

Validation of stock assessment models. Is it me or my model talking?

Laurence T. Kell

Centre for Environmental Policy, Imperial College London, Weeks
Building, 16-18 Princes Gardens, London
SW7 1NE, UK
laurie@seaplusplus.co.uk

Rishi Sharma

Food and Agricultural Organization, Fishery and Aquaculture Policy and
Resources Division, Rome, Lazio,
00153, Italy
rishi.sharma@fao.org

Toshihide Kitakado

Department of Marine Biosciences, Tokyo University of Marine Science
and Technology, 4-5-7 Konan, Minato, Tokyo
108-8477, Japan
e-mail address

Henning Winker

Joint Research Centre (JRC), European Commission, TP 051, Via
Enrico Fermi 2749, 21027 Ispra, VA,
Italy.
e-mail address

Iago Mosqueira

Wageningen Marine Research, Haringkade 1, 1976CP IJmuiden,

The Netherlands.
iago.mosqueira@wur.nl

Massimiliano Cardinale
Swedish University of Agricultural Sciences, Department of Aquatic
Resources, Institute of Marine Research, Lysekil,
Sweden
massimiliano.cardinale@slu.se

Dan
Indian Ocean Tuna Commission, Le Chantier Mall, Po Box 1011,
Victoria – SEYCHELLES'

July 15, 2020

Abstract

The adoption of the Precautionary Approach requires a consideration of uncertainty, which is commonly addressed by the use of alternative stock assessment model structures or by fixing key parameters. Evaluating model fits in such cases is difficult, however, using traditional goodness-of-fit diagnostics based on likelihoods and model residuals. While retrospective analysis based on model outputs can not be used for validation as this requires a reference set of observations. Furthermore these methods merely tells us how well we describe the past, but little about how well we can predict the future under alternative management actions. Therefore, we use hindcasting to estimate prediction skill for three alternative model structures using the Indian Ocean yellowfin tuna assessment as a case study.

The approach can be used to develop robust advice, either by estimating current stock status for an ensemble of models or by weighting Operating Models when conducting Management Strategy Evaluation. For example backcasting is used in risk modelling to evaluate the performance of alternative strategies. This requires simulating past conditions which is simple with the hindcast. Conducting Management Strategy Evaluation as part of a backcast allows the impact of feedback that affect historical catches and stock status to be evaluated. Hindcasting also provides insights that may not be available when models and strategies are tested on simulated data.

Keywords: cross-validation, diagnostics, hindcast, retrospective analysis, stock assessment, validation

- link to metrics
- Sidney quote

1 Introduction

There are various definitions of stock assessment (e.g. Hilborn 2003; Cadrin et al. 2014). 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" (Holt pers comm.). Our reasoning is this explicitly recognises that the main aim of a stock assessment is to provide the basis for the long-term sustainable management of fisheries resources. Stock assessment, therefore, requires making and validating probabilistic estimates of stock status and forecasts of the consequences of management actions.

Stock assessment is a critical element of fisheries management, and diagnostic tests are essential for determining the robustness of model estimates (Carvalho, Winker, et al. 2020). Particularly since the adoption of the Precautionary Approach to fisheries management (PA, Garcia 1996) requires a formal consideration of uncertainty, which is often addressed by the use of alternative modelling frameworks, assumptions and datasets. It can be challenging, therefore, to use diagnostics based on likelihoods and model residuals to compare models. An alternative is to use hindcasting to evaluate prediction skill (Huschke et al. 1959) by comparing model estimates to known values, i.e. by withholding recent data and predicting for the period omitted. Hindcasting can either be model based by comparing model estimates, or model-free by generating pseudo observations. The pseudo data are then compared to a reference set of observations (Jin et al. 2008; Weigel et al. 2008; Balmaseda et al. 1995). Prediction skill can help to validate models by identify model misspecification, data conflicts and where models need to be extended. As a worked example, we compare three model families used for the assessment of Indian Ocean yellowfin tuna stock, namely a full integrated statistical model (SS), an age-structured production model (ASPM), and a Bayesian state-space biomass dynamic model (JABBA).

Model validation is essential in many fields, e.g. in energy and climate modelling (A. Kell et al. 2019), as this increases confidence in the outputs of a model and leads to an increase in trust amongst the public, stake and asset-holders and policymakers (Saltelli et al. 2020). Therefore, after a model structure has is agreed and parameters estimated, it is crucial to validate the model. Validation 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 the model cannot be falsified with the available data. A different question from asking is the model fit for a given purpose, which depends on the intended use of the model. For example to evaluate whether an assessment model can help achieve maximum sustainable yield (MSY) requires conducting Management Strategy Evaluation (MSE, André E. Punt et al. 2007); see Sharma et al. 2020 for a review of current practice in the tuna Regional Fisheries Management Organisations (tRFMOs).

Model validation serves a purpose complementary 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, and model validation examines if the model family should be modified or extended. For models to be valid, they must satisfy four prerequisites (Hodges et al. 1992). Namely, the situation modelled must: i) be observable and measurable; ii) be possible to collect sufficient data informative 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 (e.g. for highly migratory stocks such as 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 Harley et al. 2001 found strong evidence that commercial catch per unit effort (CPUE) is likely to remain high while abundance declines. Prerequisite (iii) ensures that the model has prediction skill for the same conditions under which the validation tests were conducted, and prerequisite (iv) ensures that the model will still be valid under conditions that differ from those in the validation tests.

A standard tool in stock assessment to check the stability of model estimates is retrospective analysis (Hurtado-Ferro et al. 2015), which involves sequentially removing data from the most recent years, refitting the model and comparing the time series of spawning stock biomass (SSB) and fishing mortality. Stability of historical estimates, however, can be achieved at the expense of the accuracy and precision of future forecasts, for example by shrinking terminal estimates towards recent historical values. We, therefore, extend the retrospective analysis by projecting forward for the reported catch over the years removed. The use of model-based quantities, however, is not sufficient to fulfil prerequisite i). We, therefore, conduct model-free hindcasts to estimate prediction skill. This is the main objective of the paper, as current literature (e.g. Methot and Wetzel 2013, Prager 1991) primarily focuses on the past, and how well models do fitting these data. Historical performance, however, is no indicator of how well a model may perform in the future, and the hindcasting tool developed here demonstrates how to test alternative models for prediction skill.

2 Material and Methods

Indices of abundance are a key 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 log-space is the measure of fitness. When comparing models, however, the SSE is problematic because complex models tend to have many parameters to allow flexibility when fitting, which may result in a low SSE due to overfitting. Therefore, information criteria, such as AIC, have been developed to aid in model selection. AIC is only a relative measure of the appropriateness of models, however, and additional diagnostic tests are required for model validation. This is of particular importance for stock assessment models where only a single historical data set exists, and the system can not be observed directly.

The objective of this study therefore is to develop a procedure to validate and compare different families of models based on past events. To do this we extend retrospective analysis to conduct model-based and model-free hindcasts, by adding the additional step of projecting over the truncated years. Comparing model outputs with observations allows prediction skill to be estimated (L. T. Kell, Kimoto, et al. 2016), defined as any measure of the accuracy of a forecasted value compared to the actual (i.e. observed) value that is not known by the model (Glickman et al. 2000).

2.1 Materials

For our example we use the stock assessment of yellowfin tuna (*Thunnus albacares*) conducted by the Indian Ocean Tuna Commission (IOTC 2019). Yellowfin tuna supports one of the largest tuna fisheries in the Indian Ocean, with catches currently exceeding 400,000t. They are harvested by a variety of gears, from small-scale artisanal fisheries, to large gillnetters, and industrial longliners and purse seiners (Fiorellato et al. 2019).

The main assessment is conducted using Stock Synthesis (SS, Methot et al. 2013), although other methods are also employed. SS implements an age and spatially structured model that reflects the complex population and fishery dynamics of the stock. Model development has focused on spatial structure to account for the differences in regional exploitation patterns, incorporating seasonal movement dynamics, resolving data conflicts, and exploring non-stationary in selectivity and catchability (Urtizberea et al. 2019). The data used includes time series of total catch and four CPUE indices based on the long-line fisheries, spatially stratified in four regions (figure 1)

The most recent assessment established a base case as a reference model for diagnostics along with scenarios to capture a range of uncertainties (Fu et al. 2018). The assessment indicates that the stock has declined substantially since 2012, and spawning stock biomass in 2017 is now estimated to be close to the historical lowest level. The stock is estimated to be overfished, and so the IOTC has implemented a rebuilding plan to reduce overall

fishing pressure.

The base case is spatially disaggregated into two tropical regions that encompass the main year-round fisheries and two austral, subtropical regions where the long-line fisheries occur more seasonally (Langley 2015), with reciprocal movement assumed to occur between adjacent regions. The SS assessment is based on a quarterly time step to approximate the continuous recruitment and rapid growth seen in the stock. Twenty-five fisheries were defined based on fishing gear, region, time period, fishing mode and vessel type. Most 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. The population comprised 28 quarterly age-classes with an assumed unexploited equilibrium initial state in each region.

Recruitment occurs in the two equatorial regions with temporal deviates in the regional distribution and was assumed to follow a Beverton and Holt stock recruitment relationship (with a steepness of 0.8 and recruitment standard deviation of 0.6). Growth was parameterised using age-specific deviates on the k growth parameter to mimic the non-von Bertalanffy growth of juvenile and the near linear growth of adults. Natural mortality is variable with age, with the relative trend in age-specific natural mortality based on the values applied in the Pacific Ocean (M. Maunder et al. 2012).

The data used for fitting are catch and length composition data, long-line CPUE indices, tagging recaptures, and environmental data. The length composition was weighted such that they were sufficient to provide reasonable estimates of fishery selectivity and recruitment trends but not directly influence the trends in stock abundance. Regional environmental indices (current and sea temperature) allows seasonal and temporal variations to be incorporated in the estimation of fish movement.

The CPUE indices represent the primary source of information on abundance and is based on a composite long-line index from the main distant water fleets. Indices in each region were standardised using generalised linear models that accounted for differences in targeting practices and catchability amongst fleets, based on gear configurations and species composition. 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 ability of the model to estimate the distribution of biomass by region.

Tag release/recovery data collected from the main phase of the Indian Ocean large-scale tuna tagging programme were integrated into the model to inform estimates of fishing mortality, abundance, and movement.

2.2 Assessment Methods

There has been a recent trend in stock assessment toward the use of integrated analysis that combines several sources of data into a single model by a joint likelihood for the observed data (e.g. Doubleday 1976; Fournier et al. 1982; Mark N Maunder and André E Punt 2013). Datasets include records of catches and landings, indices of abundance based on catch per unit (CPUE) or from research surveys, and length and age compositions based on samples. An example of commonly used integrated assessment method is SS3 that can be configured in multiple ways, allowing for a range of scenarios to be developed to reflect uncertainty.

For example Mark N Maunder and Piner 2015 proposed a deterministic implementation of an age-structured production model (ASPM) as a diagnostic of process dynamics. Selectivity in ASPM is parameterised based on the selectivity estimated by a "full" SS model. The model is then fitted to the abundance indices, assuming a deterministic spawning recruitment relationship, and without the size composition data contributing to the likelihood function. This enables an evaluation of whether the observed catches alone can explain trends in the index of abundance. If the ASPM is able to 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 catch data alone cannot explain the trends in the indices. This can have several causes, namely (i) stock dynamics are recruitment-driven, (ii) the stock has not yet declined to the point at which catch is a major factor influencing abundance; (iii) the indices of relative abundance are not proportional to abundance; (iv) the model is incorrectly specified; or (v) the data are incorrect

The ASPM has been shown (Carvalho, André E Punt, et al. 2017) to be the best method for detecting misspecification of the key systems-modeled processes that control the shape of the production function. This is a problem since many of the required parameters in integrated assessments are difficult to estimate (e.g. Lee et al. 2011; Lee et al. 2012) and have to be fixed or priors used.

An alternative to an integrated assessment is to use a biomass dynamic model, based on an explicit production function, that requires the estimation and fixing of fewer parameters. An example is JABBA, an open source package that presents a unifying, flexible framework for biomass dynamic modelling, runs quickly, and generates reproducible stock status estimates (Winker et al. 2018). The model uses a Pella Tomlinson production function that allows the shape of the production function to be varied, and alternative assumptions about productivity, stock status and reference points to be evaluated.

The base case SS assessment conducted by the Indian Ocean Tuna Commission (IOTC) for Yellowfin Tuna, was reconfigured as an ASPM and as a biomass dynamic assessment. In the later case in order to mimic the dynamics of the base case assess-

ment, the production function parameters were tuned to the base case Stock synthesis assessment.

Indices of abundance are a key contributor to the overall likelihood when fitting stock assessment models to data. The Sum of Squared Errors (SSE) between observed and predicted indices in log-space is the measure of fitness. When comparing models, however, the SSE is problematic because complex models tend to have many parameters to allow flexibility when fitting, which may result in a low SSE due to overfitting. Therefore, information criteria, such as AIC, have been developed to aid in model selection. AIC is only a relative measure of the appropriateness of models, and additional diagnostic tests are required for model validation. This is of particular importance for stock assessment models where only a single historical data set exists, and the system can not be observed directly. Therefore, Retrospective analysis (Hurtado-Ferro et al. 2015) is commonly used to evaluate the stability of stock assessment estimates.

2.3 Retrospective Analysis

In retrospective analysis observations are sequentially removed from the terminal year, the model is then refitted to the truncated series to see if there are any systematic pattern within a model. The retrospective bias is then evaluated using the so-called Mohn’s rho (Mohn 1999).

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

where \hat{y} denotes in general a model based quantity such as estimated biomass, or predicted abundance index. The value with suffix $\hat{y}_{(1:T),t}$ means a value estimated at time t from the full series from 1 to T and $\hat{y}_{(1:t),t}$ is the value using a retrospective data window from 1 to $t(\leq T)$.

In this paper, we use a variant of the original ρ as the mean: namely

$$\rho_{Mr} = \frac{1}{n} \sum_{t=T-n}^{T-1} \frac{\hat{y}_{(1:t),t} - \hat{y}_{(1:T),t}}{\hat{y}_{(1:T),t}} \quad [\text{rho for retro-bias}], \quad (2)$$

This metric is an average of relative differences at the final time of each window. Therefore it 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 the directions of positive and negative biases are regarded as being equivalent. There are problems with the use of this metric, since for reference values which are low relative to the alternative, there is no upper limit, while in the reverse case the error cannot exceed 1.0. Therefore when using this metric it is usual to use a lower bound of -0.15 and an upper bound of 0.20 to identify acceptable performance.

2.4 Hindcast

Hindcasting, the primary focus in this paper, is an extension of retrospective analysis which projects several steps forward beyond the retrospective data window to quantify the prediction skill of a model. Theoretically, the projection period is to the end of the historical time period. However, in practice, the step size is one or several years ahead reflecting the time horizon required for robust management advice; considering non-small process stochasticity in fishery population dynamics and non-ignorable extents of observation uncertainty. For evaluating prediction skill, there are several metrics for model-dependent and model-free validations.

Hindcasting like traditional retrospective analysis involves fitting a model using a tailcutting procedure, where data are deleted sequentially for n years, i.e. from year Tn through to the last year, the additional step in the hindcast is that then the data from $T - n - 1$ are used to make predictions of what will happen in years $T - n$ through T .

Since assessment cycles are typically for three years with advice (Fricker et al. 2013) we projected the truncated estimates for three years. Three years was chosen as that is essentially the time-step between assessments in most tuna Regional Fisheries Management Organisations.

Algorithm 1 Hindcast

```

1: Fit model up to and including terminal year  $T$ 
2: for  $t = T - 3$  to  $n$  do
3:   fit model to data up to time  $t$ 
4:   for  $i = t$  to  $t + 3$  do
5:     Estimate  $\hat{y}_i$ 
6:   end for
7: end for

```

We define ‘retro-period’ and ‘hc-period’ as ‘the period of shrunken data set for retrospective model fitting’ and ‘future time period with a certain projection step (say $S \geq 1$) for hindcasting after retro-period’. Let $\hat{y}_{(1:t),t+S}$ be a projected value at time $t + S$ in an hc-period based on the conditioned model with data in a retro-period $(1, t)$.

Modified Mohn’s rho for prediction bias and absolute error:

$$\rho_p = \frac{1}{n - S + 1} \sum_{t=T-n}^{T-S} \frac{\hat{y}_{(1:t),t+S} - \hat{y}_{(1:T),t+S}}{\hat{y}_{(1:T),t+S}} \quad (3)$$

This is a simple extension of Mohn’s rho to evaluate the prediction skill of a model because all the values are produced under the model assumption. In this sense, it is a model-dependent consistency check of prediction skill.

To evaluate the absolute prediction error for the following can be used

$$|\rho_p| = \frac{1}{(n - S + 1)} \sum_{t=T-n}^{T-S} \frac{|\hat{y}_{(1:t),t+S} - \hat{y}_{(1:T),t+S}|}{\hat{y}_{(1:T),t+S}}. \quad (4)$$

When conducting the hindcast it is assumed that modelled variables are observable, processes exhibit constancy of structure in time, including those not specified in the model, and that collection of accurate and sufficient data is possible (Hodges et al. 1992).

2.4.1 Prediction Skill

The use of model based quantities means that bias can not actually be quantified. For example a reduction in both relative error (a measure of bias) and mean squared error (a measure of variance) can be achieved by shrinking terminal estimates towards recent historical values, at the expense of prediction skill. The absence of retrospective patterns in model based quantities, therefore, while reassuring is not sufficient for model validation, and model-free validation using prediction residuals should be used as well.

2.4.2 Metrics

For the evaluation of model-based quantities, we use Mohn’s rho (ρ_M ρ_{Mr} ?, equation 2) and a modified version (ρ_{Mr} ρ_p ?, equation 3). ρ_p is also a measure of prediction skill but ”model based”, however, for validation prerequisite i of Hodges et al. 1992 requires that the system be observable and measurable and so model free quantities should be used to estimate prediction skill.

We used the mean absolute scaled error (MASE, equation 11) to evaluate prediction skill. The best statistical measure to use depends, however, on the objectives of the analysis and using more than one measure can be helpful in providing insight into the nature of observation and process error structures (L. T. Kell, Kimoto, et al. 2016). Therefore we also use root mean squared error (E'^2 , equation 7), correlation (ρ) and the standard deviation (σ).

E' is a commonly used metric, as the square root of a variance it can also be interpreted as the standard deviation of the unexplained variance, lower values indicate better fits. E' is sensitive to outliers, however, and favours forecasts that avoid large deviations from the mean and cannot be used to compare across series. The correlation (ρ) in contrast is unaffected by the amplitude of the variations, insensitive to biases and errors in variance, and can be used to compare across series. E'^2 and ρ are related by the cosine rule i.e.

$$E'^2 = \sigma_o^2 + \sigma_f^2 - 2\sigma_o\sigma_f\rho \quad (5)$$

Where the reference set (o) are the observations not included in the retrospective

assessment and the values (f) are their estimates.

This means that E' , ρ and σ_f can be summarised simultaneously in a single diagram (Taylor 2001) providing a concise statistical summary of how well patterns match each other and are therefore especially useful for evaluating multiple aspects or in gauging the relative skill of different models (Griggs et al. 2002).

3 Results

The retrospective analysis and the three step ahead predictions are shown in figure 3. These show the estimates of SSB and F relative to their maximum sustainable yield (MSY) reference points B_{MSY} and F_{MSY} . The retrospective analysis shows that SS base case and JABBA assessments are biased, and bias increases for the projection. For SS, there is over estimation of SSB and underestimation of F, while JABBA shows a strong negative retrospective pattern in F and negative bias in biomass. Although for JABBA $F : F_{MSY} \leq 1$ stock biomass declines below B_{MSY} .

The analysis is summarised in Tables 1 for Monh's ρ . Mohn's ρ has to be in the range $[-0.15, 0.2]$ for an assessment to be accepted then (Hurtado-Ferro et al., 2014). All the assessments apart from the base case estimates of F pass the Mohn's ρ test for the retrospective analysis. When the 3 year projection is considered, however, all models other than the ASPM fail. Relative error is harder to interpret.

The results from the model-free hindcasts are shown in Figures 4 and 5 for the one and three year ahead predictions respectively. The background colour indicates whether $MASE \leq 1$, see Table 2 for the MASE values. For the base case and JABBA, prediction skill is poor for CPUE indices 2 and 3; the ASPM also performs poorly for Area 2. Prediction skill further deteriorates for the three step ahead projection, particularly for the base case and JABBA; although for ASPM, CPUE indices 1 and 3 still have good prediction skill. Area 1 is the centre of the stock distribution (move this sentence to methods).

The fits are summarised in Figures 6 in the form of Taylor diagrams. Although for the one year projection most models appear to have prediction skill, for the three year hindcast the base case and JABBA perform poorly. The model and prediction residuals are summarised in Figure 7 over a five year horizon, these show that SS is imprecise and biased and that while JABBA is more precise, it is also biased.

From our analysis the following can be discerned with respect to how well each models with respect to the current status quo in evaluating models, i.e. retrospective analysis. We then present the model free evaluating algorithms with hindcast for one and three year ahead projections.

The ASPM appears not to show patterns in the projections, CPUE index 2 performs poorly across all models. The difference between the base case is that it uses length com-

position data, while JABBA does not model the different regions. It therefore, appears that the length compositions add noise and that area effects (JABBA) are important.

4 Discussion

This paper aimed to support the definition of stock assessment as the description of the characteristics of a 'stock' so that its biological reaction to being exploited can be rationally predicted and the predictions tested. There are many aspects of stock dynamics and productivity about which there is little information in stock assessment datasets (e.g. Lee et al. 2011; Lee et al. 2012; Jiao et al. 2012; Simon et al. 2012; Pepin et al. 2015; Cury et al. 2014). A motivation was, therefore, to ensure that stock assessment advice is robust to uncertainty, hypotheses about different plausible states of nature are represented by alternative model structures, fixed parameters, and weighting of data components (Sharma et al. 2020).

Model selection, however, based on traditional methods like Akaike's Information Criteria (AIC, Akaike 1998) is only suitable for comparing frameworks based on the same input data. Meanwhile, there is a danger, with diagnostics based on the inspection of residuals, of "hypothesis fishing" or "p-hacking" (Wasserstein et al. 2016; Head et al. 2015), i.e. finding a pretext for excluding an index or adding extra parameters (e.g. Schirripa et al. 2017) to improve the fit. If multiple true hypotheses are tested, some of them will likely be rejected falsely. Thus, it is valuable to reserve part of the data for 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). Therefore, we propose the use of hindcasting with cross-validation to evaluate prediction skill and show how to extend retrospective analysis to do this.

As an example, three structurally different model families were used to assess the yellowfin tuna stock in the Indian Ocean. It was found that the model with the best prediction skill was ASPM. The poor performance of the integrated assessment model is likely related to large, sampling error in the length composition data, which therefore introduce noise rather than information in the estimated recruitment deviations. 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 Eastern Indian Ocean Region. It appears that it is important to model both the age structure and spatial dynamics for this stock, while the quality of length samples needs to be improved. This illustrates the value of the approach since in the tRFMOs current practice is to run multiple models to evaluate uncertainty and provide advice on, while the evaluation of the value-of-information, the weighting and extension of models is still a research topic L. T. Kell, Levontin, et al. 2016.

If models are used for forecasting, then they should be validated, which requires that model predictions are compared to observable and measurable properties of the system (Ianelli et al. 2016). The accuracy and precision of the predictions depend on the validity of the model, the information in the data, and how far ahead we wish to predict (i.e. the prediction horizon). If a model is not validated then, although it may not be possible to use it for prediction, there are still other uses, which include scenario modelling or to use the model as part of a feedback management system. An example might be conducting MSE to develop harvest strategies that are robust to the impacts of environmental forcing or species interactions. To do this requires Operating Models, however, to be conditioned on assumptions other than those used in stock assessment models. A reason for this is because trends and fluctuations in populations are determined by complex interactions between extrinsic forcing and intrinsic dynamics (Bjoernstad et al. 2004; Botsford et al. 2014). Such low-frequency fluctuations can potentially mimic or cloak critical variation in abundance linked to environmental change, over-exploitation or other types of anthropogenic forcing. As advice frameworks move towards ecosystem-based fisheries management (EBFM), a broader range of processes will need to be considered (Ianelli et al. 2016). Which will require a greater number of plausible models to describe the system. Therefore, as well as methods for identifying uncertainties and agreeing on scenarios (Leach et al. 2014) there is a need for methods to weight, reject, and extend models to include alternative hypotheses, and to evaluate the value-of-information associated with different datasets.

5 Conclusions

The use of integrated assessment models has meant that fitting to all available data has become commonplace, as scientists seek to use these models to capture all knowledge about stock size and productivity (Hilborn 2003). Key questions, as when fitting any model, are (i) is my model valid? and (ii) do I have sufficient data and knowledge to fit it? In other words, is it me or my model talking? This is of particular importance for stock assessment, which generally relies on existing stock assessment packages or model frameworks. Stock assessment is generally different from other situations where a model is built from first principles, and diagnostics are applied at each stage of development. In our experience in stock assessments, when someone asks if it is a "good" assessment, they are more likely to be asking about the catch forecast than the goodness of fit diagnostics. We therefore used hindcasting to estimate prediction skill, which in the case study clearly showed that spatial structure is required, but that the length composition sampling error is too large to estimate year-class dynamics, and thus length composition data should be either omitted from the model or revisited.

Hindcasting enables objective comparison among models with different structures and

datasets, which is challenging to do with conventional diagnostics based on likelihoods and model residuals. The hindcast procedure was based on a commonly used diagnostic, retrospective analysis, but these were then extended by adding a forecast. Retrospective analysis allows the temporal stability of advice to be evaluated using a reference series of model estimates based on the most recent assessment. It is not possible, however, to estimate bias associated with prediction skill in this way by using model-based and thus latent quantities. Instead, model estimates need to be compared to observations for objective model validation. The retrospective hindcast was therefore modified to evaluate prediction skill by cross-validating indices of abundance not used in fitting against pseudo observations, i.e. their model estimates. Helping to identify over-fitting and allows exploration of how models can be improved based on alternative structural assumptions without the risk of ‘hypothesis fishing’.

The approach can also be used with an ensemble of models to develop robust advice, either for estimates of current stock status or by conducting MSE. For example backtesting is the process of evaluating the relative performance of alternatives strategies by comparing these to historical results, and is used in financial risk modelling to assess the performance of a trading or investment strategy. This requires simulating past conditions which is simple with the hindcast. While conducting MSE as part of the backtest allows the impact of feedback on historical catches and stock status to be evaluated. A problem is that it is possible to find a strategy that would have worked well in the past, but will not work well in the future. Therefore, although a backtest MSE is useful, particularly as it allows stakeholders to see what the consequences would have been if a different strategy had been employed, it is not sufficient to ensure the robustness of a strategy to be applied in the future. Despite this limitation, backtesting provides insights that may not be available when models and strategies are tested on simulated data alone and can be performed before conducting MSE for future years.

Another promising field of applications is to use the hindcast procedure to weight multi-model ensembles using simple skill-based weighting. Here, a prediction skill score can be used to assign more weight on the better performing models as this has been found to improve forecasts (e.g. Casanova et al. 2009). This can be done to weight estimates of current status relative to reference points, or weight operating models. Testing the approach on an actual case study first was informative as it provides the insight necessary to set up a study on synthetic data.

References

- Akaike, Hirotugu (1998). “Information theory and an extension of the maximum likelihood principle”. In: *Selected papers of hirotugu akaike*. Springer, pp. 199–213.
- Arlot, Sylvain, Alain Celisse, et al. (2010). “A survey of cross-validation procedures for model selection”. In: *Statistics surveys* 4, pp. 40–79.
- Balmaseda, Magdalena A, Michael K Davey, and David LT Anderson (1995). “Decadal and seasonal dependence of ENSO prediction skill”. In: *J. Climate* 8.11, pp. 2705–2715.
- Bjoernstad, O.N, R.M. Nisbet, and J-M. Fromentin (2004). “Trends and cohort resonant effects in age-structured populations”. In: *Journal of Animal Ecology* 73.6, pp. 1157–1167.
- Botsford, Louis W et al. (2014). “Cohort resonance: a significant component of fluctuations in recruitment, egg production, and catch of fished populations”. In: *ICES Journal of Marine Science: Journal du Conseil* 71.8, pp. 2158–2170.
- Cadrin, Steven X and Mark Dickey-Collas (2014). “Stock assessment methods for sustainable fisheries”. In: *ICES Journal of Marine Science* 72.1, pp. 1–6.
- Carvalho, Felipe, André E Punt, et al. (2017). “Can diagnostic tests help identify model misspecification in integrated stock assessments?” In: *Fisheries Research* 192, pp. 28–40.
- Carvalho, Felipe, Henning Winker, et al. (2020). “A Cookbook for Using Model Diagnostics in Integrated Stock Assessments”. In: *Submitted* x.xx, pp. xxx–xxx.
- Casanova, Sophie and Bodo Ahrens (2009). “On the weighting of multimodel ensembles in seasonal and short-range weather forecasting”. In: *Monthly weather review* 137.11, pp. 3811–3822.
- Cox, David R and E Joyce Snell (1968). “A general definition of residuals”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 30.2, pp. 248–265.
- Cury, Philippe M et al. (2014). “Resolving Hjort’s Dilemma How Is Recruitment Related to Spawning Stock Biomass in Marine Fish?” In: *Oceanography* 27.4, pp. 42–47.
- Doubleday, WG (1976). “A least squares approach to analyzing catch at age data”. In: *Int. Comm. Northwest Atl. Fish. Res. Bull* 12.1, pp. 69–81.
- Fiorellato, F., L. Pierre, and J. Geehan (2019). “Review of The Statistical Data And Fishery Trends For Tropical Tunas”. In: IOTC-2019-WPTT21-08.33.
- Fournier, David and Chris P Archibald (1982). “A general theory for analyzing catch at age data”. In: *Canadian Journal of Fisheries and Aquatic Sciences* 39.8, pp. 1195–1207.
- Fricker, Thomas E, Christopher AT Ferro, and David B Stephenson (2013). “Three recommendations for evaluating climate predictions”. In: *Meteorological Applications* 20.2, pp. 246–255.

- Fu, D et al. (2018). “Preliminary Indian ocean yellowfin tuna stock assessment 1950-2017 (Stock Synthesis)”. In: IOTC-2018-WPTT20.33.
- Garcia, SM (1996). “The precautionary approach to fisheries and its implications for fishery research, technology and management: an updated review”. In: *FAO Fisheries Technical Paper*, pp. 1–76.
- Glickman, Todd S and Walter Zenk (2000). *Glossary of meteorology*. American Meteorological Society.
- Griggs, David J and Maria Noguer (2002). “Climate change 2001: the scientific basis. Contribution of working group I to the third assessment report of the intergovernmental panel on climate change”. In: *Weather* 57.8, pp. 267–269.
- Harley, Shelton J, Ransom A Myers, and Alistair Dunn (2001). “Is catch-per-unit-effort proportional to abundance?” In: *Canadian Journal of Fisheries and Aquatic Sciences* 58.9, pp. 1760–1772.
- Head, Megan L et al. (2015). “The extent and consequences of p-hacking in science”. In: *PLoS Biol* 13.3, e1002106.
- Hilborn, Ray (2003). “The state of the art in stock assessment: where we are and where we are going”. In: *Scientia Marina* 67.S1, pp. 15–20.
- Hodges, James S, James A Dewar, and Arroyo Center (1992). *Is it you or your model talking?: A framework for model validation*. Santa Monica, CA: Rand.
- Hurtado-Ferro, Felipe et al. (2015). “Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models”. In: *ICES Journal of Marine Science* 72.1, pp. 99–110.
- Huschke, Ralph E et al. (1959). “Glossary of meteorology”. In:
- Hyndman, Rob J and Anne B Koehler (2006). “Another look at measures of forecast accuracy”. In: *International journal of forecasting* 22.4, pp. 679–688.
- Ianelli, James et al. (2016). “Multi-model inference for incorporating trophic and climate uncertainty into stock assessments”. In: *Deep Sea Research Part II: Topical Studies in Oceanography* 134, pp. 379–389.
- IOTC (2019). *Report of the 21st Working Party on Tropical Tuna*. Tech. rep. IOTC-2019-WPTT21-R. Pasaia, Spain: Indian Ocean Tuna Commission.
- Jiao, Yan et al. (2012). “Modelling non-stationary natural mortality in catch-at-age models”. In: *ICES J. Mar. Sci.* 69.1, pp. 105–118.
- Jin, Emilia K et al. (2008). “Current status of ENSO prediction skill in coupled ocean–atmosphere models”. In: *Climate Dynamics* 31.6, pp. 647–664.
- Kell, Alexander John Michael, Matthew Forshaw, and A. Stephen McGough (2019). “Optimising energy and overhead for large parameter space simulations”. In: *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*. IEEE, pp. 1–8.

- Kell, Laurence T, Ai Kimoto, and Toshihide Kitakado (2016). “Evaluation of the prediction skill of stock assessment using hindcasting”. In: *Fisheries research* 183, pp. 119–127.
- Kell, Laurence T, Polina Levontin, et al. (2016). “The quantification and presentation of risk”. In: *Management Science in Fisheries: An Introduction to Simulation-based Methods*, p. 348.
- Langley, A. (2015). “Stock assessment of yellowfin tuna in the Indian Ocean using Stock Synthesis”. In: IOTC-2018-WPTT20.33.
- Lee, Hui-Hua et al. (2011). “Estimating natural mortality within a fisheries stock assessment model: an evaluation using simulation analysis based on twelve stock assessments”. In: *Fish. Res.* 109.1, pp. 89–94.
- (2012). “Can steepness of the stock–recruitment relationship be estimated in fishery stock assessment models?” In: *Fish. Res.* 125, pp. 254–261.
- Maunder, Mark N and Kevin R Piner (2015). “Contemporary fisheries stock assessment: many issues still remain”. In: *ICES Journal of Marine Science* 72.1, pp. 7–18.
- Maunder, Mark N and André E Punt (2013). “A review of integrated analysis in fisheries stock assessment”. In: *Fisheries Research* 142, pp. 61–74.
- Maunder, MN and A Aires-da-Silva (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. Tech. rep. Document YFT-01-07.
- Methot, Richard D and Chantell R Wetzel (2013). “Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management”. In: *Fisheries Research* 142, pp. 86–99.
- Mohn, R (1999). “The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data”. In: *ICES Journal of Marine Science* 56.4, pp. 473–488.
- Pepin, Pierre and C Tara Marshall (2015). “Reconsidering the impossible???linking environmental drivers to growth, mortality, and recruitment of fish 1”. In: *Canadian Journal of Fisheries and Aquatic Sciences* 72.999, pp. 1–11.
- Punt, André E. and G.P. Donovan (2007). “Developing management procedures that are robust to uncertainty: lessons from the International Whaling Commission”. In: *ICES J. Mar. Sci.* 64.4, pp. 603–612.
- Saltelli, Andrea et al. (2020). *Five ways to ensure that models serve society: a manifesto*.
- Schirripa, Michael J et al. (2017). “A hypothesis of a redistribution of North Atlantic swordfish based on changing ocean conditions”. In: *Deep Sea Research Part II: Topical Studies in Oceanography* 140, pp. 139–150.
- Sharma, Rishi et al. (2020). “Operating model design in tuna Regional Fishery Management Organizations: Current practice, issues and implications”. In: *Fish and fisheries* x.xx, pp. xxx–xxx.

- Simon, Maximilien et al. (2012). “Effects of stochasticity in early life history on steepness and population growth rate estimates: An illustration on Atlantic bluefin tuna”. In: *PloS one* 7.10, e48583.
- Taylor, Karl E (2001). “Summarizing multiple aspects of model performance in a single diagram”. In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 106.D7, pp. 7183–7192.
- Thygesen, Uffe Høgsbro et al. (2017). “Validation of ecological state space models using the Laplace approximation”. In: *Environmental and Ecological Statistics* 24.2, pp. 317–339.
- Urtizberea, A. et al. (2019). “Preliminary Assessment of Indian Ocean Yellowfin Tuna 1950-2018 (Stock Synthesis, V3.30)”. In: IOTC-2018-WPTT21-50.33.
- Wasserstein, Ronald L and Nicole A Lazar (2016). *The ASA statement on p-values: context, process, and purpose*.
- Weigel, AP, MA Liniger, and C Appenzeller (2008). “Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?” In: *Quarterly Journal of the Royal Meteorological Society* 134.630, pp. 241–260.
- Whitten, Athol R et al. (2013). “Accounting for cohort-specific variable growth in fisheries stock assessments: a case study from south-eastern Australia”. In: *Fisheries Research* 142, pp. 27–36.
- Winker, Henning, Felipe Carvalho, and Maia Kapur (2018). “JABBA: Just Another Bayesian Biomass Assessment”. In: *Fisheries Research* 204, pp. 275–288.

6 Tables

Table 1: Mohn’s (ρ_{Mr}) for retrospective analysis.

Method	Quantity	Retrospective	Projection
SSB	SS	0.04	0.32
SSB	ASPM	-0.03	-0.09
SSB	JABBA	-0.09	-0.21
F	SS	-0.15	-0.24
F	ASPM	0.03	0.08
Harvest Rate	JABBA	-0.09	-0.09

7 Figures

Table 2: Model-free results

CPUE	Quarter	MASE			RMSE			ρ			σ		
		SS	ASPM	JABBA	SS	ASPM	JABBA	SS	ASPM	JABBA	SS	ASPM	JABBA
1	1	1.04	0.56	0.98	0.46	0.26	0.42	0.45	0.74	0.34	0.45	0.27	0.41
1	2	0.80	0.44	0.95	0.34	0.20	0.38	0.69	0.90	0.44	0.34	0.21	0.39
1	3	1.33	0.99	1.08	0.50	0.39	0.35	0.48	0.32	0.47	0.42	0.38	0.30
1	4	1.13	0.58	1.23	0.39	0.23	0.36	0.51	0.69	0.46	0.40	0.23	0.33
2	1	1.91	1.48	1.51	0.46	0.34	0.42	0.18	0.70	0.29	0.43	0.21	0.29
2	2	2.29	1.83	2.03	0.74	0.59	0.59	-0.33	0.14	0.06	0.57	0.36	0.32
2	3	1.50	1.10	1.05	0.46	0.32	0.30	0.33	0.34	0.16	0.44	0.28	0.25
2	4	1.60	1.97	1.63	0.50	0.49	0.40	0.44	0.61	0.30	0.39	0.24	0.26
3	1	1.65	0.71	0.68	0.52	0.22	0.23	0.27	0.73	0.66	0.40	0.22	0.24
3	2	1.43	0.96	1.36	0.46	0.31	0.36	0.42	0.63	0.82	0.41	0.31	0.19
3	3	1.53	0.86	1.34	0.38	0.26	0.31	0.49	0.64	0.76	0.35	0.24	0.20
3	4	1.03	0.75	1.07	0.52	0.34	0.41	0.70	0.84	0.86	0.39	0.33	0.42
4	1	4.02	0.96	2.99	0.75	0.21	0.48	0.47	0.84	0.90	0.44	0.22	0.16
4	2	1.74	0.66	1.28	0.82	0.34	0.47	0.50	0.81	0.78	0.56	0.35	0.37
4	3	3.45	1.04	2.13	0.86	0.27	0.49	0.33	0.69	0.73	0.55	0.28	0.21
4	4	5.79	1.28	5.12	0.82	0.20	0.51	0.50	0.88	0.90	0.53	0.18	0.15

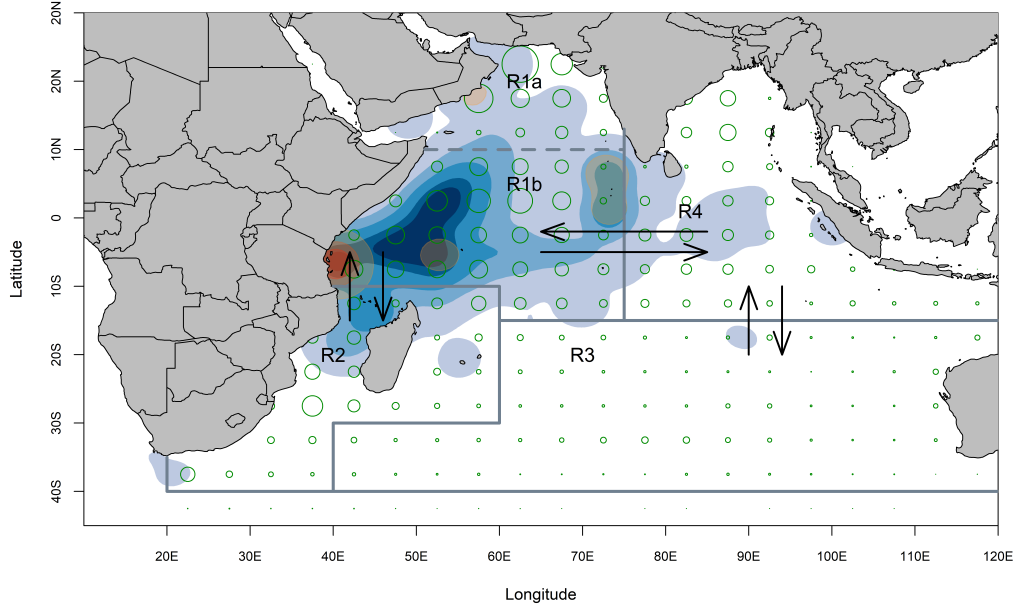


Figure 1: Spatial stratification of the Indian Ocean for the four region assessment model (R1a and R1b were treated as one model region but were retained for the fleet definition). The black arrows represent the configuration of the movement parameterization. Density contours represent of the dispersal of tag releases (red) and subsequent recaptures from Indian Ocean Regional tuna tagging programme. Green circles represent the distribution of catches from the longline fishery aggregated by 5 degree longitude and latitude for 1980 to 2017 (max. = 133 770 t).

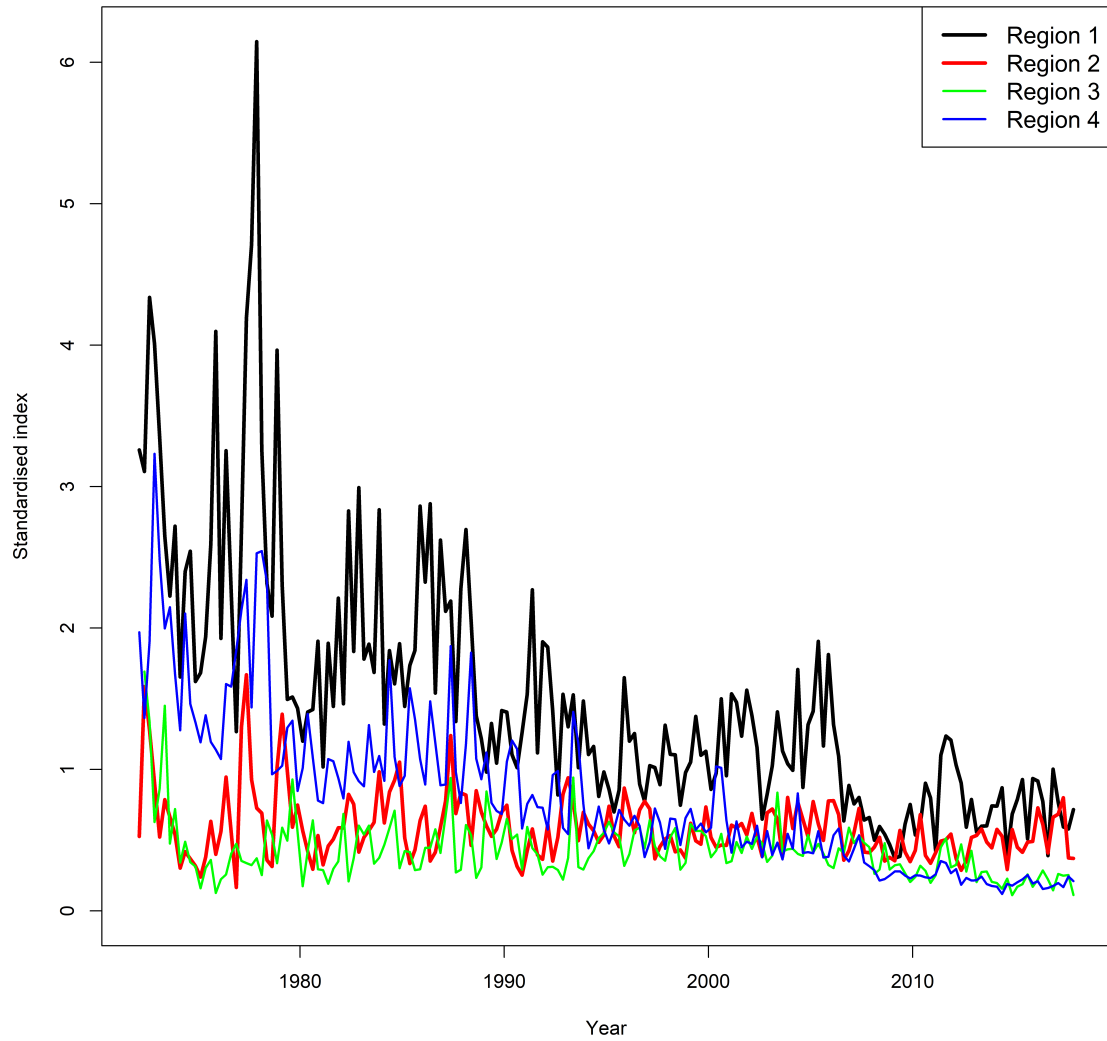


Figure 2: Regional longline CPUE indices included in the 2018 stock assessment. The difference in scales represents the relative distribution of longline vulnerable biomass amongst regions.

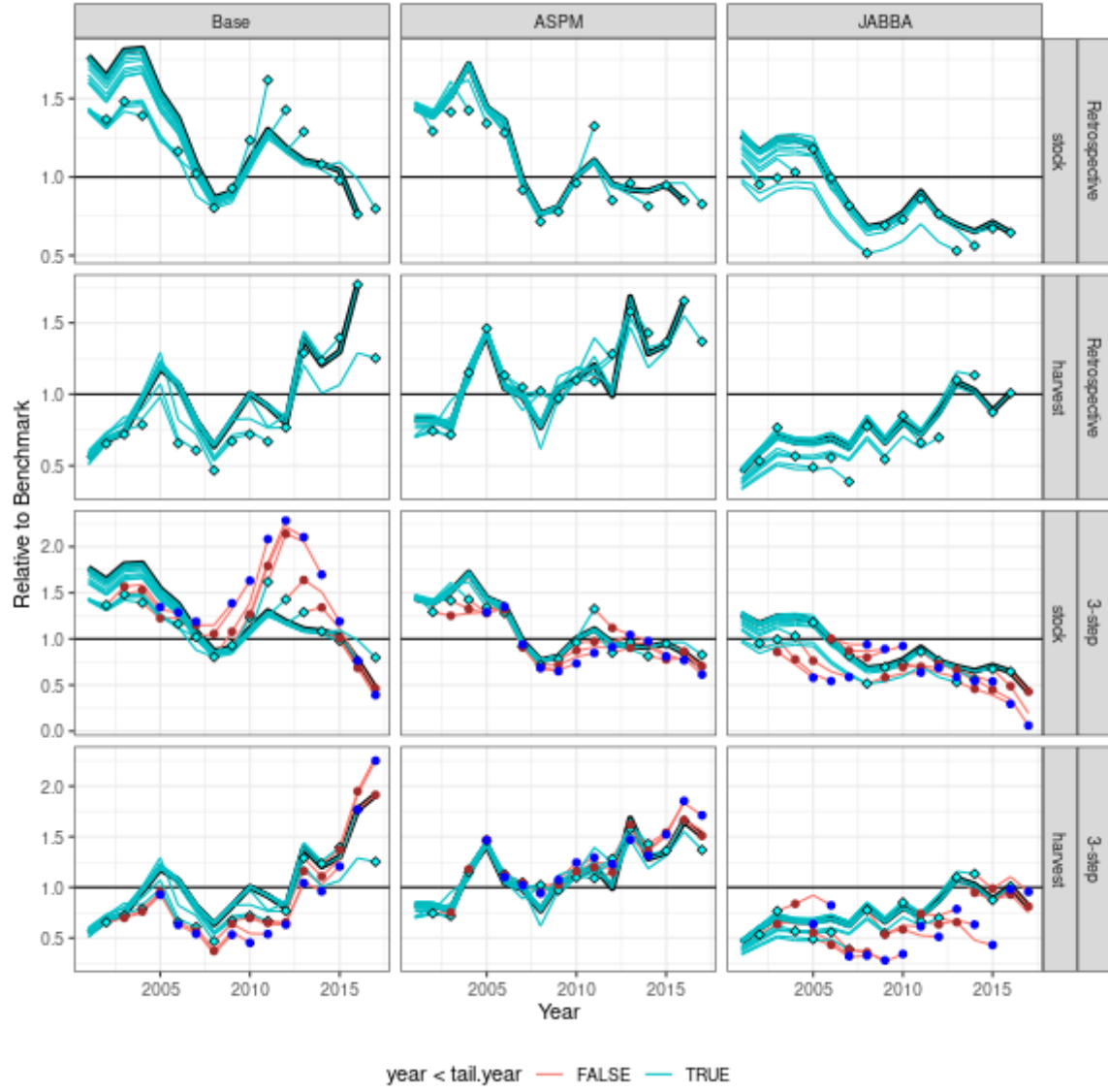


Figure 3: Retrospective analysis for the three models, points indicate the terminal years, and the thick line the assessment using all the data.



Figure 4: Hindcasts for one step ahead predictions, red dots are the observed CPUE values and lines are the fits with terminal hindcast year indicated by a point.

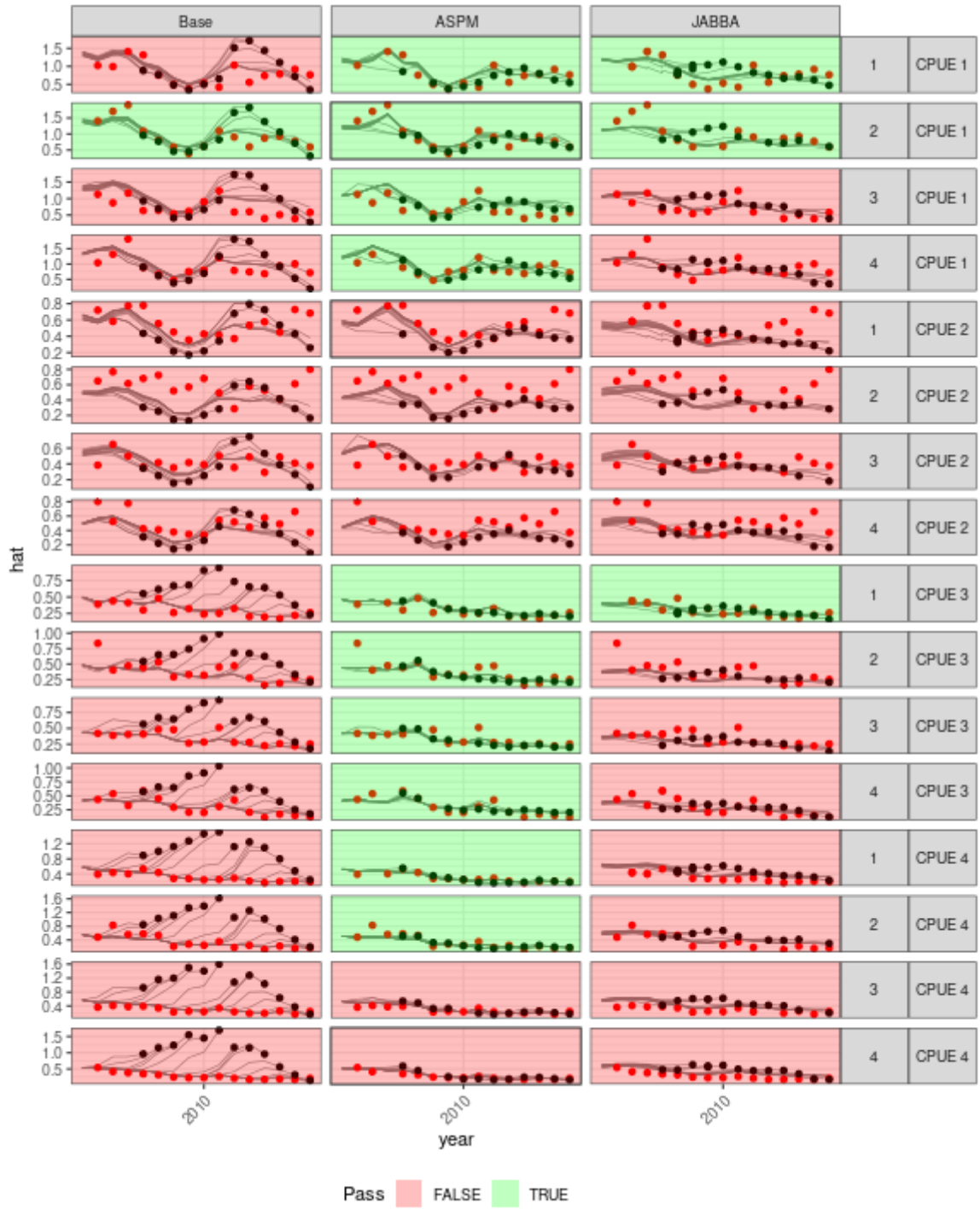


Figure 5: Hindcasts for three step ahead predictions, red dots are the observed CPUE values and lines are the fits with terminal hindcast year indicated by a point.

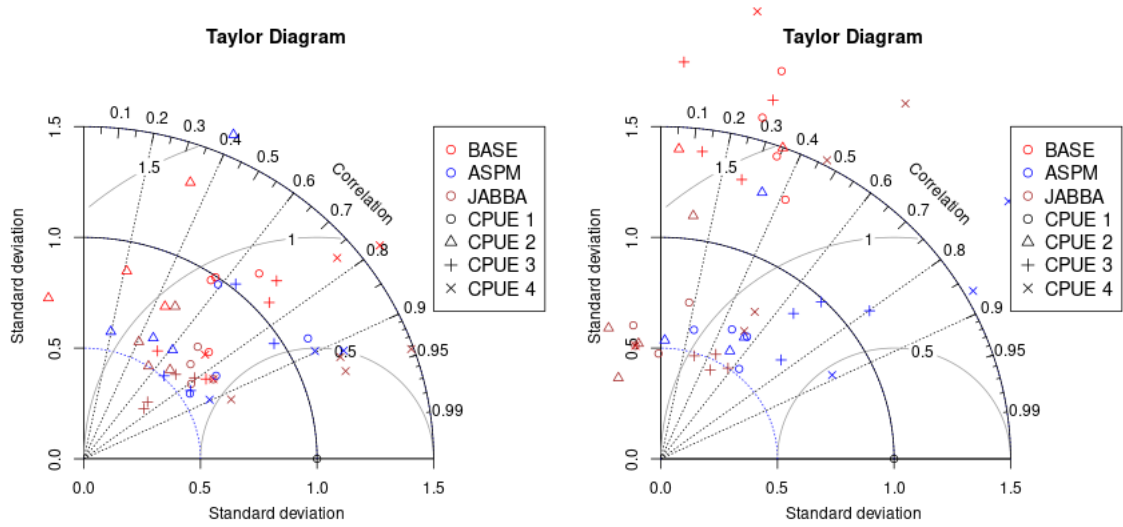


Figure 6: Taylor diagram for one and three year ahead predictions, summarising the similarity between the observed time series of CPUEs and the predicted relative stock abundance. Each point quantifies how closely predictions match observations, the angle indicates the correlation, the centred root-mean-square error difference between the predicted and observed patterns is proportional to the distance to the point on the x and the contours around this point indicate the RMSE values; the standard deviations of the predictions are proportional to the radial distance from the origin, scaled so the observed pattern has a value of 1. The open circle corresponds to a series which is identical to the reference series. The colours correspond to the model and shape to the survey.)

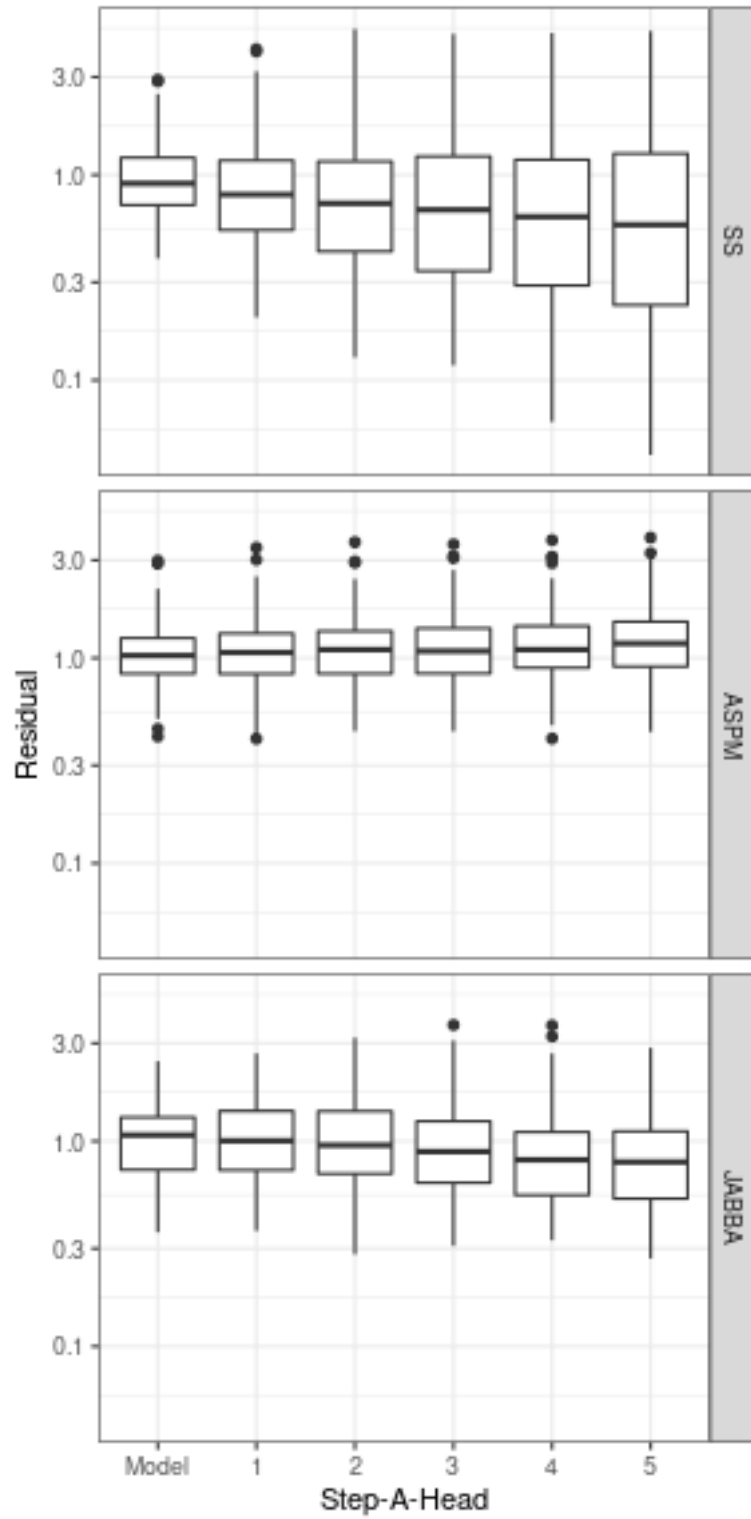


Figure 7: Residual for model (Step 0) and prediction residuals for 1,2,3,4 and 5 steps ahead.

7.1 Appendix

Metrics

The next three metrics are used for model-free validation, i.e. comparing predictions with observations. The error is defined as the difference between the predicted ($\hat{y}_{(1:t),t+S}$) and observed (y_{t+S}) values, such as the model-based predicted CPUE using a retro-period data and observed CPUE used for model fitting.

Mean Absolute Percentage Error (MAPE) for projection:

$$MAPE = \frac{1}{n - S + 1} \sum_{t=T-n}^{T-S} \frac{|\hat{y}_{(1:t),t+S} - y_{t+S}|}{y_{t+S}} \times 100 \quad (6)$$

A simple extension of the modified Mohn's rho for quantifying the relative difference between predictions and observations. This metric is also a scaled version of Mean Absolute Error (MAE). A problem with the MAE is that the relative size of the error is not always obvious. Sometimes it is hard to distinguish a big error from a small error. The MAPE can be calculated to allow forecasts of different series in different scales to be compared.

Root Mean Squared Error (RMSE) for projection error:

As an alternative measure of distance, the Mean Squared Error (MSE) is also commonly used in statistical literatures. To make comparison easier, the following squared root variant of MSE can be used:

$$RMSE = \sqrt{\frac{1}{n - S + 1} \sum_{t=T-n}^{T-S} (\hat{y}_{(1:t),t+S} - y_{t+S})^2} \quad (7)$$

In comparison to ρ_p and MAPE, RMSE is not scale-invariant and can be influenced by large discrepancies in a single data point. A useful feature, however, that the squared RMSE can, in general, be expressed, for a notational simplicity if we set S at 1, as

$$\begin{aligned} RMSE^2 &= \frac{1}{n} \sum_{t=T-n}^{T-1} (\hat{y}_{(1:t),t+1} - y_{t+1})^2 \\ &= \frac{1}{n} \sum_{t=T-n}^{T-1} (\hat{y}_{(1:t),t+1} - y_{t+1} - \bar{E})^2 + \bar{E}^2 \\ &= E'^2 + \bar{E}^2 \end{aligned} \quad (8)$$

where

$$\begin{aligned}\bar{E} &= \frac{1}{n} \sum_{t=T-n}^{T-1} (\hat{y}_{(1:t),t+1} - y_{t+1}), \\ E'^2 &= \frac{1}{n} \sum_{t=T-n}^{T-1} (\hat{y}_{(1:t),t+1} - y_{t+1} - \bar{E})^2.\end{aligned}\tag{9}$$

The centred mean squared error, E'^2 can be also expressed as

$$E'^2 = \sigma_o^2 + \sigma_f^2 - 2\sigma_o\sigma_f Cor,\tag{10}$$

where σ_o and σ_f are respectively the standard deviation of observation y_t and prediction, and Cor is the correlation between them. This means that E' , ρ and σ_f can be summarised simultaneously (Taylor 2001). Taylor diagrams provide a concise statistical summary of how well patterns match each other and are therefore useful for evaluating multiple aspects or in gauging the relative skill of different models (Griggs et al. 2002). It should be remarked that RMSE can be extended for a percentage measure as MAPE, but for the reason stated below, we use RMSE as defined above

Mean absolute scaled error (MASE) for projection:

A more robust and easier to interpret statistic for evaluating prediction skill is the MASE (Hyndman et al. 2006). MASE evaluates a model's prediction skill relative to a naïve baseline prediction, based on previous observation. A prediction is said to have skill if it improves the model forecast compared to the baseline. A widely used baseline forecast for time series is the persistence algorithm that takes the value at the previous time step to predict the expected outcome at the next time step as a naïve in-sample prediction, i.e. tomorrow weather will be the same as today. The original definition of MASE for 1-step ahead prediction is

$$MASE = \frac{\frac{1}{n} \sum_{t=T-n}^{T-1} |\hat{y}_{(1:t),t+1} - y_{t+1}|}{\frac{1}{n-1} \sum_{t=T-n+1}^{T-1} |y_{t+1} - y_t|},\tag{11}$$

and this can be extended as [actually not very much straightforward but seems as below](#)

$$MASE = \frac{\frac{1}{n-S+1} \sum_{t=T-n}^{T-S} |\hat{y}_{(1:t),t+S} - y_{t+S}|}{\frac{1}{n-S} \sum_{t=T-n+1}^{T-S} |y_{t+S} - y_t|}.\tag{12}$$

The MASE has the desirable properties of scale invariance, predictable behaviour, sym-

metry, interpretability and asymptotic normality. Compared to MAPE, which relies on the division by observations for scaling, MASE does not necessarily skew its distribution even when the observed values are close to zero. MASE is also easier to interpret as a score of 0.5 indicates that the model forecasts are twice as accurate as a naïve baseline prediction; the model thus has prediction skill. The best statistical measure to use depends on the objectives of the analysis and using more than one measure can be helpful in providing insight into the nature of observation and process error structures. Here for the evaluation of models, we will use the following metrics:

- Original Mohn's rho (ρ) for checking the retrospective bias
- Modified Mohn's rho for prediction [bias and absolute error, which? both might be meaningful though but it becomes noisy...](#) as checking model-based self-consistency check
- MASE and RMSE for model-free validation with different angles.