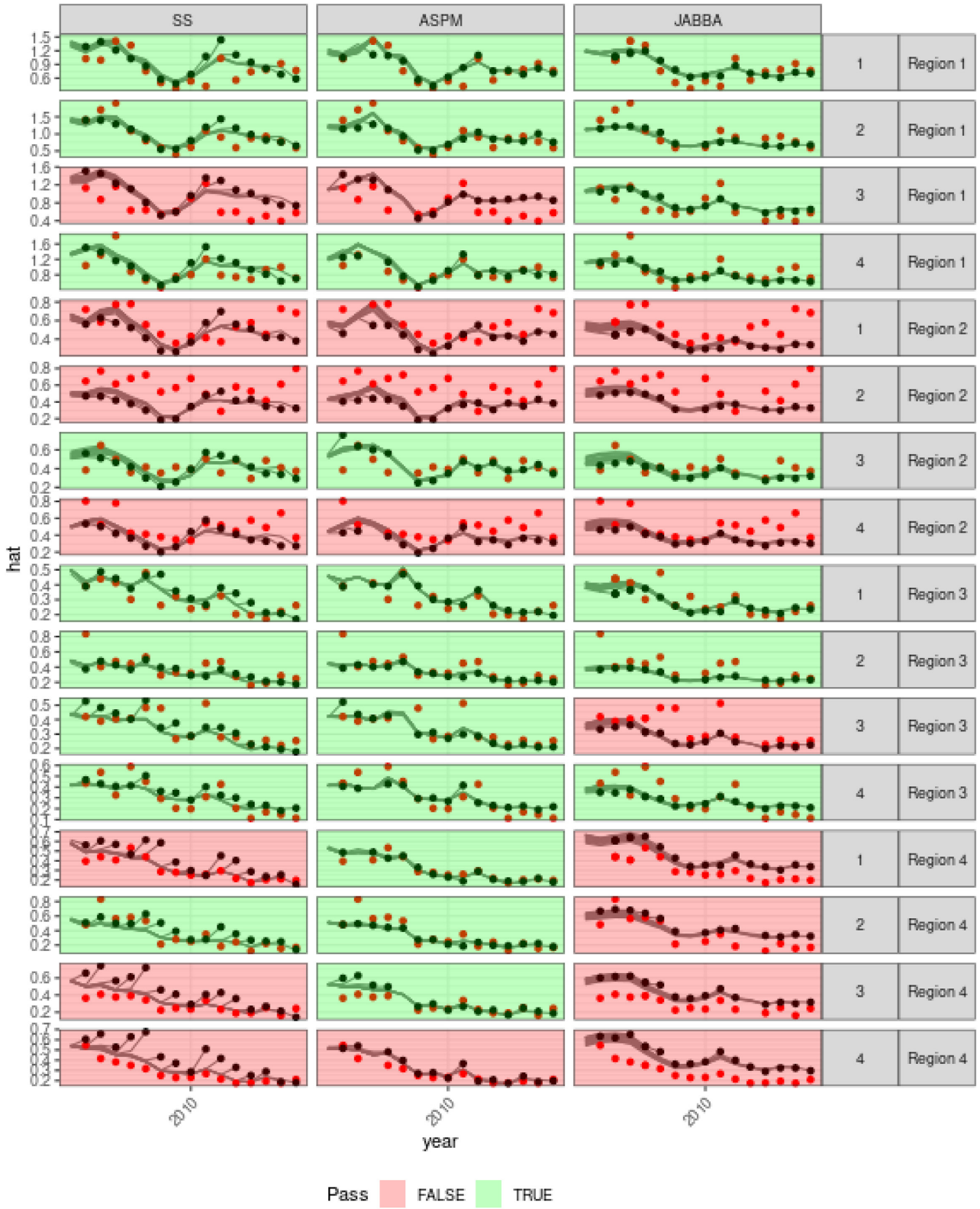

Figure 4. Hindcasts with one-step ahead for CPUE indices by region and quarter. Green backgrounds indicate that the CPUE index passes Mean Absolute Scaled Error (MASE < 1) criterion, or failed (red) otherwise. Red dots are the observed CPUE values and thin lines are the fits with terminal hincast year indicated by a solid point.

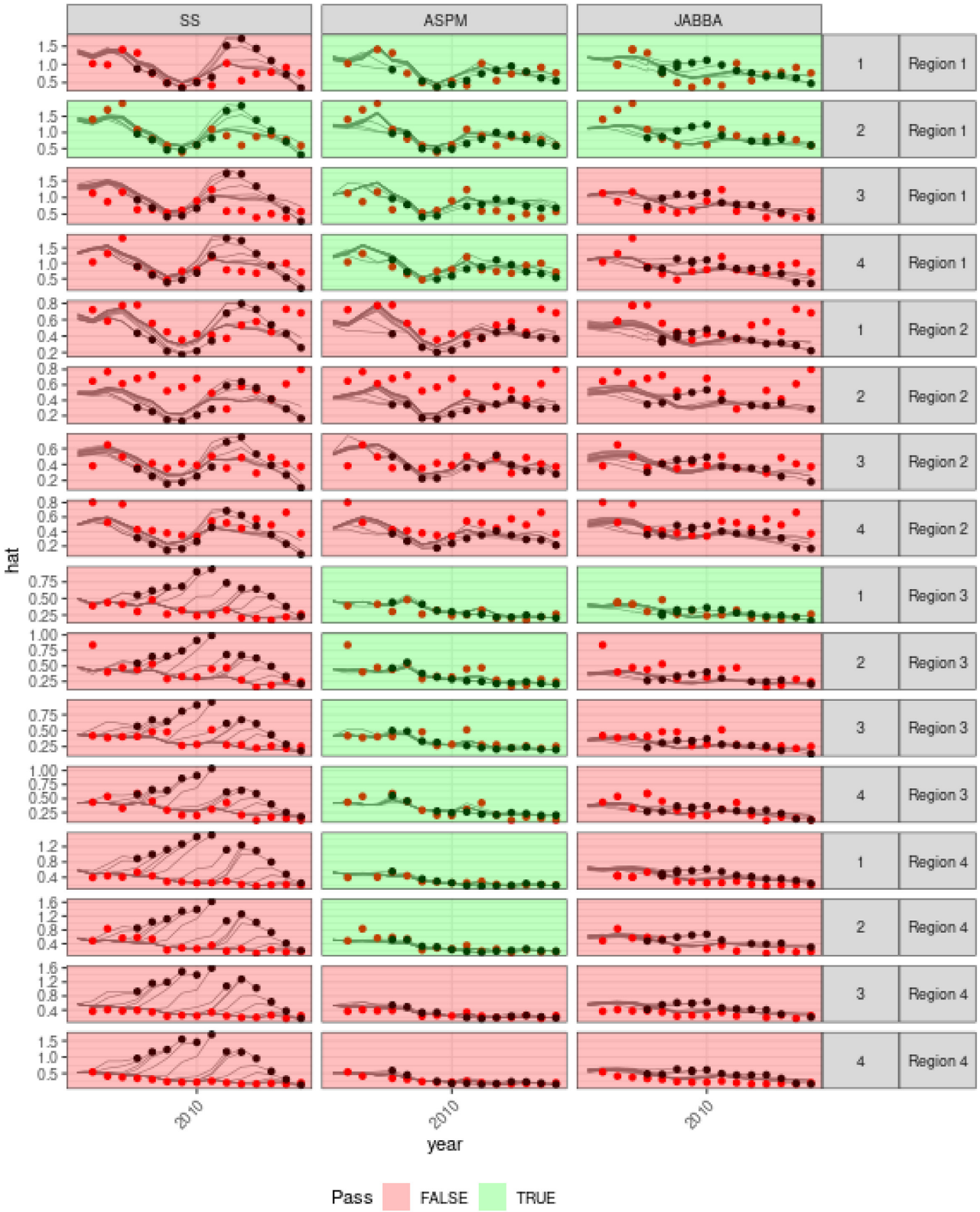

Figure 5. Hindcasts for three-step ahead for CPUE indices by region and quarter. Green backgrounds indicate that the CPUE index passes Mean Absolute Scaled Error (MASE < 1) criterion, or failed (red) otherwise. Red dots are the observed CPUE values and thin lines are the fits with terminal hincast year indicated by a solid point.

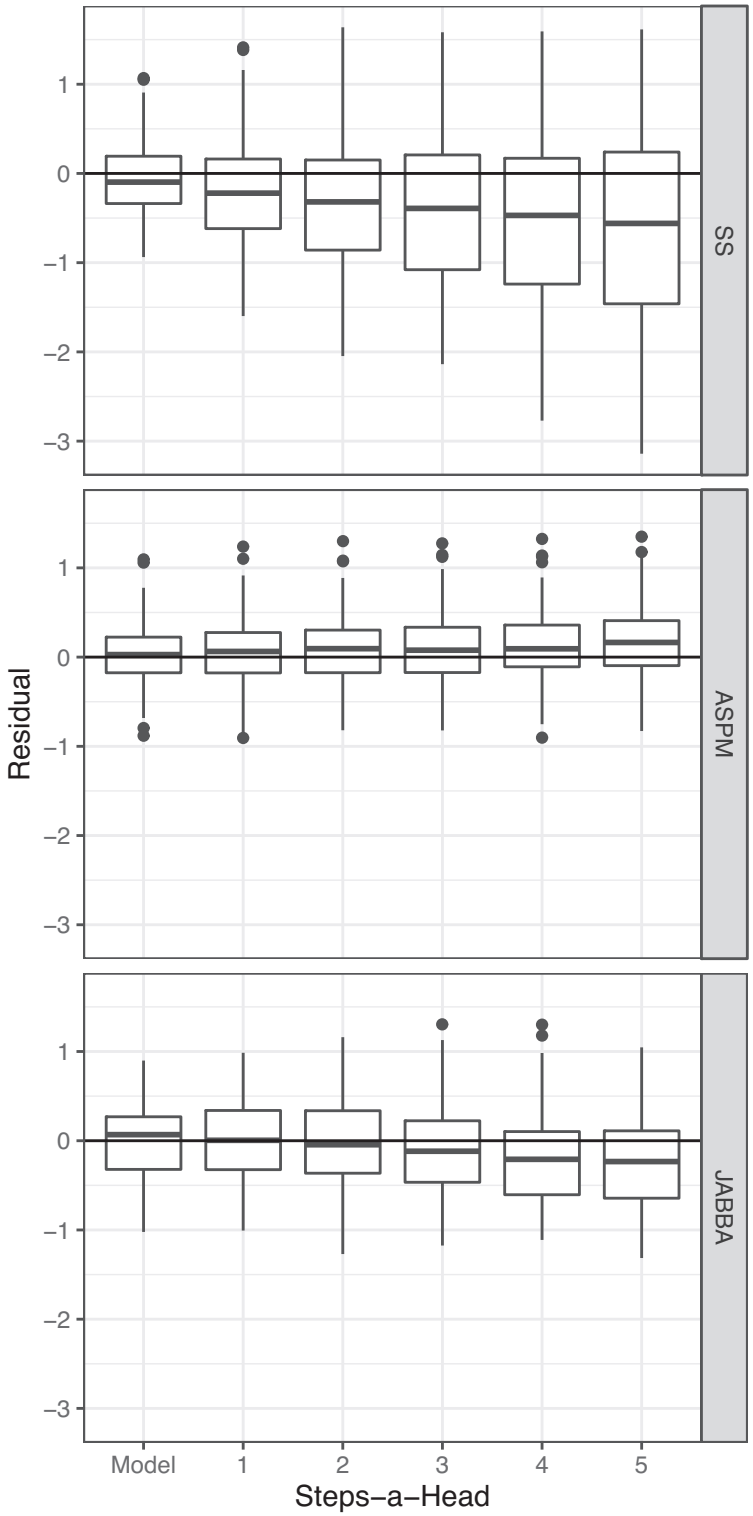

Figure 6. Boxplots showing model residuals and prediction residuals for 1,2,3,4, and 5 year ahead projections, pooled for all CPUE indices across regions and quarters.

the description of the characteristics of a 'stock' so that its biological reaction to being exploited can be rationally predicted and the predictions tested". If a model is to be used for forecasting, a model should be validated by comparing predictions to a system's observable and measurable properties (Ianelli *et al.*, 2016).

The objective of validation is not to prove that a model is correct, but to check that the model can not be falsified with the

available data. In other words, if a model has low prediction skill, then you either need more informative data or to develop alternative models. This is a step forward from hypothesis testing and model selection which is a way of rejecting rather than extending models.

Retrospective analysis evaluates the temporal stability of advice by using a reference series of model estimates based on the most recent assessment. In the Northeast US, allowing for time-varying dynamics (e.g., in natural mortality or catchability) has been used to eliminate retrospective patterns. This runs the risk of overfitting and could conceivably make estimates more biased (Brooks and Legault, 2016). It is not possible, however, to estimate bias using model-based and thus latent quantities. Instead, model predictions need to be compared to observations. Therefore, we used the hindcast to calculate prediction skill using simulated observations to identify overfitting and explore how models can be improved based on alternative structural assumptions without the risk of "hypothesis fishing".

An aim of stock assessment modelling is often to identify a 'best' model, ignoring uncertainty about model structure (Jardim *et al.*, 2020). To move beyond the best assessment to a multiple model approach, procedures need to be agreed upon for the initial selection of models and then for rejecting and weighting them. Currently, once a candidate model is agreed upon, rejection is mainly based on goodness of fit and retrospective analysis. However, this could result in overfitting, while the best way to remove a retrospective pattern is to ignore the data.

Prediction skill is an alternative and can be used to develop advice that is robust to uncertainty by weighing alternative models within an ensemble, either when providing estimates of stock status or when conducting MSE.

As the stock assessment process becomes more complex, e.g. through the increased use of integrated models, ensembles of models, and MSE, there are concerns about a lack of transparency. This is because of the many internal, implicit, and often poorly documented assumptions and a lack of access as only a few highly skilled experts can run or interrogate the models (Hilborn, 2003). Therefore to increase confidence in the outputs of a model and trust amongst the public, stake and asset-holders and policymakers, modellers need to ask, "is it me or my model talking" before others ask the question posed by Hodges and Dewar (1992) "is it you or your model talking?".

Data availability

The data underlying this article are available in the github repository flrpapers and can be found at https://github.com/flrpapers

Acknowledgements

We want to acknowledge Sidney Holt, as it was his comments in the plenary at the World Conference on Stock Assessment Methods in Boston, USA, in 2013 and subsequent email discussions that made us realize that the definition of stock assessment commonly used was inadequate. When we asked him for a definition of stock assessment, this was his reply.

"Your question deserves a serious answer, of course, and it is not easy. I used the term stock assessment because that is one being used loosely in ICES and other places — several terms are these days being used too loosely by scientists — one is 'forage fish'. I now intend to use stock assessment only in a narrow literal sense. To me, it means something like describing the characteristics of a 'stock' in such a way that its biological reaction to being exploited can be rationally predicted and the predictions tested."

This work was carried out using data provided by the International Commission for the Indian Ocean Tuna Commission (stock assessment) and reflects the information provided by several IOTC CPCs and the IOTC Secretariat. This paper's contents do not necessarily reflect the point of view of IOTC and in no way anticipate the Commission's future policy in this area.

We want to thank the ABNJ process in FAO (in particular, Alejandro Anganuzzi) for funding the original meetings across tRFMOs to discuss issues relevant to these assessments.

REFERENCES

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

Arlot, S., and Celisse, A. 2010. A survey of cross-validation procedures for model selection. Statistics surveys, 4: 40–79.

Balmaseda, M. A., Davey, M. K., and Anderson, D. L. 1995. Decadal and seasonal dependence of ENSO prediction skill. Journal of Climate, 8: 2705–2715.

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

Cadrin, S. X., and Dickey-Collas, M. 2014. Stock assessment methods for sustainable fisheries. CES Journal of Marine Science, 72: 1–6.

Carvalho, F., Punt, A. E., Chang, Y.-J., Maunder, M. N., and Piner, K. R. 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments?. Fisheries Research, 192: 28–40.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M. *et al.* 2021. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research, 240: 105959 .

Casanova, S., and Ahrens, B. 2009. On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. Monthly Weather Review, 137: 3811–3822.

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

Deroba, J., Butterworth, D., Methot, R. D., De Oliveira, J., Fernandez, C., Nielsen, A., Cadrin, S. X. *et al.* 2015. Simulation testing the robustness of stock assessment models to error: some results from the ices strategic initiative on stock assessment methods. ICES Journal of Marine Science, 72: 19–30.

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

FAO. 1996. Precautionary approach to capture fisheries and species introductions. Elaborated by the Technical consultation on the precautionary approach to capture fisheries (Including Species Introductions). FAO Technical Guidelines for Responsible Fisheries, vol. 2. Lysekil, Sweden, 1995–1996; 6–13 June.

Fiorellato, F., Pierre, L., and Geehan, J. 2019. Review of the statistical data and fishery trends for tropical tunas. IOTC-2019-WPTT21-08.

Fu, D., Langley, A., Merino, G., and Ijurco, A. 2018. Preliminary indian ocean yellowfin tuna stock assessment 1950-2017 (stock synthesis). IOTC-2018-WPTT20-33. 116 pp.

Glickman, T. S., and Zenk, W. 2000. Glossary of Meteorology. American Meteorological Society. http://glossary.ametsoc.org/wiki/Atmospheric_river.

Harley, S. J., Myers, R. A., and Dunn, A. 2001. Is catch-per-unit-effort proportional to abundance?. ICES Journal of Marine Science, 58: 1760–1772.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. 2015. The extent and consequences of p-hacking in science. PLoS Biol, 13: e1002106.

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

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

Hoyle, S. D., and Langley, A. D. 2020. Scaling factors for multi-region stock assessments, with an application to indian ocean tropical tunas. Fisheries Research, 228: 105586.

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

Hyndman, R. J., and Koehler, A. B. 2006. Another look at measures of forecast accuracy. International journal of forecasting, 22: 679–688.

Ianelli, J., Holsman, K. K., Punt, A. E., and Aydin, K. 2016. Multi-model inference for incorporating trophic and climate uncertainty into stock assessments. Deep Sea Research Part II: Topical Studies in Oceanography, 134: 379–389.

Orio, A., Karpushevskaia, A., Nielsen, A., Sundelöf, A., Berg, C.W. , Albertsen , C.M., Stralka, C. *et al.* 2019. Benchmark workshop on baltic cod stocks (wkbaltcod2). Reports 1

IOTC–WPTT21.2019. Report of the 21st working party on tropical tuna. Technical Report IOTC-2019-WPTT21-R. Seychelles. pp . 146.

Jardim, E., Azevedo, M., Brodziak, J., Brooks, E. N., Johnson, K. F., Klibansky, N., Millar, C. P. *et al.* 2020. Operationalizing ensemble models for scientific advice to fisheries management. doi:10.1093/icesjms/fsab010 .

Jiao, Y., Smith, E. P., O'Reilly, R., and Orth, D. J. 2012. Modelling nonstationary natural mortality in catch-at-age models. ICES Journal of Marine Science, 69: 105–118.

Jin, E. K., Kinter, J. L. III, Wang, B., Park, C.-K., Kang, I.-S., Kirtman, B., Kug, J.-S. *et al.* 2008. Current status of enso prediction skill in coupled ocean–atmosphere models. Climate Dynamics, 31: 647–664.

Kell, A. J., Forshaw, M., and McGough, A. S. 2020. Long-term electricity market agent based model validation using genetic algorithm based optimization. In Proceedings of the Eleventh ACM International Conference on Future Energy Systems. pp.. 1–13.

Kell, L. T., Kimoto, A., and Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries Research, 183: 119–127.

Langley, A. 2015. Stock assessment of yellowfin tuna in the indian ocean using stock synthesis. IOTC–2015–WPTT17–30. pp. 81 .

Leach, A., Levontin, P., Holt, J., Kell, L., and Mumford, J. 2014. Identification and prioritization of uncertainties for management of eastern Atlantic bluefin tuna (Thunnus thynnus). Marine Policy, 48: 84–92.

Lee, H.-H., Maunder, M. N., Piner, K. R., and Methot, R. D. 2011. Estimating natural mortality within a fisheries stock assessment model: an evaluation using simulation analysis based on twelve stock assessments. Fisheries Research, 109: 89–94.

Lee, H.-H., Maunder, M. N., Piner, K. R., and Methot, R. D. 2012. Can steepness of the stock–recruitment relationship be estimated in fishery stock assessment models?. Fisheries Research, 125: 254–261.

Mangel, M., MacCall, A. D., Brodziak, J., Dick, E., Forrest, R. E., Pourzand, R., and Ralston, S. 2013. A perspective on steepness, reference points, and stock assessment. ICES Journal of Marine Science, 70: 930–940.

Maunder, M. N., and Aires-da Silva, A. 2012. A review and evaluation of natural mortality for the assessment and management of yellowfin tuna in the eastern pacific ocean. inter-amer. trop. tuna comm. Technical Report, Document YFT-01-07.

Maunder, M. N., and Piner, K. R. 2015. Contemporary fisheries stock assessment: many issues still remain. ICES Journal of Marine Science, 72: 7–18.

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

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

Mohn, R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES Journal of Marine Science, 56: 473–488.

Pella, J., and Tomlinson, P. 1969. A Generalized Stock Production Model. Inter-American Tropical Tuna Commission. La Jolla, California, USA.

Pepin, P., and Marshall, C. T. 2015. Reconsidering the impossible—linking environmental drivers to growth, mortality, and recruitment of fish 1. ICES Journal of Marine Science, 72: 1–11.

Punt, A. E., and Donovan, G. 2007. Developing management procedures that are robust to uncertainty: lessons from the International Whaling Commission. ICES Journal of Marine Science, 64: 603–612.

Punt, A. E., Tuck, G. N., Day, J., Canales, C. M., Cope, J. M., de Moor, C. L., De Oliveira, J. A. *et al.* 2020. When are model-based stock assessments rejected for use in management and what happens then?. Fisheries Research, 224: 105465.

Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E., Espeland, W. N. *et al.* 2020. Five Ways to Ensure that Models Serve Society: a Manifesto.Nature, 582: pp. 482–484 .

Sharma, R., Levontin, P., Kitakado, T., Kell, L., Mosqueira, I., Kimoto, A., Scott, R. *et al.* 2020. Operating model design in tuna regional fishery management organizations: Current practice, issues and implications. Fish and Fisheries, 21: 940–961 .

Simon, M., Fromentin, J.-M., Bonhommeau, S., Gaertner, D., Brodziak, J., and Etienne, M.-P. 2012. Effects of stochasticity in early life history on steepness and population growth rate estimates: An illustration on atlantic bluefin tuna. PloS One, 7: e48583.

Thygesen, U. H., Albertsen, C. M., Berg, C. W., Kristensen, K., and Nielsen, A. 2017. Validation of ecological state space models using the laplace approximation. Environmental and Ecological Statistics, 24: 317–339.

Urtizberea, A., Fu, D., Merino, G., Methot, R. D., Cardinale, M., Winker, H., Walter, J. *et al.* 2019. Preliminary assessment of indian ocean yellowfin tuna 1950-2018 (stock synthesis, v3.30). IOTC-2018-WPTT21-50, 3.30: 60.

Wasserstein, R. L., and Lazar, N. A. 2016. The asa statement on p-values: Context, process, and purpose. The American Statistician, 70: 129–133.

Weigel, A., Liniger, M., and Appenzeller, C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?. Quarterly Journal of the Royal Meteorological Society, 134: 241–260.

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

Winker, H., Carvalho, F., and Kapur, M. 2018. Jabba: Just another bayesian biomass assessment. Fisheries Research, 204: 275–288.

Handling editor: Sarah Kraak